Quiz 2: Data science
Total Marks: 8 2018-04-06

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

Q1. The model bias typically tends to zero as the number of training data points tends to infinity?

a) False b) True

Reason: **No** it is not the case. Model Bias means underfitting, it does not improve as we increase the number of training examples.

Q2. Does K-mean algorithm always guarantee you the global optimum? Explain with Reasoning. Moreover how can we choose optimal number of clusters (K)?

No, k-mean algo can be trapped to a local optima. On different initialization of centroids, the k-mea algo can converge to a different solution, so does not guarantee the global optimum.

Q3. For which of the following tasks might K-means clustering be a suitable algorithm? Select all that
apply. **[1 mark]**

A) From the user usage patterns on a website, figure out what different groups of users exist.
B) Given a database of information about your users, automatically group them into different market segments.
C) Given historical weather records, predict if tomorrow's weather will be sunny or rainy.
D) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

Ans: A and B (0.5 marks for one)

Q4. Assume you have a small dataset from which a model has to be generated. Would you prefer to learn a complex model or a simple model in this case? Give reasons for your answer! [**1 mark**]

Ans: Simple model because on less training data, the complex model can overfit.

Q5. Which statements are true about Data Wrangling?

Quiz 2: Data science
Total Marks: 8 2018-04-06

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

(A) We should use imputation when we have a lot of data.
(B) Due to partial deletion, we can compromise the representativeness of our sample.
(C) Imputation is the process of approximating the missing values.
(D) Pairwise deletion is more useful when we have only few records (data).

Ans: B and C (0.5 marks for one)

Q6. [2 marks] You are a reviewer for the International conference on Learning Algorithms, and you read papers with the following experimental setups. Would you accept or reject each paper? Provide a one sentence justification. (This conference has short reviews.)

i). **accept/reject** "My algorithm is better than yours. Look at the training error rates!"

Ans: reject - because training error is not used for evaluations of a ML system. Test data should be used and test error will decide which algorithm is better.

ii). **accept/reject** "My algorithm is better than yours. Look at the test error rates. Suppose we have Choosing λ based on the test data."

Ans: reject - because λ is chosen based on test data, so the same test data cannot be used for evaluation.

Q7. Suppose you train a classifier for spam classification. In this classifier you detect an email as a 'spam' (y=1), if h(x) >= 0.5. You test and calculate the precision and recall of your trained model. Now you realize that in order to improve the performance of your system, you need to say an email as 'SPAM' only if you are really confident (e.g. h(x) > 0.9). What do you think how the recall and precision of this classifier will be changed as compared to the previous model where your h(x) >=0.5?

Ans: Recall will decrease, Precision will increase.