1 Binary Classification Prediction Procedure

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.

1.1 Binary Classification

Binary or binomial classification is the task of classifying the elements of a given set into two groups on the basis of a Classification rule. Some typical binary classification tasks are

- medical testing to determine if a patient has certain disease or not (the classification property is the presence of the disease)
- quality control in factories; i.e. deciding if a new product is good enough to be sold, or if it should be discarded (the classification property is being good enough)
- deciding whether a page or an article should be in the result set of a search or not (the classification property is the relevance of the article, or the usefulness to the user)

Statistical classification in general is one of the problems studied in computer science, in order to automatically learn classification systems; some methods suitable for learning binary classifiers include the decision trees, Bayesian networks, support vector machines, neural networks, probit regression, and logit regression.

Sometimes, classification tasks are trivial. Given 100 balls, some of them red and some blue, a human with normal color vision can easily separate them into red ones and blue ones. However, some tasks, like those in practical medicine, and those interesting from the computer science point of view, are far from trivial, and may produce faulty results if executed imprecisely.

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.

1.2 Binary Classification

Defining True/False Positives In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

Medical Testing Example:

- True positive = Sick people correctly diagnosed as sick
- False positive= Healthy people incorrectly identified as sick
- True negative = Healthy people correctly identified as healthy
- False negative = Sick people incorrectly identified as healthy.

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. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

Possible Outcomes from Classification Procedure:

TN True Negatives - correct prediction

TP True Positives - correct prediction

FN False Negatives - incorrect prediction

FP False Positives - incorrect prediction

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.

2 Model Accuracy

Prediction error refers to the discrepancy or difference between a predicted value (based on a model) and the actual value. In the standard regression situation, prediction error refers to how well our regression equation predicts the outcome variable scores of new cases based on applying the model (coefficients) to the new cases predictor variable scores.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

3 Performance of Classification Procedure

These classifications are used to calculate accuracy, precision (also called positive predictive value), recall (also called sensitivity), specificity and negative predictive value:

- **Accuracy** is the fraction of observations with correct predicted classification

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

- **Precision** is the proportion of predicted positives that are correct

$$\text{Precision} = \text{Positive Predictive Value} = \frac{TP}{TP + FP}$$

- **Negative Predictive Value** is the fraction of predicted negatives that are correct

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN}$$

- **Recall** is the fraction of observations that are actually 1 with a correct predicted classification

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

- **Specificity** is the fraction of observations that are actually 0 with a correct predicted classification

$$\text{Specificity} = \frac{TN}{TN + FP}$$

4 Performance of Classification Procedure

These classifications are used to calculate accuracy, precision (also called positive predictive value), recall (also called sensitivity), specificity and negative predictive value:

- **Accuracy** is the fraction of observations with correct predicted classification

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

- **Precision** is the proportion of predicted positives that are correct

$$\text{Precision} = \text{Positive Predictive Value} = \frac{TP}{TP + FP}$$

- **Negative Predictive Value** is the fraction of predicted negatives that are correct

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN}$$

- **Recall** is the fraction of observations that are actually 1 with a correct predicted classification

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

- **Specificity** is the fraction of observations that are actually 0 with a correct predicted classification

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Actual Class	Predicted Class	
	Negative	Positive
Negative	**TN**	**FP**
Positive	**FN**	**TP**

TN True Negatives

TP True Positives

FN False Negatives

FP False Positives

5 Recall and Precision

In a classification task, 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 elements 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).

Recall

Recall in this context is defined as the number of true positives divided by the total number of elements 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).

- **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.

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

5.1 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).

6 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.

F-Score

- 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.

$$0 \leq F \leq 1$$

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.

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

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

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

Alternatively

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

Number of cases: **100,000**

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

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

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

$$F = 2 \times \frac{0.9218 \times 0.8428}{0.9218 + 0.8428} F = 2 \times \left(\frac{0.9218 \times 0.8428}{0.9218 + 0.8428} \right) F = 2 \times \left(\frac{0.7770}{1.7646} \right) = 2 \times 0.4402 = 0.8804$$

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%