Assignment 4 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. 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 visualise 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.

2. In a popular classification algorithm it is assumed that the entire input space can be divided into axis-parallel rectangles with each such rectangle being assigned a class label. This assumption is an example of

   (a) language bias

   (b) search bias

   Sol. (a)
   The mentioned constraints have to do with what forms the classification models can take rather than the search process to find the specific model given some data. Hence, the above assumption is an example of language bias.

3. 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 100% accuracy on the training 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.

4. You are faced with a five class classification problem, with one class being the class of interest, i.e., you care more about correctly classifying data points belonging to that one class than the others. You are given a data set to use as training data. You analyse the data and observe the following properties. Which among these properties do you think are unfavourable from the point of view of applying general machine learning techniques?

   (a) all data points come from the same distribution

   (b) there is class imbalance with much more data points belonging to the class of interest than the other classes

   (c) the given data set has some missing values

   (d) each data point in the data set is independent of the other data points

   Sol. (c)
   Option (a) is favourable, since it is an implicit assumption we make when we try applying supervised learning techniques. Option (b) indicates class imbalance which is usually an issue when it comes to classification. However, note that the imbalance is in favour of the class of interest. This means that there are a large number of examples for the class which we are interested in, which should allow us to model this class well. Option (c) is of course unfavourable as it would require us to handle missing data, and given the extent of the data that is missing, could severely affect the performance of any classifier we come up with. Finally, option (d) is favourable since, once again, it is an assumption about the data that we implicitly make.

5. You are given the following four training instances:

   - $x_1 = -1, y_1 = 0.0319$
   - $x_2 = 0, y_2 = 0.8692$
   - $x_3 = 1, y_3 = 1.9566$
   - $x_4 = 2, y = 3.0343$

   Modelling this data using the ordinary least squares regression form of $y = f(x) = b_0 + b_1 x$, which of the following parameters (b_0, b_1) would you use to best model this data?

   (a) (1,1)

   (b) (1,2)

   (c) (2,1)

(d) (2,2)

Sol. (a)
Let us consider the model with parameters (1,1). We can find the sum of squared errors using this parameter setting as

$$(0 - 0.319)^2 + (1 - 0.8692)^2 + (2 - 1.9566)^2 + (3 - 3.0343)^2 = 0.12$$

Repeating the same calculation for the other 3 options, we find that the minimum squared error is obtained when the parameters are (1,1). Thus, this is the most suitable parameter setting for modelling the given data.

6. You are given the following five training instances:

- $x_1 = 1, y_1 = 3.4$
- $x_2 = 1.5, y_2 = 4.7$
- $x_3 = 2, y_3 = 6.15$
- $x_4 = 2.25, y_4 = 6.4$
- $x_5 = 4, y_5 = 10.9$

Using the derivation results for the parameters in ordinary least squares, calculate the values of b_0 and b_1. (Note that in the expression for b_1, the last term in the denominator is $\frac{(\sum_{i=1}^{N} x_i)^2}{N}$.)

(a) $b_0 = 7.54, b_1 = 0.57$

(b) $b_0 = 2.484, b_1 = 0.969$

(c) $b_0 = 0.969, b_1 = 2.484$

(d) $b_0 = 1, b_1 = 2.5$

Sol. (c)
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

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

Using the supplied data, we have:

$N = 5$

$\sum_{i=1}^{N} x_i = 1 + 1.5 + 2 + 2.25 + 4 = 10.75$

$\sum_{i=1}^{N} y_i = 3.4 + 4.7 + 6.15 + 6.4 + 10.9 = 31.55$

$\sum_{i=1}^{N} x_i y_i = 1 * 3.4 + 1.5 * 4.7 + 2 * 6.15 + 2.25 * 6.4 + 4 * 10.9 = 80.75$

$\sum_{i=1}^{N} x_i^2 = 1^2 + 1.5^2 + 2^2 + 2.25^2 + 4^2 = 28.3125$

$(\sum_{i=1}^{N} x_i)^2 = 10.75^2 = 115.5625$

$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{10.75}{5} = 2.15$

$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{31.55}{5} = 6.31$

Using the above calculations, we have

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

$$b_1 = \frac{80.75 - \frac{31.55*10.75}{5}}{28.3125 - \frac{115.5625}{5}} = 2.484$$

Using this value of b_1, we have

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

$$b_0 = 6.31 - 2.484 * 2.15 = 0.969$$

7. Recall the regression output obtained when using Microsoft Excel. In the third table, there are columns for t Stat and P-value. Is the probability value shown here for a one-sided test or a two-sided test?

   (a) one-sided test

   (b) two-sided test

   Sol. (b)
   Recall that the null hypothesis used here is H_0: $b_i = 0$, where b_i is the coefficient corresponding to the i^{th} variable ($i > 0$), and b_0 is the intercept.

8. In building a linear regression model for a particular data set, you observe the coefficient of one of the features having a relatively high negative value. This suggests that

   (a) this feature has a strong effect on the model (should be retained)

   (b) this feature does not have a strong effect on the model (should be ignored)

   (c) it is not possible to comment on the importance of this feature without additional information

   Sol. (c)
   A high magnitude suggests that the feature is important. However, it may be the case that another feature is highly correlated with this feature and it's coefficient also has a high magnitude with the opposite sign, in effect cancelling out the effect of the former. Thus, we cannot really remark on the importance of a feature just because it's coefficient has a relatively large magnitude.

9. Assuming that for a specific problem, both linear regression and regression using the K-NN approach give same levels of performance, which technique would you prefer if response time, i.e., the time taken to make a prediction given an input data point, is a major consideration?

   (a) linear regression

   (b) K-NN

(c) both are equally suitable

Sol. (a)
As we have seen, in K-NN, for each input data point for which we want to make a prediction, we have to search the entire training data set for the neighbours of that point. Depending upon the dimensionality of the data as well as the size of the training set, this process can take some time. On the other hand in linear regression, once the model has been learned we need to perform a single simple calculation to make a prediction for a given data point.

10. You are given the following five training instances:

- $x_1 = 2, x_2 = 1, y = 4$
- $x_1 = 6, x_2 = 3, y = 2$
- $x_1 = 2, x_2 = 5, y = 2$
- $x_1 = 6, x_2 = 7, y = 3$
- $x_1 = 10, x_2 = 7, y = 3$

Using the K-nearest neighbour technique for performing regression, what will be the predicted y value corresponding to the input point ($x_1 = 3, x_2 = 6$), for K = 2 and for K = 3?

(a) K = 2, y = 3; K = 3, y = 2.33
(b) K = 2, y = 3; K = 3, y = 2.66
(c) K = 2, y = 2.5; K = 3, y = 2.33
(d) K = 2, y = 2.5; K = 3, y = 2.66

Sol. (c)
When K = 2, the nearest points are $x_1 = 2, x_2 = 5$ and $x_1 = 6, x_2 = 7$. Taking the average of the outputs of these two points, we have $y = (2 + 3)/2 = 2.5$.

Similarly, when K = 3, we additionally consider the point $x_1 = 6, x_2 = 3$ to get output, $y = (2 + 3 + 2)/3 = 2.33$.

Weka-based programming assignment questions

The following questions are based on using Weka. Go through the tutorial on Weka before attempting these questions. You can download the data sets used in this assignment here.

Data set 1

This is a synthetic data set to get you started with Weka. 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 arff format which can directly be used in Weka.

Tasks

For this data set, you will need to apply linear regression with regularisation, attribute selection and collinear attribute elimination disabled (other parameters to be left to their default values). Use 10-fold cross validation for evaluation.

Data set 2

This is a modified version of the prostate cancer data set from the ESL text book in the arff format. It contains nine numeric attributes with the attribute lpsa being treated as the target attribute. This data set is provided in two files. Use the test file provided for evaluation (rather than the cross-validation method).

Tasks

For this data set, you will need to apply linear regression. First apply linear regression with regularisation, attribute selection and collinear attribute elimination disabled. Next enable regularisation and try out different values of the parameter and observe whether better performance can be obtained with suitable values of the regularisation parameter. Note that to apply regularisation you should normalise the data (except the target variable). First normalise the test set and save the normalised data. Next open the training data, normalise it and then run the regression algorithm (with the normalised test set supplied for evaluation purposes).

Data set 3

This is the Parkinsons Telemonitoring data set taken from the UCI machine learning repository. Given are two files, one for training and one for testing. The files are in csv format and need to be converted into the arff format before algorithms are applied to it. The last variable, PPE, is the target variable.

Tasks

For this data set, first apply linear regression with regularisation, attribute selection and collinear attribute elimination disabled. Note the performance and compare with the performance obtained by applying K-nearest neighbour regression. To run K-NN, select the function IBk under the lazy folder. Leave all parameters set to their default values, except for the K parameter (KNN in the interface).

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

    (a) y = 5*x1 + 6*x2 + 4*x3 + 12

    (b) y = 3*x1 + 7*x2 - 2.5*x3 - 16

    (c) y = 3*x1 + 7*x2 + 2.5*x3 - 16

    (d) y = 2.5*x1 + 7*x2 + 4*x3 + 16

    Sol. (b)
    The data set used here is generated using the function specified by option (b) and with no noise added, hence running linear regression on this data set should allow us to recover the exact parameters and observe 100% accuracy.

12. Which of the following ridge regression parameter values leads to the lowest root mean squared error for the prostate cancer data (data set 2) on the supplied test set?

    (a) 0

    (b) 2

    (c) 4

    (d) 8

    (e) 16

    (f) 32

Sol. (e)
Among the given choices of the ridge regression parameter, we observe best performance (lowest error) at the value of 16.

13. If a curve is plotted with the error on the y-axis and the ridge regression parameter on the x-axis, then based on your observations in the previous question, which of the following most closely resembles the curve?

    (a) straight line passing through origin
    (b) concave downwards function
    (c) concave upwards function
    (d) line parallel to x-axis

    Sol. (c)
    In general, as the ridge regression parameter is increased, the error begins to decrease, but after a certain point, further increase in the parameter drives up the error.

14. Considering data set 3, is the performance of K-nearest neighbour regression (where the value of the parameter K is varied between 1 and 25) comparable to the performance of linear regression (without regularisation)?

    (a) no
    (b) yes

    Sol. (b)
    At K = 5, the performance is even better.