Machine Learning and Computational Statistics

Michalis Titsias, Lecturer, AUEB, mtitsias@aueb.gr

Overview

The course introduces Machine Learning to a postgraduate level. Machine Learning aims at getting computers to automatically learn from data so that to perform complex prediction tasks and discover hidden patterns. Machine Learning has emerged as one of the most exciting and fundamental areas of Data Science that combines traditional statistical learning and pattern recognition with modern computing algorithms. In this course, students will learn advanced machine learning methods and they will gain practical experience in using these methods for real Data Science problems.

Key Outcomes

By completing the course the students will be able:

- To fully understand classical statistical learning methods such as linear regression, logistic regression, K-nearest neighbors classification and others.
- To know more advanced techniques such as neural networks and support vector machines.
- To know how to train models using optimization algorithms and perform model selection using cross validation.
- To know several unsupervised learning methods such as data clustering and probabilistic latent variable modeling.
- To visualize high dimensional data by performing dimensionality reduction techniques.
- To gain practical intuition about how to select an appropriate Machine Learning algorithm in order to solve a certain Data Science problem.

Requirements and Prerequisites

The course requires a good knowledge in applied maths as it can be obtained from undergraduate programs in maths, computer science or engineering. Especially good knowledge in multivariate calculus, matrix algebra and probability theory are essential for this course.

Required Course Material

There is no required textbook. All course materials will be provided in class and will be available for downloading.

Students will often need to bring their laptops in class in order to try out interactively several Machine Learning algorithms that will be presented.

Books

There are many books on the subject. Some recent books based on which the material of this course is based on are:

- Pattern Recognition and Machine Learning by Christopher M. Bishop, Springer (2006).
- Bayesian Reasoning and Machine Learning by David Barber, Cambridge University Press, (2012).
- Machine Learning: a Probabilistic Perspective by Kevin Murphy, MIT Press, (2012).
- Elements of Statistical Learning by Hastie, Tibshirani and Friedman, Springer, (2009).

Software/Computing requirements

Students will be able to run and work with most of the course material on their own computers. To to so, they should download and install the Anaconda Python distribution, available for all platforms at: https://www.continuum.io/downloads.

Grading

Students will be graded on their performance in the final exam and a project assignment in Python. More precisely, the grading is divided as follows:

1. Final exam will count 50% towards the final grade.
2. Two homework assignments will count 20% towards the final grade.
3. Project assignment will be announced after the sixth unit and will ount the remaining 30% of the final grade.

Course Syllabus

The course comprises ten units of three hours each.

Unit 1: Introduction to the basics of Machine Learning

We first define what is Machine Learning and discuss few examples of real applications. Then, we explain what is supervised and unsupervised learning and introduce the concepts of training and test data. We finally discuss major important concepts related to many Machine Learning systems, such as overfitting, underfitting and the fact that there is no “Free Lunch”, i.e. in order to learn from data we need to make assumptions by introducing models and prior knowledge or by applying regularization.

Unit 2: Model selection and probabilistic modeling

When learning from data we always face the problem of uncertainty about “which model is the best?”. This is due to many reasons such as finiteness of the training data set, noise in the data generation process and our own lack of understanding about what model should we use. Thus, the need of model selection naturally arises and in this course we will explain the technique of cross validation for model selection. To rigorously deal with uncertainty when building Machine Learning systems we use probabilistic modeling and the second part of this course will introduce the main concepts of probabilistic modeling and statistical estimation using maximum likelihood and regularization.

Unit 3: Linear regression and logistic regression

Linear regression is the simplest supervised learning method that dates back to Gauss and Legendre (who were applying Data Science to data in astronomy more than two centuries ago!). Here, we will explain the least squares method for fitting a linear function to data and provide also a probabilistic interpretation. Then, we will describe logistic regression which is the simplest supervised learning algorithm for binary classification. Generalizations of logistic regression to the multi-class case will also be discussed.

Unit 4-5: Non-linear models and neural networks

In modern applications the underlying functions we wish to learn from data can be highly complex and non-linear, thus non-linear regression and classification methods are needed in practice. Here, by generalizing the linear models explained previously, we present the linear-parameter models, also called radial basis functions, that add non-linearities by applying fixed transformations to the initial input data. Then, we present the current most powerful technique for supervised learning that is based on neural networks. We will also discuss scalable training of neural networks to Big Data using stochastic gradient descent optimization.

Unit 6: K-nearest neighbors and naive-Bayes

The supervised classification techniques presented previously are based on discriminative models. Here, we describe an alternative, called generative, approach to classification based on modeling the density of the observed input data. We start with a simple non-parametric technique called K-nearest neighbors that is based on memorizing the training set. Then, we present the general framework of Bayes classification by modeling the class conditional densities. Finally, we focus on naive Bayes which is a simple and scalable Bayes classifier with numerous applications such as Spam filtering and text mining.

Unit 7: Support vector machines

Here, we describe support vector machines which is one of the most successful classification methods in Machine Learning. We explain in detail the concept of margin and margin maximization and discuss the kernelized version of support vector machines, based on the so-called kernel trick, that deals with non-linear decision boundaries.

Unit 8-9: Data clustering

Here, we describe data clustering algorithms which consist of unsupervised learning methods. We start by explaining the k-means algorithm and then we introduce probabilistic clustering using mixture models. We discuss in detail the special case of mixtures of Gaussians and mixtures of Bernoullis and show how to train these models using the Expectation-Maximization (EM) algorithm. We also discuss the EM algorithm as a more general optimization tool for training complex probabilistic models with latent variables.

Unit 10: Dimensionality reduction

Here, we present the second major class of unsupervised learning methods that learn how to reduce the dimensionality of high dimensional data. Dimensionality reduction can be useful for data compression, data visualization or as a preprocessing feature extraction step when building a supervised learning system. We will detail the classical linear algorithms such as principal component analysis (PCA), probabilistic PCA and factor analysis. We will also discuss advanced non-linear techniques based on neural networks and autoencoders.