Decision Tree Classification of 2 Datasets

Lauren Howard - Presentation 18

Datasets

Mushroom dataset

- 8124 examples
- Features are categorical
- 22 features, 117 feature categories
- Classification is whether a mushroom is edible or not
- Example features: (cap-shape, cap-color, stalk-shape, etc.)

Iris dataset

- 150 examples
- Features are continuous
- 4 features
- Classification is iris type (Setosa, Versicolor, or Virginica)
- Features: (sepal-length, sepal-width, petal-length, petal-width)

Methodology

Source code available at: https://github.com/lorenmh/coen240-project

Run over 10 iterations with random test/train data, 25% test data, 75% train data.

Using Python 3 with Pandas, Numpy and Scikit-Learn to classify. sklearn.tree.DecisionTreeClassifier used to perform classification.

Scikit-Learn's DecisionTreeClassifier expects features which are continuous. Because the mushroom dataset has categorical string features, these need to be vectorized by converting the string categories into integer categories (LabelEncoder), and then using One-Hot Encoding (OneHotEncoder) to convert the 22 features into a 117 feature vector with values of either 1 or 0.

Results

No observed difference between Gini Impurity or Information Gain based decision tree classifiers.

According to the paper 'Theoretical Comparison between the Gini Index and Information Gain Criteria', Gini Impurity and Information Gain (entropy) will only differ in 2% of all cases which is likely why the results are the same.

100% prediction accuracy for all runs for the mushroom dataset.

~89%-100% prediction accuracy for the iris dataset.