Quiz 2: Data science
Total Marks: 10 2018-04-06

Name: ---------- Registration #: ---------- Section: ------

Q1: If your model is under-fitting, do you think adding the more training examples will help to make it better? Explain your answer with Reasoning.
Ans: No because increasing number of examples does not help to reduce underfitting. Underfitting occurs due to simplicity of the model (model is too simple or relevant features are not included). As we increase the number of examples, the CV error does not keep on reducing.

Q2. Suppose we have three cluster centroids $\mu_1 = \begin{bmatrix}1\\2\end{bmatrix}$, $\mu_2 = \begin{bmatrix}-3\\0\end{bmatrix}$, $\mu 3 = \begin{bmatrix}4\\2\end{bmatrix}$.
Furthermore, we have training example $\mathbf{x^{(i)}} = \begin{bmatrix}1\\-2\end{bmatrix}$. After a cluster assignment step, what will $\mathbf{C^{(i)}}$ be?
(a) $\mathbf{C^{(i)}} = 1$ (b) $\mathbf{C^{(i)}} = 3$ (c) $\mathbf{C^{(i)}}$ not assigned
(d) $\mathbf{C^{(i)}} = 2$

Solution:

a) **However, Please verify it.**

Q3. When we want to select the best model from more than one available models, why do we need to divide the data into three portions (training, cross validation, and testing) instead of two portions (training, and testing)?
Ans: On training data we train the models and get the optimal parameter values. If Among these trained models, one of them is selected based On Test data. And we use the same test data to answer the question that it generalizes well on new data or not? It will be not correct because u r using the same data for model selection and same data for answering how well it generalizes to unseen data.

That's why we divide it into 3 portions.

Among these trained models, one of them is selected based On cross validation (CV) data. The selected one is with minim CV error. Then test data is used to answer that how well it generalizes to unseen data.

Q4. [2 points] **Diagnosing bias vs. variance: Answer the following questions:**

Quiz 2: Data science
Total Marks: 10 2018-04-06

Name: ---------- Registration #: ---------- Section: ------

(1). If $J_{cv}(\theta)$ and $J_{train}(\theta)$ are high such that ($J_{cv}(\theta) \approx J_{train}(\theta)$). Is it a bias problem or variance problem?

Ans: Bias

(2). If $J_{train}(\theta)$ is low and $J_{cv}(\theta)$ **>>** $J_{train}(\theta)$. Is it a bias problem or variance problem?

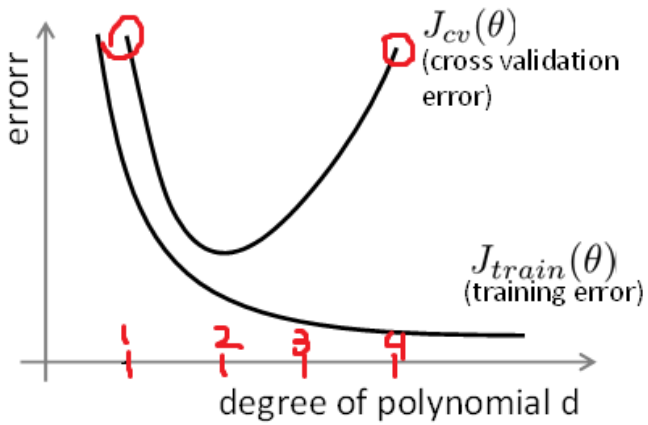
Ans: Variance

(3). For what value of d (degree of polynomial), the problem is underfit?

Ans: D=1 (or lower degree)

(4). For what value of d (degree of polynomial), the problem is overfit?

Ans: D=4 (or 3 to 4)



Q5. [3 marks] Data Wrangling: Suppose you want to train a model to predict the sale price of a house, given its size, number of stories and number of bedrooms. In order to build the prediction model, you need data for training, which you have collected from local property dealers.

X_1 (Size of house in Marla's)	X_2 (number of stories)	X_3 (number of bed rooms)	Y (Price in millions)
10	2	4	100

Quiz 2: Data science
Total Marks: 10 2018-04-06

Name: ---------- Registration #: ---------- Section: ------

20	2	8	150
30	1	6	200
1	1		10
100	3		400

1) After looking at the dataset, what do you think, what are problems in data set?

Ans: missing values of X_3

2) Suggest the appropriate approach to fix them. Explain your choice with reasoning. Give advantages and disadvantages of each apprach.

Ans: **Imputation**: Can try 2 kinds of imputations

1) Mean Imputation" if we take avg of values of X_3 thats not good because avg is 6.
   Adv: teh avg will be same.
   Disadv: but we can see that it will diturb the correlation with size of house and number of stories. How can 1 marls house has 6 bedrooms?
2) Imputation using Linear Regression: It will be good because we will train a model and then basedt on trained model we will do prediction for the missing values, that will be more close to reality. It will also not distrurb the correction betwen variables.