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

Supervised learning is the most relied upon class of techniques that enable causal inference but also deployed precision policy. How does changing one variable independently impact another variable? In We begin to introduce basic regression analysis, correlation coefficients, ordinary least squares, and the relationship between the concepts. Note that this is a very cursory review, and the deep assumptions are not tested or expounded upon.

Lecture objectives

Overview

Three classification problems in public policy:

Classifiers

What are they?

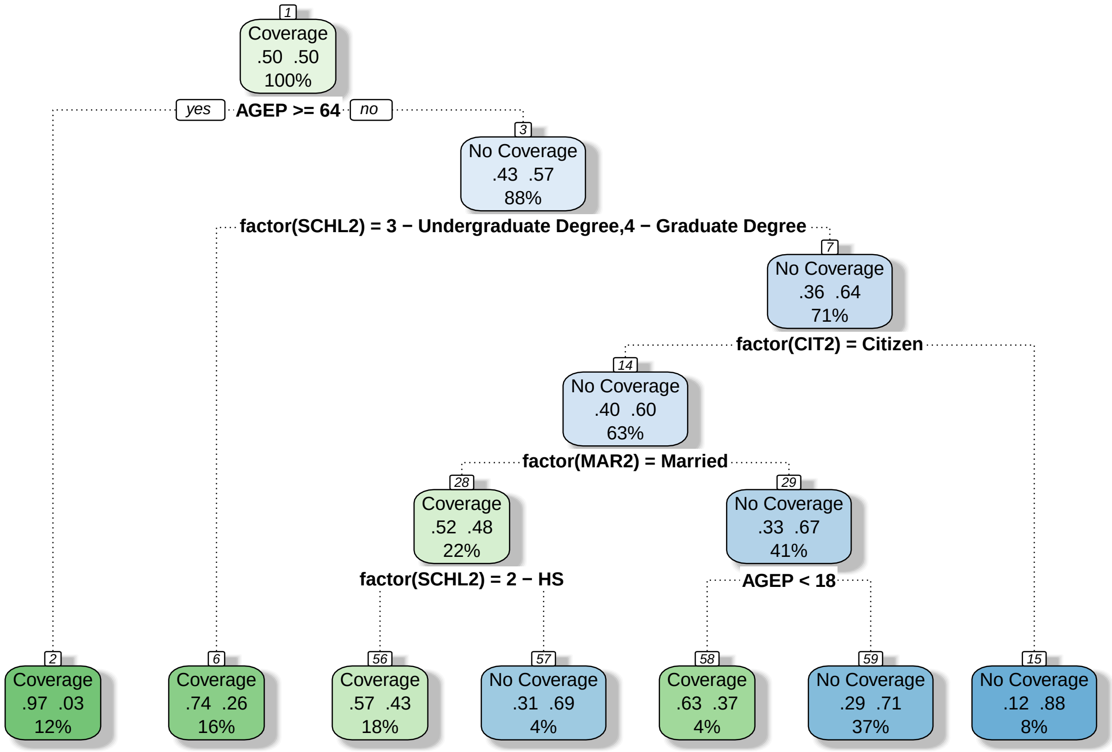
[text goes here]

Decision Trees

In everyday policy operations, a decision tree is a common tool used for communicating processes, whether it's how an actor moves through a complex system or how a population can be described based on a set of criteria.

For example, 14% of Georgians were without healthcare coverage in 2015. This is helpful to know, but ultimately, a degree of specificity would be required to use this knowledge as a tool for outreach. Using decision trees, it's possible to develop discrete profiles of the population based on observable characteristics such that discrete if-else criteria identify smaller subpopulations with similar membership. In this case, membership means health coverage. Based on this decision tree, we can infer that non-citizens under the age of 64 without a college education are 89% likely to not have coverage, and citizens between 16 and 64 who are not married have a 71% chance of not having health coverage.

Whether its failure analysis of engineering mechanisms or developing customer profiles of program participation, decision trees can help characterize intricate, non-linear patterns in data.



The Gist. Decision trees rely on basic tenants of information theory, namely the idea of *information gain*. Given a labeled set of data that contains input features, the structure of decision trees can be likened to branches of a tree: moving from the base of the tree upwards, the tree trunk splits into two or more large branches, which then in turn split into even smaller branches, eventually reaching even small twigs with leaves. Using this physical analogy, decision trees are a representation of information, subset into smaller, more homogeneous subsamples. By recursively splitting input features into two subsamples at a time with the goal of achieving greater homogeneity of labeled examples in each subsample. This splitting process is continued on each subsample until all subsamples contain only one class or a stopping criteria is met. The point at which a feature is split is known as a decision node, the trunk of the tree from which all branches spring is known as the root node, and the termini of the tree with the most homogeneous subsamples are known as leafs.

There are a number of forms decision trees are *grown*, the most commonly implemented algorithm is the C4.5.

```
C4.5 (Examples, Target, Input Features)
  Create root node
  Check levels of input features for "pure" or "nearly pure" subgroups
```

Concepts

- Recursively looks for which input feature to split based on a statistical criterion Typically uses *entropy*, which is a core measure form information theory. Entropy $= \sum -p_{Y=1} log_2(p_{Y=1}) + -p_{Y=1} log_2(p_{Y=1})$ [Information gain]
- Pruning

Random Forests

- statistical assumptions and mechanics, risks/strengths, implementation, non-technical explanation

```
#
```

Support Vector Machines

- statistical assumptions and mechanics, risks/strengths, implementation, non-technical explanation

```
#
```

Logistic Regression

- statistical assumptions and mechanics, risks/strengths, implementation, non-technical explanation

```
#
```

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]