DSFA

Spring 2021

Lecture 33

Nearest Neighbor

Announcements

- Prelim 2 regrade requests by today, 5PM
- Project 3 out, Part 1 due 5/7, Part 2 due 5/14.
- Labs the week of 5/3 and 5/10 will be dedicated to Project 3.
- Final May 22, 1:30PM, in Baker Lab 200

Announcements

Guest speakers Friday 5/7, Monday 5/10, Wednesday 5/12

- Yes, participation in these will still count (so please show up) and questions about the materials covered will be fair game for the final.
- Friday: Cornell COVID modeling team
- Monday: Professor Karen Levy, Information Science
- Wednesday: Dr. Ehi Nosakhare, Microsoft

Classification

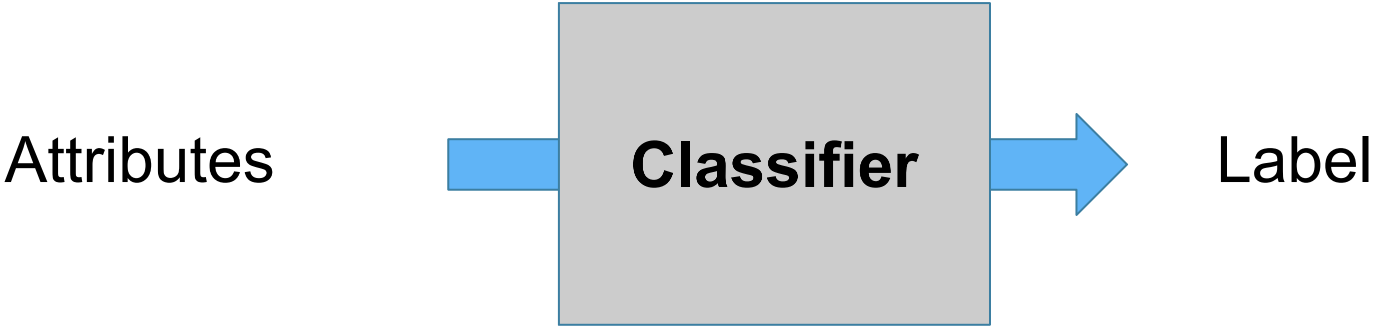
- Our study of regression:
  - One quantitative variable (x)
  - Predicts another quantitative variable (y)

- Now, classification:
  - Many quantitative variables
  - Predict a categorical variable

Classification Terminology

- Response variable: the categorical variable we try to classify
- Classes or labels: possible values of response variable
- Binary response: 0 or 1
- Attributes: variables used to make classification

Classifier



(Demo)

Nearest Neighbor

How to classify a new individual:

- Find their nearest neighbor: the individual closest to them in the data set
- Assign the new individual the **same** label as that nearest neighbor

(Demo)

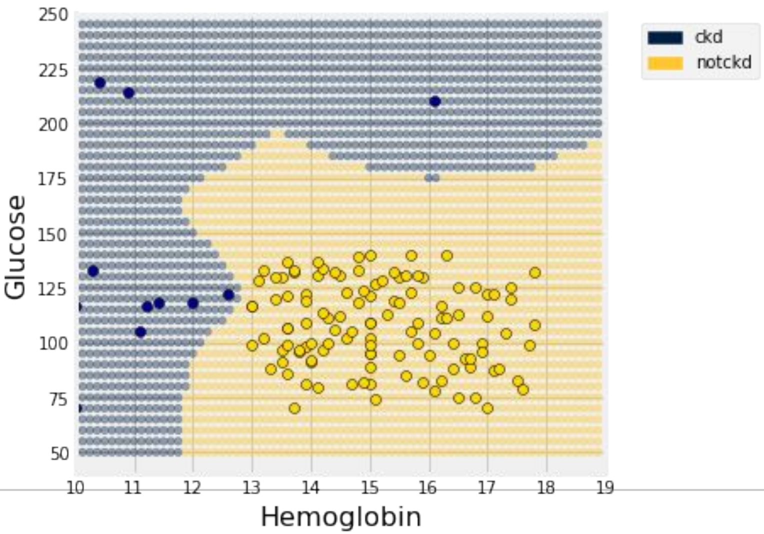
Nearest Neighbor recap

How to classify a new individual:

- Find their nearest neighbor: the individual closest to them in the data set
  - (We put data in standard units because scale of one attribute was so different than the other attribute--you will **not** need to do that on your proj3)
  - Compute table of distances from that individual to all other individuals
  - Sort by distance, so that closest is in the first row
- Assign the new individual the **same** label as that nearest neighbor

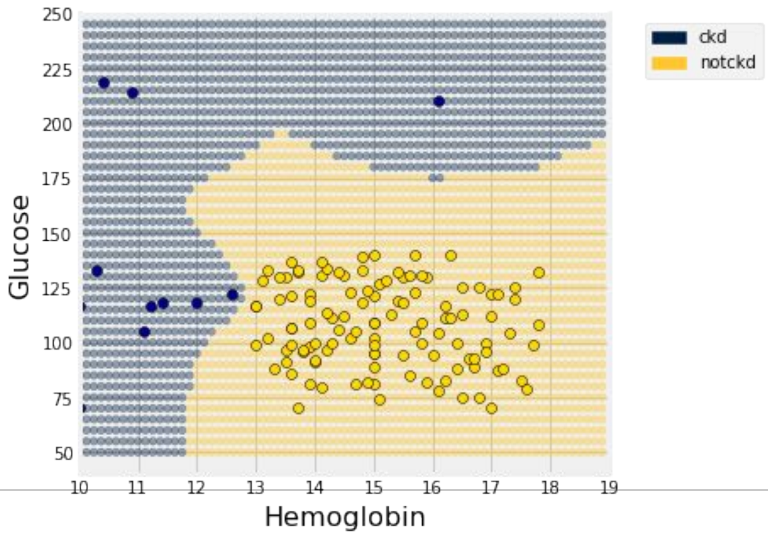
Decision Boundary

- Partition between the two classes
- Computer figured out that boundary, instead of humans having to “hard code” it: machine learning

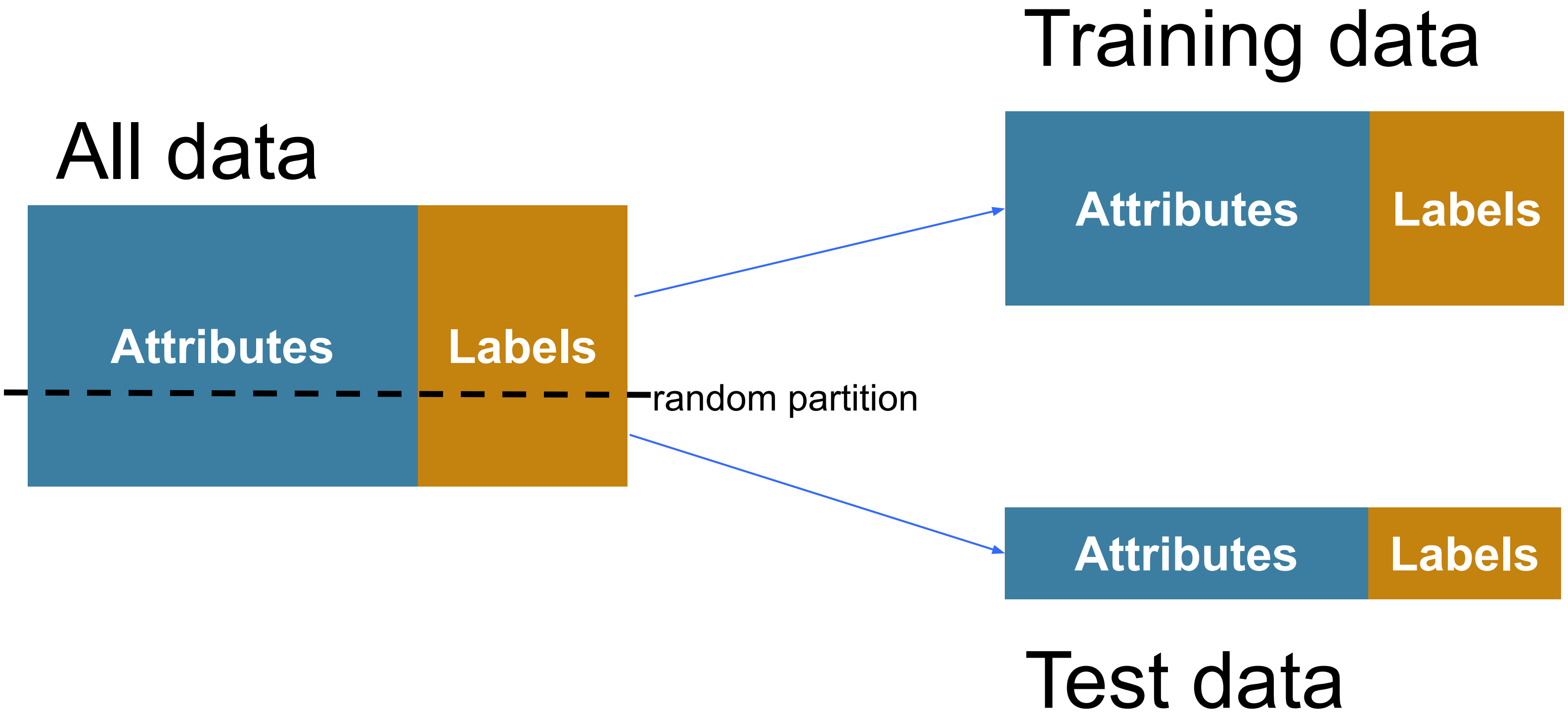


Evaluating a Classifier

How do we evaluate whether classifier is doing a good job on all those points where we have no data?



Train vs. Test



Train vs. Test

- Use training data to create the classifier
- Use test data to evaluate the finished classifier

- **Never** allow classifier to see test data until the very end: think of classifier as a cheater who would be happy to just memorize the answers

(Demo)

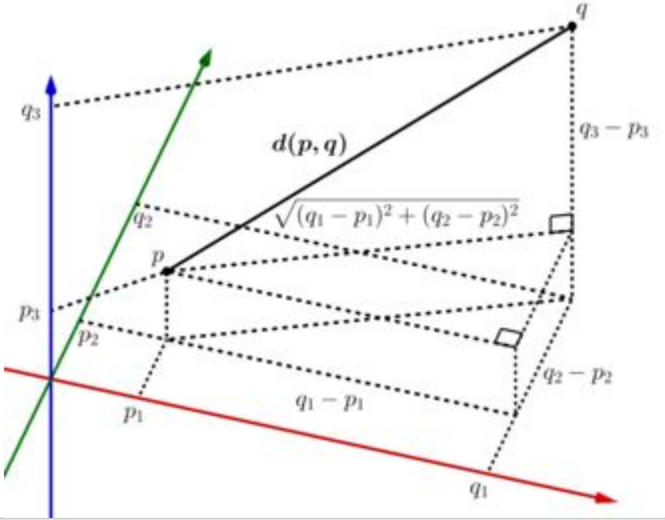
Multiple Neighbors

- If data are noisy, asking just the closest neighbor might not be ideal for accuracy
- Instead, ask the *k* closest neighbors, and take the majority label

(Demo)

Multiple Attributes

- We’ve used 2 attributes so far
- But nothing special about 2, just have to compute distances in higher dimensional spaces



(Demo)