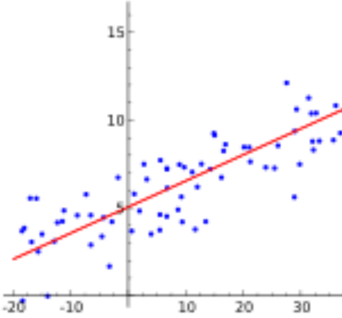
Review quizzes

- Illustrate an example in which using a traditional algorithm may be more appropriate than using a data-driven approach
- Illustrate an example in which using a data-driven approach may be more appropriate than using a traditional algorithm
- Explain kNN
- Why features are usually scaled before using kNN?
- Explain k-means

The Fizz Buzz joke

- Source: http://joelgrus.com/2016/05/23/fizz-buzz-in-tensorflow/
- Fizz Buzz
  - Print numbers between 1 and 100, except that if the number is divisible by 3 print “fizz”, if the number is divisible by 5 print “buzz”, and if it’s divisible by 15 print “fizzbuzz”
- The interviewee uses a “data driven” approach to solve the problem
  - Modeled as a classification problem
  - Collect training data
  - Prediction: classify a number into “fizz”, “buzz”, “fizzbuzz”, or a number itself

Parameters vs hyper-parameters



- Parameters: the unknown variables your models need to learn
  - E.g., *y*=*ax*+*b*, *a* and *b* are parameters
- Hyper-parameter: the variables your models cannot learn (need to specified manually)
  - E.g., the variable *k* in the knn model
- Sometimes we may abuse these terms
  - E.g., let's fine tune the parameters to improve the performance…

Dealing with ties in KNN

- KNN is based on one simple intuition: closer data points should be similar
- How to deal with ties? ⍰ You could do what ever you believe reasonable to break the tie
- Common practices include
  - Randomly select a class
  - Gradually decrease *k* by one until you break the tie

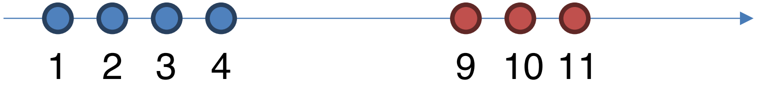
KNN with different weights

- In previous courses, we introduced knn with equal weights
  - All *k* closest neighbors are weighted equally
- However, we may assign different weights to different neighbors
- One common approach is to weight the neighbors based on the inverse of their distance

How to “evaluate” clustering result? (1/3)

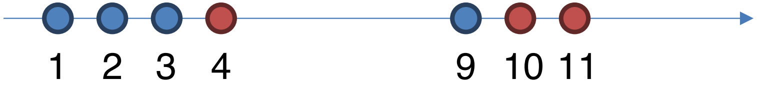
- No “correct” answer of clustering
  - Is “evaluating clustering result” reasonable?
- If really want a quantified measurement, one possible way is to compute the ratio of inter-cluster distance and intra-cluster distance
  - Large ratio probably means better clustering
- Various ways to define inter-cluster and intra-cluster distances

How to “evaluate” clustering result? (2/3)



- Averaged inter-cluster distance:
- Averaged intra-cluster distance:
- Ratio

How to “evaluate” clustering result? (3/3)



- Averaged inter-cluster distance:
- Averaged intra-cluster distance:
- Ratio

How to do knn or k-means with categorical features?

- If the definition of “distance” between categories is vague, consider one-hot encoding
- Example:
  - Values of the “nationality” feature: “UK”, “Japan”, “Mexico”
  - We may encode UK as “1,0,0”, Japan as “0,1,0”, and “Mexico” as “0,0,1”
  - “Nationality feature” becomes three features: “UK or not”, “Japan or not”, “Mexico or not”

How to do knn or k-means with categorical features?

- If the definition of “distance” between categories can be somewhat defined, perhaps use the definition
- Example:
  - Values of “height”: tall, average, short
  - We may encode tall as 3, average as 2, short as 1