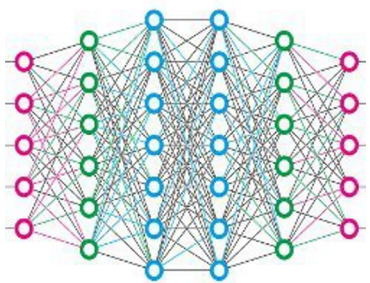
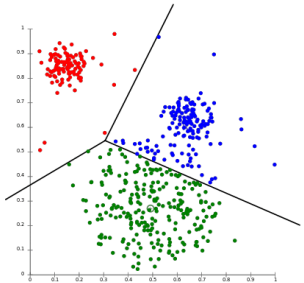
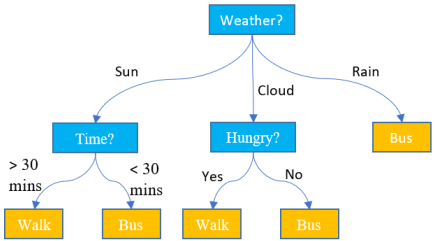
Welcome to CS 445
Introduction to Machine Learning

Feature Characterization and SkLearn's Trees

Instructor: Dr. Kevin Molloy







Announcements

- Quiz 1 on Canvas was due yesterday at 11:59 pm
- PA 0 is due this coming Monday.
- Continue to work on PA 1.

Learning Objectives From Earlier

- Define and give an example of nominal and ordinal categorical features
- Define and give an example of interval and ratio numeric features.
- Utilize a decision tree to predict class labels for new data.
- Define and compute **entropy** and utilize it to characterize the impurity of a set
- Define an algorithm to determine split points that can be used to construct a decision tree classifier.

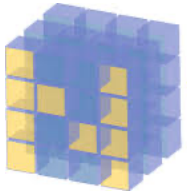
Muddiest Points

- Purpose of determining categorical (nominal, ordinal) and numeric (interval/ratio).

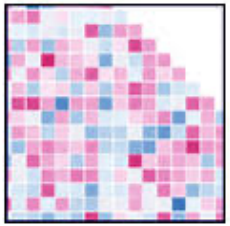
- Video posted (see Canvas's module section for this class):
  - Clarifies computing entropy for the parent node of the tree
  - Splits on continuous data
  - Definition and need for an impurity measure/formula

Learning Objectives for Today

- Utilize **NumPy and Seaborn** to visual and interpret the distribution of values for individual features in plot and whisker-plot format



- Utilize **Scikit-Learn's** DecisionTree and for classification and regression.



- Utilize accuracy, error rate, sum of squared errors (SSE) and mean squared error (MSE) to characterize model performance.

Lab on Data Investigation and SkLearn Decision Trees

Lab Today will use Jupyter Notebooks.

Download the lab for today's class from the class website and save it on your desktop. After completion, you will submit this notebook to Canvas.

Plan for Today

- Complete Lab Activities 1 – 2 (groups of 2 to 3 people)
- Discussion

- Complete Lab Activities 4
- Discussion

- Complete Lab Activity 5
- Discussion

- Complete Lab Activity 6 and 7
- **Submit** completed PDF to Canvas

Plan for Today

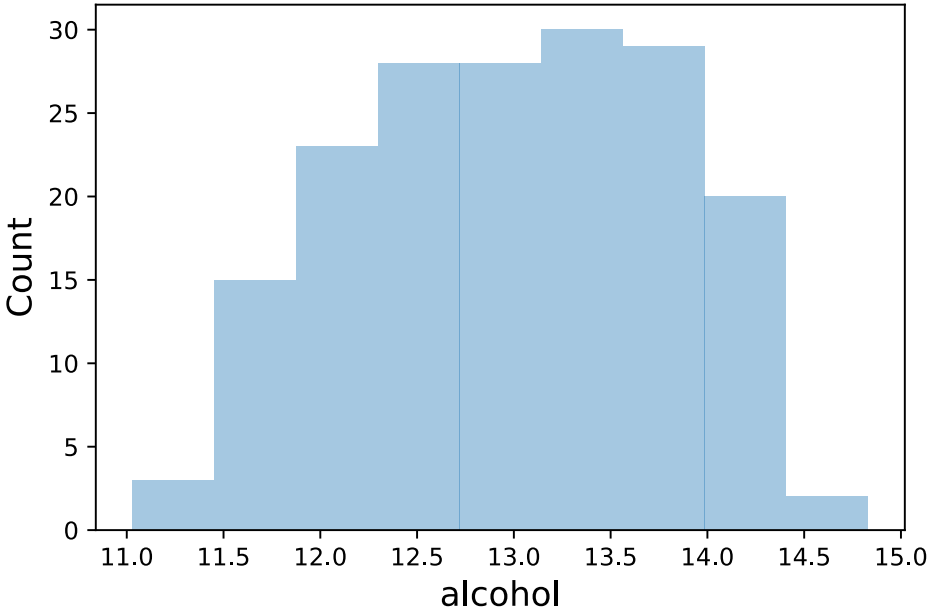
- Complete Lab Activities 1 – 3 (groups of 2 to 3 people)
- Discussion

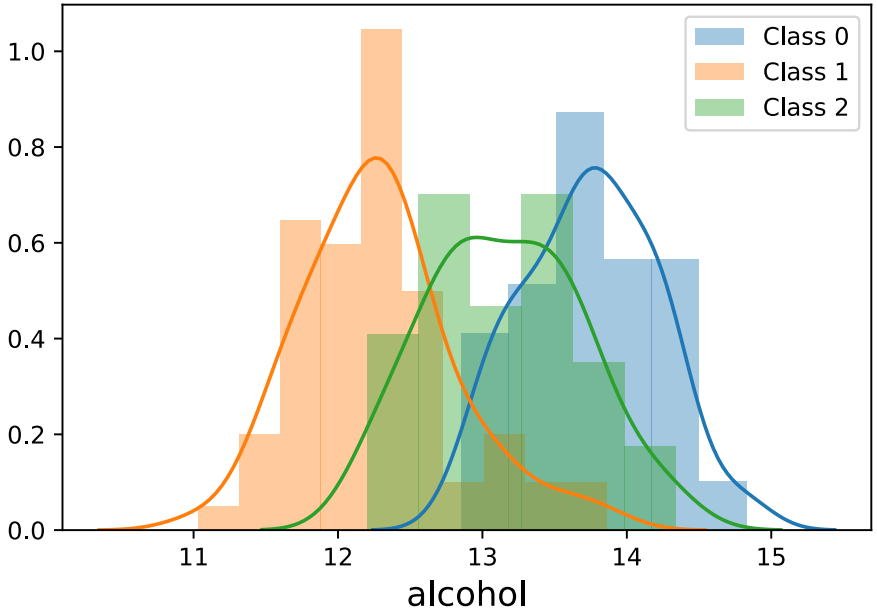
- Complete Lab Activities 4
- Discussion

- Complete Lab Activity 5
- Discussion

- Complete Lab Activity 6 and 7
- **Submit** completed PDF to Canvas

Feature Distributions

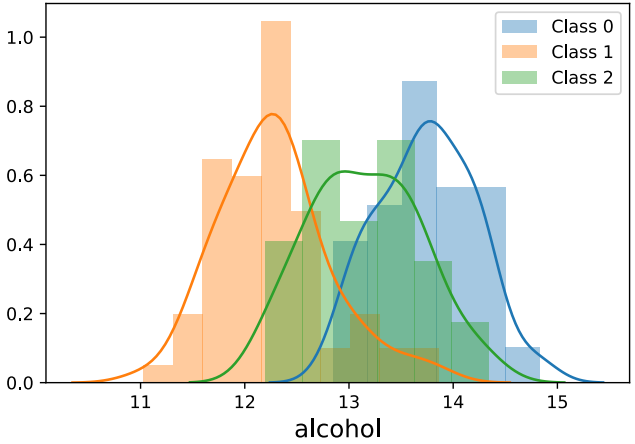


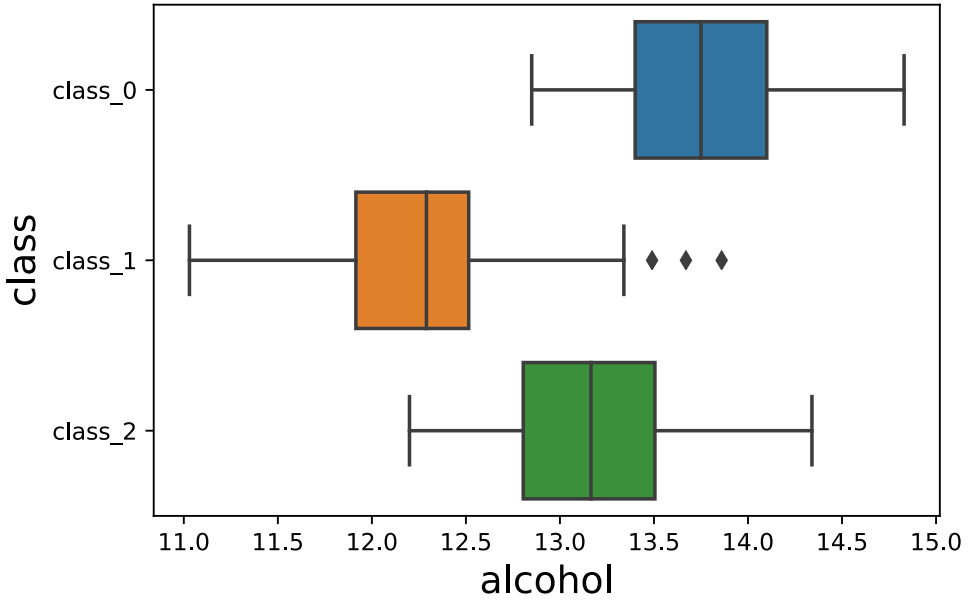


Being able to visualize the separation of the classes with respect to a feature's distribution provides valuable insight.

What would be a good split point for this feature? Would it work well?

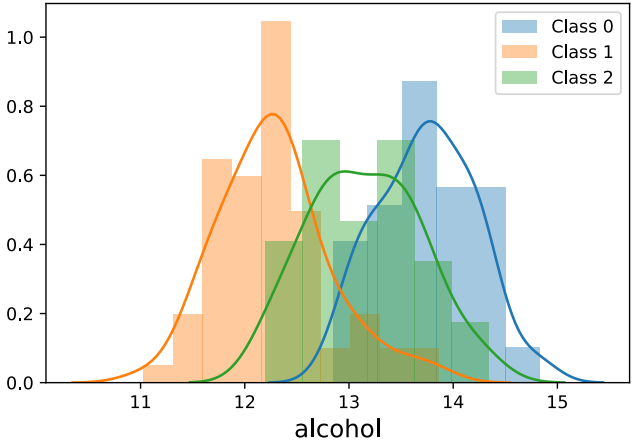
Feature Distributions

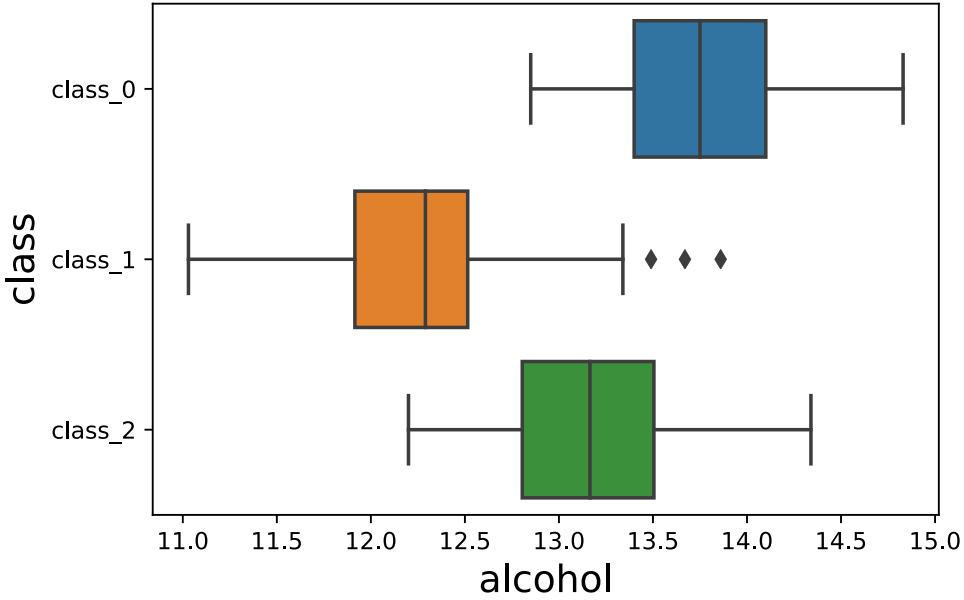




Class	count	mean	std	min	25%	50%	75%	max
class_0	59	13.7	0.46	12.9	13.4	13.75	14.1	14.83
class_1	71	12.28	0.54	11.0	11.9	12.3	12.5	13.9
class_2	48	13.2	0.5	12.2	12.8	13.2	13.5	14.3

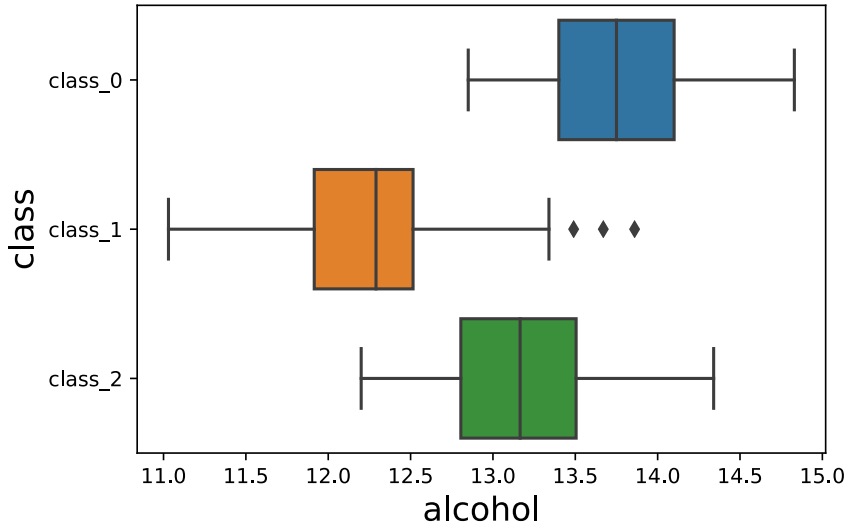
Feature Distributions



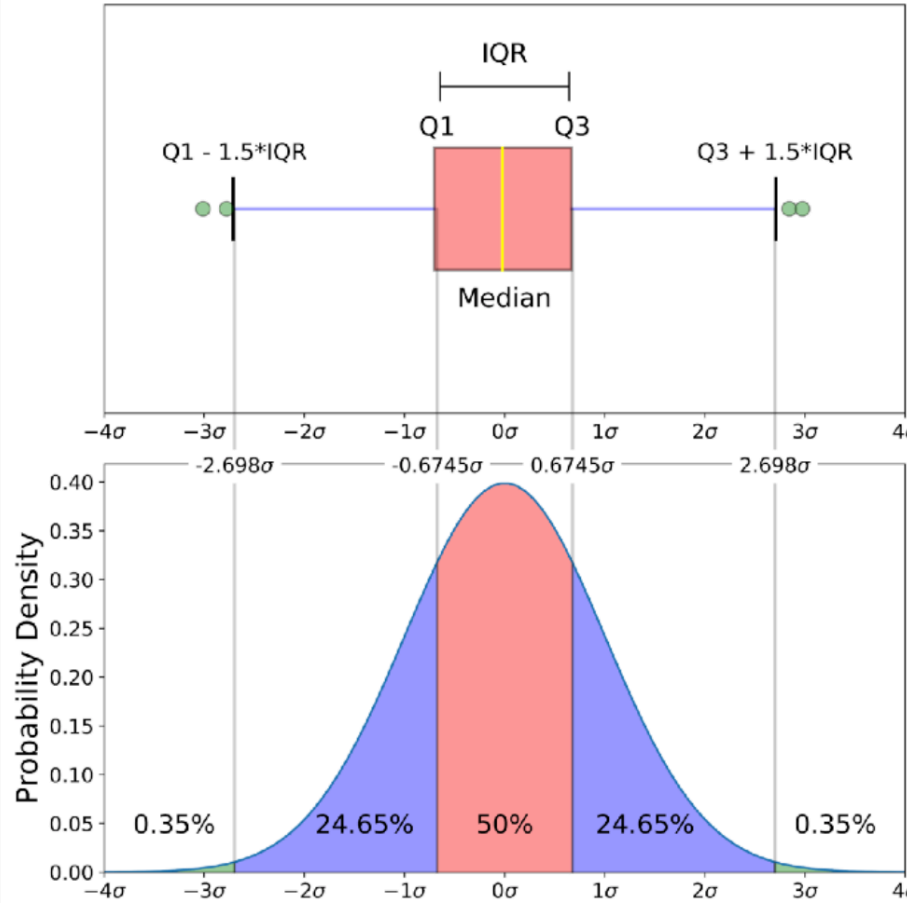


Class	count	mean	std	min	25%	50%	75%	max
class_0	59	13.7	0.46	12.9	13.4	13.75	14.1	14.83
class_1	71	12.28	0.54	11.0	11.9	12.3	12.5	13.9
class_2	48	13.2	0.5	12.2	12.8	13.2	13.5	14.3

Feature Distributions

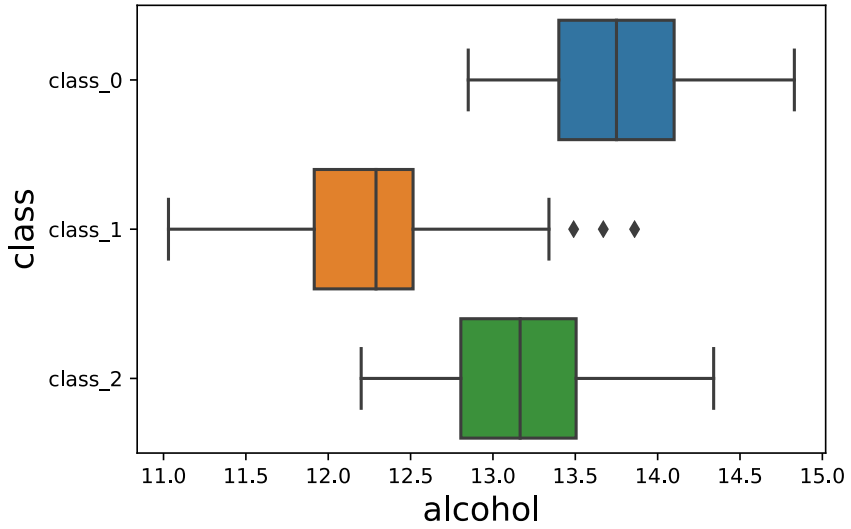


Class	count	mean	std	min	25%	50%	75%	max
0	59	13.7	0.46	12.9	13.4	13.8	14.1	14.8
1	71	12.3	0.54	11.0	11.9	12.3	12.5	13.9
2	48	13.2	0.5	12.2	12.8	13.2	13.5	14.3

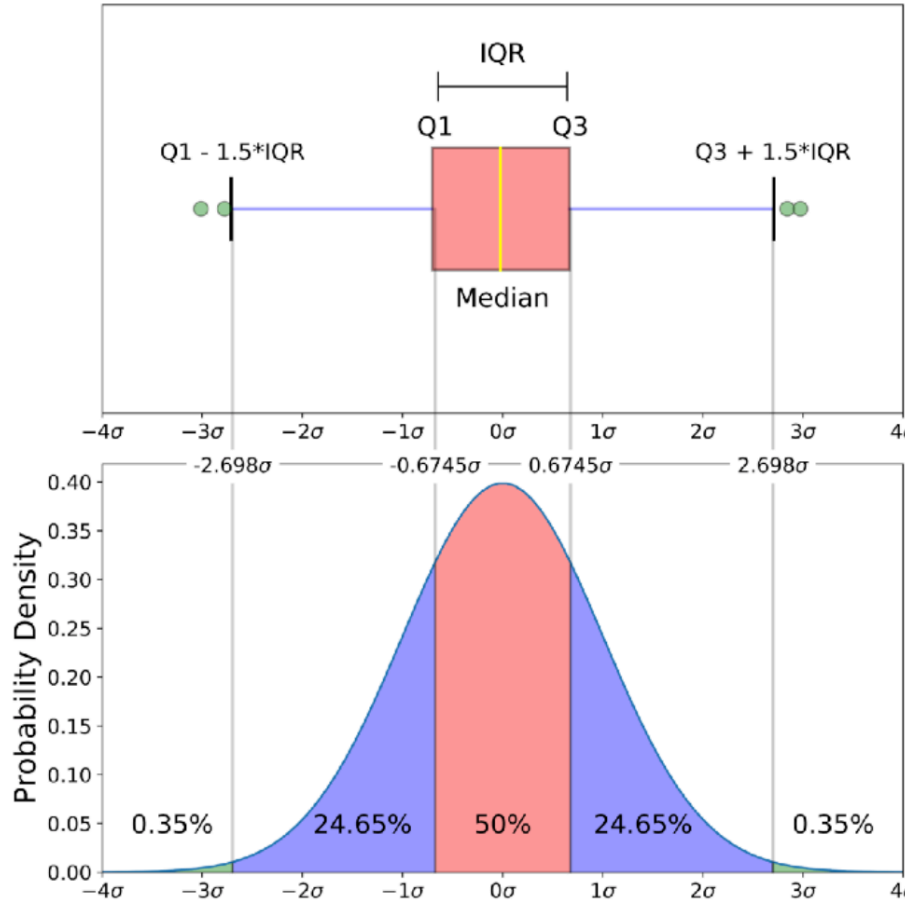


Plot from Michael Galarnyk

Is the Mean Meaningful?



Class	count	mean	std	min	25%	50%	75%	max
0	59	13.7	0.46	12.9	13.4	13.8	14.1	14.8
1	71	12.3	0.54	11.0	11.9	12.3	12.5	13.9
2	48	13.2	0.5	12.2	12.8	13.2	13.5	14.3



Plot from Michael Galarnyk

Plan for Today

- Complete Lab Activities 1 – 3 (groups of 2 to 3 people)
- Discussion

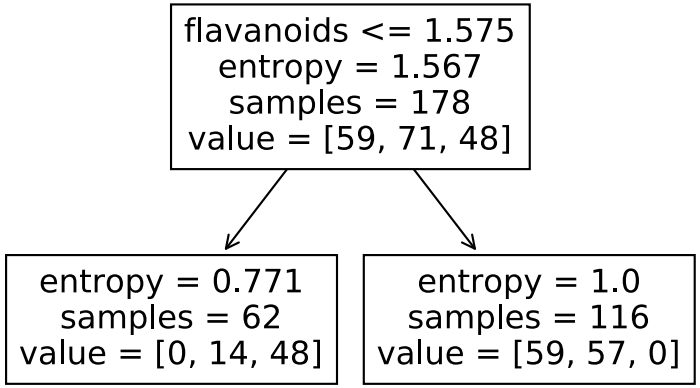
- Complete Lab Activities 4
- Discussion

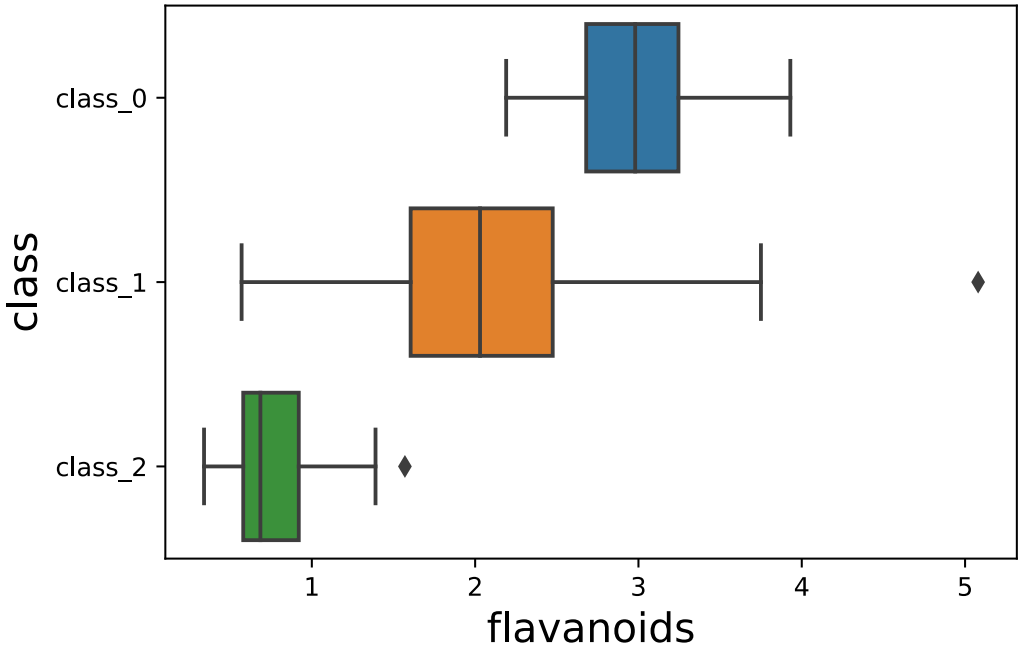
- Complete Lab Activity 5
- Discussion

- Complete Lab Activity 6 and 7
- **Submit** completed PDF to Canvas

Building a Decision Tree

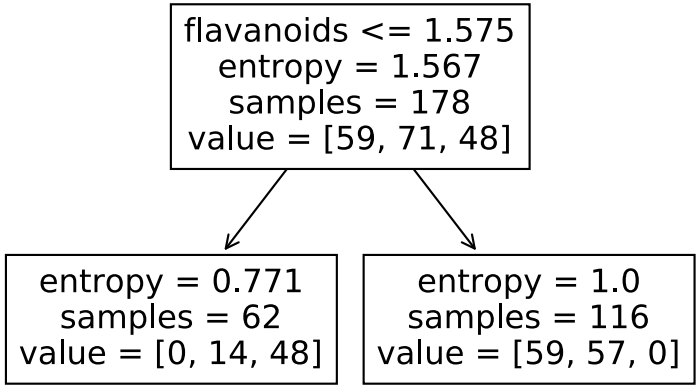
```
1. classifier = \
    tree.DecisionTreeClassifier(criterion='entropy', max_depth=1)

2. classifier.fit(wine_data['data'], wine_data['target'])
```





Predicting with the Tree

```
y_pred = classifier.predict(X)
```



What accuracy did you get back from the tree for the training data?

Plan for Today

- Complete Lab Activities 1 – 3 (groups of 2 to 3 people)
- Discussion

- Complete Lab Activities 4
- Discussion

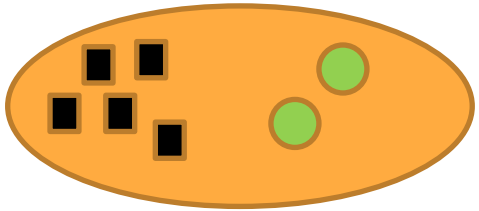
- Complete Lab Activity 5
- Discussion

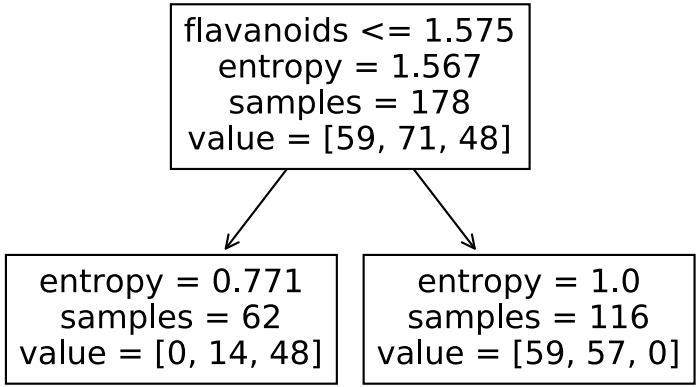
- Complete Lab Activity 6 and 7
- **Submit** completed PDF to Canvas

Regression with Decision Trees

Classification tree:

- Splits reduce entropy (best info gain)
- Prediction is the majority class in the leaf

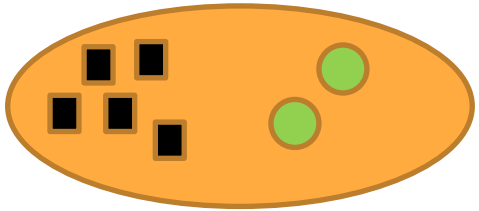


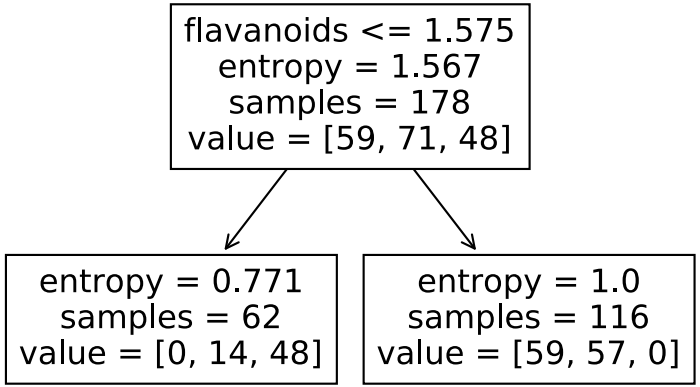


Regression with Decision Trees

Classification tree:

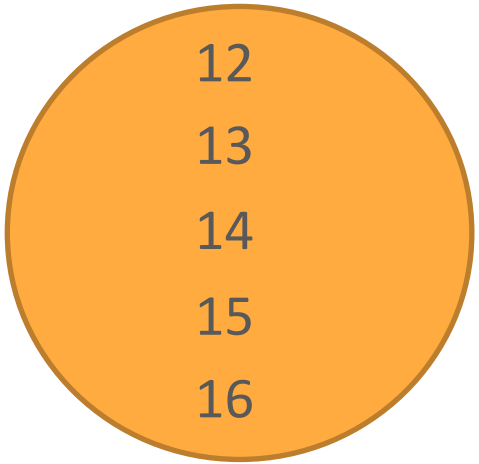
- Splits reduce entropy (best info gain)
- Prediction is the majority class in the leaf





Modifications (from a classification tree):

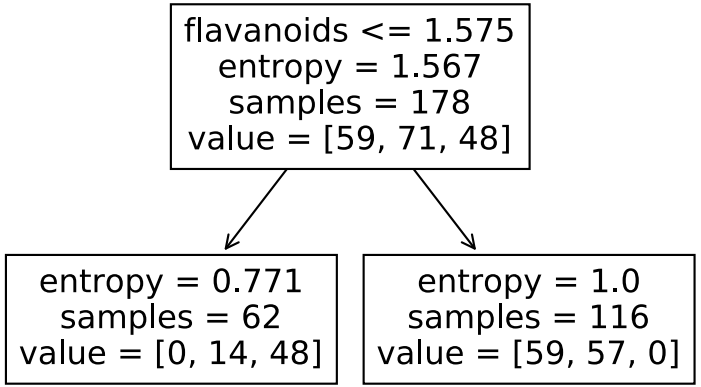
- MSE measures the "quality" of a potential split
- SSE = $\sum_{i=0}^{n-1}(y_i - f(x_i))^2$
- MSE = $\frac{1}{n}$SSE
- Prediction is the average of the examples in a leaf



Quantifying Tree Performance

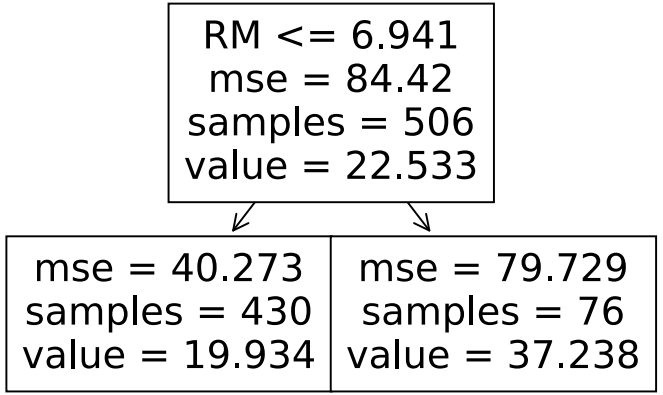
Classification

- Accuracy
- Error Rate
- Confusion Matrices



Regression:

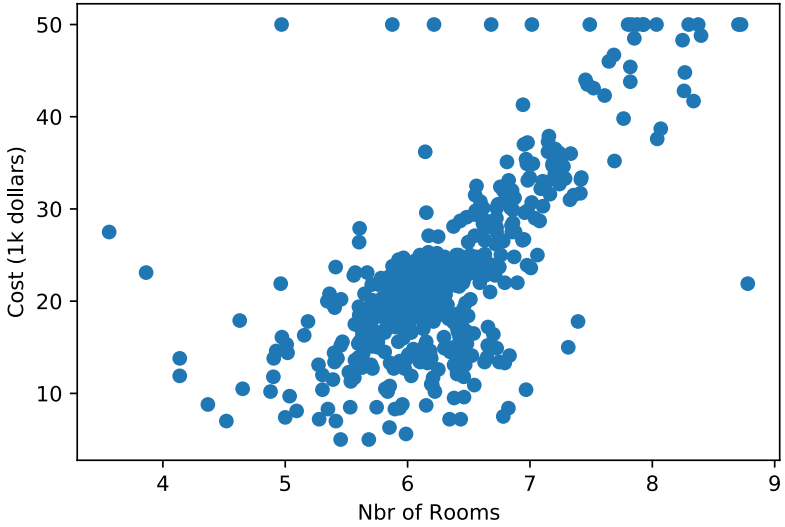
- SSE = $\sum_{i=0}^{n-1}(y_i - f(x_i))^2$
- **MSE = $\frac{1}{n}$ SSE**



Regression with Decision Trees

Boston Housing Dataset

- 14 attributes
- 506 datapoints

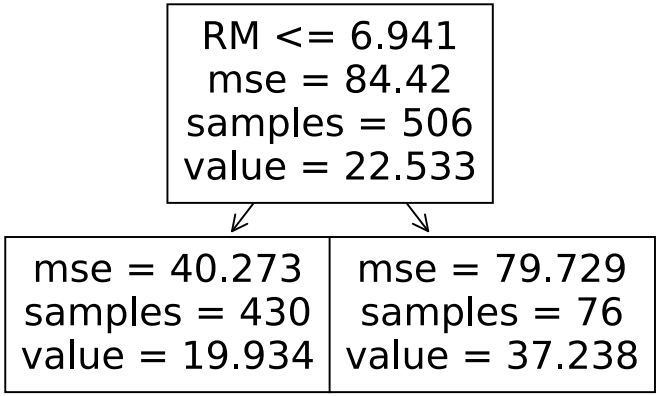


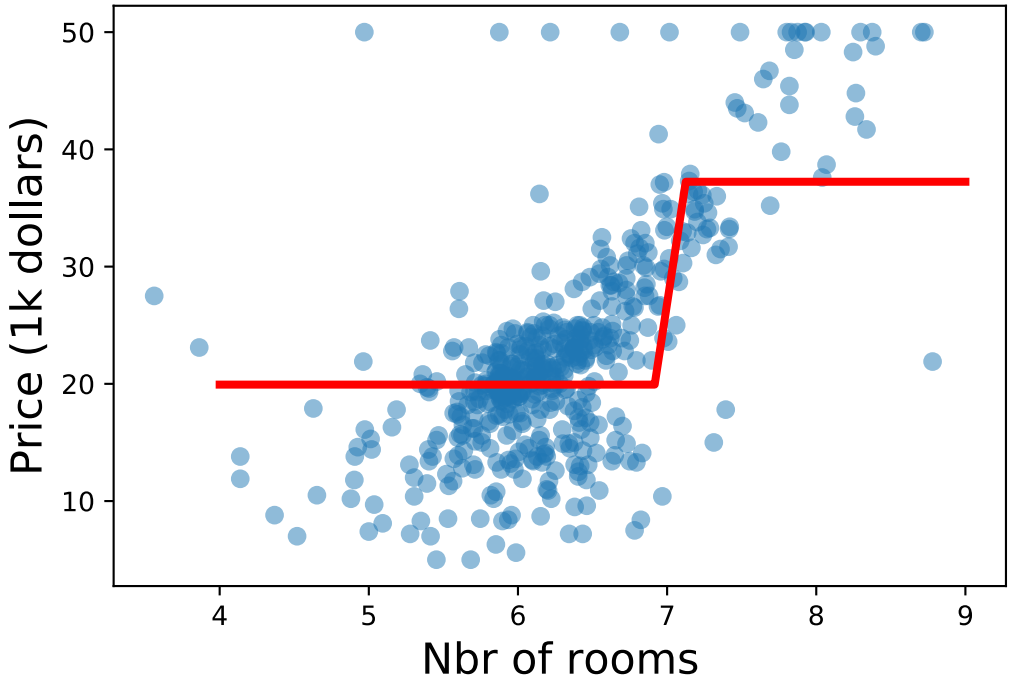
Objective: To predict the price of a home.

This is not a discrete set, but rather a value.

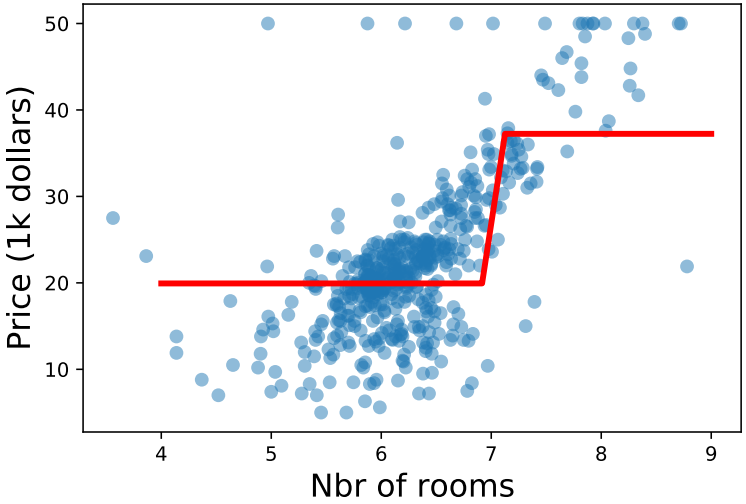
This makes this a **regression** problem.

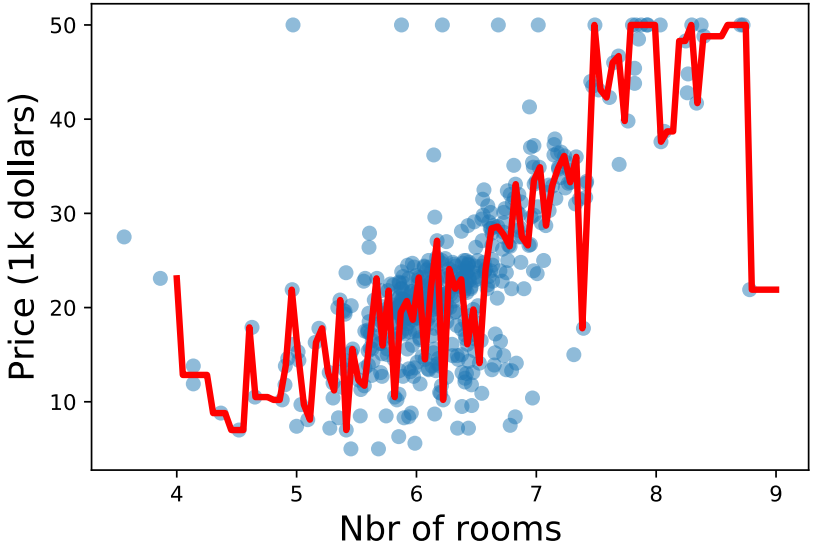
Regression with Decision Trees





Regression with Decision Trees





Tree with 2 leaves (stump)

MSE: 46.2 on training data

Tree with 441 leaves

MSE: 4.4 on training data

For Next time

Homework:

- Complete lab and submit to Canvas by Fri at 9 PM.
- Complete PA 0 and submit to Autolab by 11:59 PM Monday
- Work on PA 1

Reading: IDD Sections 2.1 and 3.3

Next Class: Lab on Model Selection and Validation