Statistical machine learning

Obozinski Guillaume

<table>
<thead>
<tr>
<th>Cursus</th>
<th>Sem.</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial engineering</td>
<td>MA1, MA3</td>
<td>Opt.</td>
</tr>
<tr>
<td>Ing.-math</td>
<td>MA1, MA3</td>
<td>Opt.</td>
</tr>
<tr>
<td>Mathématicien</td>
<td>MA1, MA3</td>
<td>Opt.</td>
</tr>
<tr>
<td>Robotics, Control and Intelligent Systems</td>
<td></td>
<td>Opt.</td>
</tr>
<tr>
<td>Statistics</td>
<td>MA1</td>
<td>Opt.</td>
</tr>
</tbody>
</table>

Contact language: English

Credits: 5

Session: Winter

Semester: Fall

Exam: Written

Workload: 150h

Weeks: 14

Hours:
- Lecture: 2 weekly
- Exercises: 2 weekly

Number of positions

Summary
A course on statistical machine learning for supervised and unsupervised learning

Content
- Introduction: supervised and unsupervised learning, loss functions, train and test errors, bias-variance tradeoff, model complexity and overfitting, linear regression, k-nearest neighbors.
- Regression: linear regression, model selection, ridge and Lasso.
- Local averaging methods: k-nearest neighbor, Nadaraya-Waston estimators
- Simple validation, cross-validation and leave-one-out.
- Classification: linear discriminant analysis, logistic regression.
- Nonparametric regression: smoothing splines, reproducing kernel Hilbert spaces.
- Support vector machines and kernelized linear regression.
- Tree-based methods: classification and regression trees, bagging, random forests.
- Boosting: AdaBoost, boosted trees.
- Deep learning: introduction to neural networks (architecture, backpropagation, CNNs)
- Unsupervised learning: principal component analysis, k-means, Gaussian mixtures and the EM algorithm.

Learning Prerequisites

Required courses
- Analysis, Linear Algebra, Probability and Statistics, Linear Models

Important concepts to start the course
This course introduces machine learning with a statistical and mathematical perspective. Prior to following this course, the student must have very good knowledge of basic probability and statistics (statistical modeling and inference, linear regression). Some of the concepts used in course: abstract conditional expectations, multivariate linear regression, convexity, strong convexity, vector space, Hilbert space.

Learning Outcomes
By the end of the course, the student must be able to:
- Formulate appropriate models for empirical data
- Estimate the parameters of a statistical model
• Interpret the fit of a model to data
• Justify the choice of a model/technique to analyze empirical data
• Implement statistical learning algorithms
• Explain the mathematical/statistical mechanisms of most common machine learning algorithms
• Assess / Evaluate underfitting / overfitting of ML algorithms

Transversal skills
• Write a scientific or technical report.
• Make an oral presentation.
• Demonstrate the capacity for critical thinking
• Take feedback (critique) and respond in an appropriate manner.
• Demonstrate a capacity for creativity.

Teaching methods
Ex cathedra lectures, exercises and computer practicals in the classroom and at home, and a project in a group of 3 students.

Assessment methods
Written final exam (70%) + Project of implementation or application on real data of a model/algorithm based on a classical research paper describing an important method from the literature. (30%)

Resources
Bibliography
• James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning, with Applications in R. Springer.

Ressources en bibliothèque
• Applied Predictive Modeling / Kuhn & Johnson
• Pattern Recognition and Machine Learning / Bishop
• Understanding machine learning (electronic version)
• Elements of Statistical Learning (electronic version)
• Introduction to Statistical Learning, with Applications (electronic version)
• Computer Age Statistical Inference / Efron & Hastie

Moodle Link