Data 8, Lab 11

Classification and k-Nearest Neighbours

Hubert Luo

Fall 2019

22 November 2019

Berkeley
UNIVERSITY OF CALIFORNIA

Classification

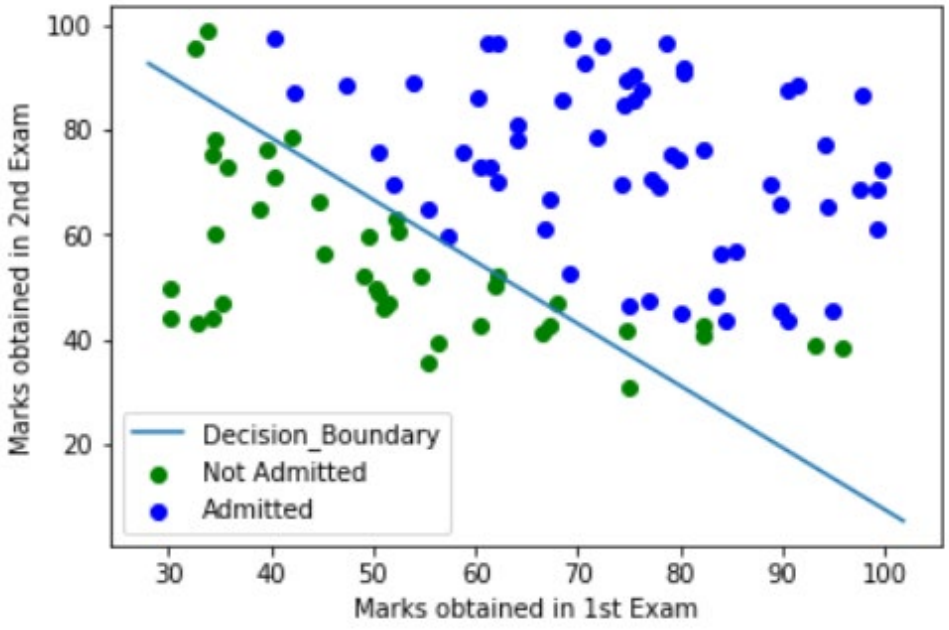
- Goal: Predict categorical data
- In this course, we will use known, labelled data points to predict the label of unknown data points
- Given a set of attributes for a data point, what label do we predict the data point to have?
- Examples:
  - Predict whether or not a patient has cancer
  - Predict the year of a student at Cal
  - Predict if an email is spam

Training vs Testing Data

- Training Data: All known, labelled data points
  - We use our training data to **create** our model
- Test Data: All unknown data points whose labels we are trying to predict
  - We use our test data to evaluate how well our model does
  - Allows us to ask how well our model **generalizes** to unknown data our model hasn't seen yet

Decision Boundary

- Decision Boundary: The curve that divides all the data based on the **predicted** label
  - May have points on the wrong side if the predictions are wrong!



K-Nearest Neighbours

Idea: Use the labels of the known data points (training set) closest to an unknown data point (test set) to predict its label

1. Find the distance between the unknown data point and each known data point in the training set
2. Sort all the data points based on the calculated distance
3. Take the k closest data points ("neighbours") and find their labels
4. The predicted label for the unknown data point is the majority of the labels of the k closest neighbours

Standardizing Data

- Before starting to classify data, we often have to **standardize data first**
- This is especially true if data is on completely different scales!
  - How do we calculate a distance that involves both the number of people in a town and the area of a town in kilometers squared?
  - These two variables clearly have completely different scales!
- $$Standardized\ Data = \frac{Origial\ Data - Average\ of\ Original\ Data}{Standard\ Deviation\ of\ Original\ Data}$$

Announcements

- Project 3 checkpoint 1 is due today (11/22)
  - Final project deadline is 12/6.
- HW12 will be released today and due on 12/6.
- Next week is American Thanksgiving:
  - Monday lecture will be held as normal
  - No lecture on Wednesday-Friday, and no lab, office hours, tutoring sections, or assignment deadlines