Machine Learning Lab 1

Di Wu

dw4y17@soton.ac.uk

1 My interpretation of the purpose of this lab

As far as i am concerned, this lab help me to understand how to generate a Gaussian distribution through other distribution, the relationship between shapes of data and covariance matrix in the situation of Multivariate Gaussian(two dimensions) and the geometric interpretation of PCA.

2 The way to get an Gaussian distribution from other distribution

2.1 Central Limit Theorem

Central Limit Theorem has different version of formulas. But they all describe the same meaning that the normalised sum of independent same random variables is properly tends toward a normal distribution.
In my opinions, the most suitable formula in this lab is

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{D} N(0,1)$$

the random variable x1 generated in matlab function is distribution to uniform distribution in area of 0 to 1, the mean of x1 is 0.5 and the variance is 1/12. For the convenience, choose n=12, so sum(rand(12,1))-6 is approximately distribution to a normal distribution. And the same situation as sum(rand(12,1))-6-(sum(rand(12,1))-6)=sum(rand(12,1))-sum(rand(12,1))

3 The shapes of Multivariate Gaussian(two dimensions) and the relationship between data shape and covariance matrix

The two dimensional data X1 is a data called white data which was a combination of two normal distribution variables and the covariance is $\begin{pmatrix}10\\01\end{pmatrix}$so there are no correlation between x1 and x2.Then after a transformation by matrix A with the condition that $A^*A = C$,we can get y with the covariance matrix $\begin{pmatrix}21\\12\end{pmatrix}$.The reason why we can get y with this covariance $y \sim \mathcal{N}(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}21\\12\end{pmatrix})$ is according to the theorem in the first introduction of this lab or more deeply as SVD(singular value decomposition).

The relationship between covariance matrix and shape of data is that the different covariance matrix define different eigenvalues and respective eigenvectors. This all togehter determine the shape of data, like $C1 = \begin{pmatrix}21\\12\end{pmatrix}$ the shape is a right upper ellipse because the covariance of $\sigma(y1, y2) = 1$. So they are positive related.if $C2 = \begin{pmatrix}2 & -1\\-1 & 2\end{pmatrix}$,the shape will be left upper. And the larger the covariance is ,the more narrow the shape is in orthogonal direction of the data shape.

4 The geometric interpretation of PCA

In the last part of experiment, we explore in which direction, yp (the projection of y) has the largest value of variance. And this is the main task of PCA. In the dimensionality reduction task, we want the lost of information is least so we assume that when the variance of projected data is largest it lost the least information. Given this, we explore in which direction projected data has the largest variance. the answer is that the largest direction is the same as the direction of eigenvector (response to the largest eigenvalue) and the projected data variance is eigenvalue. the result is $\sin(2x)+2$. The result graphes are below: