PRAMOD RAO - 2581268
YASHASWINI KUMAR - 2581[redacted]353

Sapna
PAGE NO.
DATE / /

ESL ASSIGNMENT - 3 (WE)

Problem 1

1. When the response variable is has finite number of outcomes (values). In other words, they are categorical in nature, logistic regression is used eg: Red, Blue [redacted]. & True / False.

Linear regression is not applicable since it can produce responses below 0 and beyond 1, i.e., it is not bounded between 0 & 1. ~~This is not~~
In other words it produces continious response which is not deriable for handling ~~to~~ categorical responses.

2. Logistic regression is used to model probability of an outcome based on the observations.
The relationship between independent variables & probability is:
If $X = (X_1, X_2 \dots X_p)$, then

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 \dots \beta_p X_p}}$$

where $p(x)$ is the probability obtained for the given independent variables.

3. Odds refers to the ratio between the total number of favorable outcomes to the ~~not~~ total number of unfavorable outcomes.

If 'p' is the probability of an event occuring, then 1-p is the probability of that event not occuring.

Then the odds is given by

$$\text{Odds} = (p/1-p)$$

Consider, probability of a rain to be 0.8, the the probability of not raining is 0.2. The odds of it rain on the particular day is $0.8/0.2 = 4$. In terms of ration it is 4:1, thus we can say the chances of raining is 4 times than not raining.

4. Equation 4.2 :=

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Eqn 4.4 :=

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

In consistent with definations of the text book,

Consider eqn 4.2,

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

taking log,

$$\log p(x) = \beta_0 + \beta_1 x$$

$$\frac{p(x)}{1-p(x)} = \frac{\frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}}{1 - \frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}} = \frac{\frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}}{\frac{1}{1+e^{\beta_0 + \beta_1 x}}}$$

$$\frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x}$$

taking log, we get

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \log\left(e^{\beta_0 + \beta_1 x}\right) = \beta_0 + \beta_1 x.$$

the above equation is 4.4. hence equivalent to 4

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$1 - p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$1 - p(x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$

assuming $p(x) \neq 0$, divide the above eqn by $p(x)$ we get,

$$\frac{1 - p(x)}{p(x)} = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)} \cdot \frac{1}{p(x)}$$

$$\frac{1 - p(x)}{p(x)} \cdot \frac{1}{1 + \exp(\beta_0 + \beta_1 x)} \times \frac{1 + \exp(\beta_0 + \beta_1 x)}{\exp(\beta_0 + \beta_1 x)}$$

$$\therefore \frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta_1 x)$$

taking log, we get

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \log(\exp(\beta_0 + \beta_1 x))$$

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x. \quad \text{— (1)}$$

the above equation is equivalent to 4.2.

(b) logit is defined as the log of odds i.e. $\log\left(\frac{p(x)}{1 - p(x)}\right)$

from (1) which is equivalent to logistic regression $p(x) = e^{\beta_0 + \beta_1 x} / (1 + e^{\beta_0 + \beta_1 x})$, We can say that logistic regression has a logit that is linear in X.

5) $$Odds(x) \equiv \frac{p(x)}{1-p(x)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 - \beta_i x_i)$$

$$Odd(x_i) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 - - + \beta_i x_i)$$

$$Odd(x_i + \Delta) = \exp(\beta_0 + \beta_1 (x_i) + \beta_2 x_2 + - - \beta_i (x_i + \Delta)$$

Hence

$$\frac{Odd(x_i + \Delta)}{odd(x_i} = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 - - - \beta_i x_i + \beta_i \Delta)}{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 - - - \beta_i x_i)}$$

$$= \exp(\beta_0 + \beta_1 x_1 - - \beta_i x_i + \beta_i \Delta - \beta_0 - \beta_1 x_1 - - + \beta_i x_i)$$

$$\boxed{\frac{odd(x_i + \Delta)}{odd(x_i)} = \exp(\beta_i \Delta)}$$

6) $$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$ If $p(x) = 1/2$, then

$$\frac{1}{2} = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$1 = \exp(\beta_0 + \beta_1 x)$$

$$0 = \beta_0 + \beta_1 x$$

$$\therefore x = \frac{-\beta_0}{\beta_1}$$

By choosing $x = -\beta_0/\beta_1$ we get probability to be 0.5.

The probability of 0.5 means, the chances of Y occurance is 0.5. If the problem is binary, then we can say the occurance of Y is no better than happening by chance.

7) ~~Multivariate logit~~

The conditional probability and ~~multi~~ logistic regression for k-classes can be given by

$$P(Y=1|x) = \frac{\exp(\beta_{10} + \beta_1 x)}{1 + \sum_{j=1}^{k-1} \exp(\beta_{j0} + \beta_j x)} \qquad k=1$$

$$P(Y=2|x) = \frac{\exp(\beta_{20} + \beta_2 x)}{1 + \sum_{j=1}^{k-1} \exp(\beta_{j0} + \beta_j x)} \qquad k=2$$

$$P(Y=k-1|x) = \frac{\exp(\beta_{k-1,0} + \beta_{k-1} x)}{1 + \sum_{j=1}^{k-1} \exp(\beta_{j0} + \beta_j x)} \qquad k=k-1$$

~~b) the model uses last class in den~~

$$P(Y=k|x) = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(\beta_{j0} + \beta_j x)}$$

~~c) the model use~~

Since, the model uses last class in the denominator we get, $P(Y=k|x) = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(\beta_{j0} + \beta_j x)}$

Clearly, $\sum_{j=1}^{k} P(Y=j|x) = 1$. Thus satisfying the probability.

Taking log odds, and using the result from (6) problem 1.5, i.e.,

$$\frac{odd(x_i + \Delta)}{odd(x_i)} = \exp(\beta_i \Delta)$$

We get model of the form,

$$\log\left(\frac{P(Y=1|x)}{P(Y=k|x)}\right) = \beta_{10} + \beta_1 x$$

$$\log\left(\frac{P(Y=2|x)}{P(Y=k|x)}\right) = \beta_{20} + \beta_2 x$$

$$\log\left(\frac{P(Y=k-1|x)}{P(Y=k|x)}\right) = \beta_{k-1,0} + \beta_{k-1} x$$

Let the entire parameter set be given by

$$\theta = \{\beta_{10}, \beta_1 \dots \beta_{(k-1)0}, \beta_{k-1}\}$$

then we can represent the predictions made for an input x as following

$$\hat{p}_1(x, \theta_1) = \frac{\exp(\hat{\beta}_{10} + \hat{\beta}_1 x)}{1 + \sum_{j=1}^{k-1} (\hat{\beta}_{j0} + \hat{\beta}_j x)}$$

$$\hat{p}_{k-1}(x, \theta_{k-1}) = \frac{\exp(\hat{\beta}_{(k-1)0} + \hat{\beta}_{k-1} x)}{1 + \sum_{j=1}^{k-1} (\hat{\beta}_{j0} + \hat{\beta}_j x)}$$

then the class predicted is according to

$$\hat{f}(x) = \underset{j=1,\dots k}{\arg\max}\ \hat{p}_j(x, \theta_j)$$

Problem 2

(1) Linear Discriminant Analysis

Let us consider π_K as a prior probability that an observation that is chosen randomly belongs to 'kth class

Let us consider a density function $f_K(x)$ which denotes a particular observation has come from class K.

$$\therefore \quad f_K(x) = Pr\,(x = x \mid Y = K)$$

Now, we know Bayes Theorem states as below :-

$$Pr\,(Y = k \mid X = x) = \frac{\pi_K f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \longrightarrow (1)$$

which means probability that x belongs to K^{th} class by total probability.

We have $P_K(x)$ which is posterior probability, which is equal to $Pr\,(Y = K \mid X = x)$ and substitute the values of π_K and $f_K(x)$

$$\text{i.e} \quad P_K(x) = Pr\,(Y = k \mid X = x)$$

Before we estimate $f_K(x)$ we assume it is normal/Gaussian

So, the normal density function takes the form

$$f_K(x) = \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left(-\frac{1}{2\sigma_K^2}(x - \mu_K)^2\right)$$

where, σ_K^2 is the variance of K^{th} class

μ_K is the mean of K^{th} class.

In this case, we assume for LDA, all variances are same

$$\text{i.e} \quad \sigma_1^2 = \sigma_2^2 = \cdots\cdots = \sigma_K^2$$

Hence, we consider it as σ^2

Substitute the values of $f_K(x)$ in eqn (1), we get.

$$P_{K(x)} = \frac{\pi_K f_K(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$$= \frac{\pi_K \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left(-\frac{1}{2\sigma_K^2}(x - \mu_K)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x - \mu_l)^2\right)}$$

we assumed all variance to be σ^2, so ....>

$$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}.$$

On the above equation, we take all term which do not vary across k as c because they do not contribute in maximisation.

$$\therefore \frac{\pi_k \left(\frac{1}{\sqrt{2\pi}\sigma}\right) \exp\left(-\frac{1}{2\sigma^2}(x^2 + \mu_k^2 - 2x\mu_k)\right)}{\sum_{l=1}^{K} \left(\pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)\right)}$$

$$\therefore \text{Here, C value} = \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x^2)\right)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

So, we have

$$P_k(x) = C \cdot \pi_k \exp\left(-\frac{1}{2\sigma^2}(\mu_k^2 - 2x\mu_k)\right)$$

Take log on both sides,

$$\log P_k(x) = \log\left(c \cdot \pi_k \cdot \exp\left(-\frac{1}{2\sigma^2}(\mu_k^2 - 2x\mu_k)\right)\right)$$

$$= \log(c) + \log \pi_k + \log\left(\exp\left(-\frac{1}{2\sigma^2}(\mu_k^2 - 2x\mu_k)\right)\right)$$

$$= \log(C) + \log \pi_k + -\frac{1}{2\sigma^2}(\mu_k^2 - 2x\mu_k)$$

$$= \log(C) + \log(\pi_k) + \left(-\frac{\mu_k^2}{2\sigma^2} + \frac{2\mu_k x}{2\sigma^2}\right)$$

$$\therefore \log P_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) + \log(c)$$

Here, $\log(c)$ is a constant as it dos not have any effect in maximization.

Therefore, the above equation is equivalent to the disriminant function $\delta_k(x)$.

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

which shows, maximizing the disriminant function $\delta_k(x)$ is same as maximizing the orginal function $P_k(x)$ because the positive constant ($\log c$) has no effect on maximization.

(2) Quadratic discriminant Analysis.

This case is similar to that of linear discriminant analysis but here we have a class specific mean and variance. We already know the density function is given by

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right).$$

We also know the posterior probability function $P_k(x)$

$$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)}$$

Taking log on both sides

$$\log P_k(x) = \frac{\log\left(\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)\right)}{\log\left(\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)\right)}$$

$$\log P_k(x) + \log\left(\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)\right) = \log\left(\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)\right).$$

$$\Rightarrow \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) + \log\left(\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)\right)$$

$$\Rightarrow \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) + \left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

$$\therefore \log P_k(x) + \log\left(\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)\right) = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) + \left(-\frac{x^2}{2\sigma_k^2} + \frac{\mu_k^2}{2\sigma_k^2} - \frac{x\mu_k}{2\sigma_k^2}\right)$$

The total probability remains constant as it does not have any effect in maximizing and let it be C.

$$\log P_k(x) \cdot C = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) - \frac{x^2}{2\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \frac{x\mu_k}{\sigma_x^2}.$$

The above equation is also equivalent to the discriminant function $\delta_k(x)$.

where, $\delta_k(x) = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) + -\frac{1}{2\sigma_k^2}(x-\mu_k)^2$

Therefore, in the equation below, x appears as a quadratic term in the classifier. which shows the case here, Bayes classifier is not linear but in fact quadratic

$$\log P_k(x) \cdot C = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) - \frac{x^2}{2\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \frac{x\mu_k}{\sigma^2}$$

Therefore, there is the x^2, hence quadratic.

3) Given that there are k iid samples, x_i, $i = 1, 2 \ldots k$ with a positive pairwise correlation ρ and $Var(x_i) = \sigma^2$ for $i = 1, 2 \ldots k$

We know that by definition of mean, the mean of 'k' iid random variables is given by

$$\bar{x} = \frac{1}{k} \sum_{i=1}^{k} x_i \quad -(1)$$

consider,

$$Var(aX) = E\left[(aX - a\mu)^2\right] \quad \text{where } \mu \text{ is the mean,}$$
$$= E\left[a^2 (X-\mu)^2\right]$$
$$= a^2 E\left[(X-\mu)^2\right] = a^2 Var(X)$$

$\therefore$ (1) becomes,

$$Var(\bar{x}) = Var\left(\frac{1}{k} \cdot \sum_{i=1}^{k} x_i\right) = \frac{1}{k^2} \cdot Var\left(\sum_{i=1}^{k} x_i\right) \quad -(2)$$

~~In the above equation, $x_1 \ldots x_k$ are random var~~

Now, we can write,

$$Var\left(\sum_{i=1}^{n} x_i\right) = E\left(\left[\sum_{i=1}^{k} x_i\right]^2\right) - \left(E\left[\sum_{i=1}^{k} x_i\right]\right)^2$$

Note $\left(\sum_{i=1}^{k} a_i\right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j$ viz equivalent to $(a_1 + a_2 \ldots a_n)(a_1 + a_2 \ldots a_n)$

$\therefore E\left[\left(\sum_{i=1}^{k} x_i\right)^2\right] = E\left[\sum_{i=1}^{k} \sum_{j=1}^{k} x_i x_j\right] = \sum_{i=1}^{k} \sum_{j=1}^{k} E(x_i x_j)$

$$\text{and } \left[E\left(\sum_{i=1}^{k} x_i\right)\right]^2 = \left[\sum_{i=1}^{k} E(x_i)\right]^2 = \left[\sum_{j=1}^{k} \sum_{i=1}^{k} E(x_i) \cdot E(x_j)\right]$$

$$Var\left(\sum x_i\right) = \sum_{i=1}^{k} \sum_{j=1}^{k} E(x_i x_j) - E(x_i) E(x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(x_i, x_j)$$

$\therefore$ by definition of covariance.

Using the above result, (2) becomes,

$$Var(\bar{x}) = \frac{1}{k^2} \cdot \sum_{i=1}^{k} \sum_{j=1}^{k} Cov(x_i, x_j) \quad -(3)$$

The above covariance can be written as

$$cov(x_i, x_j) = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & & \vdots \\ \vdots & & \ddots & \rho\sigma^2 \\ \rho\sigma^2 & \cdots & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

where ρ is pairwise correlation.

i.e. if $i=j$, then $cov(x_i, x_j) = \sigma^2$ & $cov(x_i, x_j) = \rho\sigma^2$

There are k^2 number of elements in $cov(x_i, x_j)$ & $i \neq j$ there are k elements where $i=j$. Hence the rest is given by $k^2 - k$.

$$\therefore \quad Var(\bar{X}) = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} cov(x_i, x_j) = \frac{1}{k^2}\left(k\sigma^2 + (k^2-k)\rho\sigma^2 \right)$$

$$= \frac{\sigma^2}{k} + \rho\sigma^2 - \frac{1}{k}\rho\sigma^2$$

$$= \rho\sigma^2 + \frac{\sigma^2}{k}(1-\rho)$$

Problem 4 (P, 20 Points)

1. (2P) Download and load the phoneme data set (phoneme.csv) from the course website. Split the dataset into training and test set according to the speaker column. Be sure to exclude the row number, speaker and response columns from the features.

   Solution. Refer to section 4.1 in the code. ■

2. (3P) Fit an LDA model, compute and report train and test error.

   Solution. Refer to section 4.2 in the code.

   The train error is **0.055** and the corresponding test error is given by **0.080**. ■

3. (3P) Plot the projection of the training data onto the first two canonical coordinates of the LDA using the plot() function. Investigate the data projected on further dimensions using the *dimen* parameter.

   Solution. Refer to section 4.3 in the code. Plot the projection of the training data onto the first two canonical coordinates of the LDA using the phonemes dataset is given by **Figure1**.

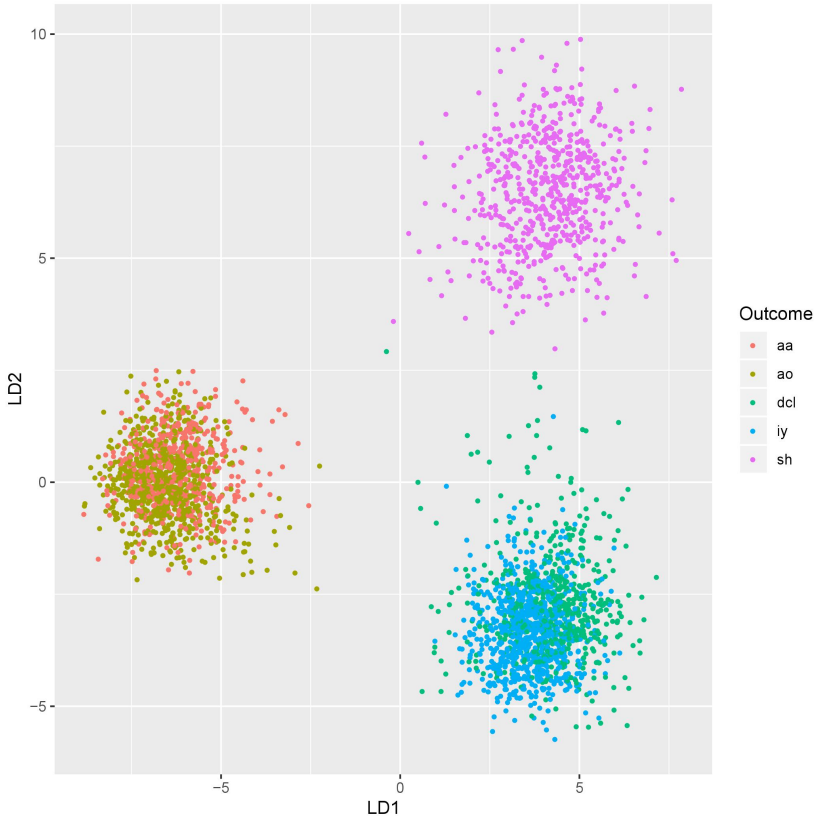

Figure 1: First two canonical coordinates of the LDA

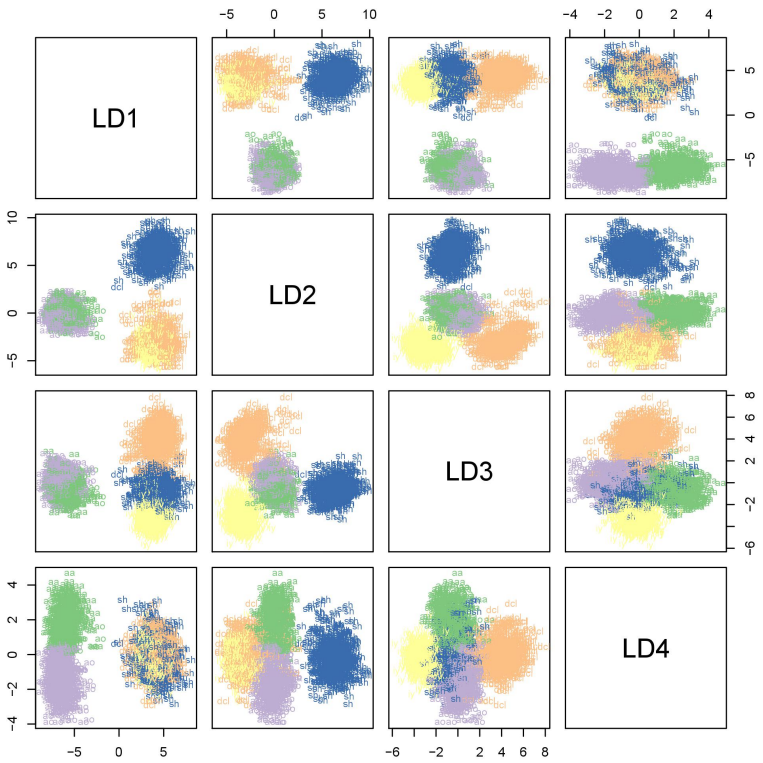

Figure 2: Plot between all the canonical coordinates of the LDA

We can see the data projected across further dimensions in **Figure2**

■

4. Select the two phonemes aa and ao. Fit an LDA model on this data set and repeat the steps done in (2).

   Solution. Refer to section 4.4 in the code.

   For the reduced dataset, the LDA train error is **0.106** and the test error is **0.214**.

■

5. (6P) Repeat steps (2) and (4) using QDA and report your findings. Would you prefer LDA or QDA in this example? Why?

   Solution. Refer to section 4.4 in the code.

   For the complete dataset, the QDA train error is **0.0** and the test error is **0.158**.

   For the classes "aa" and "ao", the QDA train error is **0.0** and the test error is **0.339**.

   Clearly the QDA model is overfitting to the train data and poorly generalising on the test data with respect to the QDA model. Hence, the preferred model is **LDA**.

■

6. (3P) Generate confusion matrices for the LDA and QDA model for aa and ao. Which differences can you observe between the models?

   Solution. Refer to section 4.6 in the code.

   The confusion matrices for the LDA and QDA model for aa and ao is given by **Table1** and **Table2** respectively.

Table 1: LDA Confusion Matrix

Pred\Actual	aa	ao
aa	121	39
ao	55	224

Table 2: QDA Confusion Matrix

Pred\Actual	aa	ao
aa	29	2
ao	147	261

From the tables it is evident that the QDA model tends to predict every observation to the class "ao", while the LDA model does fairly well in distinguishing between the two classes. ■