Quiz 1:

Machine Learning - Setting up the scenes

Email address *

michael.symmonds19@imperial.ac.uk

Please enter your name: *

Michael Symmonds

Lecture

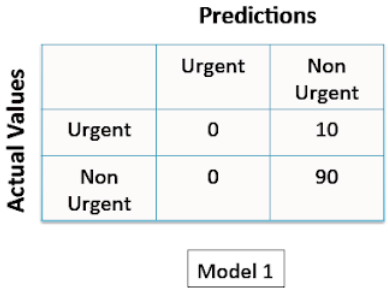
What's the difference between Supervised and Unsupervised Learning ? 3 points

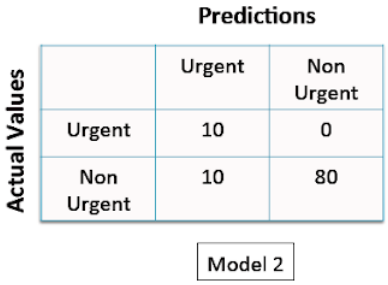
Supervised learning involves a process where we have input variables and associated output variables and we would like to use an algorithm to learn the mapping (or function) that describes the relationship between the input and output variables. The learning process is called training. Supervised learning is used when we would like to make predictions using the features (inputs) based on the associated labels (output). There are two main types of problems, classification and regression.

Unsupervised learning is the case where we have input variables but no associated output variables or labels. Thus, in this case, the problem we are trying to solve is to use the algorithm to help organise the inputs and to identify meaningful patterns therein. We are looking to describe or extract relationships in the data. Unsupervised learning is often used to cluster (identify groups) in data, estimate density (understanding the distribution of the data), to visualise the data or to reduce the dimensionality of the data for projections.

Thus the key differences are seen in the purpose that a programmer is looking to achieve (prediction vs. exploration and understanding), as well as the data that is available (input and output pairs vs only input pairs, although unsupervised learning is sometimes performed on data that has an associated output).

We want to build a Machine Learning model to classify some documents into "Urgent" and "Non Urgent". Most of the documents are non urgent. You have the following confusion matrices for two models. Compute the accuracy score for each one. Which one do you prefer ? 5 points





Model 1 Accuracy: 90%
Model 2 Accuracy: 90%
Thus, there is not a clear winner in terms of accuracy. However, since most values are actually non-urgent, we are likely to be dealing with an imbalanced dataset. Thus, let us compute the precision, recall and hence F1 score for both models.
Model 1 precision: 0 %
Model 1 recall: 0%
Model 1 F1: 0%

Model 2 precision: 50%
Model 2 recall: 100%
Model 2 F1: 66.67%

Preferred model: Model 2, since the F1 score is higher for this model. However, intuitively we can see that model 2 is better since model 1 did not predict an urgent case accurately at all whereas model 2 predicted all actual urgent cases correctly. Thus, if an actual urgent case is very important, we would want a model that would predict that accurately and hence model 2 is better.

How can we compare two models according to their ROC curve? 2 points

We can calculate the Area under the Curve (AUC) of the ROC curves for each model. A model that has a higher AUC curve is the better model. This can be understood as the true positive rate (TPR) increasing for the better model quicker or more steadily than the inferior model and hence a higher AUC. This model is seen to make fewer mistakes.

Programming Session

Did you understand the problem?

- (●) Yes
- ( ) No