Advanced Data Modeling - Week 12

Kevin O'Brien

February 9, 2018

Contents

1 Classification

1.1 What Is Classification

classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Discriminant analysis is an example of a **classification** method.

- To train (create) a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class.
- To predict the classes of new data, the trained classifier finds the class with the smallest misclassification cost.

1.2 Types I and II Error

A type I error is the incorrect rejection of a true null hypothesis. A type II error is the failure to reject a false null hypothesis. A type I error is a false positive. Usually a type I error leads one to conclude that a thing or relationship exists when really it doesn't. A type II error is a false negative.

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

1.3 False Positive and False Negative error

A false positive error, commonly called a "false alarm" is a result that indicates a given condition has been fulfilled, when it actually has not been fulfilled. A false positive error is a **Type I error** where the test is checking a single condition, and results in an affirmative or negative decision usually designated as "true or false".

A false negative error is where a test result indicates that a condition failed, while it actually was successful. A false negative error is a **Type II error** occurring in test steps where a single condition is checked for and the result can either be positive or negative.

1.4 Confusion Matrix

A **confusion matrix**, is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives.

This allows more detailed analysis than mere proportion of correct guesses (accuracy). Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced (that is, when the number of samples in different classes vary greatly).

For example, if there were 95 cats and only 5 dogs in the data set, the classifier could easily be biased into classifying all the samples as cats. The overall accuracy would be 95%, but in practice the classifier would have a 100% recognition rate for the cat class but a 0% recognition rate for the dog class.

1.5 Sensitivity and Specificity

Sensitivity and specificity are measures of the performance of a binary classification test.

- Sensitivity (also called the true positive rate, or the **recall** rate) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

$$\text{sensitivity (Recall)} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

- Specificity measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition, sometimes called the true negative rate).

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of false positives} + \text{number of true negatives}}$$

(Remark: We will use the terms **Sensitivity** and **Recall** interchangeably. Sensitivity is more commonly used in a medical context, while recall is more commonly used in data science.)

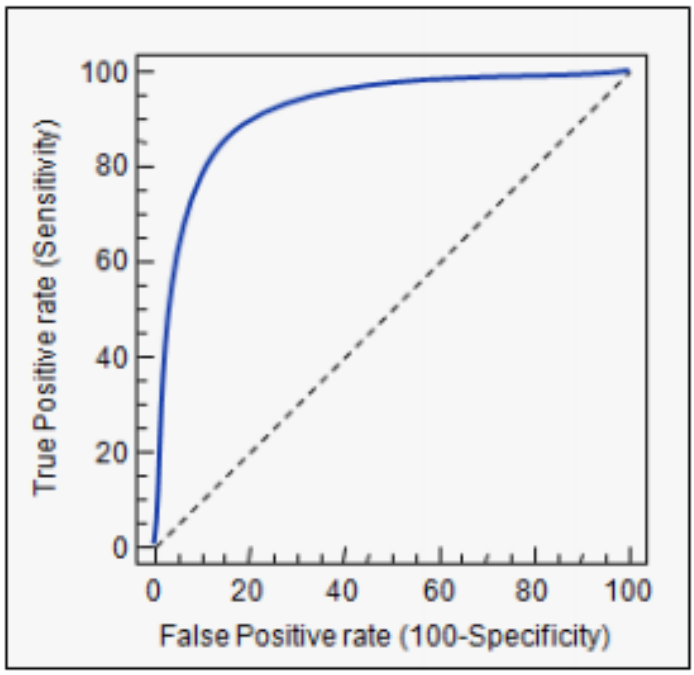

Figure 1:

1.6 Receiver Operating Characteristic (ROC) curve

In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test (Zweig and Campbell, 1993).

1.7 Properties of ROC Curves

An ROC curve demonstrates several things:

1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

2. The closer the curve follows the upper-left border of the ROC space, the more accurate the test.

3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

4. The slope of the tangent line at a cutpoint gives the likelihood ratio (LR) for that value of the test.

5. The area under the curve is a measure of accuracy.