Assignment 4 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. Cluster analysis can be employed to:

   (a) examine a firm's product offerings relative to competition.

   (b) group cities into homogeneous clusters for test marketing.

   (c) identify buyer groups sharing similar choice criteria.

   (d) segment markets.

   (e) all of the above.

   Sol. (e)
   Cluster analysis is useful for any application which requires variables to be placed into discrete groups.

2. Which among the following techniques can be used to aid decision making when those decisions depend upon some available data?

   (a) descriptive statistics

   (b) inferential statistics

   (c) predictive analytics

   (d) prescriptive analytics

   Sol. (a), (b), (c), (d)
   The techniques listed above allow us to visualize and understand data, infer properties of data sets and compare data sets, as well as build models of the processes generating the data in order to make predictions about unseen data. In different scenarios, each of these processes can be used to help in decision making.

3. You are given input data ((x,y) pairs where x and y are scalars). But you suspect that the output depends on the square of the input as well. Which of the following technique would help you to do better learning?

   (a) shuffling the data

   (b) basis expansion

   (c) subset selection

   (d) normalization of the data

Sol. (b)

4. Adding interaction terms (such as products of two dimensions) in linear regression could lead to:

   (a) Increases the bias.

   (b) Increases the variance.

   (c) Leads to lower training error.

   (d) both (a) & (b)

   (e) both (b) & (c)

   Sol. (e)

5. In a classification algorithm, consider the following two steps. In the first step, it is assumed that the input space is divided into axis-parallel rectangles, with a constraint that every rectangle must have at least one data point. In second step, one random data point from each rectangle is sampled and label of that data point is assigned to the entire region of that rectangle. Both the steps are introducing some kind of bias. Find the correct matching for the step and bias introduced by it.

   (a) first: search
   second: language

   (b) first: search
   second: search

   (c) first: language
   second: search

   (d) first: language
   second: language

   Sol. (c)
   In the first step, the mentioned constraints have to do with what forms the classification models can take. Hence it corresponds to the language bias. In second step, the constraint is on the search process to find the specific model given some data. SO the bias introduced by this step is search bias.

6. Which of the following statements is true about step-wise estimation?

   (a) a method of selecting variables for inclusion in the regression model that starts by selecting the worst variable

   (b) a method of selecting variables where independent variables are selected in terms of the incremental explanatory power they can add to the regression model

   (c) independent variables are added as long as their bi-variate correlation coefficients are statistically significant

   (d) all of the above

   (e) none of the above

   Sol. (b)

7. A residual represents which of the following?

   (a) the difference between the actual Y value and the predicted Y value for a given value of X.

   (b) the difference between the actual X value and the predicted X value for a given value of Y.

   (c) the difference between the actual Y value and the mean of Y for a given value of X

   (d) the predicted value of Y for the average X value.

   Sol. (a)
   Residual represents the difference between the actual Y value and the predicted Y value for a given value of X.

8. Suppose we trained a supervised learning algorithm on some training data and observed that the resultant model gave no error on the training data. Which among the following conclusions can you draw in this scenario?

   (a) the learned model has overfit the data

   (b) it is possible that the learned model will generalise well to unseen data

   (c) it is possible that the learned model will not generalise well to unseen data

   (d) the learned model will definitely perform well on unseen data

   (e) the learned model will definitely not perform well on unseen data

   Sol. (b), (c)
   Consider the (rare) situation where the given data is absolutely pristine, and our choice of algorithm and parameter selection allow us to come up with a model which exactly matches the process generating the data. In such a situation we can expect 100training data as well as good performance on unseen data. As an example, consider the case where we create some synthetic data having a linear relationship between the inputs and the output and then apply linear regression to model the data. The more plausible situation is of course that that we used a very complex model which ended up overfitting the training data. As we have seen, such models may achieve high accuracy on the training data but generally perform poorly on unseen examples.

9. Suppose you are given a task of predicting the temperature by using the cricket chirps. In the following table, column **X** denotes the level of cricket chirp and **Y** denotes the corresponding temperature. You decide to learn a linear regression model with the given data. What are the values of b_0 and b_1 you will get for the form $y = b_0 + b_1 x$?

   (a) $b_0 = 0.6$
   $b_1 = 4.47$

   (b) $b_0 = 17.82$
   $b_1 = 3.511$

   (c) $b_0 = 5.332$
   $b_1 = 4.013$

X	**Y**
20	88.59999847
16	71.59999847
19.79999924	93.30000305
18.39999962	84.30000305
17.10000038	80.59999847
15.5	75.19999695
14.69999981	69.69999695
17.10000038	82
15.39999962	69.40000153
16.20000076	83.30000305
15	79.59999847
17.20000076	82.59999847
16	80.59999847
17	83.5
14.39999962	76.30000305

(d) $b_0 = 25.23$
$b_1 = 3.291$

Sol. (d)
According to the derivations we saw in the lectures, we have

$$b_1 = \frac{\sum_{i=1}^{N} y_i x_i - \frac{\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i}{N}}{\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

Use these equations to find values of b_0 and b_1

10. Body fat in a human body is a function of {triceps skin-fold thickness, thigh circumference and midarm circumference}. Consider the data given in following table. The first three columns are the independent variables and fourth column is the dependent variable. Your task is to perform multiple linear regression model (minimize the mean square error) on the data and find the correct coefficients from the following choices. (coefficients in the options are given in the order: intercept, triceps skin-fold thickness, thigh circumference, midarm circumference)

Triceps	**Thigh**	**Midarm**	**Bodyfat**
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30	25.4
30.4	56.7	28.3	27.2
18.7	46.5	23	11.7
19.7	44.2	28.6	17.8
14.6	42.7	21.3	12.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51	27.5	21.1

(a) 21.8, 9.21, 8.5, 4.53

(b) 73.4, 8.19, 3.26, 4.15

(c) -86.4, 55.3, 12.1, 3.1

(d) 117.1, 4.33, -2.86, -2.19

Sol. (d)
You will get solution as:
Bodyfat = 117.1 + 4.33 Triceps - 2.86 Thigh -2.19 Mid arm

WEKA based Questions
The following questions are based on using WEKA. Go through the tutorial on WEKA before attempting these questions.
Dataset 1: Dataset 1 can be downloaded from
https://drive.google.com/file/d/0BwLesDk8tgZVc2NrOEpxaEFlNXM/view?usp=sharing
This data set contains 100 data points. The input is 3-dimensional (x1, x2, x3) with one output variable (y). This data is in the csv format which can directly be used in Weka. **Task:** You need to fit linear regression model on Dataset 1 and answer the following questions.

11. What is the best linear fit for data set 1?

(a) 00.0049 * x1 + 57.4552 * x2 + 79.0601 * x3 + 00.0301

(b) 00.0301 * x1 + 79.0601 * x2 + 57.4552 * x3 + 00.0049

(c) 23.2301 * x1 + 0.7310 * x2 + 48.3749 * x3 - 52.5001

(d) 52.5001 * x1 + 48.3749 * x2 + 00.7310 * x3 + 23.2301

Sol. (b)

12. As explained in the lecture, not all the features are equally important. For dataset 1, If we rank the features in the order of importance, which of the following rank is true?

Note: options are given in descending order of the importance. x1, x2, x3 means x1 is the most important and x3 is the least important

(a) x2, x1, x3

(b) x3, x2, x1

(c) x2, x3, x1

(d) x3, x1, x2'

(e) x1, x2, x3

(f) x1, x3, x2

Sol. (c)