**SYLLABUS OF THE DATA MINING / MACHINE LEARNING COURSE**

**Prof**

**Professors: Mauro Castelli, Assistant Professor at NOVA IMS, Universidade NOVA de Lisboa, Portugal**

**Assistant: Illya Bakurov, PhD student at NOVA IMS, Universidade NOVA de Lisboa, Portugal**

DAY 1: This class introduces basic concepts about machine learning (ML). The general idea of machine learning is introduced and a formal definition is provided. Several applications where ML is used in our daily life are presented to allow students to fully understand the breakthrough provided by this discipline. Subsequently, the concepts of training set, test set, and features are introduced and the differences between supervised and unsupervised learning are formally defined.

The second part of the class formalizes the general statement of a regression problem, where the objective is to predict a real-valued output based on one or more input values. Subsequently, some of the most commonly used regression techniques are presented. In particular, the concept of model representation is introduced by considering different ML techniques. Linear regression is the first technique introduced in this class, and the fundamental concepts of loss function and optimizer are formally described. Particular attention is given to gradient descent, considering its wide usage in ML. Finally, the advantages and limitations of linear regression are discussed.

In the last part of the class, the general statement of a classification problem is introduced, highlighting its main differences with a regression problem. Simple examples are considered to get a complete understanding of the problem and to differentiate between binary classification and multi-class classification problems.

DAY 2: this class presents logistic regression as wells as common loss functions for classification problems. Subsequently, the concept of confusion matrix is introduced, as well as different evaluation metrics (i.e., accuracy, recall, F-measure, AUROC). The selection of the evaluation metric in the case of unbalanced datasets is also discussed.

Subsequently, some of the most commonly used ML techniques for addressing classification and regression problems are presented. The class starts by presenting the K-Nearest Neighbours (K-NN) algorithm, a technique that uses the concept of similarity among the observations for addressing the underlying regression/classification problem. The algorithm is intuitively presented by using some examples and, subsequently, all the steps are analyzed and formalized. The importance of correctly tuning the value of the parameter K is discussed and the advantages and disadvantages of K-NN are analyzed.

In the final part of the class, model selection will be covered. As a first step, the concepts of approximation and estimation errors are introduced. This allows presenting an important problem that affects several ML techniques: overfitting. The concept is explained by considering different examples, and by analyzing the performance of a ML model on both training and test sets.

Subsequently, possible solutions for counteracting overfitting are presented: in particular, Lasso and Ridge regularization are taken into account, and their effect is evaluated over different problems. Finally, the concept of K-fold cross-validation is introduced, followed by a discussion about its importance for guaranteeing that robust and reliable models are generated.

DAY 3: This class introduces decision trees, a supervised ML technique that can be used to solve both regression and classification problems.

The explanation considers the case of a classification problem, but the concepts are then extended to regression tasks. A decision tree uses a tree representation to represent a solution in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. Thus, the criteria commonly used for selecting the attribute of the internal nodes are introduced: Information Gain and Gini Index. Finally, the process of building a decision tree using Information Gain as a guiding criterion is detailed.

The second part of the class discusses ensemble learning. Ensemble learning uses multiple machine learning models to try to make better predictions on a dataset. An ensemble model works by training different models (called weak learners) on a dataset and having each model make predictions individually. The predictions of these models are then combined in the ensemble model to make a final prediction. This concept is introduced by considering several real-world problems, to provide an intuitive understanding of the properties of ensemble models.

Subsequently, Random Forest, an ensemble model that consists of different decision trees, is presented. The presentation is followed by a discussion about ensemble diversity, bias/variance trade-off, and overfitting.

DAY 4: This class is dedicated to dimensionality reduction techniques. Dimensionality reduction, the process of reducing the number of variables under consideration by obtaining a set of principal variables, is presented by considering some datasets commonly used in ML literature. Dimensionality reduction is an unsupervised ML technique. Thus, a different approach with respect to the one used by the techniques introduced in previous classes of the course must be used for tackling this ML task.

After formalizing the problem, the main uses of dimensionality reduction techniques are discussed: compression to speed up learning algorithms as well as visualization of complex datasets.

Subsequently, different dimensionality reduction techniques are presented. The first technique is Principal component analysis (PCA) that is formally described by recalling mathematical concepts associated with covariance matrix and eigenvectors. This allows formulating the problem of extracting the principal components of a dataset (i.e., the most ”important” variables) as the diagonalization of a matrix.

After presenting the PCA algorithm, guidelines on how to choose the number of principal components are provided. The concepts behind PCA are explained considering both practical examples as well as by taking into account the mathematical interpretation of the whole process.

The second part of the class introduces the general idea of clustering (i.e., to group unlabeled data points such that “similar” points will be assigned to the same cluster) by considering several examples. Subsequently, the clustering problem is formalized as an optimization problem and different optimization objectives based on the concept of similarity are presented. A taxonomy of clustering techniques is outlined.

After this overview of clustering techniques, the K-means algorithm is introduced. This is the most well-known clustering algorithm and its properties and assumptions are deeply discussed. More in details, issues associated with the initialization step of the K-means algorithm are discussed and possible solutions are presented. The “elbow method”, a technique for determining a suitable value for the number of clusters to be formed is presented.

Subsequently, hierarchical clustering is presented and the differences between agglomerative and divisive algorithms are discussed.

The last part of the class is dedicated to the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. Details about the implementation are provided and a discussion of its advantages with respect to the K-means algorithm concludes the class.