Assignment 3 (Sol.)

Introduction to Machine Learning
Prof. B. Ravindran

1. 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, as was discussed in the lectures, 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.

2. The model obtained by applying linear regression on the identified subset of features may differ from the model obtained at the end of the process of identifying the subset during

   (a) Best-subset selection

   (b) Forward stepwise selection

   (c) Forward stagewise selection

   (d) All of the above

   Sol. (c)
   Let us assume that the data set has p features among which each method is used to select $k, 0 < k < p$, features. If we use the selected k features identified by forward stagewise selection, and apply linear regression, the model we obtain may differ from the model obtained at the end of the process of applying forward stagewise selection to identify the k features. This is due to the manner in which the coefficients are built in this method where at each step the algorithm computes the simple linear regression coefficient of the residual on the variable identified as having the largest correlation with the residual, and adds it to the current coefficient for that variable. Note that there will be no difference in the other two methods, because in both forward and backward stepwise selection, at each step of removing/adding a feature, linear regression is performed on the retained subset of features to learn the coefficients.

3. We have seen methods like Ridge and lasso to reduce variance among the co-efficients. We can use these methods to do feature selection also. Which one of them is more appropriate?

   (a) Ridge

   (b) Lasso

Sol. (b)
For feature selection, we would prefer to use lasso since solving the optimisation problem when using lasso will cause some of the coefficients to be exactly zero (depending of course on the data) whereas with ridge regression, the magnitude of the coefficients will be reduced, but won't go down to zero.

4. Given the following 3D input data, identify the principal component.
1 1 9
2 4 6
3 7 4
4 11 4
5 9 2
(Steps: center the data, calculate the sample covariance matrix, calculate the eigen vectors and eigen values, identify the principal component)

   (a) 0.9138
   -0.1035
   0.3926

   (b) -0.2617
   0.5891
   0.7645

   (c) -0.4205
   -0.6223
   0.2342

   (d) -0.3105
   -0.8014
   0.5112

Sol. (d)
Center the data:
-2.0000 -5.4000 4.0000
-1.0000 -2.4000 1.0000
0 0.6000 -1.0000
1.0000 4.6000 -1.0000
2.0000 2.6000 -3.0000

Find the covariance matrix $(((x-\mu)'(x-\mu))/(n-1))$:
2.5000 5.7500 -4.0000
5.7500 15.8000 -9.2500
-4.0000 -9.2500 7.0000

Solve characteristic equation to obtain the eigen values and eigen vectors
Eigen values:
0.1298
1.2415
23.9287

Eigen vectors:
0.9138 -0.2617 -0.3105
-0.1035 0.5891 -0.8014
0.3926 0.7645 0.5112

Select the principal component, i.e., the eigen vector corresponding to the largest eigen vlaue:
-0.3105
-0.8014
0.5112

5. For the data given in the previous question, find the transformed input along the first two principal components.

   (a) 6.9935 0.3021
   2.7451 -0.2727
   -0.9921 -0.4548
   -4.5081 0.0449
   -4.2383 0.3805

   (b) 6.9935 0.4003
   2.7451 -0.3876
   -0.9921 -0.4110
   -4.5081 1.6837
   -4.2383 -1.2853

   (c) 0.4003 0.3021
   -0.3876 -0.2727
   -0.4110 -0.4548
   1.6837 0.0449
   -1.2853 0.3805

   (d) 3.4894 7.2079
   -0.7590 6.4200
   -4.4962 6.3966
   -8.0122 8.4913
   -7.7424 5.5223

   Sol. (b)
   Project the mean centered data points along the first two principal components by multiplying the 5x3 mean centered data matrix with the 3x2 matrix composed of the two eigen vectors which correspond to the two highest eigen values.

6. Suppose you are only allowed to use binary logistic classifiers to solve a multi-class classification problem. Given a training set with 2 classes, this classifier can learn a model, which can then be used to classify a new test point to one of the 2 classes in the training set. You are now given a 6 class problem along with its training set, and have to use more than one binary logistic classifier to solve the problem, as mentioned before. You propose the following scheme (also known as one vs one approach in ML terminology) - you will first train a binary logistic classifier for every pair of classes. Now, for a new test point, you will run it through each of these models, and the class which wins the maximum number of pairwise contests, is the

predicted label for the test point. How many binary logistic classifiers will you need to solve the problem using your proposed scheme?

(a) 25

(b) 15

(c) 18

(d) 12

Sol. (b)
Since we need a binary logistic classifier for each pair of classes, the number of classifiers required $= \binom{6}{2} = 15$.

7. With respect to Linear Discriminant Analysis, which of the following is/are true.
(Consider a two class case)

   (a) When both the covariance matrices are spherical and equal, the decision boundary will be the perpendicular bisector of the line joining the means.

   (b) When both the covariance matrices are spherical and equal, the decision boundary will be perpendicular to the line joining the means.

   (c) When both the covariance matrices are spherical and equal and the priors $\pi_1 = \pi_2$ then the decision boundary will be perpendicular bisector of the line joining the means.

   Sol. (b) & (c)
   The first statement is not true because unequal priors can cause the decision boundary to shift away form the center of the line joining the two means.

8. For a two class classification problem, which among the following are true?

   (a) In case both the covariance matrices are spherical and equal, the within class variance term has an effect on the LDA derived direction.

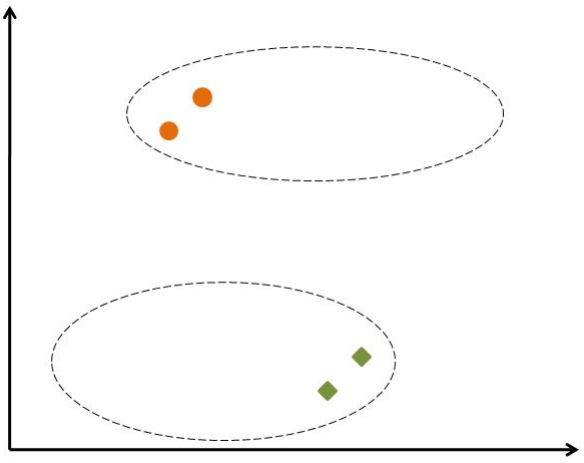
   (b) In case both the covariance matrices are spherical and equal, the within class variance term has no effect on the LDA derived direction.

   (c) In case both the covariance matrices are spherical but unequal, the within class variance term has an effect on the LDA derived direction.

   (d) In case both the covariance matrices are spherical but unequal, the within class variance term has no effect on the LDA derived direction.

   Sol. (b) & (d)
   It is easy to see that if the covariance matrix for a particular class is spherical, then the within class variance is the same along all directions, and hence we can ignore this component and focus solely on the between class variance (i.e., identify the direction along which the between class variance is maximised).

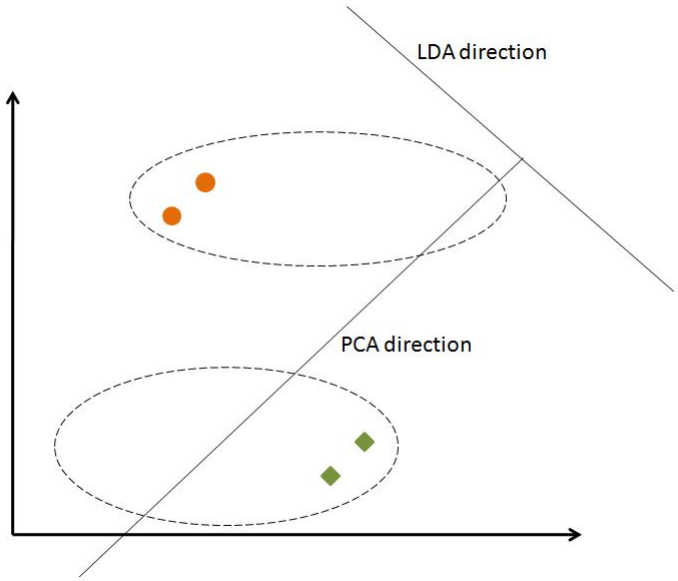
9. Suppose you are given a two dimensional two class data set as shown below with only two samples for each class. The dashed curves show the underlying (but unknown) distribution of each class. Which among the two methods for identifying a one dimensional representation of the given data would you suggest for building a classifier that will perform well on test data coming from the same underlying distributions?



(a) Linear Discriminant Analysis (LDA)

(b) Principal Components Analysis (PCA)

Sol. (b)
As shown in the following figure, the single dimension identified by PCA is superior to the one identified by LDA for the purpose of classification if we take into consideration the underlying class distributions, since the overlap among the points of the two classes would be minimal in the case of the PCA dimension.



Weka-based assignment questions

The datasets for this assignment are available here.

Dataset 1

This is a synthetic dataset to get you started with Weka. The given dataset is a 1000 point 3-dimensional data with one target variable. Full data is given to you both in csv and arff file formats. You can load the arff format directly into Weka. In addition to this, we have given the train and test spilt. Always use the test split to evaluate your model.

The variables are named $x1$, $x2$, $x3$, y. y is the variable to be regressed.

Tasks

Train a linear regressor by disabling regularizing, attribute selection and collinear attribute elimination. Note the error obtained and coefficients.

Dataset 2 (Prostate Dataset)

This is the Prostate Cancer dataset used in the ESL book. You can read the info file for more information about the dataset. We have split the data into a train and test split. You should use the test split to evaluate the model. (When training, select the "Supplied Test set" and select the test dataset.

Tasks

(a) Train a unregularized model - Disable the attribute selection, regularization, and co-linear attribute elimination and run the linear regressor. Note down the mean squared error.

(b) Ridge Regression - We have seen unregularized models till now. Weka implements ridge regression model for linear regression which we will try. We have to choose the best parameter lambda for ridge regression. In practice, we do this by searching over a range of values and see which gives the least cross-validation error. Weka has functions to this search too. We use CVParameterSearch in the meta functions, and use it to iterate over the ridge parameter(R) of the Linear Regression model. Iterate over 0 - 50 in steps of 1, use a 5 fold cross-validation. Note down the best value of lambda and corresponding error.

Dataset 3 (AutoMPG)

This dataset is the AutoMPG dataset from UCI repository. You are given the data in the arff format. You can load the dataset and see the various attributes in the dataset.You can see that there are both continuous and discrete attributes. We will now see how to handle such data. Again here you are given a pre-decided test and train split.

Tasks

(a) Drop non-predictive data - Most real world data sets come with some non-predictive attributes which one can drop. If you read the dataset description you can see that there is an attribute called the car name, which is a string. We are sure that this can be dropped. We have removed this attribute and provided the dataset in the arff format.

(b) One-hot-encoding for the discrete attributes - In the videos, sir has mentioned the issues with discrete/categorical variables. We will now convert the nominal/discrete/categorical variables to a binary representation. We can do this using the pre-processing options in Weka. Under Unsupervised instance based pre-processing options we can find the Nominal to Binary converter. After performing this, you can notice the increase in the number of attributes, it creates a new variable for each possible value of the variable.
(Note- When you perform these pre-processing steps, you should also do them on the test

set separately and save as an arff file from Weka, otherwise Weka might find train and test set incompatible.)

(c) Training - Disable the attribute selection, regularization, and co-linear attribute elimination and run the linear regressor. Note down the mean squared error.

10. What is the best linear fit for the dataset 1?

    (a) $y = 3 * x1 + 4 * x2 + 5 * x3 + 4$

    (b) $y = -1 * x1 + 2 * x2 + 5 * x3 + 5$

    (c) $y = x1 + 10 * x2 + 4 * x3 + 6$

    (d) $y = 16 * x1 + 40 * x2 + 15 * x3 + 4$

    Sol. (a)
    Using the supplied training and test data, Weka should report exactly the model described in option (a).

11. Consider the prostate cancer dataset, which of the following ridge parameters is best suited?

    (a) 0

    (b) 2

    (c) 14

    (d) 20

    (e) 100

    Sol. (c)
    On trying out the different values for the regularisation parameter you should observe that small values for the regularisation parameter do lead to a decrease in the error when compared to the model with no regularisation. As we increase the value of the regularisation parameter, the error decreases until it reaches a minimum after which it starts to increase. Among the values listed above, the lowest error is observed when the regularisation parameter is set to 14.

12. If we plot the CV-error versus the ridge parameter for the prostate dataset, what is the expected shape of the curve?

    (a) Straight line passing through origin

    (b) Inverted trough

    (c) Convex trough

    (d) Horizontal line

    Sol. (c)
    The previous problem suggests the answer. Note that you will observe similar behaviour if you consider the errors on the dedicated test set rather than the cross validation error.

13. Learn a linear model using ridge regression with lambda of 16. After looking at the learned model, which of the attribute you think is the best to be dropped?

    (a) displacement

(b) acceleration

(c) weight

Sol. (b)
We observe that acceleration has the lowest associated coefficient magnitude among the three features and hence it would be the best candidate to be eliminated without having too much of an adverse effect on the model performance (though that will need to be verified).

Note- There was an error in the automated evaluation for this question. We will correct this and update the scores accordingly.