The Report of Classification and Regression

ID(29299675)

Abstract

The report illustrates classification problems in terms of binary classification and three classification through LDA (Linear Discriminate Analysis). The second part explores Linear Regression with non-linear functions through L2 norm regulation and finally verifies the cross cross-validation to looking at the dependence of the optimal weights on training data and the scale of regularisation.

Introduction

1. The motivation of this experiment is to understand deeply key conceptual steps that underpin each algorithm through illustrating the graphical or tabular form.
2. The available resources and tools are common machine learning techniques.
3. The report will be divided into three parts below: Methodology, Results and Discussion.

1 Maths and Algorithm

1. **LDA**
   - LDA (Linear Discriminate Analysis) is used for classification, use

     $$y = W^T x$$

     , make original samples into a lower dimension, which is easy to separate. To obtain a best classification, the approach is to maximize the the covariance of between class and minimize the covariance of within class, which is maximise

     $$J(w) = \frac{S_w^T S_B w}{S_w^T S_W w}$$

     among this, S_B is between-class scatter matrix and S_W is within-class scatter matrix
2. **Linear Regression with non-linear functions via L_2 norm penalty**
   - Linear Regression with non-linear functions via L_2 norm penalty is to minimise the loss function

     $$J(\theta) = \frac{1}{2} MSE(\theta) + \frac{\lambda}{2} \sum_{i=1}^{n} \theta_i^2$$

     among this,

     $$\theta = (X^T X + \lambda I)^{-1} (X^T y)$$

     , which is called normal equation
     This can solve overfitting problem. Using polynomial regression, if the highest order of the polynomial is large, the model is prone to overfitting. The hyperparameter λ determines the strength of the model you want to regularize. The greater the regularization strength, the simpler the model will be. If $\lambda = 0$ then the ridge regression becomes a linear regression. If λ is very large, all weights are close to zero at the end, and the final result will be a horizontal straight line that traverses the data average. As λ increases, the prediction curve becomes flat (i.e., less extreme values, more general values), which reduces the variance of the model, but increases the deviation of the model.
3. **Cross Validation**

   - When the number of data sets is limited, a certain amount of data needs to be used for training, and the remaining data is used for testing.

     A more common approach to reduce any deviations caused by side-by-side sampling is to repeat the entire process, repeating multiple trainings and tests with different random samples, and averaging the error rates obtained during each iteration. A comprehensive error rate is obtained, which is the error rate estimate of the repeated side method. In the repeated side method, some data may be used as test data or training data throughout the process, which inevitably has an impact due to the inconsistent representation of the training set and the test set data. Therefore, a simple variant method was introduced, leading to cross-validation.

     Cross-validation can solve overfitting problems caused by test the parameter on the same data set

2 Results and Discussion

2.1 Separate 2 Gaussians data set for classification

$$\mu_1 = (2,2)^T, \mu_2 = (-2,-2)^T$$
$$\sigma_1 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}, \sigma_2 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix},$$

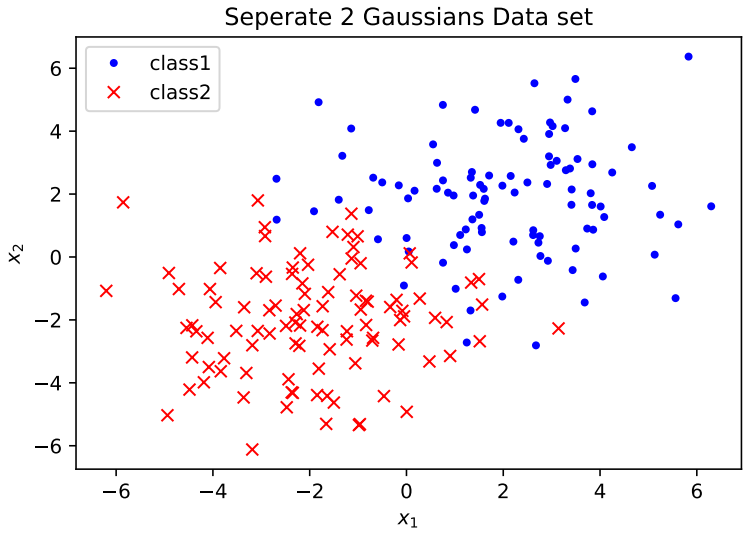

Figure 1: The plot of separate 2 Gaussians data set.

1. **Visual exploration of the consequences of projecting data onto a lower dimension.**

   - The histograms of the values y_a^n and y_b^n
   choose $w_1 = (-0.6, 0.5)^T$,$w_2 = (-3, 0.5)^T$, $w_3 = (3, -0.5)^T$,$W_4 = (-5, -0.5)^T$,$W_5 = (1.5, 1.9)^T$, the corresponding histograms for different w projection vectors are as follows: From the different choices of w, the w_5 has the best classification effect on a lower(one) dimension. From Table 1, it indicates the w_5 has the maximum of Fisher ratio, which accords with it has the best classification effect.

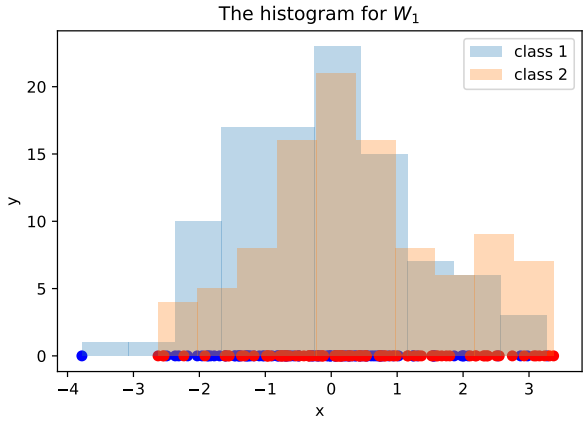

Figure 2: The Histogram for w_1

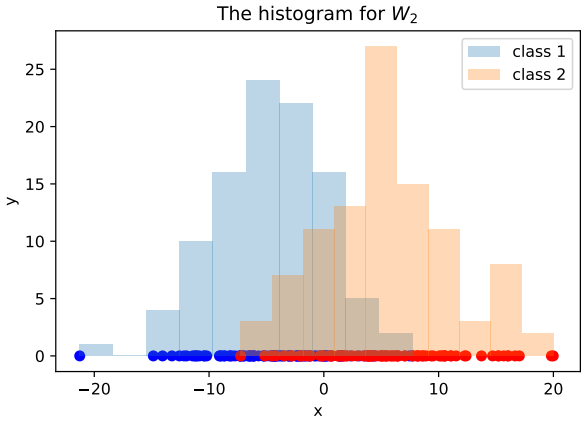

Figure 3: The Histogram for w_2

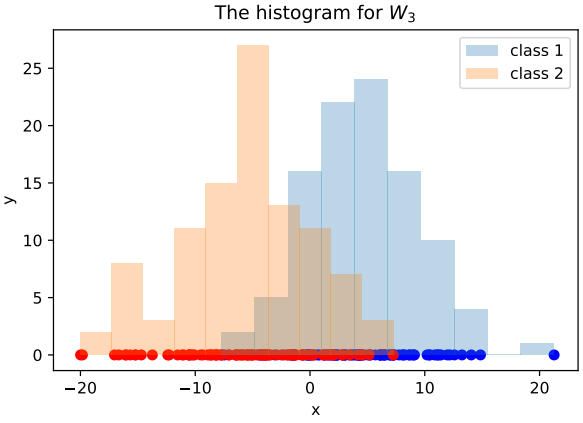

Figure 4: The Histogram for w_3

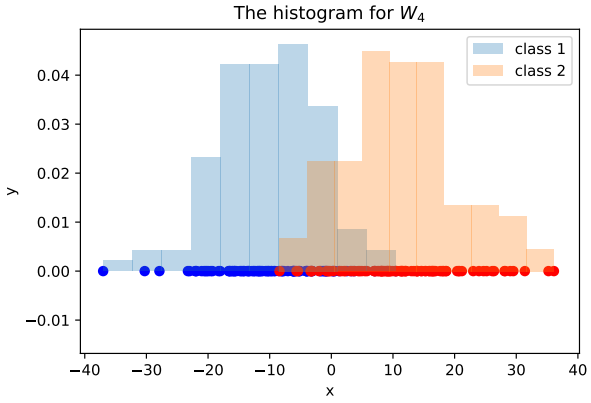

Figure 5: The Histogram for w_4

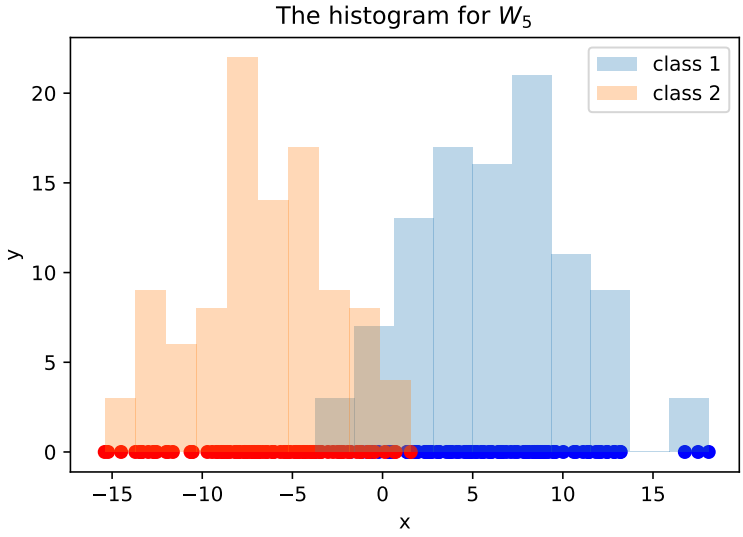

Figure 6: The Histogram for w_5

	w_1	w_2	w_3	w_4	w_5
$F(W)$	0.1549	3.6341	3.6341	6.2602	9.9198

Table 1: The fisher ratio for different w

- The dependence of $F(w)$
  For the above w, rotate a angle $\theta, \theta\epsilon[0, 2\pi)$, the variation graphic of $F(w)$ is as follows:

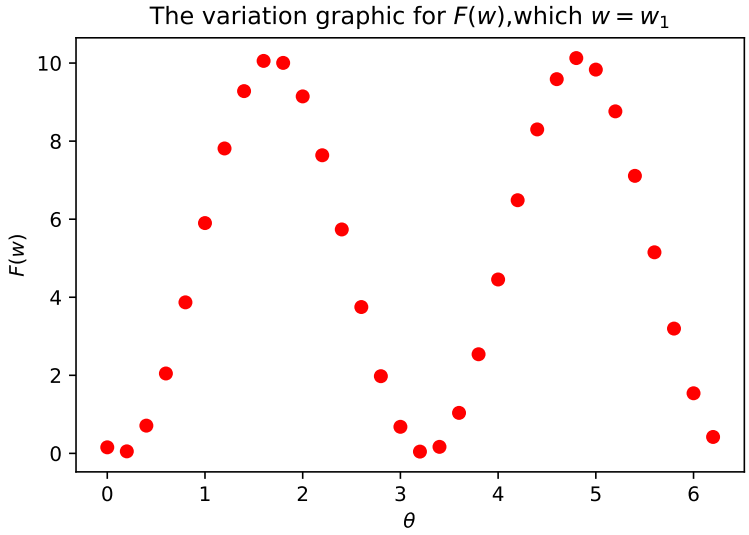

Figure 7: The variation graphic for $F(w)$,which $w = w_1$.

The changing curve of $F(W)$ is similar as $sin(\theta)$.
The optimal choice w corresponding to the maximum of $F(w)$ is $w_1 = (-0.55058181, -0.55394919)^T$. The optimal choice w corresponding to the maximum of $F(w)$ and Fisher ratio for different w is as follows:

w	$F(w)$
$w_1 = (-0.55058181, -0.55394919)^T$	10.127708730606209
$w_2 = (2.16975164, 2.13123849)^T$	10.133650068323425
$w_3 = (-2.16975164, -2.13123849)^T$	10.133650068323425
$w_4 = (-3.56219566, -3.54411647)^T$	10.130857431780456
$w_5 = (1.84757162, 1.56412254)^T$	10.102918586108007

Table 2: The w corresponding to the maximum of $F(w)$ and Fisher ratio for different w

From the above table, it shows although the norm of w is different, the normalized vector is the same which is almost $(1, 1)^T$ and the maximum of Fisher ratio is almost the same.

2. **Probability distributions visually by drawing contour plots**

- Contour lines for each class and the direction of the optimal choice vector
The theoretical decision boundary is perpendicular to the projector vector.

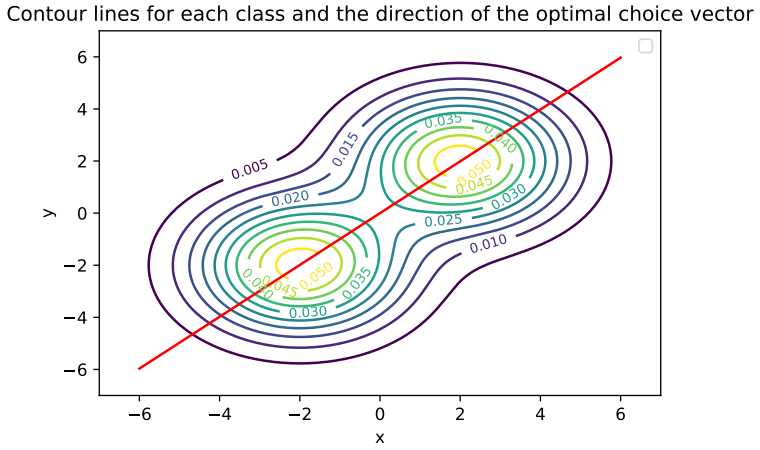

Figure 8: Contour lines for each class and the direction of the optimal choice vector

- The decision boundary for $S_a = S_b$
From Figure 11, the points with the magenta diamonds show the decision boundary, which is almost a line.

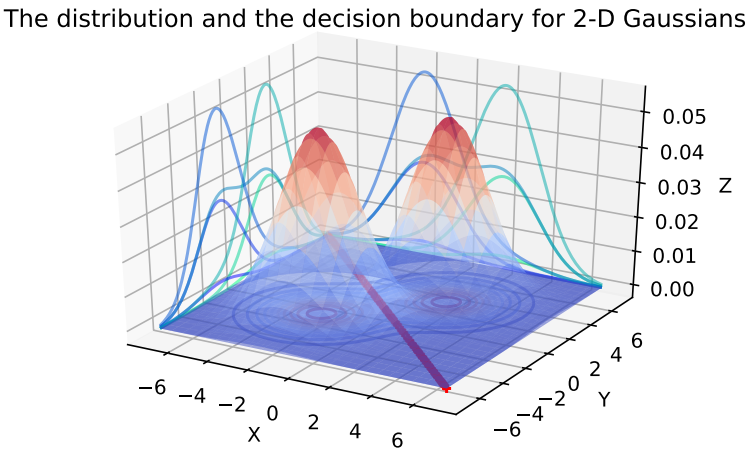

Figure 9: The distribution and the decision boundary for 2-D Gaussian

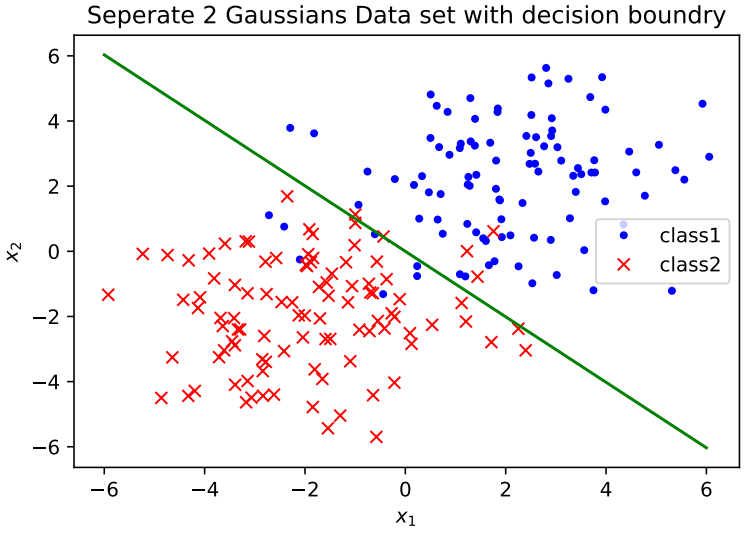

Figure 10: Separate 2 Gaussian Data set with theoretical decision boundary

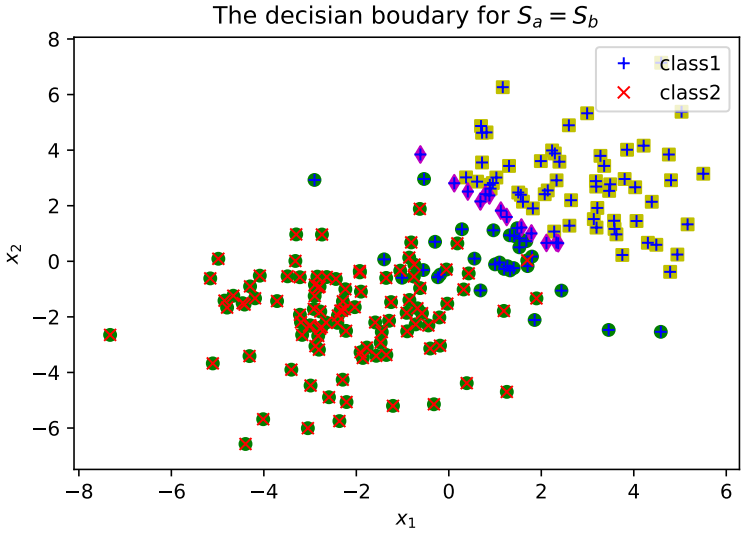

Figure 11: The decision boundary for $S_a = S_b$

- The decision boundary for $S_a \neq S_b$, with $S_a = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$, $S_b = \begin{pmatrix} 3 & 1.5 \\ 1.5 & 3 \end{pmatrix}$
  The blue points indicate the decision boundary. Ideally, it should be a quadratic curve.

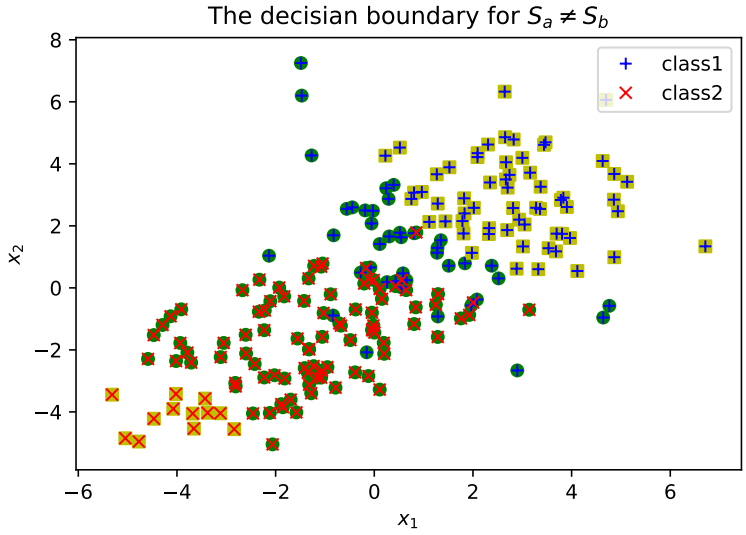

Figure 12: The decision boundary for $S_a \neq S_b$

- Using a unbalanced Fisher ratio formula for the discriminant
  Compared to balanced Fisher ratio formula, unbalanced Fisher ratio formula does not take prior probability of each class into count, which is

$$P(c = a \mid x^n) = \frac{n_a}{n_a + n_b}$$

$$P(c = a \mid x^n) = \frac{P(x^n \mid c = a)P(c = a)}{P(x^n)}$$

  The unbalanced Fisher ratio for different w is as follows:

	w_1	w_2	w_3	w_4	w_5
$F(W)$	0.03371	1.8564	1.8564	3.2803	5.2082

Table 3: The unbalanced Fisher ratio for different w

The rank order of unbalanced Fisher ratio is the same as balanced Fisher ratio, which has the same effect.

2.2 Iris data for three classification

1. Find the optimal direction w^* through the generalised eigenvalue problem
   There are four features in the data set, hence there are four eigenvectors, the eigenvector corresponding to the maximal eigenvalue(0.01963152) is the optimal direction

$$w^* = (0.25361489, -0.12043398, 0.72325561, 0.63093301)^T$$

2. The histograms of the three classes in the reduced dimensional space defined by w^*

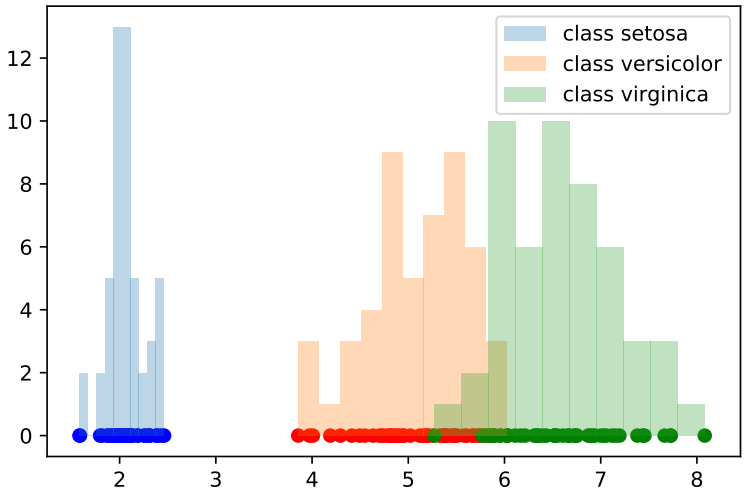

Figure 13: The histogram of the three classes

3. The histograms for eigenvectors out of the generalised eigenvectors
   $w_1 = w^* + (1, 3, 5, 7)^T$, $w_2 = w^* + (1, 1, 1, 1)^T$
   It shows eigenvectors out of the generalised eigenvectors do not have the better classification effect than generalised eigenvectors.

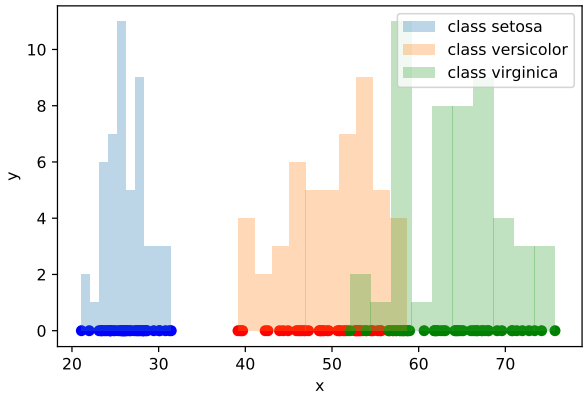

Figure 14: The Histogram for w_1

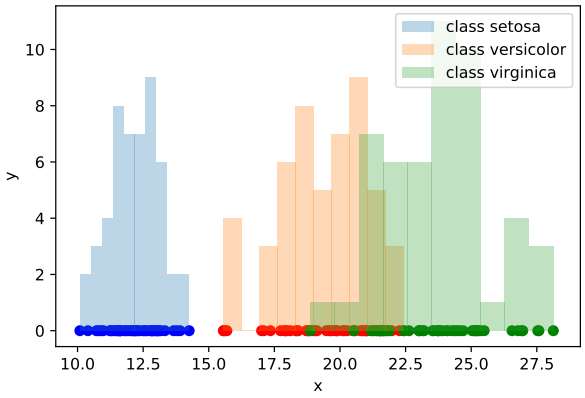

Figure 15: The Histogram for w_2

In addition, the projecting vectors that has the same classification effect as w^* is massive,which is $k * w^*, k \neq 0$. Because the generalised eigenvectors are the solution base of characteristic equation.

4. The relationship between the class separation task and the generalised eigenvector formulation
   Generalised eigenvector formulation is from LDA approach, which is find the maximum of $J(w)$. To differentiate it and set it to zero, then obtain

$$S_w^{-1} S_B w = J(w) w$$

therefore convert to solve the generalised eigenvalue problem

$$S_W^{-1} S_B w = \lambda w, where \lambda = J(w)$$

the optimal choice w is the eigenvector corresponding to the maximum eigenvalue

In terms of 2-Gaussian separation, decision boundary at equal posterior probability for either class is $Odds = 1$ i.e. $log(Odds) = 0$ and this is for 2 classification.

2.3 Linear Regression with non-linear functions

1. Linear regression through L_2 norm penalty by gradient descent
   The plots of different degrees of polynomial fitting are as follows:

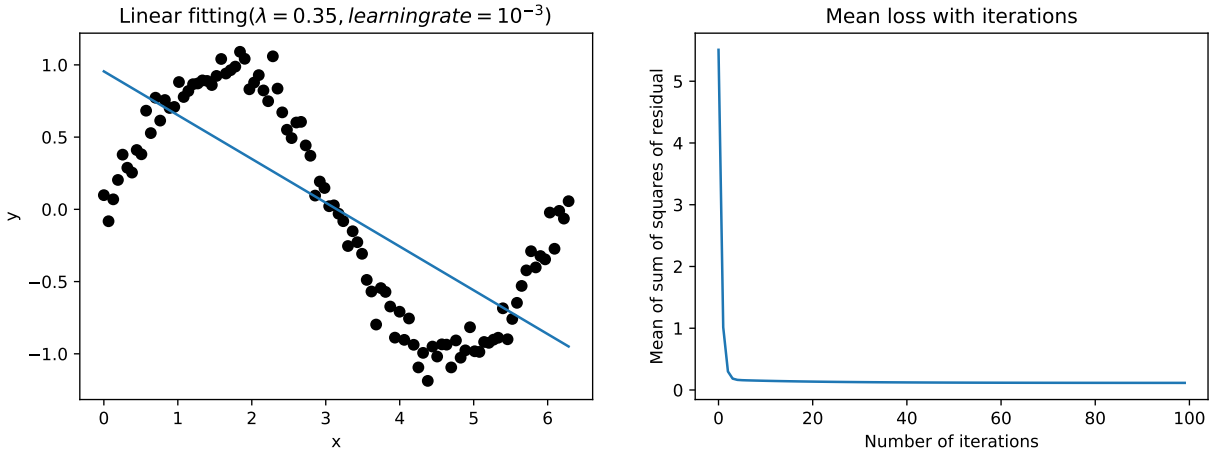

Figure 16: The plot of linear fitting

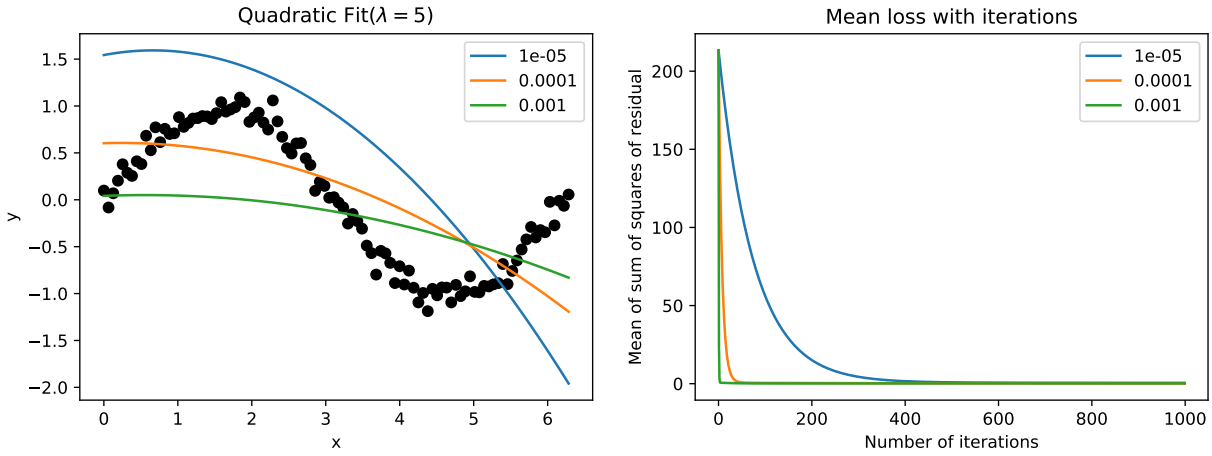

Figure 17: The plot of Quadratic fitting for different rates

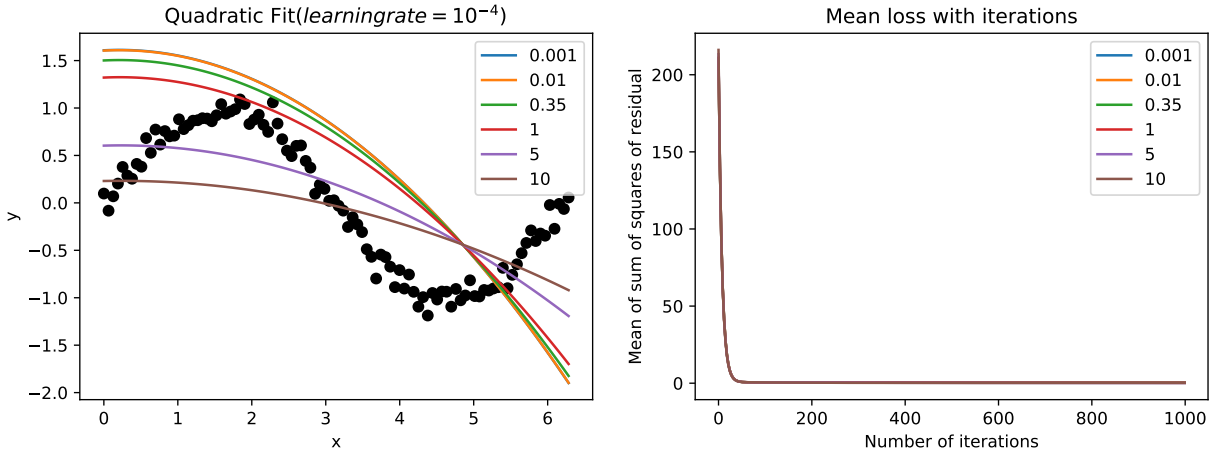

Figure 18: The plot of Quadratic fitting for different λ

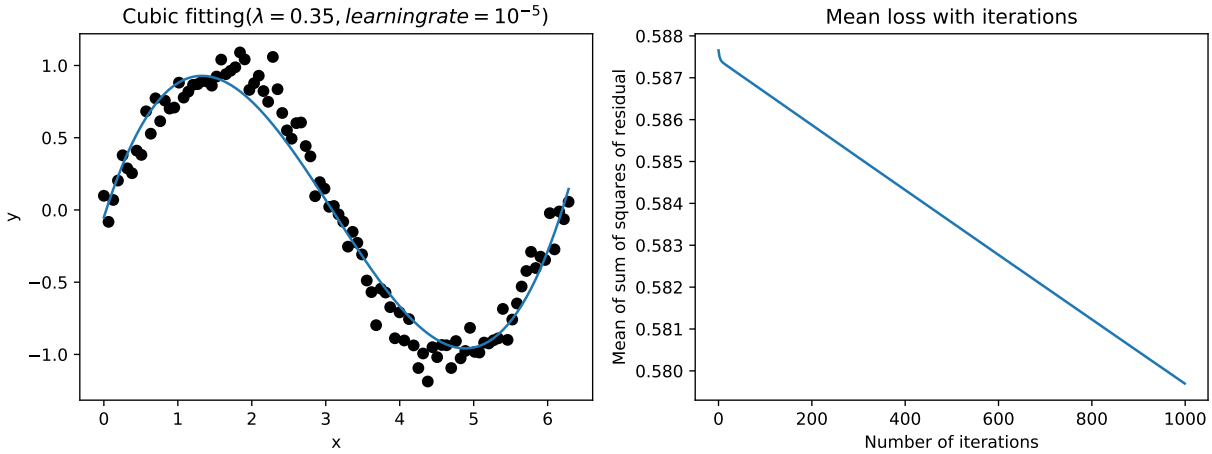

Figure 19: The plot of Cubic fitting

The figure indicates the cubic fitting can have a relative fitting effect.
The smaller the learning rate, the slower it converges. On the other hand, the larger the learning rate, it is harder to converge.
The larger the λ, the coefficients of higher degrees become smaller, the prediction curve is flatter.

2. Obtain the weights from the analytical expression
   From the below figure-fitting model curve, it shows when the hyper-parameter $\lambda > 1$, it works for fitting effect.
3. The relationship between performance metric and the degree of the polynomial and regularisation coefficient λ(training set 70%, test set 30%)
   - The degrees of the polynomial
   - The regularisation coefficient

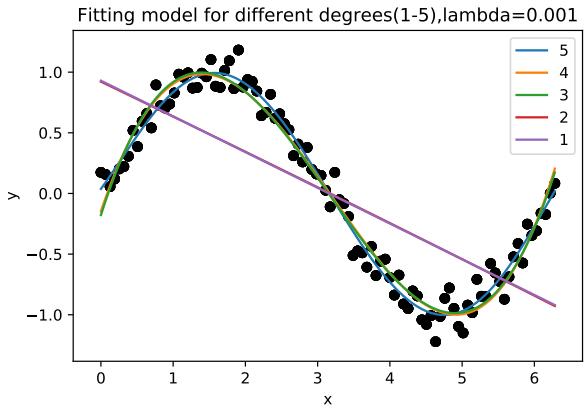

Figure 20: Fitting model for different degrees(1-5) through normal equation

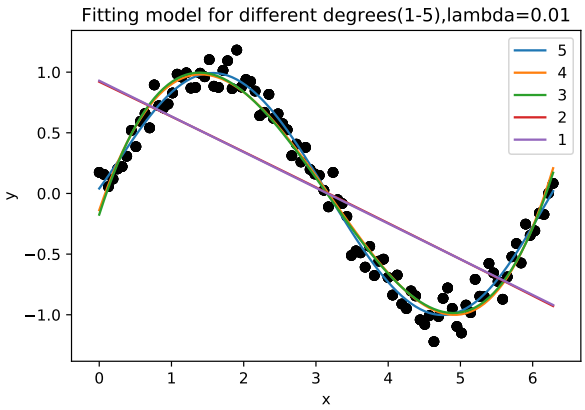

Figure 21: Fitting model for different degrees(1-5) through normal equation

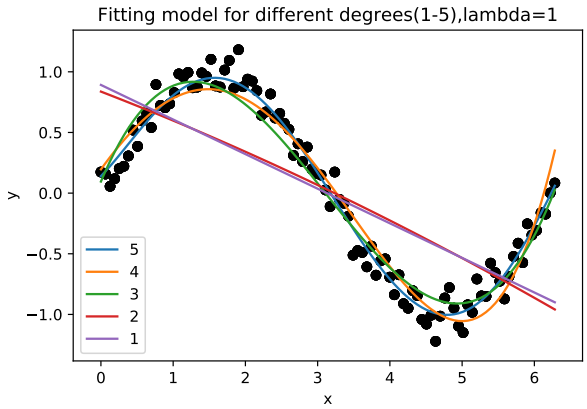

Figure 22: Fitting model for different degrees(1-5) through normal equation

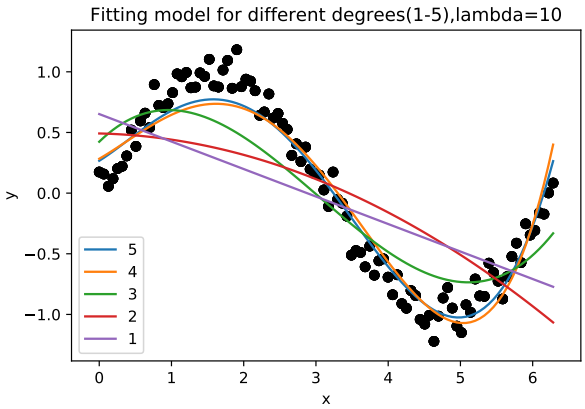

Figure 23: Fitting model for different degrees(1-5) through normal equation

The relationship figures are as follows.The figures show that the mean of the squared residuals for low degrees of polynomial is larger than high degrees of polynomial when λ is relative small. With λ increasing, the residuals for low degrees of polynomial almost keep the same, while the the residuals for high degrees of polynomial rise, especially for 3 degrees. Because the coefficients of higher degrees become smaller with the λ increasing.

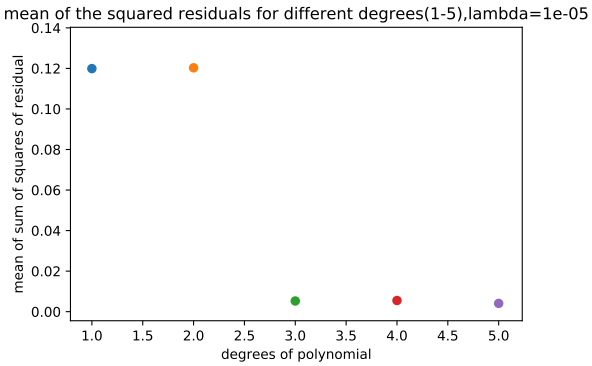

Figure 24: mean of the squared residuals on test set for different degrees(1-5)

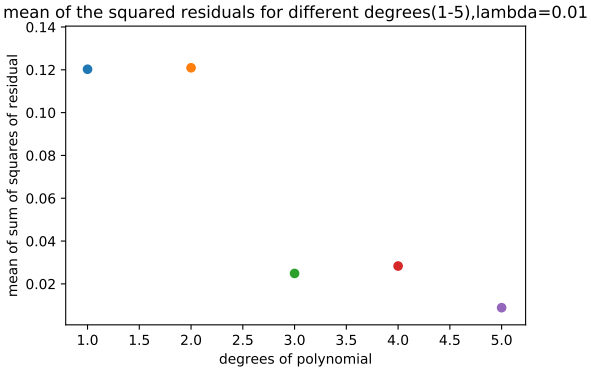

Figure 25: mean of the squared residuals on test set for different degrees(1-5)

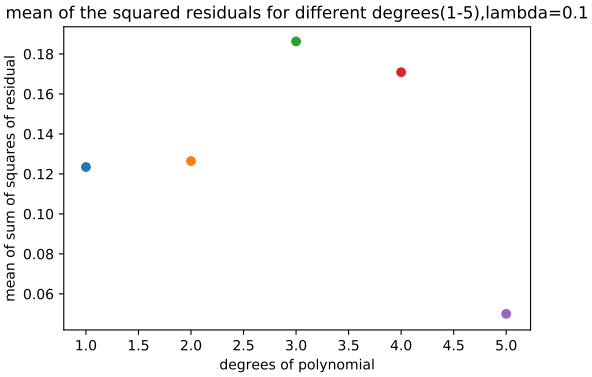

Figure 26: mean of the squared residuals on test set for different degrees(1-5)

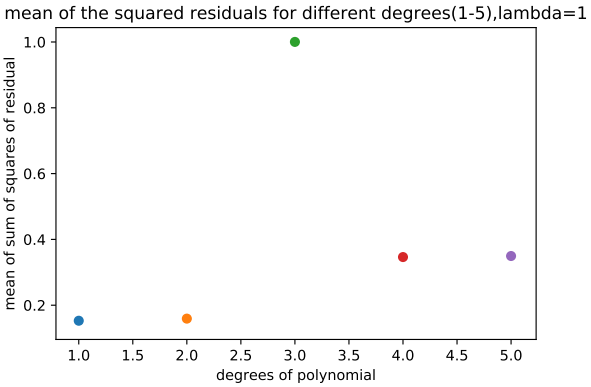

Figure 27: mean of the squared residuals on test set for different degrees(1-5)

The below figure indicates the the mean of the squared residuals increasing with the λ rising.

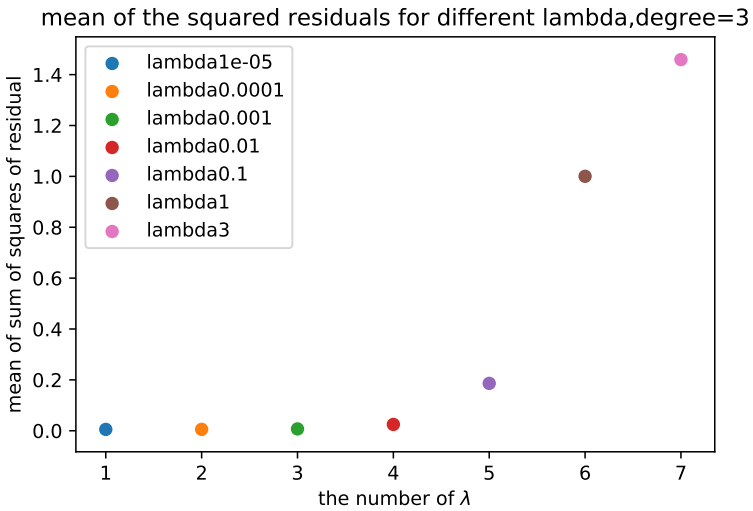

Figure 28: The plot of mean of the squared residuals on test set for different λ,degree=3

2.4 The generalisation of linear regression

Compared to the general training set and test set partition, the mean of the squared residuals of 10 fold cross-validation is a little bigger.

Overall, based on the above analysis, the model of the degree 3 of polynomial, around $\lambda = 0.1$ is relative better model, the

mean of the squared residuals is around 0.18628068498316946, compared to the maximum of true value around 1. Meanwhile, it can avoid overfitting problems caused by higher degrees of polynomials.

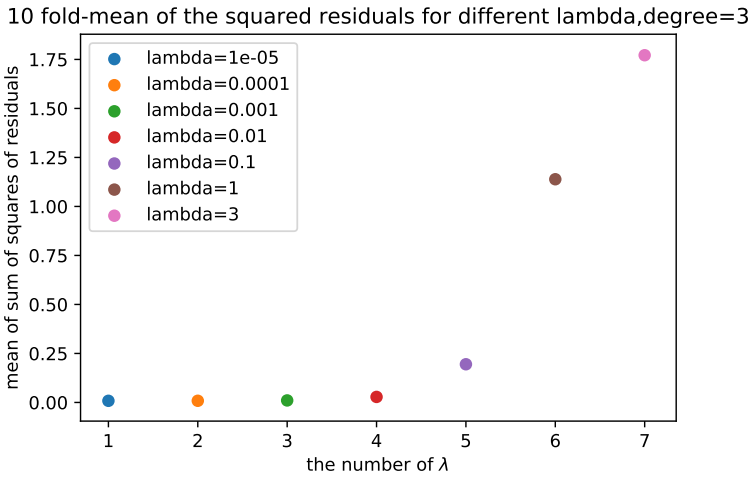

Figure 29: The plot of 10-fold mean of the squared residuals for different λ,degree=3