



APPRENTISSAGE BAYÉSIEEN ET EXPLORATION DE TEXTES

BAYESIAN MACHINE LEARNING AND TEXT MINING

Lecturers: **Alexandre SAIDI, Mohsen ARDABILIAN, Stéphane DERRODE**

| Lecturers : 12 | TC : 0.0 | PW : 0.0 | Autonomy : 0.0 | Study : 8 | Project : 0.0 | Language : FR

Objectives

We will focus on the family of Bayesian methods, which is distinguished by its optimality in the sense of certain criteria, by its reduced cost from an algorithmic point of view and by the interpretability of its results. We will also study the solutions available to the data scientist when the learning sample is small in relation to the number of parameters to be learned, or when the learning must be done in an unsupervised manner. In terms of application, we will focus on the exploration of a textual corpus to discover, for example, new customers eligible for the sale of a service/product, to predict the feelings (opinions) of customers or to understand the behaviours that predict fraud.

Keywords : Bayesian decision theory, Unsupervised learning, Hidden Markov models, Text mining, Sentiment analysis, Chatbot, Natural Language Processing, Automatic translation.

Programme

- Bayesian decision (2h)
- Gaussian mixture model (2h)
- Hidden Markov chain (2h)
- Practical work on Bayesian learning (2h)
- Computational linguistic, NLP and practical Text Mining (8h)
- Restitution of a scientific reading by group (4h)

Learning outcomes

- Select the appropriate ML method(s) for their classification problem, considering different criteria.
- Develop programs using these methods to analyze their own data.
- Implement a processing chain to interpret texts (e.g. tweet).
- Become familiar with modern text mining techniques and tools and Read recent research papers on the topics mentioned.

Independent study

Objectifs : The students, divided into groups, will have to read and understand a scientific article on one of the subjects covered in the course. The articles will be proposed by the supervisors or by the students (after agreement with the supervisors).

Méthodes : The work will be done independently and the oral presentation will take place during a group presentation session.

Core texts

M. R. Gupta and Y. Chen, • *THEORY AND USE OF THE EM ALGORITHM*, Foundations and Trends in Signal Processing, Vol. 4(3), pp. 223–296, 2011
M. Watanabe and K. Yamaguchi *THE EM ALGORITHM AND RELATED STATISTICAL MODELS*, Statistics: Dekker series of textbooks and monographs, 2004
Michael W. Berry, Jacob Kogan *TEXT MINING: APPLICATIONS AND THEORY*, Willey, 2010

Assessment

Grade = 50% knowledge + 50% know-how
Knowledge mark = 100% final exam
Know-how mark = 50% for practical and 50% scientific paper restitution