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. Binary Classification is the task of classifying the members of a given set of objects into two groups on the basis if them having a particular set of characteristics.

- 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.

Binary Classification Prediction Procedure
Binary Variabble: Positive or Negative

Four Possible Outcomes from Classification Procedure:

- **TN / True Negative**: Case was actually negative and was also predicted negative (CORRECT).
- **TP / True Positive**: Case was actually positive and was also predicted positive (CORRECT).
- **FN / False Negative**: Case was actually positive but was predicted negative (WRONG).
- **FP / False Positive**: Case was actually negative but was predicted positive (WRONG).

Remark** : We will use this notation to specify the number of cases in each category also: i.e.* ***TP = 5000 *means 5000 True Positives.*

Confusion Matrix

- The Confusion Matrix 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 classification system, 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.

	Predicted Negative	Predicted Positive
Actual State: Negative	**TN**	**FP**
Actual State: Positive	**FN**	**TP**

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.

Medical Testing example Defining true/false positives

In general, Positive = identified and negative = rejected. Therefore:

TN True negative = Healthy people correctly identified as healthy (*correctly rejected*)

FP False positive = Healthy people incorrectly identified as sick (*incorrectly identified*)

FN False negative = Sick people incorrectly identified as healthy. *incorrectly rejected*

TP True positive = Sick people correctly diagnosed as sick (*correctly identified*)

Types I and II Error (For Later)

- 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

Accuracy Rate

The accuracy rate calculates the proportion ofobservations being allocated to the **correct** group by the predictive model. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Classifications}}{\text{Total Number of Classifications}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Misclassification Rate

The misclassification rate calculates the proportion ofobservations being allocated to the **incorrect** group by the predictive model. It is calculated as follows:

$$\frac{\text{Number of Incorrect Classifications}}{\text{Total Number of Classifications}}$$

$$= \frac{FP + FN}{TP + FP + TN + FN}$$

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).

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

- **Examples:** Sensitivity (TPR), also known as recall, is the proportion of people that tested positive (TP) of all the people that actually are positive (TP+FN).
  - It can be seen as the probability that the test is positive given that the patient is sick.
  - With higher sensitivity, fewer actual cases of disease go undetected (or, in the case of the factory quality control, the fewer faulty products go to the market).
  - *(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 synonym for Precision)*

- **Examples:** Specificity (TNR) is the proportion of people that tested negative (TN) of all the people that actually are negative (TN+FP). As with sensitivity, it can be looked at as the probability that the test result is negative given that the patient is not sick.

- With higher specificity, fewer healthy people are labeled as sick (or, in the factory case, the less money the factory loses by discarding good products instead of selling them).

- The relationship between sensitivity and specificity, as well as the performance of the classifier, can be visualized and studied using the ROC curve (Which we shall see shortly).

Precision

In a binary classification procedure, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of cases labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Recall

Recall is defined as the number of true positives divided by the total number of cases that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

Accuracy, Recall and Precision: An Example

Suppose we are designing a medical diagnosis system, and we have enlisted 10000 volunteers to help us test it. Suppose there are 135 positive cases of an illness among the10,000 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.

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

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

Class Imbalance

- A data set said to be highly skewed if sample from one class is in higher number than other.
- In an imbalanced data set the class having more number of instances is called as **major class** while the one having relatively less number of instances are called as **minor class** .
- Applications such as medical diagnosis prediction of rare but important disease is very important than regular treatment.
- Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions, managing risk and predicting failures of technical equipment.
- In such situation most of the binary classification procedure are biased towards the major classes and hence show very poor classification rates on minor classes.
- It is also possible that classifier predicts everything as major class and ignores the minor class completely.
- The **Accuracy** measure is an example of an metric that is affected by this bias.
- As the **F-Score** is not computed using the True Negatives, it is less biased.

The F Score

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

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

- The F-score or F-measure is a single measure of a classification procedure's usefulness.
- The F-score considers both the ***Precision*** and the ***Recall*** of the procedure to compute the score.
- The higher the F-score, the better the predictive power of the classification procedure.
- A score of 1 means the classification procedure is perfect. The lowest possible F-score is 0.

$$0 \leq F \leq 1$$

From Before

- **Precision** is the number of correct positive results divided by the number of ***predicted positive*** results.

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

- **Recall** is the number of correct positive results divided by the number of ***actual positive*** results.

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F-score is the Harmonic mean of Precision and Recall.

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Alternatively

$$F = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

Example

Number of cases: **100,000**

	Predicted Negative	Predicted Positive
Actual State: Negative	TN = **97750**	FP = **150**
Actual State: Positive	FN= **330**	TP =**1770**

- **Accuracy** = 0.9952
- **Recall** = 0.8428
- **Precision** = 0.9218

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F = 2 \times \left(\frac{0.9218 \times 0.8428}{0.9218 + 0.8428} \right) = 2 \times \left(\frac{0.7770}{1.7646} \right) = 2 \times 0.4402$$

$$F = 0.8804$$

ROC Curves

- A receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classification system as its discrimination threshold is varied.
- The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. (The true-positive rate is also known as sensitivity in biomedicine, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as 1 - specificity).
- The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields. ROC analysis since then has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research.
- The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes.

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.

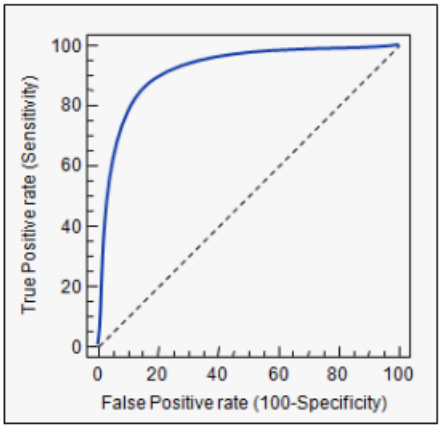

Figure 1: Receiver Operating Characteristic (ROC) curve

- 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.