



An introduction to statistical learning theory

Fabrice Rossi

TELECOM ParisTech

November 2008

Part I

Introduction and formalization

About this lecture

Main goal

Show how statistics enable a rigorous analysis of machine learning methods which leads to a better understanding and to possible improvements of those methods

Outline

1. introduction to the formal model of statistical learning
2. empirical risk and capacity measures
3. capacity control
4. beyond empirical risk minimization

2 / 154

Fabrice Rossi

Outline



Outline

Machine learning in a nutshell

Formalization

Data and algorithms
Consistency
Risk/Loss optimization

3 / 154

Fabrice Rossi



4 / 154

Fabrice Rossi



Statistical Learning

Statistical Learning = Machine learning + Statistics

Machine learning in three steps:

1. record observations about a phenomenon
2. build a model of this phenomenon
3. predict the future behavior of the phenomenon

... all of this is done by the computer (no human intervention)

Statistics gives:

- ▶ a formal definition of machine learning
- ▶ some guarantees on its expected results
- ▶ some suggestions for new or improved modelling tools

Nothing is more practical than a good theory
Vladimir Vapnik (1998)

Machine learning

- ▶ a phenomenon is recorded via observations $\Rightarrow (\mathbf{z}_i)_{1 \leq i \leq n}$ with $\mathbf{z}_i \in \mathcal{Z}$
- ▶ two generic situations:
 1. **unsupervised learning**:
 - ▶ no predefined structure in \mathcal{Z}
 - ▶ then the goal is to find some structures: clusters, association rules, distribution, etc.
 2. **supervised learning**:
 - ▶ two non symmetric component in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
 - ▶ $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$
 - ▶ modelling: finding how \mathbf{x} and \mathbf{y} are related
 - ▶ the goal is to make prediction: given \mathbf{x} , find a reasonable value for \mathbf{y} such that $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ is compatible with the phenomenon
- ▶ this course focuses on supervised learning

Supervised learning

- ▶ several types of \mathcal{Y} can be considered:
 - ▶ $\mathcal{Y} = \{1, \dots, q\}$: **classification** in q classes, e.g., tell whether a patient has hyper, hypo or normal thyroidal function based on some clinical measures
 - ▶ $\mathcal{Y} = \mathbb{R}^q$: **regression**, e.g., calculate the price of a house \mathbf{y} based on some characteristics \mathbf{x}
 - ▶ \mathcal{Y} = “something complex but structured”, e.g., construct a parse tree for a natural language sentence
- ▶ modelling difficulty increases with the complexity of \mathcal{Y}
- ▶ Vapnik’s message: “don’t built a regression model to do classification”

Model quality

- ▶ given a dataset $(\mathbf{x}_i, \mathbf{y}_i)_{1 \leq i \leq n}$, a machine learning method builds a model g from \mathcal{X} to \mathcal{Y}
- ▶ a “good” model should be such that

$$\forall 1 \leq i \leq n, g(\mathbf{x}_i) \simeq \mathbf{y}_i$$

or not?

perfect interpolation is always possible
but is that what we are looking for?

- ▶ No! We want **generalization**:
 - ▶ a good model must “learn” something “smart” from the data, not simply memorize them
 - ▶ on **new data**, it should be such that $g(\mathbf{x}_i) \simeq \mathbf{y}_i$ (for $i > n$)

- ▶ main goal: from a dataset $(\mathbf{x}_i, \mathbf{y}_i)_{1 \leq i \leq n}$, build a model g which generalizes in a satisfactory way on new data
- ▶ but also:
 - ▶ request only as many data as needed
 - ▶ do this efficiently (quickly with a low memory usage)
 - ▶ gain knowledge on the underlying phenomenon (e.g., discard useless parts of \mathbf{x})
 - ▶ etc.
- ▶ the statistical learning framework addresses some of those goals:
 - ▶ it provides **asymptotic** guarantees about learnability
 - ▶ it gives **non asymptotic** confidence bounds around performance estimates
 - ▶ it suggests/validates best practices (hold out estimates, regularization, etc.)

Machine learning in a nutshell

Formalization

Data and algorithms
Consistency
Risk/Loss optimization

Data

- ▶ the phenomenon is fully described by an **unknown** probability distribution P on $\mathcal{X} \times \mathcal{Y}$
- ▶ we are given n observations:
 - ▶ $D_n = (X_i, Y_i)_{i=1}^n$
 - ▶ each pair (X_i, Y_i) is distributed according to P
 - ▶ pairs are **independent**
- ▶ remarks:
 - ▶ the phenomenon is **stationary**, i.e., P does not change: new data will be distributed as learning data
 - ▶ extensions to non stationary situations exist, see e.g. covariate shift and concept drift
 - ▶ the independence assumption says that each observation brings new information
 - ▶ extensions to dependent observation exists, e.g., Markovian models for time series analysis

Model quality (risk)

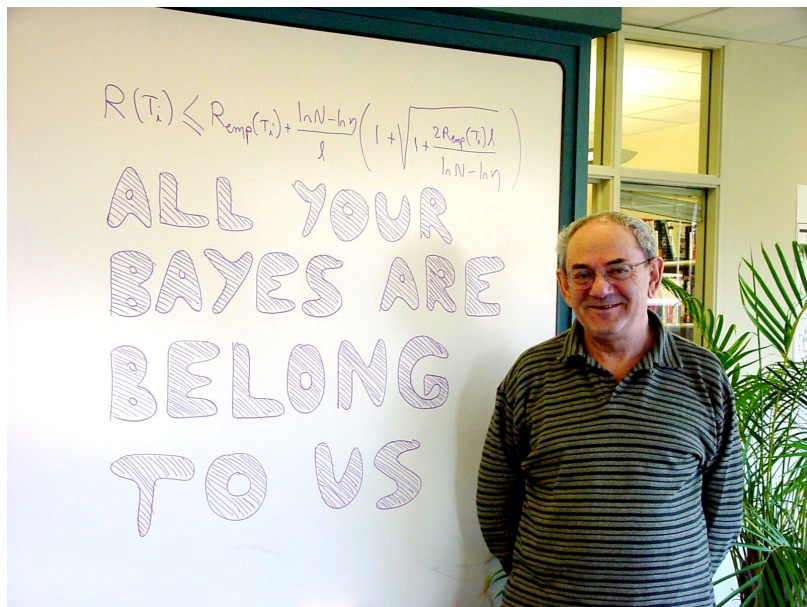
- ▶ quality is measured via a **cost function** $c: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ (a low cost is good!)
- ▶ the **risk** of a model $g: \mathcal{X} \rightarrow \mathcal{Y}$ is

$$L(g) = \mathbb{E} \{c(g(X), Y)\}$$

This is the expected value of the cost on an observation generated by the phenomenon (a low risk is good!)

- ▶ interpretation:
 - ▶ the average value of $c(g(\mathbf{x}), \mathbf{y})$ over a lot of observations generated by the phenomenon is $L(g)$
 - ▶ this is a measure of generalization capabilities (if g has been built from a dataset)
- ▶ remark: we need a probability space, c and g have to be measurable functions, etc.

We are frequentists



Cost functions

- ▶ Regression ($\mathcal{Y} = \mathbb{R}^q$):
 - ▶ the quadratic cost is the standard one:
 $c(g(\mathbf{x}), \mathbf{y}) = \|g(\mathbf{x}) - \mathbf{y}\|^2$
 - ▶ other costs ($\mathcal{Y} = \mathbb{R}$):
 - ▶ L^1 (Laplacian) cost: $c(g(\mathbf{x}), \mathbf{y}) = |g(\mathbf{x}) - \mathbf{y}|$
 - ▶ ϵ -insensitive: $c_\epsilon(g(\mathbf{x}), \mathbf{y}) = \max(0, |g(\mathbf{x}) - \mathbf{y}| - \epsilon)$
 - ▶ Huber's loss: $c_\delta(g(\mathbf{x}), \mathbf{y}) = (g(\mathbf{x}) - \mathbf{y})^2$ when $|g(\mathbf{x}) - \mathbf{y}| < \delta$
 and $c_\delta(g(\mathbf{x}), \mathbf{y}) = 2\delta(|g(\mathbf{x}) - \mathbf{y}| - \frac{\delta}{2})$ otherwise
- ▶ Classification ($\mathcal{Y} = \{1, \dots, q\}$):
 - ▶ 0/1 cost: $c(g(\mathbf{x}), \mathbf{y}) = \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}}$
 - ▶ then the risk is the probability of misclassification
- ▶ Remark:
 - ▶ the cost is used to measure the quality of the model
 - ▶ the machine learning algorithm may use a **loss** ≠ cost to build the model
 - ▶ e.g.: the hinge loss for support vector machines
 - ▶ more on this in part IV

Machine learning algorithm

- ▶ given D_n , a machine learning algorithm/method builds a model g_{D_n} ($= g_n$ for simplicity)
- ▶ g_n is a **random variable** with values in the set of measurable functions from \mathcal{X} to \mathcal{Y}
- ▶ the associated (random) risk is

$$L(g_n) = \mathbb{E} \{c(g_n(X), Y) \mid D_n\}$$

- ▶ statistical learning studies the behavior of $L(g_n)$ and $\mathbb{E} \{L(g_n)\}$ when n increases:
 - ▶ keep in mind that $L(g_n)$ is a **random variable**
 - ▶ if many datasets are generated from the phenomenon, the mean risk of the classifiers produced by the algorithm for those datasets is $\mathbb{E} \{L(g_n)\}$

Consistency

- ▶ best risk:

$$L^* = \inf_g L(g)$$

- ▶ the infimum is taken over all measurable functions from \mathcal{X} to \mathcal{Y}
- ▶ a machine learning method is:
 - ▶ **strongly consistent**: if $L(g_n) \xrightarrow{P.S.} L^*$
 - ▶ **consistent**: if $\mathbb{E} \{L(g_n)\} \rightarrow L^*$
 - ▶ **universally (strongly) consistent**: if the convergence holds for any distribution of the data P
- ▶ remark: if c is bounded then $\lim_{n \rightarrow \infty} \mathbb{E} \{L(g_n)\} = L^* \Leftrightarrow L(g_n) \xrightarrow{P} L^*$

Interpretation

- ▶ universality: no assumption on the data distribution (realistic)
- ▶ consistency: “perfect” generalization when given an infinite learning set

- ▶ strong case:

- ▶ for (almost) any series of observations $(\mathbf{x}_i, \mathbf{y}_i)_{i \geq 1}$
- ▶ and for any $\epsilon > 0$
- ▶ there is N such that for $n \geq N$

$$L(g_n) \leq L^* + \epsilon$$

where g_n is built on $(\mathbf{x}_i, \mathbf{y}_i)_{1 \leq i \leq n}$

- ▶ N depends on ϵ and on the dataset $(\mathbf{x}_i, \mathbf{y}_i)_{i \geq 1}$

- ▶ “weak” case:

- ▶ for any $\epsilon > 0$
- ▶ there is N such that for $n \geq N$

$$\mathbb{E} \{L(g_n)\} \leq L^* + \epsilon$$

where g_n is built on $(\mathbf{x}_i, \mathbf{y}_i)_{1 \leq i \leq n}$

Interpretation (continued)

- ▶ We can be unlucky:

- ▶ strong case:

- ▶ $L(g_n) = \mathbb{E} \{c(g_n(X), Y) \mid D_n\} \leq L^* + \epsilon$
- ▶ but what about

$$\frac{1}{p} \sum_{i=n+1}^{n+p} c(g_n(\mathbf{x}_i), \mathbf{y}_i)$$

- ▶ weak case:

- ▶ $\mathbb{E} \{L(g_n)\} \leq L^* + \epsilon$
- ▶ but what about $L(g_n)$ for a particular dataset?
- ▶ see also the strong case!

- ▶ Stronger result, **Probably Approximately Optimal** algorithm:

$\forall \epsilon, \delta, \exists n(\epsilon, \delta)$ such that $\forall n \geq n(\epsilon, \delta)$

$$\mathbb{P} \{L(g_n) > L^* + \epsilon\} < \delta$$

- ▶ Gives some convergence speed: especially interesting when $n(\epsilon, \delta)$ is **independent** of P (distribution free)

From PAO to consistency

- ▶ if convergence happens fast enough, then the method is (strongly) consistent

- ▶ “weak” consistency is easy (for c bounded):

- ▶ if the method is PAO, then for any fixed ϵ

$$\mathbb{P} \{L(g_n) > L^* + \epsilon\} \xrightarrow{n \rightarrow \infty} 0$$

- ▶ by definition (and also because $L(g_n) > L^*$), this means

$$L(g_n) \xrightarrow{P} L^*$$

- ▶ then, if c is bounded $\mathbb{E} \{L(g_n)\} \xrightarrow{n \rightarrow \infty} L^*$

- ▶ if convergence happens faster, then the method might be strongly consistent

- ▶ we need a very sharp decrease of $\mathbb{P} \{L(g_n) > L^* + \epsilon\}$ with n , i.e., a slow increase of $n(\epsilon, \delta)$ when δ decreases

From PAO to strong consistency

- ▶ this is based on the **Borel-Cantelli Lemma**

- ▶ let $(A_i)_{i \geq 1}$ be a series of events
- ▶ define $[A_i \text{ i.o.}] = \cap_{i=1}^{\infty} \cup_{j=i}^{\infty} A_j$ (infinitely often)
- ▶ the lemma states that

$$\text{if } \sum_i \mathbb{P} \{A_i\} < \infty \text{ then } \mathbb{P} \{[A_i \text{ i.o.}]\} = 0$$

- ▶ if $\sum_n \mathbb{P} \{L(g_n) > L^* + \epsilon\} < \infty$, then

$$\mathbb{P} \{[L(g_n) > L^* + \epsilon] \text{ i.o.}\} = 0$$

- ▶ we have $\{L(g_n) \rightarrow L^*\} = \cap_{j=1}^{\infty} \cap_{i=1}^{\infty} \cup_{n=i}^{\infty} \{L(g_n) \leq L^* + \frac{1}{j}\}$

- ▶ and therefore, $L(g_n) \xrightarrow{p.s.} L^*$

- ▶ from n i.i.d. observations $D_n = (X_i, Y_i)_{i=1}^n$, a ML algorithm builds a model g_n
- ▶ its risks is $L(g_n) = \mathbb{E} \{c(g_n(X), Y) \mid D_n\}$
- ▶ statistical learning studies $L(g_n) - L^*$ and $\mathbb{E} \{L(g_n)\} - L^*$
- ▶ the preferred approach is to show a fast decrease of $\mathbb{P} \{L(g_n) > L^* + \epsilon\}$ with n
- ▶ **no hypothesis on the data** (i.e., on P), a.k.a. **distribution free results**

- ▶ this program is quite different from classical statistics:
 - ▶ no optimal parameters, no identifiability: intrinsically non parametric
 - ▶ no optimal model: only optimal performances
 - ▶ no asymptotic distribution: focus on finite distance inequalities
 - ▶ no (or minimal) assumption on the data distribution
- ▶ minimal hypothesis
 - ▶ are justified by our lack of knowledge on the studied phenomenon
 - ▶ have strong and annoying consequences linked to the worst case scenario they imply

Optimal model

- ▶ there is nevertheless an optimal model
- ▶ regression:
 - ▶ for the quadratic cost $c(g(\mathbf{x}), \mathbf{y}) = \|g(\mathbf{x}) - \mathbf{y}\|^2$
 - ▶ the minimum risk L^* is reached by $g(\mathbf{x}) = \mathbb{E} \{Y \mid X = \mathbf{x}\}$
- ▶ classification:
 - ▶ we need the posterior probabilities: $\mathbb{P} \{Y = k \mid X = \mathbf{x}\}$
 - ▶ the minimum risk L^* (the **Bayes risk**) is reached by the Bayes classifier: assign \mathbf{x} to the most probable class $\arg \max_k \mathbb{P} \{Y = k \mid X = \mathbf{x}\}$
- ▶ we only need to mimic the prediction of the model, not the model itself:
 - ▶ for optimal classification, we don't need to compute the posterior probabilities
 - ▶ for the binary case, the decision is based on $\mathbb{P} \{Y = -1 \mid X = \mathbf{x}\} - \mathbb{P} \{Y = 1 \mid X = \mathbf{x}\}$
 - ▶ we only need to agree with the sign of $\mathbb{E} \{Y \mid X = \mathbf{x}\}$

No free lunch

- ▶ unfortunately, making no hypothesis on P costs a lot
- ▶ **no free lunch** results for binary classification and misclassification cost
- ▶ **there is always a bad distribution**:
 - ▶ given a **fixed** machine learning method
 - ▶ for all $\epsilon > 0$ and all n , there is (X, Y) such that $L^* = 0$ and

$$\mathbb{E} \{L(g_n)\} \geq \frac{1}{2} - \epsilon$$

- ▶ **arbitrary slow convergence always happens**:
 - ▶ given a **fixed** machine learning method
 - ▶ and a decreasing series (a_n) with limit 0 (and $a_1 \leq 1/16$)
 - ▶ there is (X, Y) such that $L^* = 0$ and

$$\mathbb{E} \{L(g_n)\} \geq a_n$$

Estimation difficulties

- ▶ in fact, **estimating the Bayes risk L^*** (in binary classification) **is difficult**:
 - ▶ given a **fixed** estimation algorithm for L^* , denoted \hat{L}_n
 - ▶ for all $\epsilon > 0$ and all n , there is (X, Y) such that

$$\mathbb{E} \left\{ |\hat{L}_n - L^*| \right\} \geq \frac{1}{4} - \epsilon$$

- ▶ **regression is more difficult than classification**:
 - ▶ η_n estimator for $\eta = \mathbb{E} \{ Y|X \}$ in a weak sense:

$$\lim_{n \rightarrow \infty} \mathbb{E} \{ (\eta_n(X) - \eta(X))^2 \} = 0$$

- ▶ then for $g_n(\mathbf{x}) = \mathbb{I}_{\{\eta_n(\mathbf{x}) > 1/2\}}$, we have

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E} \{ L(g_n) \} - L^*}{\sqrt{\mathbb{E} \{ (\eta_n(X) - \eta(X))^2 \}}} = 0$$



Estimation and approximation

- ▶ solution: split the problem in two parts, **estimation and approximation**
- ▶ estimation:
 - ▶ restrict the class of acceptable models to \mathcal{G} , a set of measurable functions from \mathcal{X} to \mathcal{Y}
 - ▶ study $L(g_n) - \inf_{g \in \mathcal{G}} L(g)$ when the ML algorithm must choose g_n in \mathcal{G}
 - ▶ **statistics needed!**
 - ▶ hypotheses on \mathcal{G} replace hypotheses on P and allow uniform results!
- ▶ approximation:
 - ▶ compare \mathcal{G} to the set of all possible models
 - ▶ in other words, study $\inf_{g \in \mathcal{G}} L(g) - L^*$
 - ▶ function approximation results (density), no statistics!



Consequences

- ▶ no free lunch results show that we cannot get **uniform** convergence speed:
 - ▶ they don't rule out universal (strong) consistency!
 - ▶ they don't rule out PAO with hypothesis on P
- ▶ intuitive explanation:
 - ▶ stationarity and independence are not sufficient
 - ▶ if P is arbitrarily complex, then **no generalization** can occur from a **finite** learning set
 - ▶ e.g., $Y = f(X) + \epsilon$: interpolation needs regularity assumptions on f , extrapolation needs even stronger hypothesis
- ▶ but we don't want hypothesis on P ...



Back to machine learning algorithms

- ▶ many *ad hoc* algorithms:
 - ▶ k nearest neighbors
 - ▶ tree based algorithms
 - ▶ Adaboost
- ▶ but most algorithms are based on the following scheme:
 - ▶ fix a model class \mathcal{G}
 - ▶ choose g_n in \mathcal{G} by optimizing a quality measure defined via D_n
- ▶ examples:
 - ▶ linear regression (\mathcal{X} is a Hilbert space, $\mathcal{Y} = \mathbb{R}$):
 - ▶ $\mathcal{G} = \{ \mathbf{x} \mapsto \langle w, \mathbf{x} \rangle \}$
 - ▶ find w_n by minimizing the mean squared error $w_n = \arg \min_w \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \langle w, \mathbf{x}_i \rangle)^2$
 - ▶ linear classification ($\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \{1, \dots, c\}$):
 - ▶ $\mathcal{G} = \{ \mathbf{x} \mapsto \arg \max_k (W\mathbf{x})_k \}$
 - ▶ find W e.g., by minimizing a mean squared error for a disjunctive coding of the classes



Criterion

- ▶ the best solution would be to choose g_n by minimizing $L(g_n)$ over \mathcal{G} , but:
 - ▶ we cannot compute $L(g_n)$ (P is unknown!)
 - ▶ even if we knew $L(g_n)$, optimization might be intractable (L has no reason to be convex, for instance)

- ▶ partial solution, the **empirical risk**:

$$L_n(g) = \frac{1}{n} \sum_{i=1}^n c(g(X_i), Y_i)$$

- ▶ naive justification by the **strong law of large numbers** (SLLN): when g is **fixed**, $\lim_{n \rightarrow \infty} L_n(g) = L(g)$
- ▶ the algorithmic issue is handled by replacing the empirical risk by an **empirical loss** (see part IV)

Empirical risk minimization

- ▶ generic machine learning method:

- ▶ fix a model class \mathcal{G}
- ▶ choose g_n^* in \mathcal{G} by

$$g_n^* = \arg \min_{g \in \mathcal{G}} L_n(g)$$

- ▶ denote $L_{\mathcal{G}}^* = \inf_{g \in \mathcal{G}} L(g)$
- ▶ does that work?
 - ▶ $\mathbb{E} \{L(g_n^*)\} \rightarrow L_{\mathcal{G}}^*$?
 - ▶ $\mathbb{E} \{L(g_n^*)\} \rightarrow L^*$?
 - ▶ the SLLN does not help as g_n^* **depends on n**
- ▶ we are looking for bounds:
 - ▶ $L(g_n^*) < L_n(g_n^*) + B$ (bound the risk using the empirical risk)
 - ▶ $L(g_n^*) < L_{\mathcal{G}}^* + C$ (guarantee that the risk is close to the best possible one in the class)
- ▶ $L_{\mathcal{G}}^* - L^*$ is handled differently (approximation vs estimation)

Complexity control

The overfitting issue

- ▶ controlling $L_n(g_n^*) - L(g_n^*)$ is not possible in arbitrary classes \mathcal{G}
- ▶ a simple example:
 - ▶ binary classification
 - ▶ \mathcal{G} : all piecewise functions on arbitrary partitions of \mathcal{X}
 - ▶ obviously $L_n(g_n^*) = 0$ if P_X has a density (via the classifier defined by the 1-nn rule)
 - ▶ but $L(g_n^*) > 0$ when $L^* > 0$!
 - ▶ the 1-nn rule is **overfitting**
- ▶ the **only solution** is to reduce the “size” of \mathcal{G} (see part II)
- ▶ this implies $L_{\mathcal{G}}^* > L^*$ (for arbitrary P)
- ▶ how to reach L^* ? (will be studied in part III)
- ▶ in other words, how to balance estimation error and approximation error

Increasing complexity

- ▶ key idea: choose g_n in \mathcal{G}_n , i.e., in a class that depends on the datasize
- ▶ estimation part:
 - ▶ statistics
 - ▶ make sure that $L_n(g_n^*) - L(g_n^*)$ is under control when $g_n^* = \arg \min_{g \in \mathcal{G}_n} L_n(g)$
- ▶ approximation part:
 - ▶ functional analysis
 - ▶ make sure that $\lim_{n \rightarrow \infty} L_{\mathcal{G}_n}^* = L^*$
 - ▶ related to density arguments (universal approximation)
- ▶ this is the simplest approach but also the less realistic: \mathcal{G}_n depends only on the data size

Structural risk minimization

- ▶ key idea: use again more and more powerful classes \mathcal{G}_d but also compare models between classes
- ▶ more precisely:
 - ▶ compute $g_{n,d}^*$ by empirical risk minimization in \mathcal{G}_d
 - ▶ choose g_n^* by minimizing over d

$$L_n(g_{n,d}^*) + r(n, d)$$

- ▶ $r(n, d)$ is a penalty term:
 - ▶ it decreases with n and increases with the “complexity” of \mathcal{G}_d to favor models for which the risk is correctly estimated by the empirical risk
 - ▶ it corresponds to a complexity measure for \mathcal{G}_d
- ▶ more interesting than the increasing complexity solution, but rather unrealistic on a practical point of view because of the tremendous computational cost

Regularization

- ▶ key idea: use a large class but optimize a penalized version of the empirical risk (see part IV)
- ▶ more precisely:
 - ▶ choose \mathcal{G} such that $L_{\mathcal{G}}^* = L^*$
 - ▶ define g_n^* by

$$g_n^* = \arg \min_{g \in \mathcal{G}} (L_n(g) + \lambda_n \mathbf{R}(g))$$

- ▶ $\mathbf{R}(g)$ measures the complexity of the model
- ▶ the trade-off between the risk and the complexity, λ_n , must be carefully chosen
- ▶ this is the (one of the) best solution(s)

Summary

- ▶ from n i.i.d. observations $D_n = (X_i, Y_i)_{i=1}^n$, a ML algorithm builds a model g_n
- ▶ a good ML algorithm builds a universally strongly consistent model, i.e. for all P ,

$$L(g_n) \xrightarrow{p.s.} L^*$$

- ▶ a very general class of ML algorithms is based on empirical risk minimization, i.e.

$$g_n^* = \arg \min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n c(g(X_i), Y_i)$$

- ▶ consistency is obtained by controlling the estimation error and the approximation error

Outline

- ▶ part II: empirical risk and capacity measures
- ▶ part III: capacity control
- ▶ part IV: empirical loss and regularization
- ▶ not in this lecture (see the references):
 - ▶ *ad hoc* methods such as k nearest neighbours and trees
 - ▶ advanced topics such as noise conditions and data dependent bounds
 - ▶ clustering
 - ▶ Bayesian point of view
 - ▶ and many other things...

Part II

Empirical risk and capacity measures

Outline

Concentration

Hoeffding inequality
Uniform bounds

Vapnik-Chervonenkis Dimension

Definition
Application to classification
Proof

Covering numbers

Definition and results
Computing covering numbers

Summary

Empirical risk

- ▶ Empirical risk minimisation (ERM) relies on the link between $L_n(g)$ and $L(g)$
- ▶ the law of large numbers is too limited:
 - ▶ asymptotic only
 - ▶ g must be fixed
- ▶ we need:
 1. quantitative finite distance results, i.e., PAO like bounds on

$$\mathbb{P}\{|L_n(g) - L(g)| > \epsilon\}$$

2. uniform results, i.e. bounds on

$$\mathbb{P}\left\{\sup_{g \in \mathcal{G}} |L_n(g) - L(g)| > \epsilon\right\}$$

Concentration inequalities

- ▶ in essence, the goal is to evaluate the **concentration** of the empirical mean of $c(g(X), Y)$ around its expectation

$$L_n(g) - L(g) = \frac{1}{n} \sum_{i=1}^n c(g(X_i), Y_i) - \mathbb{E}\{c(g(X), Y)\}$$

- ▶ some standard bounds:
 - ▶ Markov: $\mathbb{P}\{|X| \geq t\} \leq \frac{\mathbb{E}\{|X| \}}{t}$
 - ▶ Bienaymé-Chebyshev: $\mathbb{P}\{|X - \mathbb{E}\{X\}| \geq t\} \leq \frac{\text{Var}(X)}{t^2}$
 - ▶ BC gives for n independent real valued random variables X_i , denoting $S_n = \sum_{i=1}^n X_i$:

$$\mathbb{P}\{|S_n - \mathbb{E}\{S_n\}| \geq t\} \leq \frac{\sum_{i=1}^n \text{Var}(X_i)}{t^2}$$

Hoeffding's inequality

Hoeffding, 1963

hypothesis

- ▶ X_1, \dots, X_n , n independent r.v.
- ▶ $X_i \in [a_i, b_i]$
- ▶ $S_n = \sum_{i=1}^n X_i$

result

$$\mathbb{P}\{S_n - \mathbb{E}\{S_n\} \geq \epsilon\} \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$$
$$\mathbb{P}\{S_n - \mathbb{E}\{S_n\} \leq -\epsilon\} \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$$

Quantitative distribution free finite distance law of large numbers!

Direct application

- ▶ with a bounded cost $c(u, v) \in [a, b]$ (for all (u, v)), such as
 - ▶ a bounded Y in regression
 - ▶ or the misclassification cost in binary classification $[a, b] = [0, 1]$
- ▶ $U_i = \frac{1}{n} c(g(X_i), Y_i)$, $U_i \in [\frac{a}{n}, \frac{b}{n}]$
- ▶ **Warning:** g cannot depend on the (X_i, Y_i) (or the U_i are no longer independent)
- ▶ $\sum_{i=1}^n (b_i - a_i)^2 = \frac{(b-a)^2}{n}$
- ▶ and therefore

$$\mathbb{P}\{|L_n(g) - L(g)| \geq \epsilon\} \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

- ▶ then $L_n(g) \xrightarrow{P} L(g)$ and $L_n(g) \xrightarrow{p.s.} L(g)$ (Borel Cantelli)

Limitations

- ▶ g cannot depend on the (X_i, Y_i) (independent test set)
- ▶ intuitively:
 - ▶ $\delta = 2e^{-2n\epsilon^2/(b-a)^2}$
 - ▶ for each g , the probability to draw from P a D_n on which

$$|L_n(g) - L(g)| \leq (b-a) \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

is at least $1 - \delta$

- ▶ **but** the sets of “correct” D_n differ for each g
- ▶ conversely for a fixed D_n , the bound is valid only for some of the models
- ▶ this cannot be used to study $g_n^* = \arg \min_{g \in \mathcal{G}} L_n(g)$

The test set

- ▶ Hoeffding's inequality justifies the test set approach (*hold out estimate*)
- ▶ Given a dataset D_m :
 - ▶ split the dataset into two disjoint sets: a learning set D_n and a test set D'_p (with $p + n = m$)
 - ▶ use a ML algorithm to build g_n using only D_n
 - ▶ claim that $L(g_n) \simeq L'_p(g_n) = \frac{1}{p} \sum_{i=n+1}^m c(g_n(X_i), Y_i)$
- ▶ In fact, with probability at least $1 - \delta$

$$L(g_n) \leq L'_p(g_n) + (b-a) \sqrt{\frac{\log \frac{1}{\delta}}{2p}}$$

- ▶ Example:
 - ▶ classification ($a = 0$, $b = 1$), with a “good” classifier $L'_p(g_n) = 0.05$
 - ▶ to be 99% sure that $L(g_n) \leq 0.06$, we need $p = 23000$ (!!!)

Proof of the Hoeffding's inequality

- ▶ Chernoff's bounding technique (an application of Markov's inequality):

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{sX} \geq e^{st}\} \leq \frac{\mathbb{E}\{e^{sX}\}}{e^{st}}$$

the bound is controlled via s

- ▶ lemma: if $\mathbb{E}\{X\} = 0$ and $X \in [a, b]$, then for all s ,

$$\mathbb{E}\{e^{sX}\} \leq e^{\frac{s^2(b-a)^2}{8}}$$

- ▶ e is convex $\Rightarrow \mathbb{E}\{e^{sX}\} \leq e^{\phi(s(b-a))}$
 - ▶ then we bound ϕ
- ▶ this is used for $X = S_n - \mathbb{E}\{S_n\}$

Proof of the Hoeffding's inequality

$$\begin{aligned} \mathbb{P}\{S_n - \mathbb{E}\{S_n\} \geq \epsilon\} &\leq e^{-s\epsilon} \mathbb{E}\left\{e^{s \sum_{i=1}^n (X_i - \mathbb{E}\{X_i\})}\right\} \\ &= e^{-s\epsilon} \prod_{i=1}^n \mathbb{E}\left\{e^{s(X_i - \mathbb{E}\{X_i\})}\right\} \quad (\text{independence}) \\ &\leq e^{-s\epsilon} \prod_{i=1}^n e^{\frac{s^2(b_i - a_i)^2}{8}} \quad (\text{lemma}) \\ &= e^{-s\epsilon} e^{\frac{s^2 \sum_{i=1}^n (b_i - a_i)^2}{8}} \\ &= e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2} \quad (\text{minimization}) \end{aligned}$$

with $s = 4\epsilon / \sum_{i=1}^n (b_i - a_i)^2$

Remarks

- ▶ Hoeffding's inequality is useful in many other contexts:
 - ▶ Large-scale machine learning (e.g., [Domingos and Hulten, 2001]):
 - ▶ enormous dataset (e.g., cannot be loaded in memory)
 - ▶ process growing subsets until the model quality is known with sufficient confidence
 - ▶ monitor drifting via confidence bounds
 - ▶ Regret in game theory (e.g., [Cesa-Bianchi and Lugosi, 2006]):
 - ▶ define strategies by minimizing a regret measure
 - ▶ get bounds on the regret estimations (e.g., trade-off between exploration and exploitation in multi-arms bandits)
- ▶ Many improvements/complements are available, such as:
 - ▶ Bernstein's inequality
 - ▶ McDiarmid's bounded difference inequality

Uniform bounds

- ▶ we deal with the *estimation error*:
 - ▶ given the ERM choice: $g_n^* = \arg \min_{g \in \mathcal{G}} L_n(g)$
 - ▶ what about $L(g_n^*) - \inf_{g \in \mathcal{G}} L(g)$?

- ▶ we need uniform bounds:

$$\begin{aligned} |L_n(g_n^*) - L(g_n^*)| &\leq \sup_{g \in \mathcal{G}} |L_n(g) - L(g)| \\ L(g_n^*) - \inf_{g \in \mathcal{G}} L(g) &= L(g_n^*) - L_n(g_n^*) + L_n(g_n^*) - \inf_{g \in \mathcal{G}} L(g) \\ &\leq L(g_n^*) - L_n(g_n^*) + \sup_{g \in \mathcal{G}} |L_n(g) - L(g)| \\ &\leq 2 \sup_{g \in \mathcal{G}} |L_n(g) - L(g)| \end{aligned}$$

- ▶ therefore uniform bounds give an answer to both questions: the quality of the empirical risk as an estimate of the risk and the quality of the model select by ERM

Finite model class

- ▶ *union bound* : $\mathbb{P}\{A \text{ ou } B\} \leq \mathbb{P}\{A\} + \mathbb{P}\{B\}$ and therefore

$$\mathbb{P}\left\{\max_{1 \leq i \leq m} U_i \geq \epsilon\right\} \leq \sum_{i=1}^m \mathbb{P}\{U_i \geq \epsilon\}$$

- ▶ then when \mathcal{G} is finite

$$\mathbb{P}\left\{\sup_{g \in \mathcal{G}} |L_n(g) - L(g)| \geq \epsilon\right\} \leq 2|\mathcal{G}|e^{-2n\epsilon^2/(b-a)^2}$$

so $L(g_n^*) \xrightarrow{p.s.} \inf_{g \in \mathcal{G}} L(g)$ (but in general $\inf_{g \in \mathcal{G}} L(g) > L^*$)

- ▶ quantitative distribution free uniform finite distance law of large numbers!

Bound versus size

with probability at least $1 - \delta$

$$L(g_n^*) \leq \inf_{g \in \mathcal{G}} L(g) + 2(b-a)\sqrt{\frac{\log |\mathcal{G}| + \log \frac{2}{\delta}}{2n}}$$

- ▶ $\inf_{g \in \mathcal{G}} L(g)$ decreases with $|\mathcal{G}|$ (more models)
- ▶ but the bound increases with $|\mathcal{G}|$: trade-off between flexibility and estimation quality
- ▶ remark: $\log |\mathcal{G}|$ is the number of bits needed to encode the model choice; this is a **capacity measure**

Outline

Concentration

Hoeffding inequality
Uniform bounds

Vapnik-Chervonenkis Dimension

Definition
Application to classification
Proof

Covering numbers

Definition and results
Computing covering numbers

Summary

Infinite model class

- ▶ so far we have uniform bounds when \mathcal{G} is finite:
 - ▶ even simple classes such as $\mathcal{G}_{\text{lin}} = \{\mathbf{x} \mapsto \arg \max_k (W\mathbf{x})_k\}$ are infinite
 - ▶ in addition $\inf_{g \in \mathcal{G}_{\text{lin}}} L(g) > 0$ in general, even when $L^* = 0$ (nonlinear problems)
- ▶ solution: evaluate the **capacity** of a class rather than its size
- ▶ first step:
 - ▶ **binary classification** (the simplest case)
 - ▶ a model is then a measurable set $A \subset \mathcal{X}$
 - ▶ the capacity of \mathcal{G} measures the variability of the available shapes
 - ▶ it is evaluated with respect to the data: if $A \cap D_n = B \cap D_n$, then A and B are identical

Growth function

- ▶ abstract setting: \mathcal{F} is a set of measurable functions from \mathbb{R}^d to $\{0, 1\}$ (in other words a set of (indicator functions of) measurable subsets of \mathbb{R}^d)
- ▶ given $\mathbf{z}_1 \in \mathbb{R}^d, \dots, \mathbf{z}_n \in \mathbb{R}^d$, we define

$$\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n} = \{u \in \{0, 1\}^n \mid \exists f \in \mathcal{F}, u = (f(\mathbf{z}_1), \dots, f(\mathbf{z}_n))\}$$

- ▶ interpretation: each u describes a binary partition of $\mathbf{z}_1, \dots, \mathbf{z}_n$ and $\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}$ is then the set of partitions implemented by \mathcal{F}
- ▶ Growth function

$$\mathcal{S}_{\mathcal{F}}(n) = \sup_{\mathbf{z}_1 \in \mathbb{R}^d, \dots, \mathbf{z}_n \in \mathbb{R}^d} |\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}|$$

- ▶ interpretation: maximal number of binary partitions implementable via \mathcal{F} (distribution free \Rightarrow worst case analysis)

Vapnik-Chervonenkis dimension

- ▶ $\mathcal{S}_{\mathcal{F}}(n) \leq 2^n$
- ▶ vocabulary:
 - ▶ $\mathcal{S}_{\mathcal{F}}(n)$ is the n -th shatter coefficient of \mathcal{F}
 - ▶ if $|\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}| = 2^n$, \mathcal{F} shatters $\mathbf{z}_1, \dots, \mathbf{z}_n$

Vapnik-Chervonenkis dimension

$$VCdim(\mathcal{F}) = \sup\{n \in \mathbb{N}^+ \mid \mathcal{S}_{\mathcal{F}}(n) = 2^n\}$$

- ▶ interpretation:
 - ▶ if $\mathcal{S}_{\mathcal{F}}(n) < 2^n$:
 - ▶ for all $\mathbf{z}_1, \dots, \mathbf{z}_n$, there is a partition $u \in \{0, 1\}^n$, such that $u \notin \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}$
 - ▶ for any dataset, \mathcal{F} can fail to reach $L^* = 0$
 - ▶ if $\mathcal{S}_{\mathcal{F}}(n) = 2^n$: there is at least one set $\mathbf{z}_1, \dots, \mathbf{z}_n$ shattered by \mathcal{F}

Example

- ▶ points from \mathbb{R}^2
- ▶ $\mathcal{F} = \{\mathbb{I}_{\{ax+by+c \geq 0\}}\}$

×

×

⊗

- ▶ $\mathcal{S}_{\mathcal{F}}(3) = 2^3$
- ▶ even if we have sometimes $|\mathcal{F}_{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3}| < 2^3$
- ▶ $\mathcal{S}_{\mathcal{F}}(4) < 2^4$
- ▶ $VCdim(\mathcal{F}) = 3$

Linear models

result

if \mathcal{G} is a p dimensional vector space of real valued functions defined on \mathbb{R}^d then

$$VCdim\left(\left\{f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(z) = \mathbb{I}_{\{g(z) \geq 0\}}\right\}\right) \leq p$$

proof

- ▶ given z_1, \dots, z_{p+1} , let F from \mathcal{G} to \mathbb{R}^{p+1} be $F(g) = (g(z_1), \dots, g(z_{p+1}))$
- ▶ as $\dim(F(\mathcal{G})) \leq p$, there is a non zero $\gamma = (\gamma_1, \dots, \gamma_{p+1})$ such that $\sum_{i=1}^{p+1} \gamma_i g(z_i) = 0$
- ▶ if z_1, \dots, z_{p+1} is shattered, there is also g_j such that $g_j(z_i) = \delta_{ij}$ and therefore $\gamma_j = 0$ for all j
- ▶ consequently no z_1, \dots, z_{p+1} can be shattered

VC dimension \neq parameter number

- ▶ in the linear case $\mathcal{F} = \{f(z) = \mathbb{I}_{\{\sum_{i=1}^p w_i \phi_i(z) \geq 0\}}\}$

$VCdim(\mathcal{F}) = p$

- ▶ but in general:

- ▶ $VCdim(\{f(z) = \mathbb{I}_{\{\sin(tz) \geq 0\}}\}) = \infty$
- ▶ one hidden layer perceptron:

$$\mathcal{G} = \left\{ g(z) = T \left(\beta_0 + \sum_{k=1}^h \beta_k T \left(\alpha_{k0} + \sum_{j=1}^d \alpha_{kj} z_j \right) \right) \right\}$$

- ▶ $T(a) = \mathbb{I}_{\{a \geq 0\}}$
- ▶ $VCdim(\mathcal{G}) \geq W \log_2(h/4)/32$ where $W = dh + 2h + 1$ is the number of parameters

Snake oil warning

- ▶ you might read here and there things like

the VC-dimension of the class of separating hyperplanes with margin larger than $\frac{1}{\delta}$ is bounded by bla bla bla

- ▶ **this is plain wrong!**

- ▶ the class \mathcal{G} is **data independent**

- ▶ shattering does not take into a margin

- ▶ asking for the “VC-dimension of the class of separating hyperplanes with margin larger than $\frac{1}{\delta}$ ” is a **nonsense**

- ▶ there is an **extension** of the VC dimension called the **fat shattering dimension** for which this question as a sense, **but one cannot apply the VC bounds directly to it!**

Uniform bound

Vapnik and Chervonenkis (1971)

hypothesis

- ▶ n i.i.d. random variables Z_1, \dots, Z_n with values in \mathbb{R}^d
- ▶ \mathcal{F} is a set of measurable functions from \mathbb{R}^d to $\{0, 1\}$
- ▶ $Pf = \mathbb{E}\{f(Z_1)\}$ and $P_n f = \frac{1}{n} \sum_{i=1}^n f(Z_i)$

result

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 8S_{\mathcal{F}}(n) e^{-n\epsilon^2/8}$$

if $S_{\mathcal{F}}(n)$ grows polynomially with n , then $P_n f$ converges to Pf **uniformly** on \mathcal{F} (Borel Cantelli)

Behavior of $S_{\mathcal{F}}(n)$

Sauer's Lemma 1972 (also Vapnik and Chervonenkis, 1971)

result

If $VCdim(\mathcal{F}) < \infty$ then for all n

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^{VCdim(\mathcal{F})} \binom{n}{i}$$

bounds

$$S_{\mathcal{F}}(n) \leq n^{VCdim(\mathcal{F})} + 1$$

$$S_{\mathcal{F}}(n) \leq \left(\frac{en}{VCdim(\mathcal{F})} \right)^{VCdim(\mathcal{F})} \text{ for } n \geq VCdim(\mathcal{F})$$

- ▶ the simplest machine learning problem:
 - ▶ binary classification ($\mathcal{Y} = \{-1, 1\}$ and $\mathcal{X} = \mathbb{R}^d$)
 - ▶ cost function: $c(g(\mathbf{x}), \mathbf{y}) = \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}}$
 - ▶ model class \mathcal{G} (classifiers)

- ▶ loss class

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \rightarrow \{0, 1\} \mid \exists g \in \mathcal{G}, f(\mathbf{x}, \mathbf{y}) = \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}} \right\}$$

- ▶ for the pair f and g

$$Pf = L(g) \quad P_n f = L_n(g)$$

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq 8\mathcal{S}_{\mathcal{F}}(n) e^{-n\epsilon^2/8}$$

Warning: \mathcal{G} is fixed!

- ▶ as g takes values in $\{-1, 1\}$, we can define $\mathcal{S}_{\mathcal{G}}(n)$ and $VCdim(\mathcal{G})$
- ▶ then $VCdim(\mathcal{G}) = VCdim(\mathcal{F})$
 - ▶ if $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1), \dots, \mathbf{z}_n = (\mathbf{x}_n, \mathbf{y}_n)$ is shattered by \mathcal{F} , then $\mathbf{x}_1, \dots, \mathbf{x}_n$ is shattered by \mathcal{G} :
 - ▶ given $v \in \{-1, 1\}^n$, we define $u_i = (v_i/\mathbf{y}_i + 1)/2 \in \{0, 1\}$
 - ▶ $\exists f \in \mathcal{F}, f(\mathbf{z}_i) = u_i$
 - ▶ $f(\mathbf{z}_i) = u_i \Leftrightarrow g(\mathbf{x}_i) = (2u_i - 1)\mathbf{y}_i$ and therefore $g(\mathbf{x}_i) = v_i$
 - ▶ and then $v \in \mathcal{G}_{\mathbf{x}_1, \dots, \mathbf{x}_n}$
 - ▶ conversely if $\mathbf{x}_1, \dots, \mathbf{x}_n$ is shattered by \mathcal{G} , then there exists $\mathbf{z}_1, \dots, \mathbf{z}_n$ shattered by \mathcal{F}
 - ▶ therefore $\mathcal{S}_{\mathcal{F}}(n) = 2^n \Leftrightarrow \mathcal{S}_{\mathcal{G}}(n) = 2^n$ and $\mathcal{S}_{\mathcal{F}}(n) < 2^n \Leftrightarrow \mathcal{S}_{\mathcal{G}}(n) < 2^n$
 - ▶ leading to $VCdim(\mathcal{G}) = VCdim(\mathcal{F})$
- ▶ stronger result: $\mathcal{S}_{\mathcal{F}}(n) = \mathcal{S}_{\mathcal{G}}(n)$

Risk bounds

- ▶ with probability at least $1 - \delta$

$$\begin{aligned} |L(g) - L_n(g)| &\leq 2\sqrt{2 \frac{\log \mathcal{S}_{\mathcal{F}}(n) + \log \frac{8}{\delta}}{n}} \\ &\leq 2\sqrt{2 \frac{VCdim(\mathcal{G}) \log n + \log \frac{8}{\delta}}{n}} \end{aligned}$$

when $3 \leq VCdim(\mathcal{G}) \leq n < \infty$

- ▶ in addition

$$L(g_n^*) \leq \inf_{g \in \mathcal{G}} L(g) + 4\sqrt{2 \frac{VCdim(\mathcal{G}) \log n + \log \frac{8}{\delta}}{n}}$$

Empirical risk minimization is therefore meaningful

Discussion

- ▶ VC dimension provides a capacity measure for classes of classifiers
- ▶ geometrical and combinatorial bound: no statistics...
- ▶ ...and therefore distribution free result!
- ▶ quite difficult to compute in general
- ▶ worst case analysis
- ▶ finite VC dim \Leftrightarrow finite sampled inference is possible:
 - ▶ only for the empirical risk minimization principle
 - ▶ the VC theorem gives finite VC dim \Rightarrow consistency
 - ▶ we'll see later than consistent \Rightarrow finite VC dim

Risk bounds

- ▶ with probability at least $1 - \delta$

$$\begin{aligned} |L(g) - L_n(g)| &\leq 2\sqrt{2 \frac{\log \mathcal{S}_{\mathcal{F}}(n) + \log \frac{8}{\delta}}{n}} \\ &\leq 2\sqrt{2 \frac{VCdim(\mathcal{G}) \log n + \log \frac{8}{\delta}}{n}} \end{aligned}$$

when $3 \leq VCdim(\mathcal{G}) \leq n < \infty$

- ▶ in addition

$$L(g_n^*) \leq \inf_{g \in \mathcal{G}} L(g) + 4\sqrt{2 \frac{VCdim(\mathcal{G}) \log n + \log \frac{8}{\delta}}{n}}$$

Empirical risk minimization is therefore meaningful

Discussion

- ▶ VC dimension provides a capacity measure for classes of classifiers
- ▶ geometrical and combinatorial bound: no statistics...
- ▶ ...and therefore distribution free result!
- ▶ quite difficult to compute in general
- ▶ worst case analysis
- ▶ finite VC dim \Leftrightarrow finite sampled inference is possible:
 - ▶ only for the empirical risk minimization principle
 - ▶ the VC theorem gives finite VC dim \Rightarrow consistency
 - ▶ we'll see later than consistent \Rightarrow finite VC dim

- ▶ compare with the finite class bound:

$$L(g_n^*) \leq \inf_{g \in \mathcal{G}} L(g) + 4 \sqrt{2 \frac{VCdim(\mathcal{G}) \log n + \log \frac{8}{\delta}}{n}}$$

$$L(g_n^*) \leq \inf_{g \in \mathcal{G}} L(g) + 2 \sqrt{\frac{\log |\mathcal{G}| + \log \frac{2}{\delta}}{2n}}$$

- ▶ the VC-dim of a finite set is bounded by $\log_2 |\mathcal{G}|$
- ▶ additional $\log n$ term: we can get rid of it (with a lot of efforts!)

- ▶ a consequence of VC theorem is that there is a **universal constant** c (independent of P) such that

$$\mathbb{E} \{L(g_n^*)\} - \inf_{g \in \mathcal{G}} L(g) \leq c \sqrt{\frac{VCdim(\mathcal{G}) \log n}{n}}$$

- ▶ refined analysis (e.g., *chaining*) leads to better bounds:

$$\mathbb{E} \{L(g_n^*)\} - \inf_{g \in \mathcal{G}} L(g) \leq c' \sqrt{\frac{VCdim(\mathcal{G})}{n}}$$

using bounds of the form

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq \frac{\gamma}{\epsilon \sqrt{n}} \left(\frac{\gamma n \epsilon^2}{VCdim(\mathcal{F})} \right)^{VCdim(\mathcal{F})} e^{-2n\epsilon^2}$$

Proof of the VC result (simplified case)

Fundamental symmetrization lemma

hypothesis

- ▶ n i.i.d. random variables Z_1, \dots, Z_n with values in \mathbb{R}^d
- ▶ an independent **ghost sample** Z'_1, \dots, Z'_n
- ▶ $P'_n f = \frac{1}{n} \sum_{i=1}^n f(Z'_i)$
- ▶ $n\epsilon^2 \geq 2$

result

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| \geq \epsilon \right\} \leq 2 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| \geq \frac{\epsilon}{2} \right\}$$

Proof of the lemma

- ▶ let f^* be the (random) function that maximizes $|Pf^* - P_n f^*|$

$$\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \mathbb{I}_{\{|Pf^* - P'_n f^*| < \epsilon/2\}} = \mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon \wedge |Pf^* - P'_n f^*| < \epsilon/2\}} \\ \leq \mathbb{I}_{\{|P'_n f^* - P_n f^*| > \epsilon/2\}}$$

- ▶ compute the expectation with respect to the ghost sample

$$\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \mathbb{P} \{ |Pf^* - P'_n f^*| < \epsilon/2 \mid Z_1, \dots, Z_n \} \\ \leq \mathbb{P} \{ |P'_n f^* - P_n f^*| > \epsilon/2 \mid Z_1, \dots, Z_n \}$$

- ▶ Bienaymé-Chebyshev:

$$\mathbb{P} \{ |Pf^* - P'_n f^*| \geq \epsilon/2 \} \leq \frac{4 \text{Var} f^*(Z_1)}{n\epsilon^2} \leq \frac{1}{n\epsilon^2}$$

$$\mathbb{I}_{\{|Pf^* - P_n f^*| > \epsilon\}} \left(1 - \frac{1}{n\epsilon^2} \right) \leq \mathbb{P} \{ |P'_n f^* - P_n f^*| > \epsilon/2 \mid Z_1, \dots, Z_n \}$$

Proof of the lemma

- ▶ use $1 - \frac{1}{n\epsilon^2} \leq \frac{1}{2}$ and take the expectation with respect to the data

$$\mathbb{P} \{ |Pf^* - P_nf^*| > \epsilon \} \leq 2\mathbb{P} \{ |P'_nf^* - P_nf^*| > \epsilon/2 \}$$

- ▶ bound the right term by $2\mathbb{P} \{ \sup_{f \in \mathcal{F}} |P'_nf - P_nf| \geq \frac{\epsilon}{2} \}$ and remember that f^* maximizes $|Pf^* - P_nf^*|$
- ▶ the practical interest is to replace a maximum over an arbitrary set \mathcal{G} by a maximum over a finite set of values for $P'_nf^* - P_nf^*$ (at most 2^{2n} values)
- ▶ these values are obtained via $\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n, \mathbf{z}'_1, \dots, \mathbf{z}'_n}$ whose cardinal is given by the growth function

Back to the proof of the VC theorem

- ▶ for a fixed f , Hoeffding's inequality applied to the $U_i = \frac{1}{n}(f(Z_i) - f(Z'_i))$ gives

$$\mathbb{P} \{ |P'_nf - P_nf| > \epsilon \} \leq 2e^{-n\epsilon^2/2}$$

- ▶ and therefore

$$\begin{aligned} \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_nf| > \epsilon \right\} &\leq 2\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_nf - P_nf| \geq \frac{\epsilon}{2} \right\} \\ &= 2\mathbb{P} \left\{ \sup_{f \in \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n, \mathbf{z}'_1, \dots, \mathbf{z}'_n}} |P'_nf - P_nf| \geq \frac{\epsilon}{2} \right\} \\ &\leq 2S_{\mathcal{F}}(2n) \mathbb{P} \left\{ |P'_nf - P_nf| \geq \frac{\epsilon}{2} \right\} \\ &\leq 4S_{\mathcal{F}}(2n) e^{-n\epsilon^2/8} \end{aligned}$$

Summary

- ▶ the empirical risk minimization principle is based on the inequality

$$L(g_n^*) - \inf_{g \in \mathcal{G}} L(g) \leq 2 \sup_{g \in \mathcal{G}} |L_n(g) - L(g)|$$

- ▶ uniform distribution free bounds are needed to obtain consistency
- ▶ in the binary classification case:
 - ▶ we have

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq 8S_{\mathcal{G}}(n) e^{-n\epsilon^2/8}$$

- ▶ $S_{\mathcal{G}}$ measures the geometrical and combinatorial complexity of \mathcal{G}
- ▶ consistency is equivalent to a finite VC-dimension, a quantity that characterizes the behavior of $S_{\mathcal{G}}$

Outline

Concentration

Hoeffding inequality
Uniform bounds

Vapnik-Chervonenkis Dimension

Definition
Application to classification
Proof

Covering numbers

Definition and results
Computing covering numbers

Summary

What about regression?

- ▶ capacity measure:
 - ▶ naive idea: count the distinct values of $(f(\mathbf{z}_1), \dots, f(\mathbf{z}_n))$ when $f \in \mathcal{F}$
 - ▶ but in regression $f(\mathbf{x}) \in [0, B]$ (rather than $f(\mathbf{x}) \in \{0, 1\}$)
 - ▶ therefore $|\mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}| = \infty$ (in general)
- ▶ however if $f(\mathbf{x}) \simeq h(\mathbf{x})$, I should count only one model
- ▶ **Covering numbers:**
 - ▶ on \mathbb{R}^d , define $d(u, v) = \frac{1}{d} \sum_{i=1}^d |u_i - v_i|$
 - ▶ $A \subset \mathbb{R}^d$: an ϵ -covering of A is a finite set z_1, \dots, z_q such that $A \subset \bigcup_{i=1}^q B(z_i, \epsilon)$ with $B(u, \epsilon) = \{v \in \mathbb{R}^d \mid d(u, v) \leq \epsilon\}$
 - ▶ Covering numbers $N(\epsilon, A)$: size of the smallest ϵ -covering of A
 - ▶ Remark: covering numbers depend on the metric in the chosen space, we will see other metrics in Part IV

Uniform convergence

[Pollard, 1984]

hypothesis

- ▶ Z_1, \dots, Z_n n independent random variables
- ▶ \mathcal{F} set of functions with values in $[0, B]$
- ▶ $Pf = \mathbb{E}\{f(Z_1)\}$ and $P_n f = \frac{1}{n} \sum_{i=1}^n f(Z_i)$

result

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 8 \mathbb{E} \{ N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) \} e^{-n\epsilon^2/(128B^2)}$$

with

$$\mathcal{F}_{Z_1, \dots, Z_n} = \{u \in [0, B]^n \mid \exists f \in \mathcal{F}, u = (f(Z_1), \dots, f(Z_n))\}$$

Comments

- ▶ this is again a uniform quantitative finite distance result
- ▶ $N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n})$ replaces the shatter coefficient
- ▶ the covering number corresponds to the intuition presented before:
 - ▶ if f_1 and f_2 differ not too much, then they count only for one
 - ▶ in fact $d(u, v) = \frac{1}{d} \sum_{i=1}^d |u_i - v_i|$ exactly says that f_1 and f_2 are “equal” in the definition of $N(\epsilon, \mathcal{F}_{Z_1, \dots, Z_n})$ if

$$\frac{1}{n} \sum_{i=1}^n |f_1(Z_i) - f_2(Z_i)| < \epsilon$$

- ▶ a major difference with the shatter coefficients:
 - ▶ this is **not** distribution free
 - ▶ we take the expectation over P
 - ▶ more on this latter :-)

Proof

is done by via a symmetrization lemma:

hypothesis

- ▶ Z_1, \dots, Z_n n independent random variables with values in \mathbb{R}^d
- ▶ \mathcal{F} set of functions with values in $[0, B]$
- ▶ n i.i.d. *Rademacher* random variables $\sigma_1, \dots, \sigma_n$ with values in $\{-1, 1\}$, with $\mathbb{P}\{\sigma_i = 1\} = 1/2$
- ▶ $n\epsilon^2 \geq 2B^2$

result

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 4 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \right\}$$

- ▶ first step: standard symmetrization (by a ghost sample)

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| > \epsilon \right\} \leq 2 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| > \epsilon/2 \right\}$$

- ▶ the condition $n\epsilon^2 \geq 2B^2$ corresponds to $\text{Var}(f(Z_i)) \leq B^2/4$
- ▶ second step: use the *Rademacher* random variables

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |P'_n f - P_n f| > \frac{\epsilon}{2} \right\} = \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - f(Z'_i)) \right| > \frac{\epsilon}{2} \right\}$$

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - f(Z'_i)) \right| \leq \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| + \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z'_i) \right|$$

- ▶ therefore, by union bound

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - f(Z'_i)) \right| > \frac{\epsilon}{2} \right\}$$

$$\leq \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \right\} + \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z'_i) \right| > \frac{\epsilon}{4} \right\}$$

- ▶ and that's all
- ▶ Remark: using the *Rademacher* variables is simply a way to get rid of the ghost sample; we could proceed without this trick using the standard VC symmetrisation lemma

Proof of the main result

- ▶ let g_1, \dots, g_M be a minimal $\epsilon/8$ -covering of $\mathcal{F}_{Z_1, \dots, Z_n}$ ($M = N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n})$)
- ▶ for $f \in \mathcal{F}$, there is $g^* \in \{g_1, \dots, g_M\}$ such that

$$\frac{1}{n} \sum_{i=1}^n |f(Z_i) - g^*(Z_i)| \leq \frac{\epsilon}{8}$$

- ▶ therefore

$$\frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| \leq \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g^*(Z_i) \right| + \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (f(Z_i) - g^*(Z_i)) \right|$$

$$\leq \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g^*(Z_i) \right| + \frac{\epsilon}{8}$$

Proof of the main result

- ▶ and therefore

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\}$$

$$\leq \mathbb{P} \left\{ \max_{1 \leq j \leq M} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g_j(Z_i) \right| > \frac{\epsilon}{8} \mid Z_1, \dots, Z_n \right\}$$

- ▶ we apply Hoeffding's inequality to the $U_i = \frac{1}{n} \sigma_i g_j(Z_i)$ (for fixed values of the Z_1, \dots, Z_n) which have zero expectation:

$$\mathbb{P} \left\{ \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g_j(Z_i) \right| > \frac{\epsilon}{8} \mid Z_1, \dots, Z_n \right\} \leq 2e^{-n\epsilon^2/(128B^2)}$$

- ▶ then

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \leq N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) 2e^{-n\epsilon^2/(128B^2)}$$

Application to regression

- ▶ a quite generic setting:
 - ▶ $\mathcal{Y} = \mathbb{R}^p$
 - ▶ bounded cost function $c(B)$: for instance the quadratic cost associated to a class of bounded models and for a bounded \mathcal{Y}
 - ▶ model set \mathcal{G}

- ▶ loss class

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \rightarrow [0, B] \mid \exists g \in \mathcal{G}, f(\mathbf{x}, \mathbf{y}) = c(g(\mathbf{x}), \mathbf{y}) \right\}$$

- ▶ for the pair f, g

$$Pf = L(g)$$

$$P_n f = L_n(g)$$

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq 8 \mathbb{E} \{ N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) \} e^{-n\epsilon^2/(128B^2)}$$



In practice?

- ▶ as explained before, the result seems limited:
 - ▶ $\mathbb{E} \{ N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) \}$ is distribution dependent
 - ▶ seems quite difficult to compute

- ▶ in fact, VC results can be extended to show

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq 8 \mathbb{E} \{ |\mathcal{F}_{Z_1, \dots, Z_n}| \} e^{-n\epsilon^2/8}$$

- ▶ could we bound $\mathbb{E} \{ N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) \}$ using only geometric and combinatorial properties of \mathcal{F} ?



Pseudo dimension

- ▶ associate to $\mathcal{F} = \{ f : \mathbb{R}^d \rightarrow [0, B] \}$

$$\mathcal{F}^+ = \{ f^+ : \mathbb{R}^d \times [0, B] \rightarrow \{0, 1\} \mid \exists f \in \mathcal{F}, f^+(x, t) = \mathbb{I}_{\{t \leq f(x)\}} \}$$
- ▶ then [Pollard, 1984]

$$N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq \left(\frac{4eB}{\epsilon} \log \frac{2eB}{\epsilon} \right)^{VCdim(\mathcal{F}^+)}$$

- ▶ **packing number**: $M(\epsilon, \mathcal{F}, \mu)$ is the size of the largest collection of functions f of \mathcal{F} such that

$$\int_{\mathbb{R}^d} |f_i(x) - f_j(x)| \mu(dx) \leq \epsilon$$

$$M(2\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}, 1/n) \leq N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq M(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}, 1/n)$$



Fat shattering

However, $VCdim(\mathcal{F}^+) < \infty$ is not always needed to get a good behavior of $\mathbb{E} \{ N(\epsilon/8, \mathcal{F}_{Z_1, \dots, Z_n}) \}$

γ fat shattering

$\mathbf{z}_1, \dots, \mathbf{z}_n$ γ -shattered by \mathcal{F} if for all $u \in \{-1, 1\}^n$, there is $t \in [0, B]^n$ and $f \in \mathcal{F}$ such that

$$(f(\mathbf{z}_i) - t_i)u_i \geq \gamma$$

the γ fat-shattering dimension of \mathcal{F} , $\text{fat}_\gamma(\mathcal{F})$, is the size of the largest set γ -shattered by \mathcal{F}

if $d = \text{fat}_\gamma(\mathcal{F})$ then [Alon et al., 1997]

$$N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) < 2 \left(\frac{4nB^2}{\epsilon^2} \right)^{d \log_2(4eBn/(d\epsilon))}$$



In practice

- ▶ the pseudo dimension is generally sufficient for real world models
- ▶ some properties:
 - ▶ if $\mathcal{H} = \{h + f \mid f \in \mathcal{F}\}$ for a fixed h , then $VCdim(\mathcal{H}^+) = VCdim(\mathcal{F}^+)$
 - ▶ if h is a non decreasing function from $[0, B]$ to \mathbb{R} and if $\mathcal{H} = \{h \circ f \mid f \in \mathcal{F}\}$, then $VCdim(\mathcal{H}^+) \leq VCdim(\mathcal{F}^+)$
 - ▶ from k function classes $\mathcal{F}_1, \dots, \mathcal{F}_k$, we define the class $\mathcal{F} = \{f_1 + \dots + f_k \mid f_i \in \mathcal{F}_i\}$, then

$$N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq \prod_{j=1}^k N(\epsilon/k, \mathcal{F}_{j, \mathbf{z}_1, \dots, \mathbf{z}_n})$$

- ▶ from two classes \mathcal{F}_i ($i = 1, 2$) with respective bounds $[-B_i, B_i]$, we define the class $\mathcal{H} = \{f_1 f_2 \mid f_i \in \mathcal{F}_i\}$ and then

$$N(\epsilon, \mathcal{H}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq N(\epsilon/(2B_2), \mathcal{F}_{1, \mathbf{z}_1, \dots, \mathbf{z}_n}) N(\epsilon/(2B_1), \mathcal{F}_{2, \mathbf{z}_1, \dots, \mathbf{z}_n})$$



Example

- ▶ a general framework from [Lugosi and Zegler, 1995]:
 - ▶ $|Y| \leq L$
 - ▶ $c(u, v) = |u - v|^p$
 - ▶ $\mathcal{G} = \left\{ \sum_{j=1}^k w_j \phi_j; \sum_{j=1}^k |w_j| \leq \beta \right\}, \beta \geq L, |\phi_j| \leq 1$
- ▶ $\mathcal{F} = \left\{ f(x, y) = \left| \sum_{j=1}^k w_j \phi_j(x) - y \right|^p; \sum_{j=1}^k |w_j| \leq \beta \right\}$
- ▶ $f(x, y) \leq 2^p \max(\beta^p, L^p) \leq 2^p \beta^p$
- ▶ as $||a|^p - |b|^p| \leq p|a - b| \max(a, b)^{p-1}$,

$$\int |f_1(x, y) - f_2(x, y)| \nu(dx, dy) \leq p(2\beta)^{p-1} \int |g_1(x) - g_2(x)| \mu(dx)$$

$$\text{then with } \nu = 1/n, N(\epsilon, \mathcal{F}_{\mathbf{z}_1, \dots, \mathbf{z}_n}) \leq N\left(\frac{\epsilon}{p(2\beta)^{p-1}}, \mathcal{G}_{\mathbf{z}_1, \dots, \mathbf{z}_n}\right)$$



Example

- ▶ as \mathcal{G} is a subset of a k dimensional vector space, $VCdim(\mathcal{G}^+) \leq k$
- ▶ therefore

$$\begin{aligned} N\left(\frac{\epsilon}{p(2\beta)^{p-1}}, \mathcal{G}_{\mathbf{z}_1, \dots, \mathbf{z}_n}\right) &\leq 2 \left(\frac{e2^{p+1}\beta^p}{\epsilon/(p(2\beta)^{p-1})} \log \frac{e2^{p+1}\beta^p}{\epsilon/(p(2\beta)^{p-1})} \right)^k \\ &\leq 2 \left(\frac{ep2^{2p}\beta^{2p-1}}{\epsilon} \right)^k \end{aligned}$$



Summary

- ▶ we have in general

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| > \epsilon \right\} \leq C(n, \mathcal{G}, \epsilon) e^{-cn\epsilon^2}$$

- ▶ c is determined only by $\sup_{g \in \mathcal{G}} \|g\|_\infty$
- ▶ $C(n, \mathcal{G}, \epsilon)$ measures the **capacity** \mathcal{G} :
 - ▶ covering numbers or shatter coefficients
 - ▶ Vapnik-Chervonenkis dimension
 - ▶ pseudo-dimension and fat-shattering dimension
 - ▶ **distribution free** uniform bounds
- ▶ the estimation problem is solved!
- ▶ ...in the case of empirical risk minimization



Necessary conditions

▶ binary classification:

- ▶ $N(\mathcal{F}, Z_1, \dots, Z_n) = |\mathcal{F}_{Z_1, \dots, Z_n}|$
- ▶ $H_{\mathcal{F}}(n) = \log \mathbb{E} \{N(\mathcal{F}, Z_1, \dots, Z_n)\}$ is the VC-entropy
- ▶ a necessary and sufficient condition (N.S.C.) for the uniform convergence of the empirical risk to the risk is:

$$\frac{H_{\mathcal{F}}(n)}{n} \rightarrow 0$$

- ▶ a **distribution free** N.S.C. is: $VCdim(\mathcal{F}) < \infty$
- ▶ for regression:
 - ▶ similar results
 - ▶ the γ fat-shattering dimension must be finite for all γ
 - ▶ the target variable has to be bounded

Limitations

▶ \mathcal{G} must be **fixed** and must have **limited capacity**

- ▶ in general $\inf_{g \in \mathcal{G}} L(g) > L^*$
- ▶ the analysis does not apply to some techniques:
 - ▶ mainly when \mathcal{G} depends on X_1, \dots, X_n
 - ▶ in general the union class over all possible datasets has a too large capacity

▶ in addition, the risk must be bounded:

- ▶ this is a rather strong hypothesis in regression
- ▶ easy to check for \mathcal{G} ...
- ▶ ... but not for Y

▶ solutions:

- ▶ adapt the capacity to the data
- ▶ use data dependent bounds and/or clipping in regression

Still no free lunch!

▶ binary classification case

▶ when $VCdim(\mathcal{G}) = \infty$, we are in trouble:

- ▶ consider a fixed ML algorithm that picks up a classifier in \mathcal{G} with infinite VC dimension (using whatever criterion)
- ▶ for all $\epsilon > 0$ and all n , there is (X, Y) such that $L_{\mathcal{G}}^* = 0$ and

$$\mathbb{E} \{L(g_n)\} \geq \frac{1}{2e} - \epsilon$$

▶ **shattering an infinite dataset** brings even more troubles:

- ▶ a fixed ML algorithm picks up a classifier in \mathcal{G}
- ▶ there is an infinite set A such that for all $B \subset A$, there is $g \in \mathcal{G}$ such that $g(x) = 1$ on B and $g(x) = 0$ on $A \setminus B$
- ▶ if (a_n) is a series with limit zero and such that $a_1 \leq 1/16$
- ▶ then there is (X, Y) such that $L_{\mathcal{G}}^* = 0$ and for all n

$$\mathbb{E} \{L(g_n)\} \geq a_n$$

Still no free lunch!

▶ $VCdim(\mathcal{G}) < \infty$ is a real **capacity limit**:

- ▶ $VCdim(\mathcal{G}) < \infty$
- ▶ for all $\epsilon > 0$, there is (X, Y) such that

$$\inf_{g \in \mathcal{G}} L(g) - L^* > \frac{1}{2} - \epsilon$$

▶ even **increasing the capacity with n** is not a perfect solution:

- ▶ take a series of classes with increasing but finite capacities $VCdim(\mathcal{G}^{(j)}) < \infty$
- ▶ for any series (a_n) with limit zero, there is (X, Y) such that after a rank k

$$\inf_{g \in \mathcal{G}^{(k)}} L(g) - L^* > a_k$$

- ▶ the set of all classifiers (measurable functions from \mathcal{X} to $\{-1, 1\}$) **cannot** be represented **exactly** as a countable union of classes with finite VC dimension

Part III

Capacity control

Generic control

Binary classification
Regression

Data-driven control

Structural Risk Minimization
Validation
Regularization

Estimation and approximation

- ▶ back to our program:

$$L(g_n) - L^* = \underbrace{L(g_n) - \inf_{g \in \mathcal{G}} L(g)}_{\text{estimation}} + \underbrace{\inf_{g \in \mathcal{G}} L(g) - L^*}_{\text{approximation}}$$

- ▶ empirical risk minimization and capacity control give a solution to the **estimation** part
- ▶ but they request a fixed \mathcal{G} for which the approximation part cannot be zero in a distribution free way
- ▶ how to fix this problem?
- ▶ central idea: allow \mathcal{G} to depend on the data
 - ▶ either in a generic way (i.e., \mathcal{G}_n);
 - ▶ or more ambitiously in more direct way \mathcal{G}_{D_n}

Increasing complexity/capacity

for binary classification

hypothesis

- ▶ $(\mathcal{G}^{(j)})_j$ with increasing finite $VCdim(\mathcal{G}^{(j)}) < \infty$
- ▶ asymptotically perfect: $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}^{(j)}} L(g) = L^*$
- ▶ $k_n \rightarrow \infty$ et $\frac{VCdim(\mathcal{G}^{(k_n)}) \log n}{n} \rightarrow 0$

result

the classifier defined by $g_n^* = \arg \min_{g \in \mathcal{G}^{(k_n)}} L_n(g)$ is universally strongly consistent

$$L(g_n^*) \xrightarrow{p.s.} L^*$$

Remark: it's impossible to get distribution free convergence speed (no free lunch)

In practice

- ▶ a central question: are those hypotheses realistic?

- ▶ a simple example:

- ▶ consider $\mathcal{X} = [0, 1]$ and

$$\mathcal{G}^{(j)} = \left\{ g \mid g(\mathbf{x}) = \text{sign} \left(a_0 + \sum_{k=1}^j (a_k \cos 2k\pi \mathbf{x} + b_k \sin 2k\pi \mathbf{x}) \right) \right\}$$

- ▶ $\mathcal{G}^{(j)}$ is a $2j + 1$ dimensional vector space then $VCdim(\mathcal{G}^{(j)}) \leq 2j + 1$
 - ▶ as the $\mathcal{G}^{(j)}$ correspond to truncated Fourier series, we have $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}^{(j)}} L(g) = L^*$ (by approximation of the conditional expectation $\mathbb{E}\{Y|X\}$)
 - ▶ for $k_n = n^\alpha$ with $0 < \alpha < 1$, the limit conditions are fulfilled
- ▶ more generally, those hypotheses are reasonable and backed up by universal approximation results
- ▶ **Warning:** the classes $\mathcal{G}^{(j)}$ are fixed *a priori*, they cannot depend on the data

Proof

$$L(g_n^*) - L^* = \left[L(g_n^*) - \inf_{g \in \mathcal{G}^{(k_n)}} L(g) \right] + \left[\inf_{g \in \mathcal{G}^{(k_n)}} L(g) - L^* \right]$$

- ▶ the first term is under control:

$$L(g_n^*) - \inf_{g \in \mathcal{G}^{(k_n)}} L(g) \leq 2 \sup_{g \in \mathcal{G}^{(k_n)}} |L_n(g) - L(g)|$$

$$\begin{aligned} \mathbb{P} \left\{ L(g_n^*) - \inf_{g \in \mathcal{G}^{(k_n)}} L(g) \geq \epsilon \right\} &\leq \mathbb{P} \left\{ \sup_{g \in \mathcal{G}^{(k_n)}} |L(g) - L_n(g)| > \epsilon/2 \right\} \\ &\leq 8\mathcal{S}_{\mathcal{F}^{(k_n)}}(n) e^{-n\epsilon^2/32} \\ &\leq 8(n^{VCdim(\mathcal{G}^{(k_n)})} + 1) e^{-n\epsilon^2/32} \end{aligned}$$

- ▶ Borel Cantelli gives the almost sure convergence
- ▶ the second term is under control via the hypothesis

Regression

- ▶ similar principle

- ▶ asymptotically optimal: $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}^{(j)}} L(g) = L^*$
 - ▶ increasing capacity, controlled via e.g. the pseudo-dimension $VCdim(\mathcal{G}^{(j)+})$
 - ▶ bounded function $\sup_{g \in \mathcal{G}^{(j)}} \|g\|_\infty < \infty$, but with growing bounds
 - ▶ bounded target $|Y| < \infty$

- ▶ removing the bound constraint on the target is easy:

- ▶ [Lugosi and Zegler, 1995]
 - ▶ for $c(u, v) = |u - v|^p$ and $\mathbb{E}\{|Y|^p\} < \infty$
 - ▶ central idea: clipped target $Y_L = \text{sign}(Y) \min(|Y|, L)$ with growing clipping value

Neural networks

Multi-layer perceptrons [Lugosi and Zegler, 1995]

- ▶ sigmoid function σ from \mathbb{R} to $[0, 1]$, non decreasing and such that $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ et $\lim_{x \rightarrow \infty} \sigma(x) = 1$
- ▶ for instance $\sigma(x) = 1/(1 + e^{-x})$
- ▶ model class

$$\mathcal{G}(k, \beta) = \left\{ g(\mathbf{x}) = \sum_{i=1}^k c_i \sigma(\langle \mathbf{a}_i, \mathbf{x} \rangle + b_i) + c_0; \sum_{i=1}^k |c_i| \leq \beta \right\}$$

- ▶ if $k_n \rightarrow \infty$ and $\beta_n \rightarrow \infty$, then $\bigcup_n \mathcal{G}(k_n, \beta_n)$ is dense in $L^p(\mu)$ (for any probability distribution μ , [Hornik et al., 1989]) and therefore $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}(k_n, \beta_n)} L(g) = L^*$ ($c(u, v) = |u - v|^p$)
- ▶ capacity control: $\frac{k_n \beta_n^{2p} \log(k_n \beta_n)}{n} \rightarrow 0$

Generic control

Binary classification
Regression

Data-driven control

Structural Risk Minimization
Validation
Regularization

101 / 154

Fabrice Rossi

Data-driven control



Comments

- ▶ the trade off between estimation error and approximation error is handled via the capacity estimation
- ▶ basically, this corresponds to adding the confidence bound induced by the VC results to the empirical risk
- ▶ beautiful result but not very practical:
 - ▶ empirical risk minimization for binary classification is intractable
 - ▶ this has to be repeated for a large number of classes
- ▶ More important limitation: the classes $\mathcal{G}^{(j)}$ cannot depend on the data

103 / 154

Fabrice Rossi

Data-driven control



Structural Risk Minimization

- ▶ in the previous solution, the capacity control is independent from the data
- ▶ it would be nice to explicitly balance the estimation error $L(g_n^*) - \inf_{g \in \mathcal{G}} L(g)$ and the approximation error $\inf_{g \in \mathcal{G}} L(g) - L^*$
- ▶ central idea: add to $L_n(g)$ a measure of the capacity of \mathcal{G}
- ▶ **Structural Risk Minimization** (binary classification):
 - ▶ asymptotically perfect: $\lim_{j \rightarrow \infty} \inf_{g \in \mathcal{G}^{(j)}} L(g) = L^*$
 - ▶ controlled capacity: $\sum_{j=1}^{\infty} e^{-VCdim(\mathcal{G}^{(j)})} < \infty$
 - ▶ capacity measure: $r(j, n) = \sqrt{\frac{8}{n} VCdim(\mathcal{G}^{(j)}) \log(en)}$
 - ▶ if g_n^* minimizes $\tilde{L}_n(g) = L_n(g) + r(j(g), n)$, where $j(g) = \inf\{k \mid g \in \mathcal{G}^{(k)}\}$
 - ▶ then $L(g_n^*) \xrightarrow{p.s.} L^*$

102 / 154

Fabrice Rossi

Data-driven control



Proof

- ▶ $g_{n,j} = \arg \min_{g \in \mathcal{G}^{(j)}} L_n(g)$ (and therefore $g_n^* = \arg \min_j \tilde{L}_n(g_{n,j})$)
- ▶ decomposition

$$L(g_n^*) - L^* = (L(g_n^*) - \inf_j \tilde{L}_n(g_{n,j})) + (\inf_j \tilde{L}_n(g_{n,j}) - L^*)$$
- ▶ the first term equals $L(g_n^*) - \tilde{L}_n(g_n^*)$ and therefore

$$\begin{aligned} \mathbb{P} \{L(g_n^*) - \tilde{L}_n(g_n^*) > \epsilon\} &\leq \mathbb{P} \left\{ \sup_j (L(g_{n,j}) - \tilde{L}_n(g_{n,j})) > \epsilon \right\} \\ &\leq \mathbb{P} \left\{ \sup_j (L(g_{n,j}) - L_n(g_{n,j}) - r(j, n)) > \epsilon \right\} \\ &\leq \sum_{j=1}^{\infty} \mathbb{P} \{ |L(g_{n,j}) - L_n(g_{n,j})| > \epsilon + r(j, n) \} \\ &\leq \sum_{j=1}^{\infty} 8n^{VCdim(\mathcal{G}^{(j)})} e^{-n(\epsilon + r(j, n))^2 / 8} \end{aligned}$$

104 / 154

Fabrice Rossi

Data-driven control



► then

$$\begin{aligned} \sum_{j=1}^{\infty} 8n^{\text{VCdim}(\mathcal{G}^{(j)})} e^{-n(\epsilon+r(j,n))^2/8} &\leq \sum_{j=1}^{\infty} 8n^{\text{VCdim}(\mathcal{G}^{(j)})} e^{-n\epsilon^2/8} e^{-r(j,n)^2/8} \\ &\leq 8e^{-n\epsilon^2/8} \sum_{j=1}^{\infty} e^{-\text{VCdim}(\mathcal{G}^{(j)})} \end{aligned}$$

- Borel Cantelli gives almost sure convergence of $L(g_n^*)$ to $\inf_j \tilde{L}_n(g_{n,j})$
- for $\epsilon > 0$, we find k such that $\inf_{g \in \mathcal{G}^{(k)}} L(g) - L^* \leq \epsilon$
- then for n large enough $r(k, n) \leq \frac{\epsilon}{2}$ (as $r(k, n)$ converges to 0 with n for any fixed k)

► therefore

$$\begin{aligned} \mathbb{P} \left\{ \inf_j \tilde{L}_n(g_{n,j}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \epsilon \right\} &\leq \mathbb{P} \left\{ \tilde{L}_n(g_{n,k}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \epsilon \right\} \\ &\leq \mathbb{P} \left\{ L_n(g_{n,k}) + r(k, n) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \epsilon \right\} \\ &\leq \mathbb{P} \left\{ L_n(g_{n,k}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) > \frac{\epsilon}{2} \right\} \\ &\leq \mathbb{P} \left\{ \sup_{g \in \mathcal{G}^{(k)}} |L_n(g) - L(g)| > \frac{\epsilon}{4} \right\} \\ &\leq 8n^{\text{VCdim}(\mathcal{G}^{(k)})} e^{-n\epsilon^2/128} \end{aligned}$$

- thus $\mathbb{P} \left\{ \limsup_{n \rightarrow \infty} \inf_j \tilde{L}_n(g_{n,j}) - \inf_{g \in \mathcal{G}^{(k)}} L(g) = 0 \right\} = 1$
- which gives the result

Model selection via validation

- a more classical use of data driven capacity control is the **validation set** method
- D_n is split in two disjoint subsets $D_m = (X_i, Y_i)_{1 \leq i \leq m}$ and $D_l = (X_i, Y_i)_{m+1 \leq i \leq n}$ ($l = n - m$)
- D_m is used to build a class of models \mathcal{G}_m
- then empirical risk minimisation is used:
 $g_n = \arg \min_{g \in \mathcal{G}_m} \frac{1}{l} \sum_{i=m+1}^n c(g(X_i), Y_i)$
- standard bounds apply to $L(g_n)$ chosen from \mathcal{G}_m
- any trick can be used on D_m **including data dependent classes**
- applications:
 - choice of the parameters of a ML method (e.g., number of neighbors)
 - more generally, apply ERM for different classes on D_m and choose the best model using D_l

Regularization

- can be seen as a generalization of the structural risk minimization
- base idea:
 - choose g_n by minimizing on \mathcal{G} the quantity

$$A_n(g) + \lambda \mathbf{R}(g)$$

- $A_n(g)$ is an empirical performance measure, a **loss** (this might be something else than L_n)
- $\mathbf{R}(g)$ is a complexity measure for g
- if g^* minimizes $A_n(g) + \lambda \mathbf{R}(g)$ then with $\mu = \mathbf{R}(g^*)$

$$g^* = \arg \min_{g \in \{g' \in \mathcal{G} \mid \mathbf{R}(g') \leq \mu\}} A_n(g)$$

- if A_n and \mathbf{R} are convex, both optimization problems are equivalent (duality): regularization corresponds to reduced model classes!

Examples

- ▶ Ridge approaches (a.k.a., weight decay):
 - ▶ take any parametric model, i.e.,
 $\mathcal{G} = \{g(\mathbf{x}) = F(\mathbf{x}, \mathbf{w}), \mathbf{w} \in \mathbb{R}^m\}$
 - ▶ use a squared difference cost function for the loss (and the risk)
 - ▶ penalize with $\mathbf{R}(F(\cdot, \mathbf{w})) = \|\mathbf{w}\|^2$
- ▶ Support vector machines:
 - ▶ in the linear case, classifier of the form $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$
 - ▶ obtained by minimizing over (\mathbf{w}, b)

$$\frac{1}{n} \sum_{i=1}^n \max(0, 1 - \mathbf{y}_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) + \lambda \|\mathbf{w}\|^2$$

- ▶ this is a ridge penalty with the **hinge loss** (different from the classification risk)



Difficulties

- ▶ Two different difficulties:
 - ▶ if A_n is not L_n , then we must prove that minimizing A_n leads to a low value of L
 - ▶ in general \mathcal{G} has infinite VC dimension: we must show that the reduced classes $\mathcal{G}_\mu = \{g \in \mathcal{G} | \mathbf{R}(g) \leq \mu\}$ have finite capacities
- ▶ In addition a practical difficulty is the choice of λ :
 - ▶ data size driven solution: λ_n with a good behavior with n
 - ▶ data driven solution: via validation like strategies



Outline

Convex Loss minimization

Motivations

Convex loss

Regularization

Kernel Machines

SVM consistency

Part IV

Beyond empirical risk minimization



Algorithmic difficulties

- ▶ for **regression problems**, optimizing L_n is “easy”
 - ▶ for instance $L_n(g) = \frac{1}{n} \sum_{i=1}^n \|g(\mathbf{x}_i) - \mathbf{y}_i\|^2$ leads to a least square problem
 - ▶ if \mathcal{G} is parametric

$$\mathcal{G} = \{g(\mathbf{x}) = F(\mathbf{x}, w), w \in W\},$$

with e.g., gradient based descent techniques

- ▶ for **classification problems**, this is more difficult
 - ▶ the risk is $\mathbb{P}\{g(X) \neq Y\}$
 - ▶ therefore the empirical risk as a discrete aspect
 - ▶ this leads to NP complete (or hard) problems in some cases



113 / 154

Fabrice Rossi

Convex Loss minimization

Squared error

- ▶ the easy case
- ▶ $\eta(\mathbf{x}) = \mathbb{E}\{Y \mid X = \mathbf{x}\}$
- ▶ given a strongly consistent algorithm:
 $\mathbb{E}\{(g_n(X) - Y)^2 \mid D_n\} \xrightarrow{p.s.} \mathbb{E}\{(\eta(X) - Y)^2\}$ and then
 $\mathbb{E}\{(g_n(X) - \eta(X))^2 \mid D_n\} \xrightarrow{p.s.} 0$
- ▶ we have
 $\mathbb{P}\{h_n(X) \neq Y \mid D_n\} - \mathbb{P}\{h^*(X) \neq Y\} \leq \mathbb{E}\{(g_n(X) - \eta(X))^2 \mid D_n\}$
where h^* is the (optimal) Bayes classifier
- ▶ therefore $L(h_n) \xrightarrow{p.s.} L^*$



115 / 154

Fabrice Rossi

Convex Loss minimization

Minimizing a loss

- ▶ standard solution: minimizing a loss rather than the risk
- ▶ for instance, a quadratic loss $a(u, v) = (u - v)^2$, which replaces $L_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{g(X_i) \neq Y_i\}}$ by

$$A_n(g) = \frac{1}{n} \sum_{i=1}^n a(g(X_i), Y_i) = \frac{1}{n} \sum_{i=1}^n (g(X_i) - Y_i)^2$$

- ▶ in general, g takes values in \mathbb{R} and the associated classifier is defined by $h(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$
- ▶ using regression related results gives bounds on $|A_n(g) - A(g)|$, where

$$A(g) = \mathbb{E}\{a(g(X), Y)\}$$

and convergence results to $A_{\mathcal{G}}^* = \inf_{g \in \mathcal{G}} A(g)$

- ▶ but we want information on $L(g)$ (or $L(h)$)!



114 / 154

Fabrice Rossi

Convex Loss minimization

A more general case

- ▶ to extend the squared error case, we need to guarantee that $L(g) - L^*$ converges to zero when $A(g) - A^*$ does
- ▶ [Steinwart, 2005]:
 - ▶ $a: \mathbb{R} \times \{-1, 1\} \rightarrow \mathbb{R}$
 - ▶ $C(t, \alpha) = \alpha a(t, 1) + (1 - \alpha) a(t, -1)$
 - ▶ $A(g) = \int C(\mathbb{P}\{Y = 1 \mid X = x\}, g(x)) P_X(dx)$
 - ▶ a is **admissible** if:
 - ▶ a is continuous
 - ▶ $\alpha < 1/2 \Rightarrow \arg \min_t C(t, \alpha) < 0$
 - ▶ $\alpha > 1/2 \Rightarrow \arg \min_t C(t, \alpha) > 0$
 - ▶ in essence, this means that to minimize A , g has to have the same sign as the regression function!
 - ▶ is this case, for all ϵ , there is δ such that $A(g) - A^* \leq \delta$ implies $L(g) - L^* \leq \epsilon$



116 / 154

Fabrice Rossi

Convex Loss minimization

Example: the hinge loss

▶ hinge loss

$$a(x, y) = \max(1 - yx, 0)$$

$$a(g(\mathbf{x}), \mathbf{y}) \geq \mathbb{I}_{\{g(\mathbf{x}) \neq \mathbf{y}\}}$$

$$C(t, \alpha) = \alpha \max(1 - t, 0) + (1 - \alpha) \max(1 + t, 0) :$$

- ▶ if $t \geq 1$, then $C(t, \alpha) = (1 - \alpha)(1 + t) \geq 2(1 - \alpha)$
- ▶ if $t \leq -1$, then $C(t, \alpha) = \alpha(1 - t) \geq 2\alpha$
- ▶ if $t \in [-1, 1]$, then

$$C(t, \alpha) = \alpha(1 - t) + (1 - \alpha)(1 + t) = 1 + (1 - 2\alpha)t$$
- ▶ if $\alpha < 1/2$, the minimum on $[-1, 1]$ is reached in -1 and equals $2\alpha \leq 2(1 - \alpha)$
- ▶ symmetrically, when $\alpha > 1/2$, the minimum is reached in 1 and equals $2(1 - \alpha) \leq 2\alpha$

▶ therefore, the hinge loss is admissible

Convex risk minimization

▶ Interesting particular case: $a(u, v) = \phi(uv)$ where ϕ is non negative

- ▶ as in the general case, $C(t, \alpha) = \alpha\phi(t) + (1 - \alpha)\phi(-t)$
- ▶ $H(\alpha) = \inf_t C(t, \alpha)$ and $H^-(\alpha) = \inf_{t(2\alpha-1) \leq 0} C(t, \alpha)$
- ▶ ϕ is calibrated if $H^-(\alpha) > H(\alpha)$ for all α

▶ for all ϕ , there is a corresponding ψ such that

$$\psi(L(g) - L^*) \leq A(g) - A^*$$

▶ ϕ is calibrated if and only if

$$\lim_{i \rightarrow \infty} \psi(\alpha_i) = 0 \Leftrightarrow \lim_{i \rightarrow \infty} \alpha_i = 0$$

▶ if ϕ is convex, then ϕ is calibrated if and only if ϕ has a derivative in 0 and $\phi'(0) < 0$

Margin

▶ when $h(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$, the binary classification cost corresponds to $\mathbb{I}_{\{\text{sign}(g(\mathbf{x})) \neq \mathbf{y}\}} = \mathbb{I}_{\{g(\mathbf{x})\mathbf{y} < 0\}}$

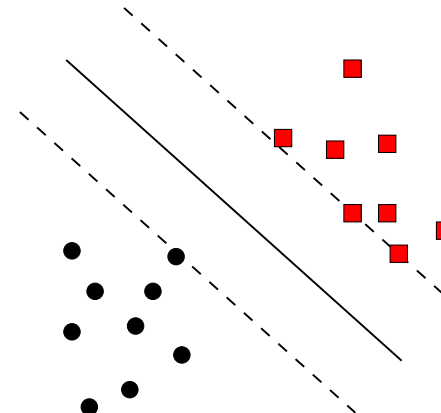
▶ intuitively:

- ▶ when $g(\mathbf{x})\mathbf{y}$ is large, g makes a robust decision (it is not sensitive to noise on \mathbf{x})
- ▶ when $-g(\mathbf{x})\mathbf{y}$ is large, g makes a big mistake the label of \mathbf{x}
- ▶ it seems interesting to try to optimize a function of $g(\mathbf{x})\mathbf{y}$

▶ linear case:

- ▶ $g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$
- ▶ the oriented distance between \mathbf{x} and $g(\mathbf{x}) = 0$ is $\frac{|\langle \mathbf{w}, \mathbf{x} \rangle + b|}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$
- ▶ keeping the target in mind: $\frac{y(\langle \mathbf{w}, \mathbf{x} \rangle + b)}{\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}}$
- ▶ with the standard normalization $\sqrt{\langle \mathbf{w}, \mathbf{w} \rangle} = 1$, the distance becomes $y g(\mathbf{x})$: this is the **margin**

Maximal margin



- ▶ linearly separable data: many linear classifiers
- ▶ if some data are close to the separator, the margin is small \Rightarrow low robustness
- ▶ choose the classifier by maximizing the margin
- ▶ **Support vector machines**

Examples of calibrated convex loss

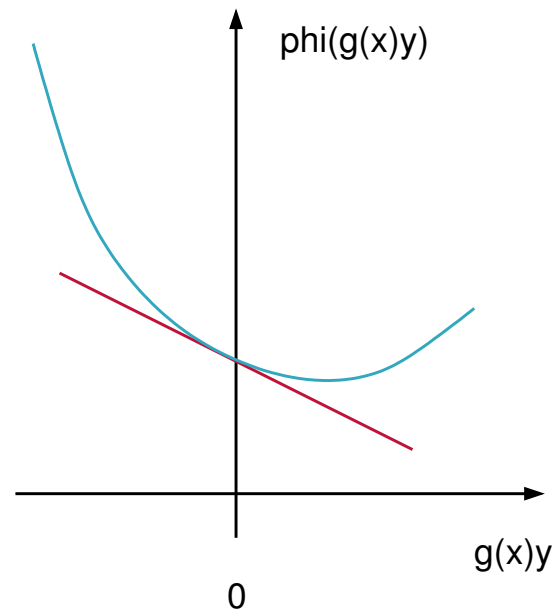
- ▶ $\phi(a) = \max(1 - a, 0)$ ($a(x, y) = \max(1 - yx, 0)$)
- ▶ $\phi(a) = e^{-a}$ ($a(x, y) = e^{-yx}$)
- ▶ $\phi(a) = |1 - a|^p$ ($a(x, y) = |1 - yx|^p$)
- ▶ **Crucial point:** if ϕ is calibrated, there is γ such that

$$\gamma\phi(a) \geq \mathbb{I}_{\{a \leq 0\}}$$

and therefore

$$\gamma a(g(x), y) \geq \mathbb{I}_{\{g(x)y \leq 0\}} = \mathbb{I}_{\{\text{sign}(g(x)) \neq y\}}$$

Calibration



$\phi'(0) < 0$ implies that around $g(x)y = 0$, the loss is higher for a negative margin than for a positive one

Summary

- ▶ in classification, algorithmic considerations lead to minimize a loss rather than the risk
- ▶ under some admissibility conditions, a minimal loss implies a minimal cost
- ▶ margin:
 - ▶ a particular case is the one of losses based on the margin $g(x)y$
 - ▶ for convex losses $\phi(g(x)y)$, admissibility (a.k.a., calibration) is equivalent to $\phi'(0) < 0$
- ▶ this (partially) justifies algorithms such as Support Vector Machines and Adaboost

Plan

Convex Loss minimization

Motivations

Convex loss

Regularization

Kernel Machines

SVM consistency

Support Vector Machines

- ▶ linear SVM: $g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ obtained by minimizing

$$\frac{1}{n} \sum_{i=1}^n \max(0, 1 - \mathbf{y}_i(\langle \mathbf{w}, \mathbf{x}_i \rangle)) + \lambda \|\mathbf{w}\|^2$$

- ▶ non linear version:

- ▶ kernel:

- ▶ $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

- ▶ K is symmetric and semi definite $\sum_i^n \sum_j^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$

- ▶ e.g., $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$

- ▶ then $g(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x})$ chosen by minimizing

$$\frac{1}{n} \sum_{i=1}^n \max(0, 1 - \mathbf{y}_i g(\mathbf{x}_i)) + \lambda \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

- ▶ **Remark:** we drop the b as it leads to technical complications

Reproducing Kernel Hilbert Space

- ▶ K generates a **Reproducing Kernel Hilbert Space**, \mathcal{H} , the completion of

$$H = \left\{ g(\mathbf{x}) = \sum_{i=1}^p \alpha_i K(\mathbf{x}_i, \mathbf{x}); p \in \mathbb{N}, \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X} \right\}$$

with respect to the inner product

$$\left\langle \sum_{i=1}^p \alpha_i K(\mathbf{x}_i, \cdot), \sum_{i=1}^m \beta_i K(\mathbf{x}'_i, \cdot) \right\rangle = \sum_{i=1}^p \sum_{j=1}^m \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{x}'_j)$$

- ▶ then if $g(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x})$

$$\|g\|_{\mathcal{H}}^2 = \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

Representer theorem

- ▶ central result in RKHS
- ▶ let's consider the problem

$$\min_{g \in \mathcal{H}} \left(U(g(\mathbf{x}_1), \dots, g(\mathbf{x}_n)) + \Omega(\|g\|_{\mathcal{H}}^2) \right)$$

where U is any function from \mathbb{R}^n to \mathbb{R} and Ω is a non decreasing function

- ▶ then there is $\alpha_1, \dots, \alpha_n$ such that $g = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \cdot)$ reaches the minimum
- ▶ furthermore, if Ω is increasing, any solution of the problem has this specific form

Kernel methods

- ▶ the representer theorem leads to a family of kernel methods:
 - ▶ choose a loss function U and a regularization function Ω
 - ▶ find $g \in \mathcal{H}$ that minimizes

$$\{ U(g(\mathbf{x}_1), \dots, g(\mathbf{x}_n)) + \Omega(\|g\|_{\mathcal{H}}^2) \}$$

by solving in \mathbb{R}^n the following optimization problem

$$\min_{\alpha \in \mathbb{R}^n} \left\{ U \left(\sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}_1), \dots, \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}_n) \right) + \Omega \left(\sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \right\}$$

- ▶ if $U = \frac{1}{n} \sum_{i=1}^n a(g(\mathbf{x}_i), \mathbf{y}_i)$ with an admissible a this could lead to a consistent classifier
- ▶ standard SVM corresponds to the hinge loss for a and to $\Omega(u) = \lambda_n u$

SVM consistency

[Steinwart, 2005]

- ▶ two difficulties:
 - ▶ \mathcal{H} has generally an infinite VC dimension
 - ▶ the optimized criterion is neither the empirical loss, nor an empirical convex loss
- ▶ Steinwart approach:
 - ▶ show that when λ goes to 0 the influence of regularization vanishes: asymptotically, we are optimizing A_n
 - ▶ use the convex loss results to show that optimizing A leads to an optimal L
 - ▶ show that the regularization term allows to control $A_n - A$



129 / 154

Fabrice Rossi

Regularization

Regularization

- ▶ let's define:

$$A_\lambda(g) = \mathbb{E} \{ \max(0, -Y(g(X))) \} + \lambda \|g\|_{\mathcal{H}}^2$$

$$A_{\lambda,n}(g) = \frac{1}{n} \sum_{i=1}^n \max(0, -\mathbf{y}_i(g(\mathbf{x}_i) + b)) + \lambda \|g\|_{\mathcal{H}}^2$$

- ▶ any minimizer of A_λ on \mathcal{H} is such that $\|g\|_{\mathcal{H}}^2 \leq \frac{2}{\lambda}$
- ▶ similarly, any minimizer of $A_{\lambda,n}$ is such that $\|g\|_{\mathcal{H}}^2 \leq \frac{2}{\lambda}$
- ▶ as a consequence, regularization allows to control $|A(g) - A_n(g)|$
- ▶ but it vanishes asymptotically as with $g_\lambda = \arg \min_{g \in \mathcal{H}} A_\lambda(g)$, we have

$$\lim_{\lambda \rightarrow 0} A_\lambda(g_\lambda) = \inf_g A(g)$$



131 / 154

Fabrice Rossi

Regularization

Some hypothesis

- ▶ X_i take values in \mathcal{X} a compact metric space
- ▶ the kernel K is such that:
 - ▶ $\phi(\mathbf{x}) = K(\mathbf{x}, \cdot)$ (from \mathcal{X} to \mathcal{H}) is continuous
 - ▶ $\{g \in C(\mathcal{X}) \mid g(\cdot) = K(w, \phi(\cdot)); w \in \mathcal{H}\}$ is dense in $C(\mathcal{X})$ (the kernel is universal)
 - ▶ e.g., the Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$ for \mathcal{X} a bounded and closed subset of \mathbb{R}^d
- ▶ Steinwart studies any admissible a and very general regularization functions Ω (as well as regression problems in [Christmann and Steinwart, 2007])
- ▶ is this lecture:
 - ▶ a is the hinge loss
 - ▶ $\Omega(\lambda_n, \|g\|_{\mathcal{H}}) = \lambda_n \|g\|_{\mathcal{H}}^2$
 - ▶ no b term



130 / 154

Fabrice Rossi

Regularization

Concentration

- ▶ $|A(g) - A_n(g)|$ is controlled via covering numbers
- ▶ for a class \mathcal{F} of functions with values in $[0, B]$

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |Pf - P_n f| \geq \epsilon \right\} \leq 2N_\infty \left(\frac{\epsilon}{3}, \mathcal{F} \right) e^{-2n\epsilon^2/(9B^2)}$$

- ▶ Remark: we use here covering numbers with respect to the sup norm
- ▶ elementary proof via a $\epsilon/3$ -covering of \mathcal{F} , f_1, \dots, f_m :
 - ▶ $|Pf - P_n f| = |Pf - Pf_i + Pf_i - P_n f_i + P_n f_i - P_n f| \leq \frac{2}{3}\epsilon + |Pf_i - P_n f_i|$
 - ▶ $\mathbb{P} \{ \sup_{f \in \mathcal{F}} |Pf - P_n f| \geq \epsilon \} \leq \mathbb{P} \{ \sup_{1 \leq i \leq m} |Pf_i - P_n f_i| \geq \frac{\epsilon}{3} \}$
 - ▶ Hoeffding's inequality gives $\mathbb{P} \{ |Pf_i - P_n f_i| \geq \frac{\epsilon}{3} \} \leq 2e^{-2n\epsilon^2/(9B^2)}$



132 / 154

Fabrice Rossi

Regularization

Concentration

- ▶ let's denote $\|K\|_\infty = \sup_{\mathbf{x} \in \mathcal{X}} \sqrt{K(\mathbf{x}, \mathbf{x})}$
- ▶ $\|g\|_{\mathcal{H}} \leq \delta$ implies $\|g\|_\infty \leq \delta \|K\|_\infty$:
 - ▶ reproducing property: $g(\mathbf{x}) = \langle g, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$
 - ▶ $|\langle g, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}| \leq \|g\|_{\mathcal{H}} \|K(\mathbf{x}, \cdot)\|_{\mathcal{H}}$
 - ▶ $\|K(\mathbf{x}, \cdot)\|_{\mathcal{H}} = \sqrt{K(\mathbf{x}, \mathbf{x})} \leq \|K\|_\infty$
- ▶ given the class

$$\mathcal{F}_\delta = \{a(g(\cdot), \cdot); \|g\|_{\mathcal{H}} \leq \delta\}$$

with $a(u, v) = \max(0, 1 - uv)$

- ▶ then for all $f \in \mathcal{F}_\delta$, $f(\mathbf{x}, \mathbf{y}) \in [0, \delta \|K\|_\infty]$ and therefore

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}_\delta} |Pf - P_n f| \geq \epsilon \right\} \leq 2N_\infty \left(\frac{\epsilon}{\delta}, \mathcal{F}_\delta \right) e^{-2n\epsilon^2 / (9\delta^2 \|K\|_\infty^2)}$$



Recapitulation

- ▶ $g_n = \arg \min_{g \in \mathcal{H}} A_{\lambda_n, n}(g)$ (when $\lambda_n \rightarrow 0$)
- ▶ $g_n^* = \arg \min_{g \in \mathcal{H}} A_{\lambda_n}(g)$
- ▶ given $\epsilon > 0$:
 - ▶ there is $\delta < 0$ such that $A(g) \leq A^* + \delta$ implies $L(g) \leq L^* + \epsilon$ (admissibility)
 - ▶ for $n \geq n_0$, $|A_{\lambda_n}(g_n^*) - A^*| \leq \delta/3$
 - ▶ if $|A(g_n) - A_n(g_n)| < \delta/3$ and $|A(g_n^*) - A_n(g_n^*)| < \delta/3$ then

$$\begin{aligned} A(g_n) &\leq A(g_n) + \lambda_n \|g_n\|_{\mathcal{H}}^2 \\ &\leq A_n(g_n) + \lambda_n \|g_n\|_{\mathcal{H}}^2 + \delta/3 \\ &\leq A_n(g_n^*) + \lambda_n \|g_n^*\|_{\mathcal{H}}^2 + \delta/3 \\ &\leq A(g_n^*) + \lambda_n \|g_n^*\|_{\mathcal{H}}^2 + 2\delta/3 \\ &\leq A^* + \delta \end{aligned}$$

- ▶ therefore for $n \geq n_0$, $\mathbb{P} \{L(g) \geq L^* + \epsilon\} \leq \mathbb{P} \{|A(g_n) - A_n(g_n)| \geq \delta/3\} + \mathbb{P} \{|A(g_n^*) - A_n(g_n^*)| \geq \delta/3\}$



Concentration

- ▶ let g_1, \dots, g_m be a minimal ϵ -covering of $\mathcal{H}_\delta = \{g \in \mathcal{H} \mid \|g\|_{\mathcal{H}} \leq \delta\}$
- ▶ $\|g_i - g_j\|_\infty \leq \epsilon$ implies $\|a(g_i(\cdot), \cdot) - a(g_j(\cdot), \cdot)\|_\infty \leq \epsilon$
- ▶ this $(a(g_i(\cdot), \cdot))_{1 \leq i \leq m}$ is a ϵ -covering of \mathcal{F}_δ
- ▶ then $N_\infty(\epsilon, \mathcal{F}_\delta) \leq N_\infty(\epsilon, \mathcal{H}_\delta)$
- ▶ and therefore

$$\mathbb{P} \left\{ \sup_{g \in \mathcal{H}_\delta} |A(g) - A_n(g)| \geq \epsilon \right\} \leq 2N_\infty \left(\frac{\epsilon}{\delta}, \mathcal{H}_\delta \right) e^{-2n\epsilon^2 / (9\delta^2 \|K\|_\infty^2)}$$

- ▶ if $g_{n, \lambda} = \arg \min_{g \in \mathcal{H}} A_{\lambda, n}(g)$ then

$$\mathbb{P} \{|A(g_{n, \lambda}) - A_n(g_{n, \lambda})| \geq \epsilon\} \leq 2N_\infty \left(\frac{\epsilon}{\delta}, \mathcal{H}_{\sqrt{\frac{\epsilon}{\lambda}}} \right) e^{-n\epsilon^2 \lambda / (9\|K\|_\infty^2)}$$



Covering numbers

- ▶ to conclude, we need a bound on $N_\infty(\epsilon, \mathcal{H}_\delta)$
- ▶ this can be obtained using operator approximation theory
- ▶ when K is a regular kernel, we have in general

$$\ln N_\infty(\epsilon, \mathcal{H}_1) \leq c\epsilon^{-\gamma}$$

for some constants $c > 0$ and $\gamma > 0$, and therefore

$$\ln N_\infty(\epsilon, \mathcal{H}_\delta) \leq c \left(\frac{\delta}{\epsilon} \right)^\gamma$$

- ▶ better results can be obtained for some specific kernels, such as the Gaussian one on \mathbb{R}^d

$$\ln N_\infty(\epsilon, \mathcal{H}_1) \leq c \left(\log \frac{1}{\epsilon} \right)^{d+1}$$



- ▶ finally, we just need to control

$$2N_\infty \left(\frac{\epsilon}{3}, \mathcal{H}_{\sqrt{\frac{2}{\lambda_n}}} \right) e^{-n\epsilon^2 \lambda_n / (9\|K\|_\infty^2)}$$

- ▶ for instance $\frac{\|K\|_\infty^2}{n\lambda_n} \ln N_\infty \left(\epsilon, \mathcal{H}_{\sqrt{\frac{2}{\lambda_n}}} \right) \rightarrow 0$ lead to strong consistency
- ▶ for a regular kernel, consistency is obtained via, e.g.,

$$n\lambda_n^{1+\gamma/2} \rightarrow \infty$$

Extensions

Rademacher averages (a.k.a. data dependent bounds)

- ▶ remember the Rademacher variables σ_i ?
- ▶ given a bounded set of vectors $A \in \mathbb{R}^p$, we define

$$R_p(A) = \mathbb{E} \left\{ \sup_{\mathbf{a} \in A} \frac{1}{p} \left| \sum_{i=1}^p \sigma_i a_i \right| \right\}$$

- ▶ using McDiarmind bounded differences concentration inequality, we can obtain data dependent bounds, that is with probability at least $1 - \delta$

$$\sup_{f \in \mathcal{F}} |Pf - P_n f| \leq 2R_n(\mathcal{F}_{Z_1, \dots, Z_n}) + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}$$

- ▶ easier way to get VC bounds

- ▶ regularization in a Reproducing Kernel Hilbert Space \mathcal{H} :
 - ▶ corresponds to searching for the classifier in a ball $\mathcal{H}_\delta = \{g \in \mathcal{H} \mid \|g\|_{\mathcal{H}} \leq \delta\}$
 - ▶ which allows to bound the difference between the empirical loss and the loss with a covering number
 - ▶ which is in turn controlled via regularity assumptions on the kernel K
- ▶ admissibility conditions on the loss a and hypothesis on the regularization function Ω lead to strong (almost) universal consistency for kernel machines
- ▶ similar results are available for regularized boosting
- ▶ the **crucial point** is the regularization: it circumvents the infinite VC of \mathcal{H}

Extensions

Rademacher averages

- ▶ the bound can be evaluated from the data
- ▶ can also be used to derive margin based bounds in binary classification
- ▶ in essence $L(g)$ can be bounded by some classical terms (capacity, etc.) added to a margin based empirical loss: to the empirical loss, we add “label errors” when the value of $g(\mathbf{x}_i)\mathbf{y}_i$ is too small
- ▶ this justifies somehow the idea of minimizing the margin

Extensions

Taking into account the variance

- ▶ Hoeffding's like inequalities do not take into account the variance of the considered variables
- ▶ in binary classification, the variance is under control
- ▶ taking this into account leads to the following bound: with probability at least $1 - \delta$ (where $V = VCdim(\mathcal{G})$)

$$L(g_n^*) - \inf_{g \in \mathcal{G}} L(g) \leq C \left(\sqrt{\inf_{g \in \mathcal{G}} L(g) \frac{V \log n + \frac{1}{\delta}}{n}} + \frac{V \log n + \frac{1}{\delta}}{n} \right)$$

- ▶ when $\inf_{g \in \mathcal{G}} L(g) = 0$, this gives a very fast rate of convergence



Extensions

- ▶ Other supervised algorithms:
 - ▶ k -nn and trees
 - ▶ boosting
 - ▶ randomized methods (e.g., Random Forests)
- ▶ Clustering consistency
- ▶ Open questions:
 - ▶ resampling methods: can we obtain non asymptotic distribution free bounds for leave-one-out, k -fold cross-validation and bootstrap?
 - ▶ optimization aspects: what happens if we can't reach $\inf_{g \in \mathcal{G}} A_n(g)$?
- ▶ and so on...



Extensions

Noise conditions

- ▶ fast rate is obtained with $\inf_{g \in \mathcal{G}} L(g) = 0$
- ▶ can we do better?
 - ▶ the complexity of binary classification is strongly related to the behavior of the regression function $\eta(\mathbf{x}) = \mathbb{E}\{Y \mid X = \mathbf{x}\}$ around $\frac{1}{2}$
 - ▶ if $|2\eta(\mathbf{x}) - 1| > h$ for all \mathbf{x} , the convergence can be even better
 - ▶ but this is a strong condition, that can be relaxed into a Mammen-Tsybakov noise condition given by

$$\exists \alpha \in [0, 1[, B > 0, \forall t \geq 0, \mathbb{P}\{|2\eta(X) - 1| > t\} \leq Bt^{\frac{\alpha}{1-\alpha}}$$

- ▶ all of this can be extended to the case when the Bayes classifier is not in \mathcal{G}



Wrapping up

What have we learned?

- ▶ statistical learning framework:
 - ▶ based on **stationarity**: future data will generated by the same process as old data
 - ▶ with maximal information coming from each observation: **independence**
- ▶ **consistency**: a good machine learning algorithm should eventually give the best possible performances on the data
- ▶ consistency mixes two different problems:
 - ▶ **estimation** problems: can we trust the empirical risk/loss?
 - ▶ **approximation** problems: given the chosen method, is it even possible to reach the best performances?
- ▶ for a fixed class of models, learnability is equivalent to **finite capacity**: the capacity is measured via a **dimension** (or a set of dimensions)



Wrapping up

What have we learned?

- ▶ the **only way** to have **universal** consistent models is to balance estimation problems and approximation problems
- ▶ this can be done via some form of **regularization**:
 - ▶ at minimum the capacity is adapted to the size of the dataset
 - ▶ preferably, it is adapted to the data themselves
- ▶ by products:
 - ▶ **concentration inequalities** can be applied to other problems, especially now that we have (very) large datasets
 - ▶ **margin** seems to be an interesting concept



145 / 154

Fabrice Rossi

Summary

Outline

Researchers

Bibliography



147 / 154

Fabrice Rossi

Part V

Appendix



146 / 154

Fabrice Rossi

Researchers

This a non exhaustive and very personal list of homepages that one should monitor for interesting papers and/or advances in statistical learning ¹

- ▶ Peter Bartlett: <http://www.stat.berkeley.edu/~bartlett/>
- ▶ Luc Devroye: <http://cg.scs.carleton.ca/~luc/>
- ▶ Gábor Lugosi: <http://www.econ.upf.edu/~lugosi/>
- ▶ David Pollard: <http://www.stat.yale.edu/~pollard/>
- ▶ Ingo Steinwart: <http://www.c3.lanl.gov/~ingo/>

¹ Please, don't be annoyed if you are not in this list





148 / 154




Fabrice Rossi

Researchers



Bibliography I

-  N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler
Scale-sensitive dimensions, uniform convergence, and learnability
Journal of the ACM, 44(4):615-631, 1997.
[Fat shattering and learnability](#)
<http://homes.dsi.unimi.it/~cesabian/Pubblicazioni/jacm-97b.pdf>
-  S. Boucheron, O. Bousquet, and G. Lugosi.
Theory of classification: a survey of some recent advances.
ESAIM; Probability and Statistics, 9:323–375, November 2005.
[State of the art and summary of recent results \(in 2005\)](#)
<http://www.econ.upf.edu/~lugosi/esaimsurvey.pdf>



Bibliography II

-  O. Bousquet, S. Boucheron, and G. Lugosi.
Advanced lectures in machine learning, volume 3176 of *LNAI*,
chapter Introduction to statistical learning theory, pages
169–207.
Springer-Verlag, 2004.
[A good introduction](#)
http://www.econ.upf.edu/~lugosi/mlss_slr.pdf
-  N. Cesa-Bianchi, and G. Lugosi,
Prediction, Learning, and Games.
Cambridge University Press, New York, 2006
[A reference book on predicting “individual sequences”](#)
-  A. Christmann and I. Steinwart,
Consistency and Robustness of Kernel Based Regression.
Bernoulli, Vol. 13, pp. 799-819, 2007.
[Consistency of kernel machines in regression](#)
http://www.c3.lanl.gov/ml/pubs/2005_regression/paper.pdf

Bibliography III

-  P. Domingos and G. Hulten,
A General Method for Scaling Up Machine Learning Algorithms
and its Application to Clustering.
*Proceedings of the Eighteenth International Conference on
Machine Learning (pp. 106-113), 2001.*
[Application of concentration inequalities to stream processing](#)
<http://www.cs.washington.edu/homes/pedrod/papers/mlc01.pdf>
-  L. Devroye, L. Györfi, and G. Lugosi.
A Probabilistic Theory of Pattern Recognition, volume 21 of
Applications of Mathematics.
Springer, 1996.
[Reference book on the subject, despite its age](#)

Bibliography IV

-  S. Kulkarni, G. Lugosi, and S. Venkatesh.
Learning pattern classification – a survey.
IEEE Transactions on Information Theory, 44(6):2178–2206,
October 1998.
[State of the art in 1998](#)
http://www.princeton.edu/~kulkarni/Papers/Journals/jl1998_klv_transit.pdf
-  K. Hornik, M. Stinchcombe, and J. White.
Multilayer feedforward networks are universal approximators.
Neural Networks, 2:359-366, 1989
[Universal approximation by MLP](#)
http://weber.ucsd.edu/~mbacci/white/pub_files/hwcv-028.pdf

Bibliography V



G. Lugosi and K. Zeger.

Nonparametric estimation via empirical risk minimization.

IEEE Transactions on Information Theory, 41(3):677–687, May 1995.

Regression modes and L^p risk

<http://www.code.ucsd.edu/~zeger/publications/journals/LuZe95-IT-Nonparametric/LuZe95-IT-Nonparametric.pdf>



G. Lugosi and K. Zeger.

Concept learning using complexity regularization.

IEEE Transactions on Information Theory, 42(1):48–54, January 1996.

Structural risk minimization

<http://www.code.ucsd.edu/~zeger/publications/journals/LuZe96-IT-Concept/LuZe96-IT-Concept.pdf>



Bibliography VI



D. Pollard.

Convergence of Stochastic Processes.

Springer-Verlag, New York, 1984.

Reference book, available online

<http://www.stat.yale.edu/~pollard/Books/1984book/pollard1984.pdf>



I. Steinwart.

Consistency of support vector machines and other regularized kernel machines.

IEEE Transactions on Information Theory, 51(1):128—142, January 2005.

As the title says

