CS 4011 : Principles of Machine Learning
Programming Assignment #1

Jahnvi Patel
CS15B046

August 30, 2017

1 Synthetic Dataset Creation

A multi-variate Gaussian distribution having p dimensions can be defined in terms of its mean μ, which is $p \times 1$ vector and its covariance matrix, cov, which is $p \times p$ positive definite, symmetrical, and invertible matrix.For the given problem, in order to generate:

(i) Mean - Randomly generate a vector of 20 features as mean vector for the first class. Now, in order to ensure sufficient overlap between the two classes, we declare a parameter δ which is a measure of distance between the two centroids. While randomly generating the mean vector for second class, we ensure that the distance between the centroids does not exceed δ.

(ii) Co-variance: There are several methods to generate a positive semi-definite matrix, and two commonly used include Cholesky decomposition ($M = AA^T$) and Singular Value Decomposition ($M = QSQ^T$, where Q is an orthonormal matrix, and S is a diagonal matrix with non-negative entries). For the given problem, Cholesky decomposition has been used, by randomly generating a 20×20 matrix A and ensuring that the resultant covariance matrix is non-spherical.

The generated parameters are specified in the params.txt file. Upon obtaining two different distributions, we generate 2000 data points corresponding to each class and sample 70% of the data set for training.

2 Linear Classification

To classify the two sets of data points, we perform linear regression on an $n \times 2$ indicator matrix, that denotes a 1 if the given data point belongs to that particular set, i.e. $Y_{ij} = 1 if X_i \in setj$. Upon obtaining Y_{pred}, we classify the test points as $Class_i = argmax(Y_{pred}[i])$.
To summarise:

(i) Data: 70% Training and 30% Test Data comprising points sampled from 2 different Gaussian distribution classes.

(ii) Model: Simple linear model solved by linear regression on indicator matrix

(iii) Parameters: The coefficients learnt, they are specified in coeffs.csv file.

(iv) Objective function: To minimise the mean squared error.

The results obtained are in results.txt file. Also, it should be noted that changing the distance between centroids of two classes alters the scores.
The variation in scores obtained is as follows:
For $\delta = 0.65$:

(i) Accuracy score: 0.734

(ii) Precision score: 0.727

(iii) Recall score: 0.721

(iv) F measure: 0.724

For $\delta = 0.7$:

(i) Accuracy score: 0.835

(ii) Precision score: 0.834

(iii) Recall score: 0.828

(iv) F measure: 0.831

For $\delta = 0.9$:

(i) Accuracy score: 0.975

(ii) Precision score: 0.976

(iii) Recall score: 0.974

(iv) F measure: 0.975

Thus, we observe that as we move the classes away from each other, the accuracy increases.

3 k-NN classifier

K-Nearest neighbours method considers k nearest points in space to estimate the result. kNN being a more complex model, produces better estimates than linear regression. However, a complex model is prone to over-fitting. So, we observe the impact of scores on change of k:
For $k = 10$:

(i) Accuracy score: 0.509

(ii) Precision score: 0.5

(iii) Recall score: 0.415

(iv) F measure: 0.454

For $k = 60$:

(i) Accuracy score: 0.564

(ii) Precision score: 0.552

(iii) Recall score: 0.589

(iv) F measure: 0.570

For $k = 500$:

(i) Accuracy score: 0.521

(ii) Precision score: 0.511

(iii) Recall score: 0.563

(iv) F measure: 0.536

We observe that for smaller values of k, the model would be unstable because of over-fitting. As we increase k, bias of the model starts increasing, at the same time model becomes more stable. It would reach an optimum point for a certain k, in this case around 55, post which the accuracy of model reduces because of increase in bias. This can be clearly explained using the bias-variance trade-off.

The corresponding score for linear regression for same set of training and test data produced the following result:

(i) Accuracy score: 0.835

(ii) Precision score: 0.834

(iii) Recall score: 0.828

(iv) F measure: 0.831

We also observe that although kNN model fits the training data better than linear regression, it performs poorly on test data. This is because of over-fitting by kNN.

4 Data Imputation

Imputation by mean is one of the most common methods used. It's advantage is that it considers the overall nature of the data set. At the same time, it has several disadvantages like:

(i) It does not consider the correlation among sample parameters.

(ii) Also, imputation by mean under-estimates the mean square error leading to wrong assumptions about the model.

(iii) spike in data because of some anomaly would strongly bias mean and hence wrongly assign the imputed values too.

Other alternative methods:

1. Imputation by median/mode has several disadvantages like a spiked value. i.e. an incorrect observation ending up as median/mode.

2. A better form of imputation would be by kNN or Fuzzy k means that estimates the values depending on the neighboring values. However, such an imputation would lead to over-fitting of the model.

3. Yet another better way is regression imputation, i.e. perform regression on other observed available values and use the estimates for imputation. However, this method over-estimates correlation between the features.

4. Maximum Likelihood Estimate (MLE) Imputation: substitutes the estimates producing the highest log-likelihood, i.e. most probable values for the given distribution. This method of imputation also takes the model into account, along with the observed data values.

5. Multiple Imputation is yet another form of regression estimate that repeats the estimation on different datasets and combines their predictions to impute the values.

Thus, it is observed the kind of imputation underestimates the standard error and generally depends on model, and the kind of dataset that the model would be working on.

5 Linear Regression

To estimate the crime rate, we eliminate the non-predictive features and perform linear regression on predictive data features. To summarise:

(i) Data: 80% Training and 20% Test Data comprising points sampled the imputed data set.

(ii) Model: Simple linear model solved by linear regression on crime rate.

(iii) Parameters: The coefficients learnt, they are specified in coeffs.csv file.

(iv) Objective function: To minimize the mean squared error.

The results obtained are in results.txt file. Residual Sum of Squares is defined as the product of mean squared error and number of test entries for which the computation is done. The average RSS value is: 7.576.

6 Regularized Linear Regression

The accuracy of data model is controlled by the bias-variance trade-off. Therefore, we do not want the model complexity to shoot up. To avoid this, several regularization techniques like LASSO, and ridge are used to regularize and control the coefficients. Ridge regression is characterized by α which is the tuning parameter for model complexity.
So, we observe the impact of scores on change of α:
Residual Sum of Squares for $\alpha = 0.01$ is: 7.524
Residual Sum of Squares for $\alpha = 1$ is: 7.219

Residual Sum of Squares for $\alpha = 100$ is: 7.647
For low values of α, the model will over-fit and therefore, RSS value is high. Upon increasing α, model complexity decreases and so does RSS. However, after certain values, any increase in α will lead to under-fitting and hence, increase in RSS. It is also observed that smaller the value of α, higher is the magnitude of the coefficients. This is because L-2 regularization brings down the value of high parameters. The improvement in data model is evident from the reduced RSS values as compared to a normal linear regression. The corresponding score for linear regression for same set of training and test data produced the following RSS value: 7.576.
However, ridge regression has a disadvantage, it will contain predictors corresponding for all features in contrast to LASSO that zeroes down the value of relatively unimportant parameters. However, ridge regression ensures that all non-predictive features get a zero value. So, we can remove these features. For feature selection, LASSO should be used.

7 References

http://www.theanalysisfactor.com/
https://liberalarts.utexas.edu