Binary Classification

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.

- 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.

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 doesnt. 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

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
- 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.

Confusion Matrix

- The confusion table is a table in which the rows are the observed categories of the dependent and the columns are the predicted categories.
- A confusion matrix 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.

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).

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}$$

  - *(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.)*

- **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).

$$\text{Specificity} = \frac{TN}{TP + FN}$$

  - *(Remark: Not commonly used in Data Sciences, and NOT a synoym for Precision)*

Receiver Operating Characteristic (ROC) curve

[h!]

- 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.

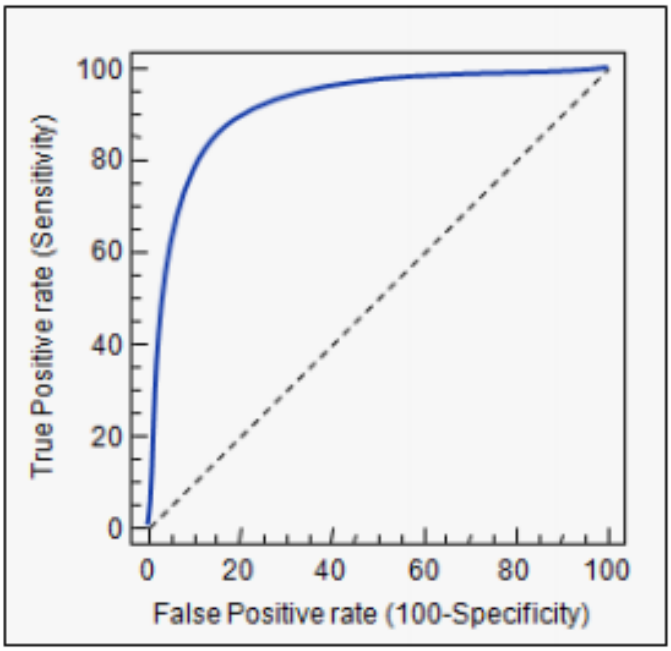
True Positive rate (Sensitivity)
100
80
60
40
20
0
0
20
40
60
80
100
False Positive rate (100-Specificity)

Figure 1:

Accuracy, Recall and Precision: An Example

Calculating precision and recall is actually quite easy. Imagine there are 135 positive cases among 10,000 cases. You want to predict which ones are positive, and you pick 265 to have a better chance of catching many of the 135 positive cases. You record the IDs of your predictions, and when you get the actual results you sum up how many times you were right or wrong. There are four ways of being right or wrong:

- TN / True Negative: case was negative and predicted negative
- TP / True Positive: case was positive and predicted positive
- FN / False Negative: case was positive but predicted negative
- FP / False Positive: case was negative but predicted positive

Now count how many of the 10,000 cases fall in each category:

	Predicted Negative	Predicted Positive
Negative Cases	TN: 9,700	FP: 165
Positive Cases	FN: 35	TP: 100

- What percent of your predictions were correct?
  The accuracy was (9,760+60) out of 10,000 = 98.00%
- What percent of the positive cases did you catch?
  The recall was 100 out of 135 = 74.07%
- What percent of positive predictions were correct?
  The precision was 100 out of 265 = 37.74%
- What percent of negative predictions were correct?
  The specifity was 9700 out of 9735 = 99.64%

The F Score

The F-score or F-measure is a measure of a classification procedures accuracy. It considers both the precision and the recall to compute the score.

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Cross Validation

When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications. The cross validated set of data is a more honest presentation of the power of the discriminant function than that provided by the original classifications and often produces a poorer outcome. The cross validation is often termed a jack-knife classification, in that it successively classifies all cases but one to develop a discriminant function and then categorizes the case that was left out. This process is repeated with each case left out in turn.This is known as leave-1-out cross validation. This cross validation produces a more reliable function. The argument behind it is that one should not use the case you are trying to predict as part of the categorization process.

Error Rates

- We can evaluate error rates by means of a training sample (to construct the discrimination rule) and a test sample.
- An optimistic error rate is obtained by reclassifying the training data. (In the training data sets, how many cases were misclassified). This is known as the apparent error rate.
- The apparent error rate is obtained by using in the training set to estimate the error rates. It can be severely optimistically biased, particularly for complex classifiers, and in the presence of over-fitted models.
- If an independent test sample is used for classifying, we arrive at the true error rate.
- The true error rate (or conditional error rate) of a classifier is the expected probability of misclassifying a randomly selected pattern. It is the error rate of an infinitely large test set drawn from the same distribution as the training data.

Misclassification Cost

As in all statistical procedures it is helpful to use diagnostic procedures to asses the efficacy of the discriminant analysis. We use cross-validation to assess the classification probability. Typically you are going to have some prior rule as to what is an acceptable misclassification rate. Those rules might involve things like, what is the cost of misclassification? Consider a medical study where you might be able to diagnose cancer. There are really two alternative costs. The cost of misclassifying someone as having cancer when they dont. This could cause a certain amount of emotional grief. Additionally there would be the substantial cost of unnecessary treatment. There is also the alternative cost of misclassifying someone as not having cancer when in fact they do have it. A good classification procedure should

- result in few misclassifications
- take prior probabilities of occurrence into account
- consider the cost of misclassification

For example, suppose there tend to be more financially sound firms than bankrupt firm. If we really believe that the prior probability of a financially distressed and ultimately bankrupted firm is very small, then one should classify a randomly selected firm as non-bankrupt unless the data overwhelmingly favor bankruptcy. There are two costs associated with discriminant analysis classification: The **true misclassification cost per class**, and the **expected misclassification cost** (ECM) per observation.

Suppose there we have a binary classification system, with two classes: class 1 and class 2. Suppose that classifying a class 1 object as belonging to class 2 represents a more serious error than classifying a class 2 object as belonging to class 1. There would an assignable cost to each error. c(i—j) is the cost of classifying an observation into class j if its true class is i. The costs of misclassification can be defined by a cost matrix.

	Predicted Class 1	Predicted Class 2
Class 1	0	c(2—1)
Class 2	c(1—2)	0

Expected Cost of Misclassification (ECM)

Let p1 and p2 be the prior probability of class 1 and class 2 respectively. Necessarily p1 + p2 = 1.

The conditional probability of classifying an object as class 1 when it is in fact from class 2 is denoted p(1—2). Similarly the conditional probability of classifying an object as class 2 when it is in fact from class 1 is denoted p(2—1).

$$ECM = c(2|1)p(2|1)p1 + c(1|2)p(1|2)p2$$

(In other words: the sum of the cost of misclassification times the (joint) probability of that misclassification. A reasonable classification rule should have ECM as small as possible.