Subject: Machine Learning: theory and applications (ML)
Code: 32419
Institution: Escuela Politécnica Superior
Degree: Master's program in Research and Innovation in Information and Communications Technologies (i^2-ICT)
Level: Master
Type: Elective [Computational Intelligence]
ECTS: 6

COURSE GUIDE: Machine Learning: Theory and applications (ML)

Academic year: 2017-2018

Program: Master's program in Research and Innovation in Information and Communications Technologies (i^2-ICT)
Center: Escuela Politécnica Superior
University: Universidad Autónoma de Madrid

Last modified: 2015/05/19

1. ASIGNATURA / COURSE (ID)

Aprendizaje Automático: teoría y aplicaciones
Machine Learning: theory and applications (ML)

1.1. Programa / program

Máster Universitario en Investigación e Innovación en Tecnologías de la Información y las Comunicaciones (i^2-TIC)

Master in Research and Innovation in Information and Communications Technologies (i^2-ICT) [Officially certified]

1.2. Course code

32419

1.3. Course areas

Computer Science and Artificial Intelligence

1.4. Tipo de asignatura / Course type

Optativa [itinerario: Inteligencia computacional]
Elective [itinerary: Computational Intelligence]

1.5. Semester

First semester

1.6. Credits

6 ECTS

1.7. Language of instruction

The lectures are mostly in Spanish. Some of the lectures and seminars can be in English. Most of the course materials and lecture notes will be in English.

1.8. Recommendations / Related subjects

Knowledge of linear algebra, probability and statistics at an introductory level, as well as basic programming skills, are useful to follow the course.

Related courses are:

- Procesamiento de información temporal [Temporal information processing]
- Métodos bayesianos aplicados [Applied Bayesian Methods]
- Recuperación de Información [Information Retrieval]
- Minería Web [Web Mining]

1.9. Lecturers

Add @uam.es to all email addresses below.

Dr. Manuel Sánchez-Montañés Isla (Coordinator)
Computer Science Department
Escuela Politécnica Superior
Office: B-303
Phone: +34 914972290
e-mail: manuel.smontanes
Web: http://www.eps.uam.es/~msanchez

Dr. Luis F. Lago Fernández
Computer Science Department
Escuela Politécnica Superior
Office: B-307
Phone: +34 914972211
e-mail: luis.lago
Web: http://www.eps.uam.es/~lflago

Dra. Ana González Marcos
Computer Science Department
Escuela Politécnica Superior
Office: B-332
Phone: +34 914972234
e-mail: ana.marcos

1.10. Objetivos de la asignatura / Course objectives

El objetivo de esta asignatura es proporcionar formación al estudiante en paradigmas avanzados de aprendizaje automático y su utilización en aplicaciones prácticas. Se estudiarán las metodologías necesarias para llevar a cabo los diferentes pasos necesarios para resolver satisfactoriamente problemas reales: formalización del

problema, pre-procesamiento de los datos, identificación de datos relevantes, construcción del clasificador automático, cuantificación de su rendimiento, y validación y estimación de su precisión en datos futuros.

The goal of this course is to provide the student with training in advanced machine learning paradigms, and their use in practical applications. The student learns the methodologies needed to carry out the required steps to effectively solve real problems: problem formalization, data pre-processing, identification of relevant data, construction of automatic classifiers, assessment of their performance, as well as validation and estimation of their accuracy on unseen data.

At the end of each unit, the student should be able to:

UNIT BY UNIT SPECIFIC OBJECTIVES	
PART I: Introduction	
UNIT 1.- Introduction to Machine Learning	
1.1.	Present an overview of the field, showing related areas and applications.
1.2.	Identify classification, pattern recognition, optimization, decision and clustering problems.
1.3.	Identify the basic workflow for obtaining a predictor model.
1.4.	Validate and select the most adequate models for a particular problem.
UNIT 2.- Data pre-processing	
2.1.	Represent the data set in the convenient manner.
2.2.	Identify main concepts such as types of attributes, outliers, noise, duplicate records, missing values.
2.3.	Apply the basic pre-processing workflow: data cleaning, data integration, data transformation and data reduction.
UNIT 3.- Dimensionality reduction	
3.1.	Understand the concept of "curse of dimensionality".
3.2.	Understand the differences between feature extraction and feature selection, and between filter and wrapper methods.
3.3.	Apply multiple testing techniques.
3.4.	Express the dissimilarities between the two main methods of feature extraction: unsupervised geometry-driven and supervised task-driven approaches (PCA and LDA).
3.5.	Understand the principles of non-linear dimensionality reduction.
PART II: Supervised Learning	
UNIT 4.- Classification and regression	
4.1.	Understand the concept and assumptions of supervised learning.
4.2.	Identify which practical problems can be addressed with supervised learning techniques, and formalize the problems accordingly.
4.3.	Describe the assumptions, derivation, use and limitations of linear and logistic regression techniques.

4.4.	Explain the concepts of underfitting and overfitting, and the standard techniques to avoid them.
4.5.	Describe problems of classification and regression as particular cases of the GLM framework.
UNIT 5.- Bayesian decision theory	
5.1.	Understand the concepts and assumptions underlying Bayesian decision theory: a priori and a posteriori probabilities, costs.
5.2.	Apply them to derivate discriminant functions and the likelihood ratio test, and relate them to MAP and ML criteria.
5.3.	Apply this theory to derive different supervised learning techniques: generative learning algorithms, Gaussian discriminant analysis, Näive Bayes.
UNIT 6.- Density estimation	
6.1.	Understand the concept of density estimation, and its relationship with supervised learning.
6.2.	Understand the concept of Maximum Likelihood parameter estimation in density fitting, and derive the solution for the Gaussian case.
6.3.	Use different techniques of non-parametric density estimation: histograms, Parzen windows, kernels.
UNIT 7.- Decision trees	
7.1.	Explain the concept and elements of a decision tree.
7.2.	Derive a general strategy for constructing a decision tree depending on a "division criterion".
7.3.	Understand and formalize the requirements that an appropriate division criterion should satisfy, and review the practical differences between different criteria.
7.4.	Understand overfitting in a decision tree, and how to deal with it using pruning.
7.5.	Describe the differences between pre- and post- pruning.
UNIT 8.- Ensemble methods	
8.1.	Define the concept of ensemble learning, and understand the assumptions under which ensemble learning makes sense.
8.2.	Understand the bias-variance dilemma and its consequences.
8.3.	Apply bagging and boosting to make ensembles of classifiers.
UNIT 9.- Support Vector Machines	
9.1.	Understand the concept of VC dimension and compute it for simple examples.
9.2.	Express in an intuitive way why maximum margin linear separators are optimal, and know the tools to go deeper in the field on his/her own.
9.3.	Use Support Vector Machines to solve classification problems and interpret the results (support vectors, margin, decision line, etc.).

PART III: Unsupervised Learning	
UNIT 10.- Parametric models	
10.1.	Understand and describe the EM algorithm.
10.2.	Apply the EM algorithm to fit a mixture of Gaussians to a data set.
10.3.	Validate the different fits using AIC and BIC.
UNIT 11.- Non-parametric models	
11.1.	Express the differences between parametric and non-parametric models for unsupervised learning.
11.2.	Apply the k-means algorithm to cluster a data set.
11.3.	Apply the fuzzy c-means algorithm to cluster a data set.
11.4.	Apply agglomerative clustering methods using different distance measures, visualize and interpret the results using dendrograms.
11.5	Discuss other clustering approaches.
UNIT 12.- Cluster validation	
12.1.	Describe in simple terms the problematic of cluster validation.
12.2.	Use different cluster validity indices to validate and compare the outcomes of clustering algorithms.

1.11. Course contents

PART I: Introduction

1. Introduction to Machine Learning
    - Basic concepts (definitions, types of approaches)
    - Machine Learning applications
    - Components and stages in a Machine Learning project
    - Evaluation, regularization and model selection (error measures, confusion matrix, ROC analysis, PR, validation)
    - Introduction to the Machine Learning software Weka

2. Data pre-processing
    - Normalization and standardization
    - Discretization
    - Processing of special variables: dates, series, categorical and nominal attributes
    - Missing values and outliers

3. Dimensionality reduction
    - The curse of dimensionality
    - Variable selection (filter and wrapper methods)
    - Variable construction (LDA, PCA, spectral methods)

PART II: Supervised Learning

4. Classification and regression
    - Linear regression (LMS)
    - Logistic regression and classification
    - Underfitting and overfitting
    - Generalized Linear Models

5. Bayesian decision theory
    - Likelihood ratio test
    - Bayes risk
    - MAP and ML criteria
    - Generative Learning Algorithms, Gaussian Discriminant Analysis
    - Naïve Bayes

6. Density estimation
    - Parametric (Maximum Likelihood, EM)
    - Non-parametric (histograms, Parzen Windows, kernels)

7. Decision trees
    - Definitions
    - Decision tree construction. Division criteria (expected error, Gini index, information gain)
    - Pruning

8. Ensemble methods
    - Definition of ensemble learning. Precision and diversity assumptions
    - Statistical, numerical and representational problems of single classifiers
    - Bagging and boosting
    - The bias-variance tradeoff

9. Support Vector Machines
    - The Vapnik-Chervonenkis dimension
    - Structural Risk Minimization
    - Maximum margin classifier, optimal separating hyperplane
    - Separable and non-separable cases
    - Nonlinear problems. Cover's theorem. Kernels
    - The SMO algorithm

PART III: Unsupervised Learning

10. Parametric models
    - Model based clustering and density estimation
    - Maximum Likelihood. The EM algorithm
    - Gaussian Mixtures
    - Validation criteria. AIC and BIC, cross-validation

11. Non-parametric models
    - Unsupervised classification (clustering)
    - K-means algorithm
    - Fuzzy c-means
    - Hierarchical clustering

12. Cluster validation
    - Geometric indices
    - Information and likelihood based indices
    - Negentropy based indices

1.12. Course bibliography

1. T. Hastie, R. Tibshirani, J.H. Friedman. The Elements of Statistical Learning. Springer 2009
2. C.M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006
3. R.O. Duda, P.E. Hart. D.G. Stork; Pattern Classification; Wiley, 2000
4. T.M. Mitchell. Machine Learning. McGraw-Hill, 1997

1.13. Coursework and evaluation

The course involves lectures, lab assignments and a final exam.

- In the ordinary exam period, the evaluation will be made according to the following scheme:

    - 50 % Lab assignments
    - 50 % Final exam

    It is necessary to have a pass grade (≥ 5) in both the lab assignments and the final exam to pass the course. This applies also to the extraordinary exam period.

    The grades of the individual parts are kept for the extraordinary exam period.

- In case of a fail grade in the ordinary exam period, in the extraordinary exam period the student has the opportunity to turn in all the lab assignments with corrections.

  The grade will be determined by:

  - 50 % Lab assignments [if the lab assignments are not turned in, the grade obtained in the ordinary exam period will be used]
  - 50 % Final exam [mandatory]