Megan Cusey

Summer 2019 – MSD 534

Week 7 Discussion

The examples supplied uses the Titanic Data Tree Induction Model in Rapid Miner where the maximum number of nodes =8, produces the following confusion matrix:

**Chapter 5 – Overfitting and Its Avoidance**

The general concepts for this chapter was the theme that all models will overfit the training data. How much overfitting is ok? When does it become a problem? How do you recognize overfitting and how to avoid it?

The goal for creating a model is to find patterns in the data to that will generalize to the population of the data, not just the training set. The more complicated the model, the more the model will tend to pick up the characteristics of the training set, not necessarily the characteristics of the population. As seen in a fitting chart, the more complicated the graph, the better accuracy will be on the training data. This is misleading because the same model applied to the holdout data will not be as accurate. Finding the “sweet spot” where the graph is not too complex that it can produce the most accurate results that it finds the general patterns of the whole data set, not just the training data set, will produce a better model.

In tree induction, model complexity occurs when too many nodes are in the model. For linear regression, too many data sets or dimensions will increase complexity. The cross-validation method provides a metric for how well the model can perform without too much overfitting. The method splits the training data and applies it to the model multiple times and measures performance such as the mean and variance. The average of the test performance results indicates a realistic accuracy measurement that can be expected of a model without too much overfit.

**Chapter 7 – What is a good model?**

This chapter provides some thoughts on how you should evaluate a model. Often, the measure of overall accuracy is not enough to decide if the model is sufficient. In classification, a false positive or false negative can have a different impact. The book used an example of a medical test. If somebody had a false positive diagnosis, further testing can be done to determine the true negative. However, if a false negative was the result of the test, then the impact is greater because the person doesn’t get the medical care he or she needs for the condition. The overall accuracy of a model doesn’t consider the nature of the data.

The Confusion Matrix provides further information to evaluate the classifiers of a model.

Example:

Overall Accuracy: 78.53%

|  |  |  |  |
| --- | --- | --- | --- |
|  | True False | True True | Class Precision |
| Predicted False | 157 | 39 | 80.10% |
| Predicted True | 28 | 88 | 75.86% |
| Class Recall | 84.86% | 69.29% |  |

The confusion matrix allows the analyst to decipher the model’s results as a function of the classifiers. With this information, the unequal balance of the classifiers can be considered.

Another obstacle in building models is determining the correct measurements or target label to answer a business question. Does the model truly address the problem specified? The book introduces an analytical framework called Expected Value. In summary, the expected value framework breaks down the data product into three parts: The structure of the problem, elements of the analysis to be extracted from the data, and the elements of the analysis that need to be gathered from other sources.

**Chapter 8 – Visualizing Model Performance**

This chapter expands on the ideas of expected value. After computed the expected value of all instances, we can apply a threshold to help apply a particular classifier that articulates the cost/benefits. Typically, the threshold is set when the profit is positive.

A Profit Curve is a visualization that displays the profit for test instances as a function of the percent of test instances by decreasing score. This allows the analyst to see which classifier would produce the highest profit when selecting the top-n% of the population. Profit Curves are a good tool to use if both class priors and cost-benefit estimates are known and stable.

Receiver Operating Characteristics (ROC) graphs can be used if there are uncertain conditions. The ROC graphs plots a classifier’s false positive rate and true positive rate. Each point corresponds to a confusion matrix. The area under the ROC curve is a useful metric when a single number is needed to summarize performance.

An alternative to ROC and Profit Curves are Cumulative Response Curves/lift curve which graph the percent of positives correctly classified and the percentage of the population targeted.