**Foundations of Machine Learning Coursework**

Name: Wenhao Wang Student ID: 30624061 E-mail: ww2g18@soton.ac.uk

**Introduction**

Firstly, non-linear regression could fit functions when data looks non-linear. To be pacific, a quadratic function fit better than a linear line, because combination of different degrees of polynomials have a flexible curve to fit non-linear data.

Secondly, LDA(Linear Discriminant Analysis) is an efficient tool to reduce dimensional space of data. It is usually used in project data into lower discriminant dimensions. After projection, it become easier to classify different classes because LDA projects data into most discriminant way.

Thirdly, The remaining part of this report is implementation and conclusion.

**Implementation**

**Classification**

LDA focus on within – class scatter matrix and between-class scatter matrix. It is intended to find a lower dimension that minimise scatter within a class and maximise scatter between classes.

**Separate two gaussians**

1.(a) In this part, I generate two gaussian distribution with and project them into , when fisher ration is the maximum value.

First, choose three illustrative vectors . In another word, they are Y axis, X axis and . Projected histogram shown below.

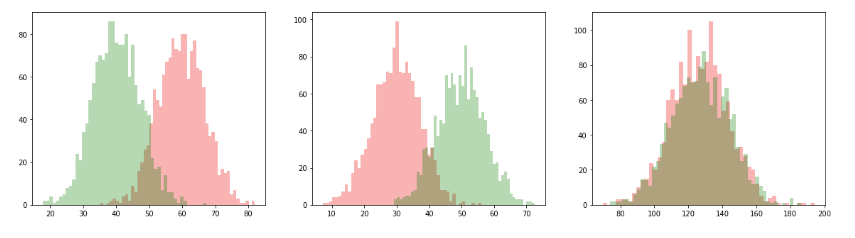


Fig.1 Histogram of projected data

Second, calculate Fisher ratio under index of , find the maximum of and relative . Then use get the optimal choice vector w\*.

1.(b) & 2.(a) Fig.2 shows the dependence of . In this case, , , . By calculation, maximum fisher ratio is 25.979, when equals to 2.352. The blue line in Fig.3 is optimal w vector, contour lines are also shown in this figure.

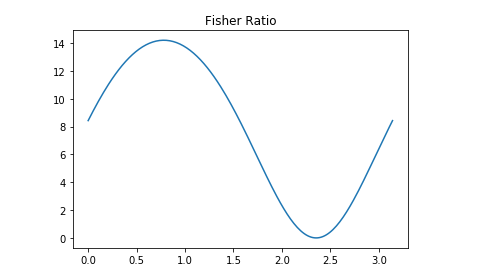
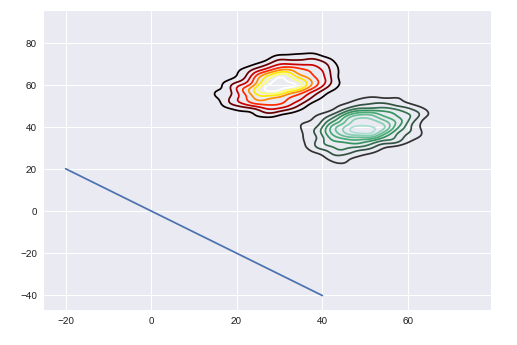
 

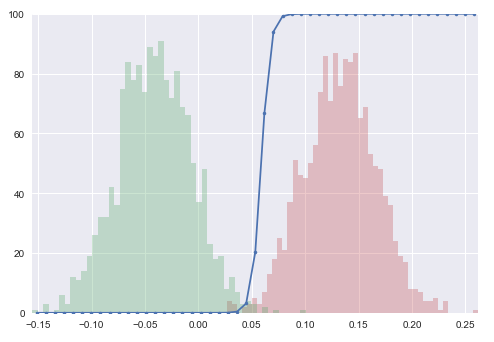
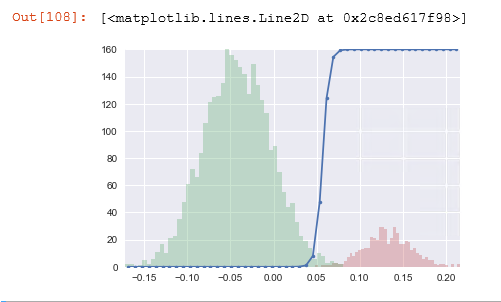
Fig.2 Dependence of F(w) Fig.3 w\* vector and contour line

2.(b) After project data into one dimension, we need classification method to classify these two gaussians. Logistic regession is a method based on log-odds. We can write down log-odds as folow equations.

Logistic regression treats the as the target of linear regression.

Thus, inverse we could have .

Following figures are classification using theory from above. Fig.4a is the classification under , dicision boundary is the blue line. Fig. 4b is when , .

a. Decision boundary b. Decision boundary

Fig.4 Log-odds classification

2.(c) Here, we redefine Fisher ratio as . However, after calculation of unbalanced fisher ratio, optimal w\* is still the same(when equals to 2.352). According to Bayes’ rule the probability of current point is , where is probability of class c, is probability of point x, is probability of presence of point x in class c. In this equation, and is set by class number and points number. Thus, determine . In another word, LDA focus on scatter within and between class (to be pacific, fisher ratio focus on and ), these scatters decided fisher ratio, thus without in denominator does not account for the different fractions of data in each class.

**Iris data**

LDA could also applied in multiple classes. In Iris dataset, there are 3 classes and 4 features in this dataset. For multiple classes LDA, we need to computing scatter matrices.

Within-class scatter matrix

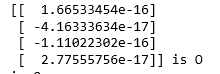
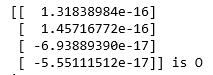
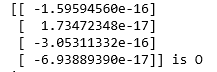
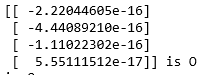
Between-class scatter matrix

is generalised eigenvalue. In *scipy* library, *scipy.linalg. eigh* function will return sorted eigenvalues and corresponding eigenvectors.

1. By using *scipy* function, I got

Eigenvalue = [-4.03197199e-17, 2.99331166e-17, 1.85044576e-03, 2.15146385e-01]

Then verify these eigenvalues using

Thus, these generalised eigenvalues fit the condition of

2. Normally, top k sorted eigenvalue and its eigenvector will be the optimal w\*. In another word, the eigenvectors with the largest eigenvalues bear the most information about the distribution of the data. In this dataset, we found:

eigenvalue 1: -0.000%

eigenvalue 2: 0.000%

eigenvalue 3: 0.853%

eigenvalue 4: 99.147%

Thus, the last two eigenvalues especially the last weights the distribution of data. As shown in Fig.5, in 2-D distribution data split discriminant in direction of LD1. Moreover, Fig.6 shows that projection on second eigenvector is not discriminant.

Thus, we could project data into the largest eigenvalue direction. As shown in the left figure of Fig.6.

3. As shown in Fig.7 and Fig.8, when (here a = np.ones((4,1))\*0.1). Consequently, direction of projection changed, and data distribution became less discriminative.

Adding eigenvalue into w\* does not influence distribution due to the largest eigenvalue weights 99.4% of discriminant.

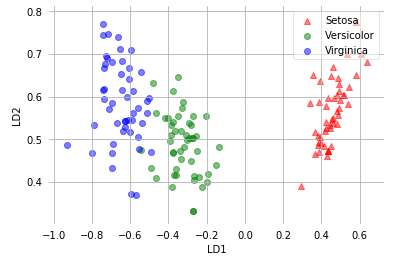


Fig.5 2-D distribution

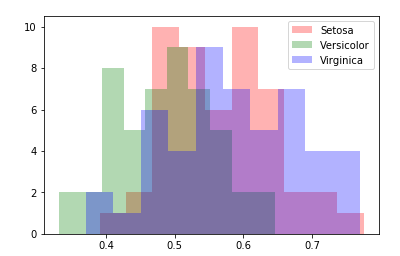
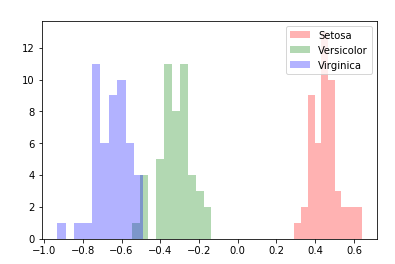


Fig.6 1-D distribution

(left figure is the largest eigenvalue; right figure is the second large eigenvalue)

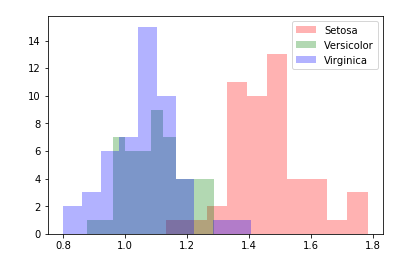


Fig.7 Direction is:

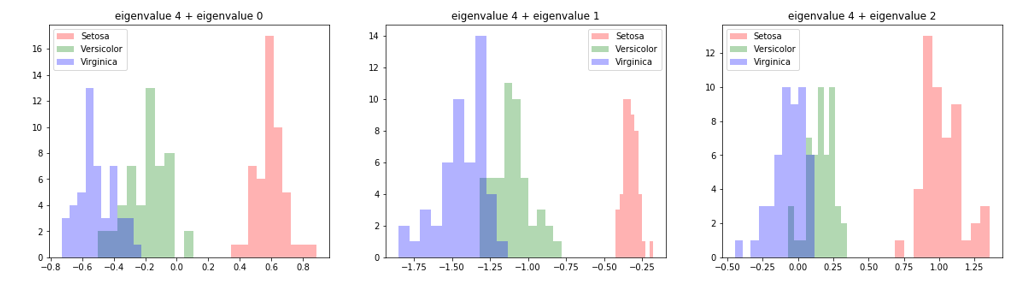


Fig.8

4. In compare with separate two gaussians I conclude that, these two works are all concentrate on scatter between and within classes. As fisher ratio function definition above it is mean square of subtract of two classes divided by sum of variance of two classes. The second solution is using eigenvalues to find optimal w\*, which is a solution for multiply classes classification.

**Part 2 Linear Regression with non-linear functions**

**Performing linear regression**

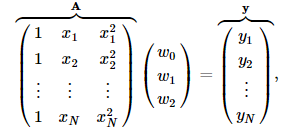
1. L2 norm is defined like this:

Regularised linear regression should be

Thus, regularised gradient decent is:

Designed polynomial function shown below:

For non-linear function, we assume . Matrix **A** is called design matrix. We could put **A** into regularised gradient decent function as linear regression,



2. According to the analytical expression from coursework.

We could compute weight using equation below. In addition, predicted curves are shown in Fig.9.

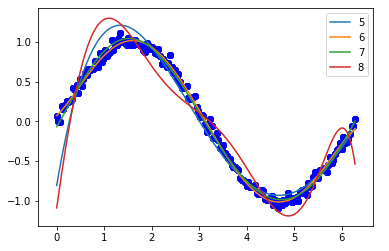


Fig.9 The analytical expression

3.a In this coursework I choose N = 4, 5, 6 (In my coursework it means degree = 5, 6, 7). There are 180 points, 2/3 of points were divided into training points and the rest 1/3 are testing points. Strength of was get from 3.2b

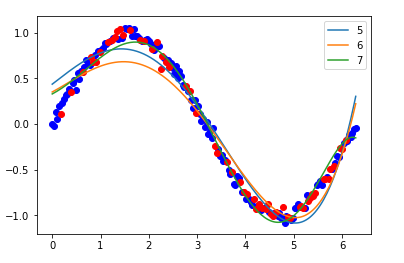
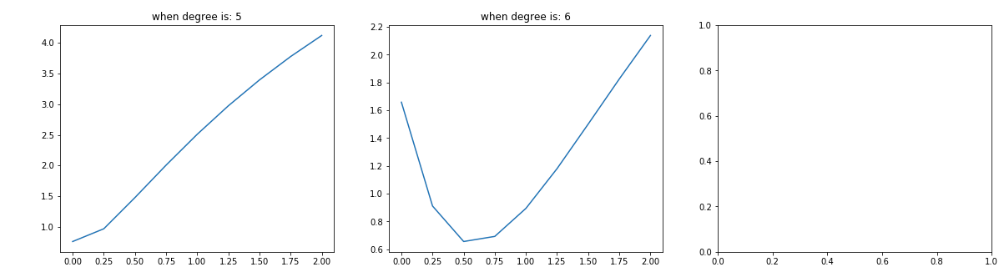


Fig.10 Non-linear regression

3.2b In order to find the best strength of , I tried different strength of and plot them in two figures.(I only test N = 4 and 5, because N=6 need too much computing time.)

When degree = 5, = 0, degree = 6, = 0.5.



**How does linear regression generalise?**

In this part I divided points into 10 folds and train different folds in a loop. After the loop, find the best model that have minimum residual on generalisation performance.

