Data 8, Lab 10

Classification, k-Nearest Neighbours, and Conditional Probability

Hubert Luo

Spring 2020

1 May 2020

Berkeley
UNIVERSITY OF CALIFORNIA

Classification

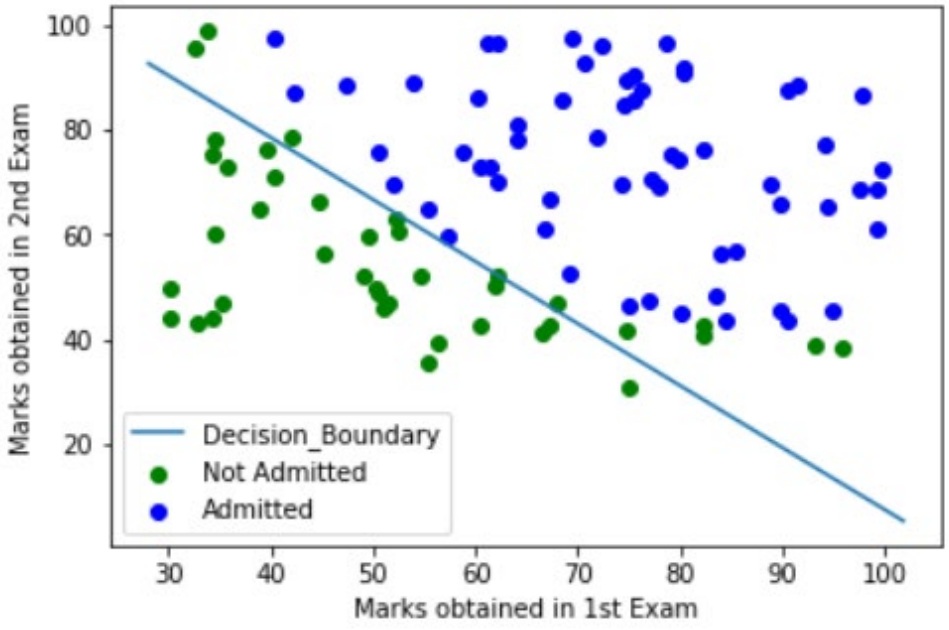
- Goal: Predict labels of categorical data
- In Data 8: Use known, labelled data points to predict the label of unknown data points in a supervised manner
- Given a set of attributes for a data point, what label do we predict the data point to have?
- Examples:
  - Predict whether or not a patient has 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
    a. From closest (smallest distance) to farthest (largest distance)
3. Take *k* closest data points ("neighbours") and get 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 we classify data, we usually **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!
- $$Standardized\ Data = \frac{Origial\ Data - Average\ of\ Original\ Data}{Standard\ Deviation\ of\ Original\ Data}$$

Confusion Matrix Example

	True Label: Positive	True Label: Negative
Predicted Label: Positive	True Positive	False Positive
Predicted Label: Negative	False Negative	True Negative

- The above is an example of a **confusion matrix** used for classification with two labels, Positive and Negative.
- Example: A medical screening is conducted to predict whether or not a patient has cancer
  - True Positive: Patient has cancer and screening says they do
  - False Negative: Patient has cancer but screening says they don't
  - False Positive: Patient doesn't have cancer but screening says they do
  - True Negative: Patient doesn't have cancer and screening says they don't

Conditional Probability

- Let C and D be events, P denote the probability
- $\mathrm{P}(C\ Happening\ Given\ D\ Happened) = \frac{P(C\ and\ D\ both\ happening)}{P(D\ happening)}$
  - Out of scope for Data 8, but in probability theory we write the above as P(C|D) = $\mathrm{P}(C|D) = \frac{P(CD)}{P(D)}$
- Note that P(D happening) = P(C and D both happening) + P(D happens but C does not happen)

Conditional Probability Example (Q2b)

After implementing his classifier with a different k, Gregory runs the classifier on 1000 customers and finds that:

- 501 of the A customers were classified correctly
- 208 of the B customers were classified correctly
- 104 of the A customers were classified incorrectly
- 187 of the B customers were classified incorrectly

Question: Given that a customer was classified incorrectly, the likelihood that they are a B type customer

P(Type B customer given classified incorrectly)

$$= \frac{P(Type\ B\ customer\ classified\ incorrectly)}{P(Classified\ incorrectly)}$$

$$= \frac{P(Type\ B\ customer\ classified\ incorrectly)}{P(Type\ A\ customer\ classified\ incorrectly) + P(Type\ B\ customer\ classified\ incorrectly)}$$

$$= \frac{\frac{187}{1000}}{\frac{104}{1000} + \frac{187}{1000}} = \frac{187}{104+187} = \frac{187}{291}$$

Berkeley
UNIVERSITY OF CALIFORNIA

Announcements

- Project 3 due Friday 5/1
- Final Topical Review Labs Next Week! See Piazza for schedule, slides, and topics