EE P 596 Conceptual Assignment 1:
Due by 11:59pm Thursday, January 13

Qingchuan Hou

January 13, 2022

1. You run gradient descent for 15 iterations, with learning rate $\alpha = 0.3$ and compute loss function $J(w)$ after each iteration. You find that the value of $J(w)$ **decreases slowly** and is still decreasing after 15 iterations. Based on this, which of the following conclusions seems most plausible?

   (A). α is large, try to decrease α=0.1 would be better.
   (B). α is small, try to increase α=1.0 will make the model converge faster.
   (C). α is appropriate and no need to change.
   (D). J(w) cannot converge no matter what α you pick.

   Answer: B

2. Suppose you have m = 23 training examples with n = 5 features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is $w = (X^TX)^{-1}X^Ty$. For the given values of m and n, what are the dimensions of w, X, and y in this equation?

   (A). w is 5x1, X is 23x5, y is 23x1.
   (B). w is 5x5, X is 23x5, y is 6x1.
   (C). w is 6x6, X is 6x23, y is 6x1.
   (D). w is 6x1, X is 23x6, y is 23x1.

   Answer: D

3. Suppose you have a dataset with m = 50 examples and n = 200000 features for each example. You want to use multivariate linear regression to fit the parameters w to our data. Should you prefer gradient descent or the normal equation?

   (A). Normal equation because it is computationally efficient.
   (B). Gradient descent because it's more efficient to compute.
   (C). Normal equation because the number of examples are small.
   (D). Gradient descent because it's more accurate than normal equation.

   Answer: D

4. Which of the following is the reason for using feature scaling(data normalization)?

   (A). Because the calculation of normal equation will have no matrix inversion problems(i.e, singularity matrix).
   (B). Because solving the normal equation will be more efficient.
   (C). Because when optimizing by gradient descent, the speed of convergence will be faster.
   (D). Because solving the normal equation will be more accurate than without feature scaling.

   Answer: D

5. You are given a set of 2-D points (time, values), and you want to fit a curve to the points such thatthe curve can capture the relationship between time and values. In the 3 plots below, which of the following best describes how well the curves fit the points?

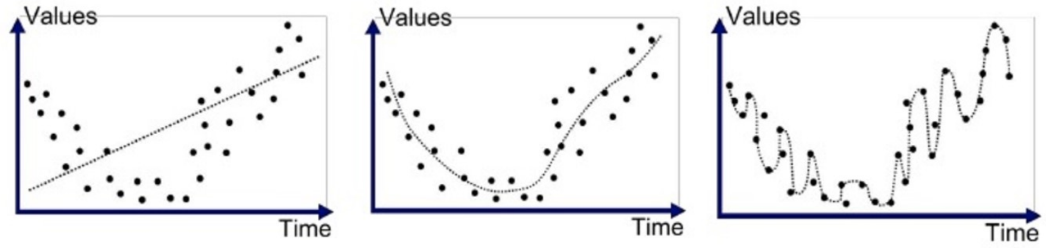

Figure 1: Fitting curve to a set of 2D points

   (A). 1st plot fits well, 2nd plot overfits, 3rd plot underfits.
   (B). 1st plot overfits, 2nd plot underfits, 3rd plot fits well.
   (C). 1st plot underfits, 2nd plot fits well, 3rd plot overfits.
   (D). 1st plot underfits, 2nd plot overfits, 3rd plot fits well.

   Answer: C

6. You trained 2 models - A and B on a dataset, the training and test error with respect to number ofiterations are shown below, which of the following is the best description for 2 models?

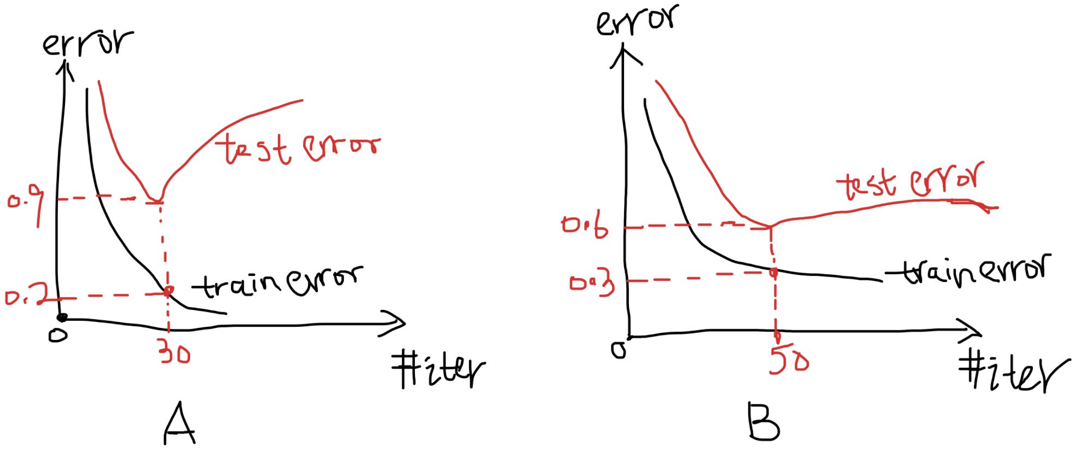

Figure 2: Train-test error plots for model A and B

(A). model A is a better fitting than B because it has lower training error.
(B). model A is performing better because it takes fewer iterations to converge.
(C). model B is less overfiting than model A.
(D). model B is performing better only because the gap between training and test error is smaller.

Answer: C