Introduction to Decision Trees: Takeaways

Syntax

- Converting a categorical variable to a numeric value:

```
col = pandas.Categorical(income["workclass"])
```

- Retrieving the numeric value of the categorical variable:

```
col.codes
```

- Using Python to calculate entropy:

```
def calc_entropy(column):
"""
Calculate entropy given a pandas series, list, or numpy array.
"""
counts = numpy.bincount(column)
probabilities = counts / len(column)
entropy = 0
for prob in probabilities:
    if prob > 0:
        entropy += prob * math.log(prob, 2)
return -entropy
```

- Using Python to calculate information gain:

```
def calc_information_gain(data, split_name, target_name):
"""
Calculate information gain given a data set, column to split on, and target
"""
original_entropy = calc_entropy(data[target_name])
column = data[split_name]
median = column.median()
left_split = data[column <= median]
right_split = data[column > median]
to_subtract = 0
for subset in [left_split, right_split]:
    prob = (subset.shape[0] / data.shape[0])
    to_subtract += prob * calc_entropy(subset[target_name])
return original_entropy - to_subtract
```

Concepts

- Decision trees are a powerful and popular machine learning technique. The decision machine learning algorithm enables us to automatically construct a decision tree that tells us what outcomes we should predict in certain situations.
- Decision trees can pick up nonlinear interactions between variables in the data that linear regression cannot.
- A decision tree is made up of a series of nodes and branches. A node is where we split the data based on a variable, and a branch is one side of the split. The tree accumulates more levels as the data is split based on variables.
- A tree is levels deep where is one more than the number of nodes. The nodes at the bottom of the tree are called terminal nodes, or leaves.
- When splitting the data, you aren't splitting randomly; there is an objective to make a prediction on future data. To meet complete our objective, each leaf must have only one value for our target column.
- One type of algorithm used to construct decision trees is called the ID3 algorithm. There are other algorithms like CART that use different metrics for the split criterion.
- A metric used to determine how "together" different values are is called entropy, which refers to disorder. For example, if there were many values "mixed together", the entropy value would be high while a dataset consisting of one value would have low entropy.

- The formula for entropy is where is a unique value in a single column, is the probability of the value occurring in our data, is the base of our logarithm, and is the number of unique values in a single column.
- You can use information gain to tell which split will reduce entropy the most.
- The formula for information gain is the following:

is information gain, is our target variable, is the variable you are splitting on, and is the number of times a unique value is the target variable.

Resources

- Pandas categorical class documentation
- Information Theory
- Entropy