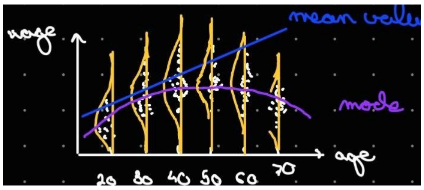
First class essay Raul Adell

This essay may contain ideas given in class and complemented by the book given on the references of the course. Book section: 2.1 What Is Statistical Learning?

Machine learning can be defined in many ways, specially influenced by the background of the one who does it. However, one way of doing that is the field of study that uses algorithms and statistical models to enable computer systems to learn from data, without being explicitly programmed. Thus, given the data x we aim to infer a relationship which can be written by: $y = f(x) + \varepsilon$, being f the structure we want to determine, y the desired prediction and ε the error, which takes into account model precision and effects of important unmeasured inputs. x can be named input variables, predictions, independent variables or features and y can be named output variables, response, dependent variables or target.

Supervised and unsupervised learning are two common approaches to machine learning that are used to build models that can recognize patterns in data.
Supervised learning is a type of machine learning where the algorithm is trained using labeled data. In other words, the algorithm is provided with a set of inputs and their corresponding outputs. The algorithm then uses this data to learn the relationship between the inputs and outputs, and can make predictions for new data that it has not seen before. There are two main types of supervised learning algorithms: classification and regression. In classification, the algorithm is trained to predict a discrete output. In class we discussed the prediction of the stock exchange dynamics (either up or down) based on the % of change of the previous days. Flipped histograms come handy in this kind of problem. In regression, the algorithm is trained to predict a continuous output, such as a numerical value. For example, the wage of a person given his age, education level and years of experience. Histograms are key in the intuition of the desired relationship $f(x)$, one must decide a criterion. One option could be designing $f(x)$ such that it goes through all the mode points in the different histograms over the different features. Nonetheless this is a mathematical challenging approach and building $f(x)$ such that it goes through all mean values of the different histograms is a more feasible idea.



Unsupervised learning, on the other hand, is a type of machine learning where the algorithm is trained on unlabeled data. The goal of unsupervised learning is to discover patterns or structure in the data, without being told what to look for. This can be useful in situations where the data is unstructured or the underlying relationships are unknown. Unsupervised learning is used in many applications, such as clustering, dimensionality reduction, and anomaly detection. Finding structure in data can be very important since when things get very high dimensional they may get scarcely interpretable and quite complex.

The trade-off between prediction accuracy and model interpretability was also quickly discussed in class. On one hand, highly complex models with many parameters have the potential to achieve high levels of accuracy in predicting outcomes. However, these models can be difficult to interpret, making it challenging to gain insights into how the model arrived at its predictions. On the other hand, simpler models with fewer parameters may be more interpretable, allowing for a better understanding of the relationships between the input variables and the output variable. However, these models may sacrifice some level of accuracy in their predictions. One has to consider the situation in order to go towards one side or the other. For example, decision trees are used in bank loans because the decision must be clearly justified.

The class concluded with some ideas to take into consideration with our framework for defining what is the $f(x)$.

- Unexplained noise ε is important for theoretical limitations.
- Sensibility of the relationship. How the output changes when the input is perturbed and compatibility with saturation and rapid growth.
- Ranking of variables. How to order variables by their importance/predictive power.
- How to measure the performance of a ML model. Some examples are RMS, MAD, % of mistakes, R^2. The selection between these may depend on the type of problem being addressed and the nature of the data.