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 95but a 01 Predicted Predicted Negative 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 doesnt.
- A type II error is a false negative. Null hypothesis (H0) is true Null hypothesis (H0) is false Reject Type I error Correct Outcome null hypothesis False positive True positive Fail to reject Correct Outcome Type II error null hypothesis True negative 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: Accuracy = Number of Correct Classifications Total Number of Classifications Accuracy = T P + T N T P + F P + T N + F N 2 Misclassification Rate The misclassification rate calculates the proportion ofobservations being allocated to the incorrect group by the predictive model. It is calculated as follows: Number of Incorrect Classifications Total Number of Classifications = F P + F N T P + F P + T N + F N 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). Sensitivity (Recall) = T P T P + F N

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). Specificity = T N T P + F N (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).

0.1 Precision

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). Precision = T P T P + F P (1) 3

0.2 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). Recall = T P T P + F N (2)

0.3 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 = 98.002. What percent of the positive cases did you catch?

The recall was 100 out of 135 = 74.073. What percent of positive predictions were correct?

The precision was 100 out of 265 = 37.744. What percent of negative predictions were correct?

The specifity was 9700 out of 9735 = 99.64

0.4 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. 4

0.5 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 = 2 precision recall precision + recall

The F-score or F-measure is a single measure of a classification procedures 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 F 1 From Before

Precision is the number of correct positive results divided by the number of predicted positive results. Precision = T P T P + F P

Recall is the number of correct positive results divided by the number of actual positive results. Recall = T P T P + F N The F-score is the Harmonic mean of Precision and Recall. F = 2 1 Recall + 1 Precision Alternatively F = 2
Precision Recall Precision + Recall 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 Precision Recall Precision + Recall F = 2
0.9218 0.8428 0.9218 + 0.8428 = 2
0.7770 1.7646 = 2 0.4402 F = 0.8804 5

1 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

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

2 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
$$\text{Specificity} = \frac{TN}{TN + FP}$$

2.1 Definitions (From Week 1)

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.

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

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

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

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

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

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

2.2 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.

2.3 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.

2.4 Definitions (From Week 1)

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.

Accuracy Rate

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

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

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