COMPLEX TRAJECTORIES

Educational Data Mining

May 2023

Co-funded by the Erasmus+ Programme of the European Union

Author:

Vera Miguéis

The following documents are embedded in a whole MOOC course on Longitudinal analysis techniques developed in the framework of the same Complex Trajectories project. Those interested in doing the MOOC should consider the following procedure:

1. Access the AULAbERTA space through the link: https://aulaberta.uab.pt/

2. Register in the platform a. Select one of the available languages (Portuguese or English) b. Follow the instructions given in the platform

3. After creating the account, they should access the MOOC Longitudinal Analysis through the link: https://aulaberta.uab.pt/course/view.php?id=94

Co-funded by the Erasmus+ Programme of the European Union

Educational data mining

Vera Migueis | Uporto (vera.migueis@fe.up.pt)

Co-funded by the Erasmus+ Programme of the European Union

Welcome to the educational data mining module within unit 2!

Contents

- Introduction
- Literature review
- Data Mining Techniques | Predictive Techniques
  - K-NN
  - Decision tree
  - Random Forests
  - Naïve Bayes
  - Support Vector Machines
  - Neural Networks
  - ....
- Case studies

Co-funded by the Erasmus+ Programme of the European Union

All over this module we will pass through a brief introduction to this field, a literature review on the topic and then we will get to know some data mining techniques, and Ill be mainly focusing on the predictive techniques, namely k-nearest neighbor, decision trees, random forests, naïve bayes, svm and neural networks. I'll finish with a description of two case studies in the scope of educational data mining.
I hope this module can motivate you for this quite recent world of data analytics.

Educational data mining

Co-funded by the
Erasmus+ Programme
of the European Union

Introduction

- **Technology innovation** is transforming the world
- The development of **learning technologies has enabled to track student learning** (Dimicet al., 2019)

DATA » INSIGHTS » ACTION

Co-funded by the Erasmus+ Programme of the European Union

We know that educational institutions in general are now collecting data. For example. institutions are using technology to support our students and professors and from these technologies they can obtain data. So, the development of these learning Technologies are enabling as professors and managers to monitor students learning. Indeed, the data that is being collected may enable us to collect insights that then may be used to support institutions actions.
For example, with the data collected we can get insights on the reasons that promote dropout out, and having identified the motivations, we may design strategies to mitigate them. For example, we may need to redefine the experience provided to certain students.

Introduction

Educational data Mining

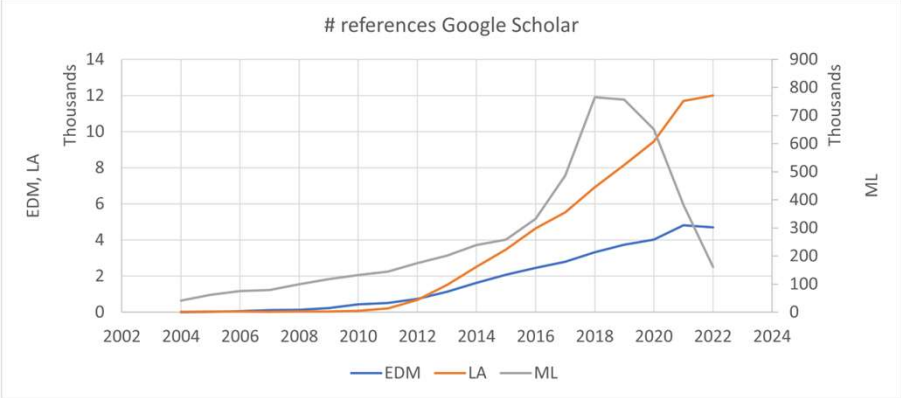
Learning Analytics

- EDM and LA aim to extract useful **knowledge** for **supporting teaching and learning** (Siemens and Baker, 2012)
- EDM and LA have been the focus of **intensive research in the past decade** (Du et al, 2020)

Co-funded by the Erasmus+ Programme of the European Union

The two fields in the literature that are exploring students' data are educational Data Mining and learning Analytics. There are some authors that distinguish these two concepts, but I belong to the group of researchers that do not see that much differences in the two. In the end both the educational mining and learning analytics are about getting some knowledge to support the teaching and learning processes. These two topics have drawn huge attention in the last decade, and this can be seen in this graph.

Literature review

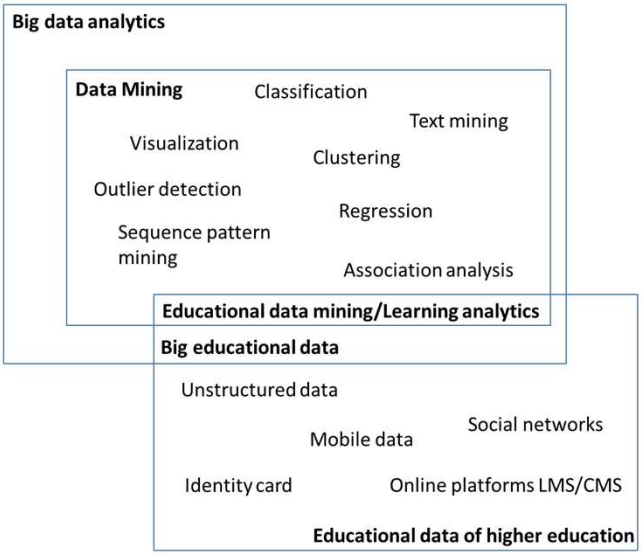


Papers on EDM and LA in January 2023 (obtained from google scholar)

Co-funded by the Erasmus+ Programme of the European Union

The number of studies on EDM in blue and the number of studies on learning analytics has indeed grown a lot in the last 20 years.

Literature review

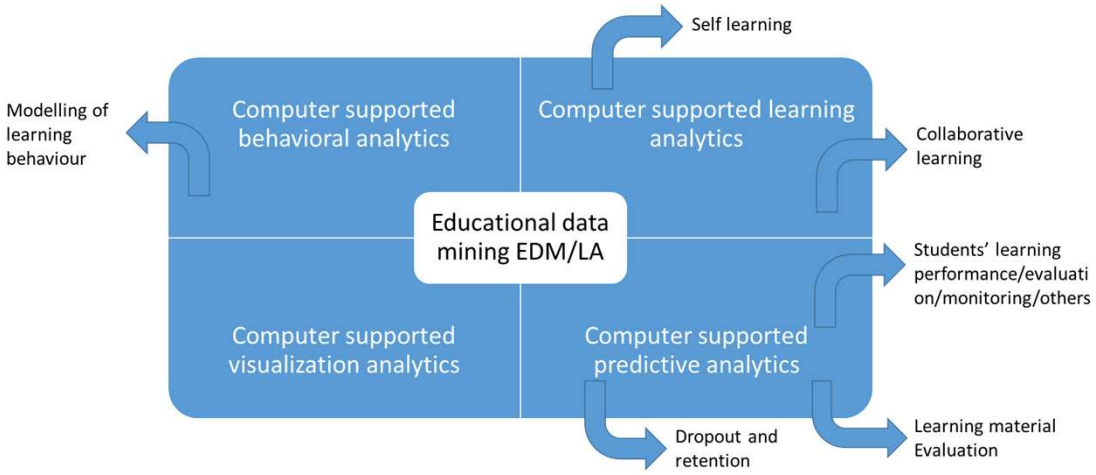


Data mining (EDM and LA) use in the educational field (adapted from Aldowah et al, 2019)

Co-funded by the Erasmus+ Programme of the European Union

This picture tries to frame educational data mining and learning analytics into domains that are quite trendy right now. When talking about huge volumes of data, we can refer to big data analytics. Within the umbrella of big data analytics, we find data mining that accommodates a big set of techniques data support the knowledge acquisition. We may be talking about clustering techniques, visualization, outlier detection and text mining.
In what regards the educational data of higher education, that supports these techniques, we can refer to unstructured data, such as text, social networks data and data from navigation in the LMS, as well as on structured data.

Literature review



Co-funded by the Erasmus+ Programme of the European Union

Data mining schemes in higher education (adapted from Aldowah et al, 2019)

In what regards the main dimensions of educational data mining and LA, we have computer supported behavioural analytics, computer supported learning analytics, visualization and then the predictive analytics. This last have been deserved particular attention, namely in topics such as dropout and retention, prediction of performance.

Thank you!

vera.migueis@fe.up.pt

Educational data mining

Vera Miguéis | Uporto (vera.migueis@fe.up.pt)

Co-funded by the Erasmus+ Programme of the European Union

Data mining models

Co-funded by the
Erasmus+ Programme
of the European Union

Lets move now to the data mining models, with a more technical perspective.

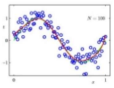
Data mining techniques

Predictive

Classification

Learns a method for predicting the observation class from pre-labeled (classified) observations

Regression



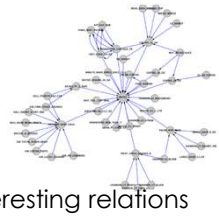
Learns a method for predicting a continuous attribute

Descriptive

Clustering

Finds natural "natural" grouping of observations given un-labeled data

Association



Method for discovering interesting relations between variables in large databases

Co-funded by the Erasmus+ Programme of the European Union

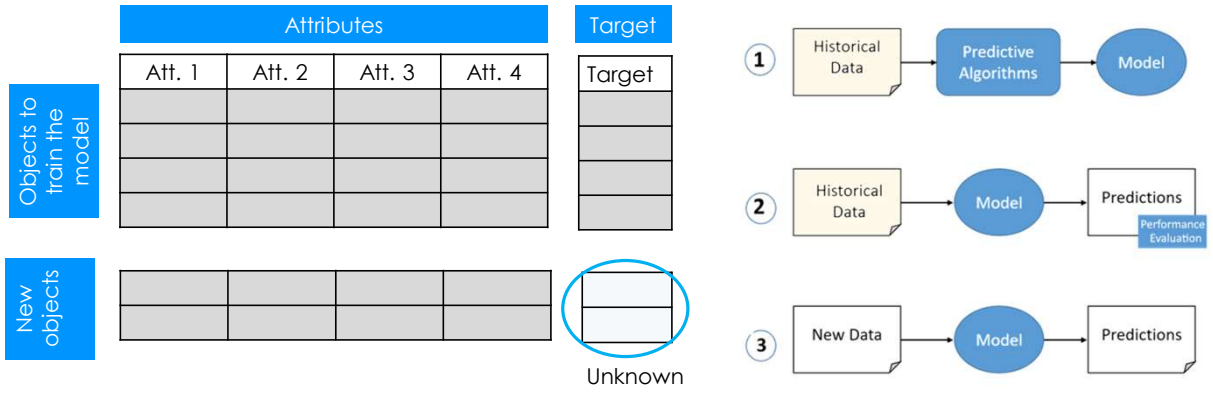
Data mining techniques that may support educational mining, are usually classified as predictive and descriptive. Focusing on the descriptive, we have clustering and association analysis. In the clustering analysis, we are interested in grouping observations given their similarities. It may mean grouping students, professors and even courses. This can be used to design targeted actions to each segment of clusters.

Regarding association analysis, we have for example the identification of links between events. For example, someone that fail certain course also fails another one.

Then, moving to the predictive components, we have mainly classification and regression. In both, we are interested in predicting the values of target variables given the values of certain explanatory variables. For example, we may be interested in predicting the grade a student may have given age, gender and so. This is a regression problem because our target variable is numeric. A classification problem may be predicting if a student will droupout or not (since the target variable is categorical, this is called a classification problem).

I believe that most of you already know regression, but perhaps not with this perspective of predicting. So, when doing a regression usually in statistics we are focused in obtaining the best fits to our data. In data mining context we are not interested in getting the best fit but to be able to predict the value for new observations. So we are more focused on generalization.

Predictive models



Co-funded by the Erasmus+ Programme of the European Union

My purpose in this module it show you how predictive models work.
So, lets suppose we have a very simple problem in which we are interested in estimating the propensity of students to dropping out in the first year. So, dropping out would be our target variable. For that purpose, any analysts should define the variables they believe may impact churn event. Let's suppose we consider age gender; high school grade, and how far students lives from the university as predictors.
Note that these variables should be identified by an experts on the field.

In order to create a predictive model, we need to look at past data with some students having dropped out and some who didn't. Suppose we have this table already filled in, the idea of any predictive model is to learn with past data to establish a link between the attributes and the target. Having established this relationship, we want to use it to predict dropout for new observation, I mean new students!

Obviously, that before adopting such models, we need to evaluate its quality. For this, we need to picking again some historical data, charactering other students and apply our model and compare the predictions of our model with what actually has happened.

In case the model presents good performance, it is used for predicting dropout for new students .

Predictive models

Training Data

id	Gender	Age	Approved?
Student A	Male	20	no
Student B	Male	18	yes
Student C	Female	22	yes
Student D	Male	17	yes
Student E	Male	19	no
Student F	Male	19	no

Prediction Algorithm

If Gender="Male" and Age>=19
THEN Approved="No"

Co-funded by the Erasmus+ Programme of the European Union

This was very general, that's why I bring here a specific example. Let's suppose that we are interested in predicting if someone will get approached or not in a course. For that we will use age and gender.
To create a prediction model may involve only defining a rule that can identify those who were not approved. A rule can be something like this: if male and older than 19, then he/she won't get approved.

So I have the model but now I will check either this is or not a good model.

Predictive models

Model

If Gender="Male" and Age>=19 THEN Approved="No"

Testing Data

id	Gender	Age	Approved?	Prediction	
Student G	Female	22	yes	yes	☺
Student H	Male	17	yes	yes	☺
Student I	Male	19	yes	no	☹
Student J	Male	22	no	no	☺

Accuracy=3/4=75%

Unseen Data

id	Gender	Age	Approved?
Student K	Male	18	Yes

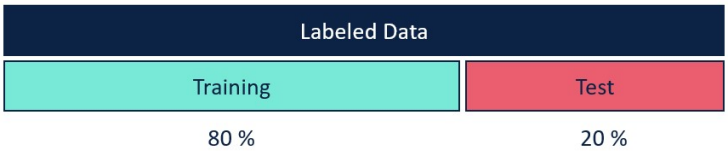
Co-funded by the Erasmus+ Programme of the European Union

So based on our model for example the first students would be predicted as approved. Only the third would be predicted as not approved.
Then we can compare the actual values with the predicted ones. And in this case we could conclude that this is quite good because in three out of four times the model predicted correctly. And obviously then I can use it for any other new data, I mean a new student that is coming I don't know if it was approved or not and I will make up my prediction. So since this is male and 18 I will say that the student will get approved.

This is the basic procedure for any the scientists to train data mining models. We need historical data to create a model and to test it and then apply to new data.

Predictive models evaluation

- Holdout approach:



Co-funded by the Erasmus+ Programme of the European Union

Moving ahead, just to elaborate a bit more on training and test set concepts. Usually, we have historical data and we split it into 80% for training and 20% for testing.
Some analysis use 70% for training and 30% for testing.
This is a basic approach and there are much more sophisticated ones.

Predictive models evaluation

- Metrics:
  - Confusion matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

  - Most widely-used metric:

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

Co-funded by the Erasmus+ Programme of the European Union

Before moving to the algorithms, I would like to show some of the performance measures that we analysts use to evaluate the quality of a model.
Perhaps you already heard about the contingency table. In this table we compare what actually happened with what the model predicts as happening. So, A value in this table corresponds for example to the number of students that were predicted as dropping out and indeed dropped out. C corresponds to those predicted as dropping out and in practice did not. These are called the false positive.

Having estimated these scores, we can easily get the accuracy: those predicted correctly (TP + TN) divided by the total number of observations.

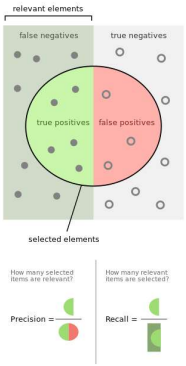
Predictive models evaluation

- Metrics:
  - Precision: Proportion rate of relevant within positive predicted; Measures the exactness

$$\text{Precision} = \frac{TP}{TP + FP}$$

  - Sensitivity (or recall): **True Positive** recognition rate; Measures the completeness

$$\text{Sensitivity} = \frac{TP}{P}$$



Co-funded by the Erasmus+ Programme of the European Union

Two other popular performance measures are precision and recall. In what regards the precision, it estimates the proportion of true positives among those predicted as positive.
The recall of sensitivity is the proportion of true positives among those positives.

Thank you!

vera.migueis@fe.up.pt

Educational data mining

Vera Miguéis | Uporto (vera.migueis@fe.up.pt)

Co-funded by the
Erasmus+ Programme
of the European Union

Examples of techniques

Co-funded by the
Erasmus+ Programme
of the European Union

Let’s now explore some techniques used for prediction purposes.

K-nearest neighbor

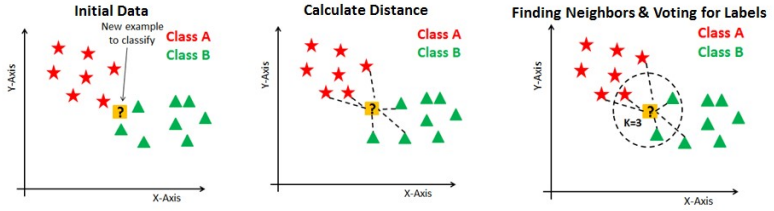
- "If it walks like a duck, quacks like a duck, and looks like a duck, then it's probably a duck."
  - Find the *k* training observations that are closest to the observation to be predicted, and classify the observation as the majority class of the k nearest training observations
  - A distance measure is used to evaluate how similar the observations are
  - k is defined by the data analyst

Co-funded by the Erasmus+ Programme of the European Union

Regarding the data mining techniques, we will start with k-nearest neighbor. Its basic idea is: If it walks like a duck, quacks like a duck, and looks like a duck, then it's probably a duck. Thus, the goal of this technique is to find the most similar historical observations in order to infer what may happen.
For example, when trying to predict students' performance, we look back at historical data to find similar students and from these we make our predictions. Again, please consider what are similar observations depends on the context. For example, when estimating performance, it can be the entrance grades, gender, age, etc.

It is important to highlight the need to set k value. This is, we may want to infer similarities based only in one historical observations or many. There are methods that support this choice, but this is out of the scope of this module.

K-nearest neighbour



- Problems:
  - Computationally expensive
  - Easily biased by irrelevant attributes

Co-funded by the Erasmus+ Programme of the European Union

This picture aims at illustrating how the algorithm works. Imagine that you want to infer if that observation in yellow will be part of class A or B. Considering only those two dimensions, we need to estimate the distances between any of the historical observations and the question mark. Let's suppose that we define K=3, then the class assigned will be class B because it’s the majority within the neighborhood.

This algorithm is very easy to understand but has some disadvantages. It is computationally expensive and is quite sensitive to the inclusion of attributes that are not impactful in determining the target variable.

K-nearest neighbour

- Suppose k=1:

Objects to train the model

Attributes		Target
Age	Enrolment grade	Attrition
26	10.3	1
18	12	0
18	14	0
18	11	1

Objects to test

Attributes		Target	
Age	Enrolment grade	Attrition	Prediction
25	10	1	1
18	14	1	0
18	14	0	0

☺ ☹ ☺

Accuracy=2/3

New objects

Age	Enrolment grade	Attrition
18	15	0
15	11	1

Co-funded by the Erasmus+ Programme of the European Union

In this example, if we assume k=1, we will, for each observation in the test set, to estimate the closest neighbour. For example, for the first element the closest element is the first. This is why first prediction is one.
Doing the same for the other observations, the accuracy is 2/3.
Thus, assuming that this is a good model we can estimate the chance of dropping out for the other observations.

Decision tree

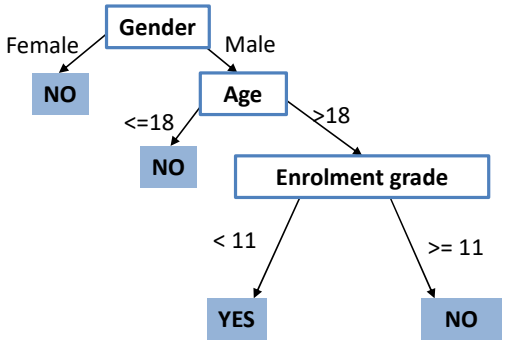
- Based on **recursive partitioning** of the data space
- **Tree-shaped structures** that represent sets of decisions which **generate rules** for the classification of a dataset
- This is an interpretable algorithm, i.e. provides insights for decision making

Co-funded by the Erasmus+ Programme of the European Union

An alternative algorithm for k-nn is decision trees. This algorithm is based on a recursive partitioning procedure. It leads to a tree shaped structure that represents a set of decisions which generate rules for the classification of observations. Decision trees have the particularity of being interpretable, because they result in rules that may provide insights for the decision maker.

Decision tree

- Ex:

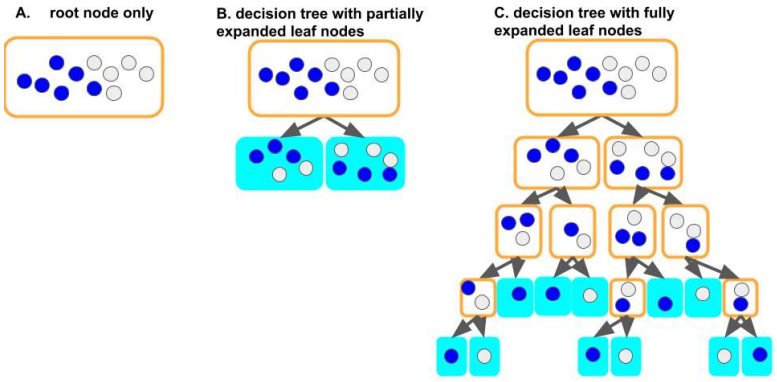


- A **decision tree is constructed automatically**, based on a splitting criterion, such as the **decrease in entropy**.

Co-funded by the Erasmus+ Programme of the European Union

This picture shows the output of the decision tree algorithm. Looking at this decision tree we can infer that if a student is female, she won't dropout. If it's male, and is less than 18 years old, he won't dropout. But if he/she is older than 18 and the enrolment grade is lower that eleven he will dropout.

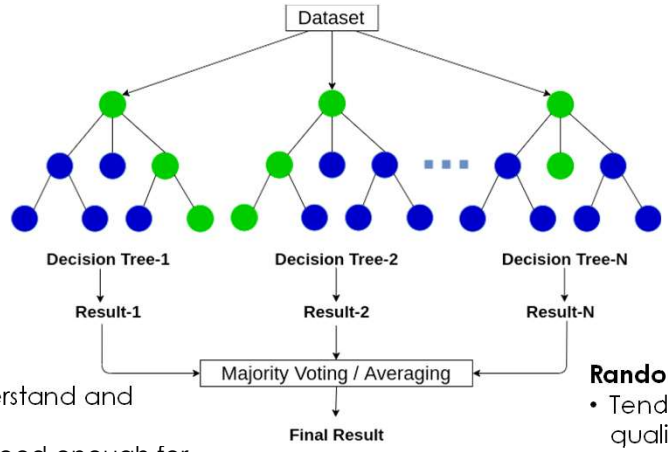
Decision tree



Co-funded by the Erasmus+ Programme of the European Union

How are the rules created? Based on data partitioning. We start with all the observation together but then we split them based on an attribute that is chosen based on a statistic criteria, in order to guarantee lower entropy in the subsamples. In the last picture it was the gender it was chosen. The partition is done recursibly until the entropy obtained in the end is minimum.

Random forests



Decision Tree:
- Easy to understand and implement
- Not power good enough for complex data

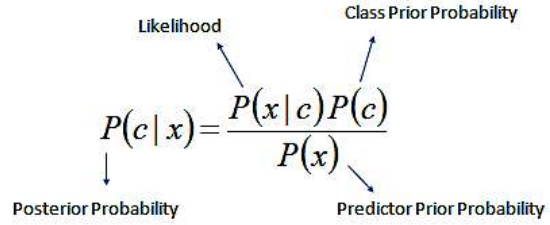
Random Forests:
- Tend to result in very high quality results
- Not easy to understand results

Co-funded by the Erasmus+ Programme of the European Union

Another alternative algorithm for prediction purposes is random forests. This algorithm is based on a set of decision trees, each being constructed from a sample of observations and based on different attributes.
The final predictions are based on majority voting, considering the outputs of all trees. This algorithm tends to provide very high quality results, although its not easy to understand their results, because you don't get some final rules as you get in decision trees.

Naïve Bayes

- Simple probabilistic classifier which is **based on Bayes theorem**
- Strong assumptions regarding independence



Objects to train the model

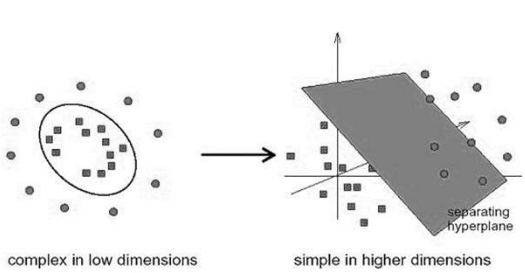
Attributes		Target
Age	Enrolment grade	Attrition
26	10.3	1
18	12	0
18	14	0
18	11	1

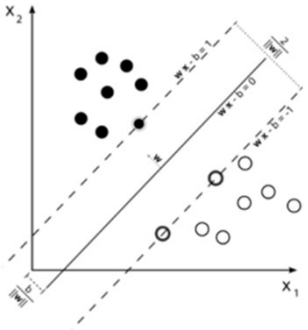
$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Co-funded by the Erasmus+ Programme of the European Union

Another classification algorithm is naïve bayes. It is based on conditional probabilities and is supported by the formula here presented, referring to the theorem of bayes. Let's consider that c represents our targets, so, for example, attrition or dropout. So, we need to estimate the probability of dropout or not given the attribute values.
So, for the example, here shown, we will need to estimate the probability of being 26 given the dropouts event. The same for the probability of being 18 years old and dropout. So based on historical data we estimate all these probabilities and obviously the probabilities are multiplied by the probability of dropping out and not and divide by the probability of X. So we'll have this posterior probability and leads to predicting a student as dropping out or not.

Support vector machines

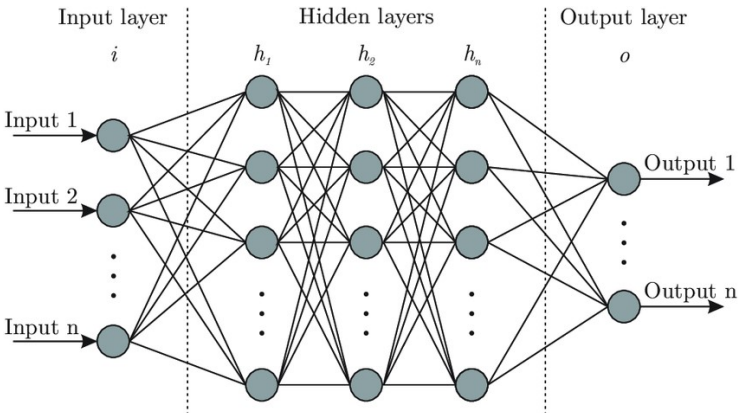




Co-funded by the
Erasmus+ Programme
of the European Union

Another algorithm is Supported Vector Machines. The rationale behind is quite simple: is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane, and it's estimated based on certain algebraic procedure.

Neural networks



- Can handle extremely complex tasks
- Almost impossible to interpret predictions

Co-funded by the Erasmus+ Programme of the European Union

Finally, the neural networks. So, this last algorithm that I decided to briefly present consist of thousands of neurons (or nodes) that are densely connected. In most neural network models, neurons are organized into layers. This includes an input layer, which includes neurons for all the provided predictor variables, then we have a hidden layer(s), and an output layer. The hidden layers of a neural network effectively transform the inputs into something that the output layer can interpret. The output layer returns a target label, or anything that you'll like to predict.

Thank you!

vera.migueis@fe.up.pt