Data 621 HW 2 Comments & Supplements

Dennis Pong, Euclid Zhang, Jie Zou, LeTicia Cancel, Joseph Connolly

2022-03-16

Supplemental Documentation for Homework 2

- The purpose of this document is to provide more in-depth explanation about the topics discussed in this homework assignment

Confusion Matrix

- Confusion matricies are used to display the performance of a classification algorithm. It’s comprised of 4 values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FP)

Accuracy

- Accuracy, one of the most commonly calculated values from a classification algorithm, is calculated by the following:

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

- It is important to note, that this value can be misleading if working with an imbalanced dataset, such as one that includes a favorable amount of positive or negative values.

Classification Error Rate

- The classification error rate is the number of observations which are misclassified over the sample size. It’s denoted as:

$\frac{1}{n}\sum_{i=1}^{n} I(\hat{y} \neq y)$

where $I(\hat{y} \neq y) = 1 \ \ if \hat{Y}_i \neq y_i, \ and \ 0 \ otherwise$

It should be noted that $\hat{y}$ be a binary (0 or 1) vector of the predicted class and y also be a binary vector (0 or 1) of the observed class labels. This can be classified by

```
y.hat <- 2
y <- 3

mean(abs(y.hat - y))
```

```
## [1] 1
```

Precision

- Precision is a measure of the closeness of multiple measures relative to each other. For instance:

$precision = \frac{TP}{TP+FP}$

Sensitivity

- Sensitivity is a probability measure for a true positive; also the ratio of the proportion of a population that is valued as a True Positive, over the sum of True Positives and False Negatives. In the context of a disease, sensitivity can be thought of those who test positive over the total population of patients that have the disease.

$sensitivity = \frac{TP}{TP+FP}$

Specificity

- Unlike sensitivity, specificity is a probability measure for a true negative. It is the ratio of the counts of True Negatives over the sum of True Negatives and False Positives

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

F_1 Score

- Also known as the F-Score, this a measure of model’s accuracy on the dataset in the context of binary classifications (positive and/or negative, 0,1, etc.)

- This combines precision and recall of the model, and is also known as the harmonic mean of the model’s precision and recall.

- The F Score is represented by the following formula:

$$F_\beta = (1+\beta)\left(\frac{precision * recall}{\beta^2 * precision + recall}\right)$$

…where β is a value indicating more importance for recall which is β times more important than precision. In the instance of calculating the F_1 score, this implies that $\beta = 1$. Thus the formula for the F_1 score is obtained:

$$F_1 = 2\left(\frac{precision \ast recall}{precision + recall}\right)$$

ROC Curve

- The ROC curve, also known as the receiver operating characteristic curve, is utilized to illustrate how well the model functions. It’s built by plotting the true positive rate against the false positive rate. The ROC curve is interpreted by referring to the straight diagonal line that touches end to end of the curve; the rule of thumb is that the closer the curve is to the line, the less accurate the model is.

AUC Curve

- AUC, or Area Under the Curve, is used to summarize the perfomance of other classifiers. It’s the same as the probability that a randomly chosen positive event is ranked higher than a randomly chosen negative event. This, in practice, is a comprable measure of predictive accuracy.