Announcements

● Project 3 will be released on Wednesday. Get ready to classify song lyrics.

● Homework due Wed/Thurs as usual.

● Monday 2-5 office hours in 3106 Etcheverry from now on.
Regression

- Estimating or predicting one numerical variable $y$ based on other variables
- Because $y$ is numerical, you can make predictions like “$y$ will be between 13.8 and 15.1”.

- But what if $y$ were categorical? How would you predict it?
Classification

- Response variable is categorical; values are **classes**
- **Binary response**: Only two classes, 0 and 1

- Try to **classify** the response into one of the categories, based on:
  - Values of **predictor variables**, called **attributes**
  - **Training set** of data in which the classes of the individuals are known
Nearest Neighbor Classifier

- New individual, unknown class
- Find individual in training set “closest” to this new individual
  - That’s the new individual’s “nearest neighbor”
- Assign the new individual the same class as the nearest neighbor

(Demo)
**k-Nearest Neighbor Classifier**

- New individual, unknown class
- Find the $k$ closest individuals in the training set
  - They are the new individual’s “$k$ nearest neighbors”
- Assign the new individual the same class as the majority of the $k$ nearest neighbors ($k$ is usually taken to be an odd number)

(Demo)
By the Numbers

- Binary response
- Multiple attributes
- $k$-nearest neighbor classifier
Accuracy of Classifier

● What fraction of individuals does it classify correctly?

● Need to compare:
  ○ Classifier’s predictions
  ○ True classes of individuals

● For this, need to know the true classes. But we only know those for the training set. So now what?
The Test Set

- Split original training set at random into two sets

- Use one of the sets for training:
  - Explore as much as you want
  - Develop classifier

- Use the other set (test set) to compare the classifier’s predictions and the true classes (Demo)
Rows of Tables

- Each row contains all the data for one individual
- `tbl.row(i)` evaluates to i\textsuperscript{th} row of \texttt{tbl}
- Type: “row object”; not all elements are of same type
- `tbl.row(i).item(j)` evaluates to item indexed \( j \) of \texttt{tbl} row indexed \( i \)
- If all elements of a row \texttt{my_row} are of the same type (e.g. all numerical), then \texttt{np.array(my_row)} evaluates to an array containing the elements of \texttt{my_row}