Lecture 8: Classifiers

Intro to Data Science for Public Policy, Spring 2016

by Jeff Chen & Dan Hammer, Georgetown University McCourt School of Public Policy

Contents

The above diagram was produced using a decision tree algorithm, which is one of many forms of supervised learning known as *classifiers* or *classification algorithms*. Classifiers take on many forms. Some use recursive partitioning to break a population into many, more homogeneous subpopulations. Others estimate a series of equations to fit a line or plane between two or more classes. Others will average the results of an ensemble of models to predict membership. Each class of model is defined with mathematical scenarios in mind. This chapter is organized into three sections. Section 1 describes common considerations in classification models. Section 2 provides an overview of a number of classifiers, including kNN, logistic regression, decision trees, as well as random forests, support vector machines, and ensemble methods. Section 3 describes the types of applications of these methods.

Measure	Description	Interpretation
Receiving Operating Characteristic (ROC) Curve	ROC curves plotpairs of TPRs and FPRs that correspond to varied discriminant thresholds between 0 and 1. By systematically testing thresholds. For example, TPRs and FPRs are calculated and plotted given probability thresholds $p = 0.2$, $p = 0.5$, and $p = 0.8$.	Once plotting the curve with TPR as Y and FPR as X, the area under the curve (AUC) represents robustness of the model, ranging from 0.5 (model is as good as a coin toss) to 1.0 (perfectly robust model). The AUC statistic is sometimes referred to as the "concordance".
F_1 Score	The score is formulated as $F_1 = 2 \times \frac{precision \times recall}{precision + recall} = 2 \times \frac{PPV \times TPR}{PPV + TPR}$ where precision or PPV = $\frac{TP}{TP+FP}$ and recall or TPR = $\frac{TP}{TP+FN}$	The measure is bound between 0 and 1, where 1 is the top score indicating a better model.

Applications of classifiers

Appropriate uses of classification techniques

[text goes here]

```
#
```

Scoring

[text goes here]

```
#
```

prediction and prioritization

[text goes here]

```
#
```

Propensity score matching

[text goes here]

```
#
```

Exercise Data

- [Labor and wage analysis]