Data spaces

01RLPOV

A.A. 2018/19

Course degree
Master of science-level of the Bologna process in Computer Engineering - Torino

Course structure

Teaching	Hours
Lezioni	40
Esercitazioni in aula	20

Teachers

Teacher	Status	SSD	h.Les	h.Ex	h.Lab	h.Tut	Years teaching
Vaccarino Francesco (/pls/portal30/sviluppo.scheda_pers_swas.show?m=11485)	Ricercatore	MAT/03	40	0	0	0	2

Teaching assistant
Espandi

Context

SSD	CFU	Activities	Area context
MAT/03 SECS-S/01	4 2	F - Altre (art. 10, comma 1, lettera f) F - Altre (art. 10, comma 1, lettera f)	Altre conoscenze utili per l'inserimento nel mondo del lavoro Altre conoscenze utili per l'inserimento nel mondo del lavoro

Statistiche superamento esami (/pls/portal30/esami.superi.grafico?p_cod_ins=01RLPOV&p_a_acc=2019)

Anno accademico di inizio validità

2018/19

Course description

The main objective of this course is to provide the students with solid mathematical bases of the major techniques used in supervised and unsupervised statistical (machine) learning with a special focus on their geometrical aspects.

Expected Learning Outcomes

- Knowledge and understanding of the main learning techniques (detailed knowledge of the mathematics behind the most popular learning techniques; be acquainted of the limitations of the various techniques; awareness of the structural problem as e.r. the curse of dimensionality)

- Practical application of the acquired knowledge (ability to identify the applicability domain of the various techniques with respect of the nature of data; ability to extract information from real and simulated data by applying the learned techniques via software application or development).

Prerequisites

The students are assumed to know the topics covered by standard courses in mathematics given in the Bs.D. in Engineering. Furthermore, a knowledge in basic probability and statistics is required: pdf, normal, expectation, mean, variance – covariance. SVD will be explained along the course.

Course syllabus

GENERALITIES ON DATA REPRESENTATION.
Metric and topological spaces. Coordinatization. Curse of dimensionality; The Law of Large Numbers; the Geometry of High Dimensions: properties of the Unit Ball; Generating Points Uniformly at Random from a Ball; Gaussians in High Dimension; Random Projection and Johnson-Lindenstrauss Lemma.

STATISTICAL LEARNING
What Is Statistical Learning? Why Estimate f? How Do We Estimate f? The Trade-Off Between Prediction Accuracy and Model Interpretability. Supervised Versus Unsupervised Learning. Regression Versus Classification Problems. Assessing Model Accuracy. Measuring the Quality of Fit. The Bias-Variance Trade-Off.

LINEAR REGRESSION
Simple Linear Regression. Multiple Linear Regression.

CLASSIFICATION
An Overview of Classification. Logistic Regression. Multiple Logistic Regression. Logistic Regression for >2 Response Classes. Linear Discriminant Analysis. Quadratic Discriminant Analysis. Comparison of Classification Methods. K-Nearest Neighbours.

RESAMPLING METHODS
Cross-Validation. Leave-One-Out Cross-Validation. k-Fold Cross-Validation. Cross-Validation on Classification Problems. The Bootstrap. Mathematical justification of these methods.

TREE-BASED METHODS
The Basics of Decision Trees. Regression Trees. Classification Trees. Advantages and Disadvantages of Trees. Bagging, Random Forests, Boosting.

SUPPORT VECTOR MACHINES. Classification Using a Separating Hyperplane. The Maximal Margin Classifier. Construction of the Maximal Margin Classifier. The Non-separable Case. Support Vector Classifiers. Support Vector Machines. SVMs with More than Two Classes. OVO and OVA. Relationship to Logistic Regression. Kernel Methods. Geometrical interpretation via Segre embedding.

UNSUPERVISED LEARNING. SVD. Principal Components Analysis. Independent component analysis. Multidimensional Scaling (MDS) as an optimization problem.

Additional information

Course structure

Lessons, exercise classes and laboratory sessions will be given. There will be three hours of lesson per week plus 1.5 one hour and half of exercises / further lessons. These latter are split into two group: one for mathematical engineering and the other for software enngineering.

Reading materials

An Introduction to Statistical Learning
with Applications in R
Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

https://www.amazon.it/Introduction-Statistical-Learning-Applications/dp/1461471370/ref=sr_1_1?ie=UTF8&qid=1474898531&sr=8-1&keywords=An+Introduction+to+Statistical+Learning+with+Applications+in+R

freely available at

http://www-bcf.usc.edu/~gareth/ISL/

Assessment and grading criteria

Modalità di esame: prova orale obbligatoria; elaborato scritto individuale;

The goal of the exam is to test the knowledge of the candidate about the topics included in the official program and to test their skills in analysing data using the methods explained in the course.

The exam consists in two parts: first the candidate will write a technical relation "tesina" on the analysis of a data set performed by using the methods taught in the course. This will be software independent i.e. one can use Orange, R, Matlab, Rapidminer, Python, C++ etc. according to their knowledge or willingness.
Once the "tesina" is approved by the professor, then the student is allowed to present it in an oral exam (about 20.min) during which the professor will also ask questions on the theoretical aspects of the methods used in the tesina.

Sample work from the previous years will be provided on the website.

CAVEAT: students following this course as a submodule of Statistical Models will give the exam according to the rules fixed thereby.