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

[text goes here]

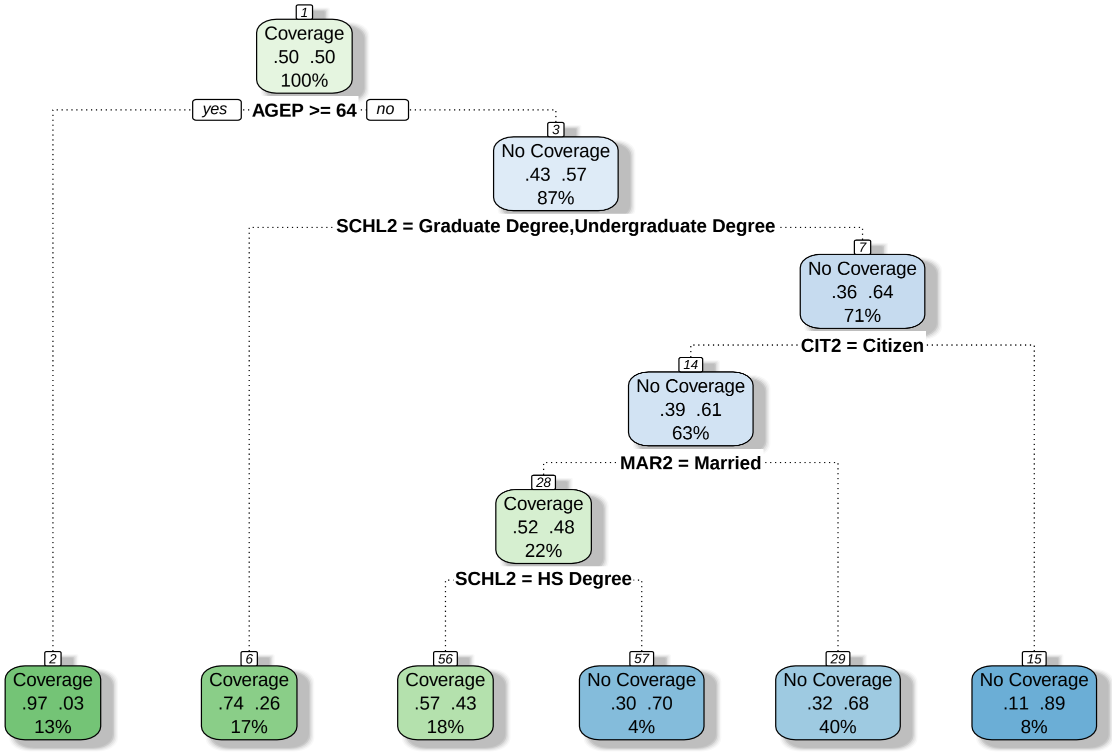
Decision Trees

In everyday policy setting and operations, decision trees are a common tool used for communicating complex processes, whether for how an actor moves through intricate and convoluted bureaucracy or how a subpopulation can be described based on a set of criteria. While the garden variety decision tree can be laid out qualitatively, supervised learning allows decision trees to be create in an empirical fashion that not only have the power to aesthetically communicate patterns, but also predict how a non-linear system behaves.

According to the US Census Bureau's American Community Survey, approximately 16.5% of Georgians were without healthcare coverage in 2015. To some degree, this is surprising, but it is important to know to support public policy and outreach. This statistic on its own, however, is not useful; it needs to be contextualized and segmented so that government can conduct outreach to ensure that all citizens have access to a modern fundamental service. 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. Below, a decision tree has been estimated following the specification: $Pr(Coverage) = f(\text{Sex, Age, Education, Marital Status, Race, Citizenship})$. Based on this decision tree, we can identify well-defined cells of Georgians who have and do not have healthcare coverage. For example, non-citizens under the age of 64 without a college education are 90% likely to not have coverage, and citizens between 16 and 64 who are not married have a 71% chance of not having health coverage.

It is important to note that these cells are not arbitrarily defined nor does a black box guide their creation. The tree was grown using fundamental principles of a discipline known as *information theory* that helps to identify when an empirical variable contains information. 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 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.



The Gist. The structure of a decision tree 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. Given a labeled set of data that contains input features, the branches of a decision tree is grown by subsetting a population into smaller, more homogeneous units. In other words, moving from the root of the tree to the terminating branches, each subsequent set of branches should contain records that are more similar, more homogeneous or purer.

How is a decision tree empirically grown? For this, we can rely on pseudocode to lay out the process for the C4.5 algorithm, which is perhaps the most commonly implemented decision tree algorithm described in Quinlan (1993):

```
C4.5 (Sample = S, Target = Y, Input Features = X)
  Screen records for cases that meet termination criteria.
      If each base case that is met, partition sample to isolate homogeneous cases.
  For each input feature X, calculate the normalized information gain IG from splitting X into two subs
  Partition S into two partitions using feature X that corresponds to the highest IG.
  Repeat process for each sub-partition until termination criteria is met.
```

There are a number key concepts that guide the decision tree algorithm, including (1) termination criteria and (2) normalized information gain.

[Termination criteria]

[Information gain] To understand information gain means to understand the concept of *entropy*, which is a measure of purity or certainty of information. Given discrete classes or “states”, entropy is defined as:

$$\text{Entropy} = \sum -p_i log_2(p_i)$$

where i is an index of states, p is the proportion of observations that are in state i, and $log_2(p_i)$ is the Base 2 logarithm of the proportion for state i.

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]