

## Institute of Statistics



### STAT2550 Data Mining

[15h+15h exercises] 5 credits

This course is taught in the 2nd semester

**Teacher(s):** Libei Chen  
**Language:** French  
**Level:** Second cycle

#### Aims

In this course, we will learn data mining methodology and techniques for knowledge discovery in large databases. We will also see how data mining differs from traditional statistics and how to treat a practical problem with an appropriate data mining tool.

#### Main themes

- Data Mining application domains
- Steps of a data mining project
- Sampling and partitioning of the data base and training and validation sets
- Data pretreatment and validation
- Preliminary variable analysis, variables reduction and transformation
- Classification and modeling tools of data mining
- Decision trees
- Neural networks
- Tools to validate and compare estimated models
- Case studies

#### Content and teaching methods

- Introduction to data mining
  - o Data and data mining systems
  - o Data mining applications
  - o Data mining process and methodology
  - o Data mining in customer relationship management (CRM)
  - o Traditional statistics versus data mining
- Data preparation for data mining
  - o Data preparation stages
  - o Data specification
  - o Data extraction and aggregations
  - o Data audit and exploration
  - o Data pre-processing
- Predictive modelling
  - o Decision trees
  - o Neural networks
  - o Model validation and assessment
- Descriptive modelling
  - o Clustering
  - o K-means
  - o Kohonen Self-Organising Map
- Case studies

**Other information (prerequisite, evaluation (assessment methods), course materials recommended readings,**

...)

## References:

1. Berry M. and G. Linoff (2000), "Mastering Data Mining, The Art and Science of Customer Relationship Management", John Wiley.
2. Bishop, C.M. (1995), Neural Networks for Pattern Recognition, Oxford.
3. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984), "Classification and Regression Trees", Wadsworth, Inc., Belmont, California.
4. Han J. and M. Kamber (2000), "Data Mining: Concepts and Techniques", Morgan Kaufmann,.
5. Hastie Tr., R. Tibshirani and J. Friedman (2001), "The Elements of Statistical Learning -Data Mining, Inference and Prediction", Springer.
6. Haykin S., "Neural Networks: A comprehensive Foundation", Prentice Hall, 1999
7. Kohonen T. (1995), "Self-Organizing Maps", Springer Series in Information Sciences, Oxford University Press.
8. Piatetsky-Shapiro G. and W. J. Frawley (1991), "Knowledge Discovery in Databases", AAAI/MIT Press.
9. Piatetsky-Shapiro G., U. Fayyad, and P. Smith (1996). "From data mining to knowledge discovery: An overview", In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press,.
10. Pyle D. (2000), "Data Preparation for Data Mining", Morgan Kaufman.
11. Richard O. Dula, Pete E. Hart and David G. Stork (2000), "Pattern Classification", John Wiley, Second edition.
12. Van Hulle M. (2000), "Faithful Representations and Topographic Maps: From Distortion- to Information-Based Self-Organization", John Willey & Sons Inc.

**Other credits in programs**

<b>ECGE3DS/MK</b>	Diplôme d'études spécialisées en économie et gestion (Master in business administration) (marketing)	(5 credits)	Mandatory
<b>ECGE3DS/SC</b>	Diplôme d'études spécialisées en économie et gestion (Master in business administration) (Supply Chain Management)	(5 credits)	Mandatory
<b>STAT21MS/DM</b>	Première année du master en statistique, orientation générale, à finalité spécialisée (data management et data mining)	(5 credits)	Mandatory
<b>STAT21MS/EA</b>	Première année du master en statistique, orientation générale, à finalité spécialisée (économie et assurance)	(5 credits)	
<b>STAT21MS/MS</b>	Première année du master en statistique, orientation générale, à finalité spécialisée (marketing et sondage)	(5 credits)	
<b>STAT22MS/DM</b>	Deuxième année du master en statistique, orientation générale, à finalité spécialisée (data management et data mining)	(5 credits)	Mandatory
<b>STAT22MS/EA</b>	Deuxième année du master en statistique, orientation générale, à finalité spécialisée (économie et assurance)	(5 credits)	
<b>STAT22MS/MS</b>	Deuxième année du master en statistique, orientation générale, à finalité spécialisée (marketing et sondage)	(5 credits)	
<b>STAT3DA/E</b>	diplôme d'études approfondies en statistique (statistique et économétrie)	(5 credits)	
<b>STAT3DA/P</b>	diplôme d'études approfondies en statistique (pratique de la statistique)	(5 credits)	