* 194 students and 122 undergrads :O
* Room change isn’t permanent – will move rooms around
* Recitation is 7:30 pm to 9:00 pm Thursday in 105 Annenberg and will usually be 45 to 60 minutes long
* Homework 1 will be released tonight and due next Friday
* 6 homeworks each worth 10% and two mini projects worth 30 % of final grade
  + Classification
  + Visualization
  + Generation project
* Final exam worth 10% but no midterm
* Up to 48 free late hours – should specify the number of late hours
* Supervised learning
* Find function that maps from input space to output space
* Take data and target signal and you get a model/function that approximates the signal y
* Unsupervised learning has no target – there’s nothing to predict
  + Learning goal is to find some low dimensional summary or reconstruction
* Bag of words is a simple way to represent words
* Each email/collection of words is represented by a long vector and each element in the vector is a word – in practice the feature vector may be millions or billions long because it has to cover all combinations of words
* Linear classifier is defined by vector w and scalar b
  + Workhorse of machine learning
* Loss function gives how off we are from the desired target
* Learning objective is to minimize the sum of losses
  + Other types are classification and squared loss
* Simplest optimization algorithm is gradient descent
  + Go in the opposite direction of the partial derivative and look for several iterations
  + Taking derivatives with respect to w
* Computing gradient for 0 1 loss:
  + Cannot compute gradient of 0 1 loss because there’s a discontinuity in the function and it’s not differentiable
  + Very difficult to optimize – the problem is intractable
  + There’s no slope that tells me which direction is better
  + Optimize smooth surrogate loss function that has desirable behavior
  + One is squared loss, but it’s not the best surrogate
* Test error is the thing you actually care about ( out of sample error)
* You want to match to the “true” distribution
* Instead we’re training on the training data and we are assuming it was sampled i.i.d from the true distribution P(x, y)
* Overfitting occurs when test error is much larger than training error
* The resulting model fs can be treated as a random variable
* The expected test error is the expected value of test error across the different possible training sets
* Variance is variability and how much it deviates from the average prediction
* The bias term is difference between target label and average prediction
* Chart with red and blue graphs – we repeatedly sample different points (blue points) identically and independently distributed
* The variability is the variation in the training models
* Bias-variance trade off
  + Black line is the average and the green line is within 1 standard deviation
  + The x axis is the value of the feature space
* High variance implies there is high overfitting
* High bias implies underfitting – with no variance the model class could have a high error
* Bias decreases with model complexity and is independent of training data size
* The data was generated with a cubic function corrupted with noise
* Can use a validation set and break it into a training set and a validation set and train the model on the training set and evaluate on the validation set
  + The problem is that the validation set may be too small
  + If this is a problem we’ll do something like 5-fold cross validation