Assignment-2
AV489 Machine Learning for Signal Processing

Subrahmanya V Bhide (SC18B030)
Indian Institute of Space Science and Technology
Department of Aerospace Engineering

(Dated: 19 March 2021)

This document is a report based on the tasks done as part of the Assignment-2 for the course Machine Learning for Signal Processing. The tasks cover Bayesian Descion Theory and Maximum Likelihood Estimation.

QUESTION 1

A modified form of the 'Optical Recognition of Handwritten Digits Dataset' from the UCI repository is used wherein only the data for digits 5 and 6 are considered to implement a binary bayesian classifier. The original 32×32 input matrix is reduced to an input matrix of size 8×8. The input for the classifier is however a vector of length 64. The distribution is assumed to be a multivariate gaussian with parameters *mean(μ)* and *covariance(Σ)*.
The MLE estimate for mean (μ) is given by Eqn. 1.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

The MLE estimate for the Covariance matrix is given by Eqn 2.

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^t \tag{2}$$

Three cases for the covariance matrix are considered, where in the first case the Σ for both classes are different ($\Sigma_5 \neq \Sigma_6$) for which the summation in the expression must be taken seperately over the training data for the two classes. In the next case we assume that the covariance for both the classes are the same ($\Sigma_5 = \Sigma_6$). Here the summation is taken over the training data for both the classes. For the third case we assume the covariance matrices to be diagonal and equal. The i^{th} diagonal element of the 64×64 matrix is given by Eqn. 3. Since the matrices are same for both the classes the summation is taken over the training data for both the classes.

$$\hat{\Sigma_{i,i}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2_{64-i,64-i} \tag{3}$$

The estimate for the priors are the ratio of instances of the training data belonging to C_i to the total number of instances used for training. Thus,

$$\pi = P(r \in C_5) = \frac{396}{777}$$

$$1 - \pi = P(r \in C_6) = \frac{381}{777}$$

After the parameters are estimated, Bayesian descion theory is used to test the classifier on data which was not used for training (i.e. testing data). For the first case of $\Sigma_5 \neq \Sigma_6$ the confusion matrix is given in Tab. I. The misclassification rate is $\frac{76}{333} = 22.82\%$.

For the second case of $\Sigma_5 = \Sigma_6$ the confusion matrix is given in Tab. II. The misclassification rate is $\frac{48}{333} = 14.4\%$.

We can observe that as compared to the first case in the second case the rate of misclassification is lower. Based on this result we can hypothesis that considering $\Sigma_5 = \Sigma_6$ and also assuming that Σ's are diagonal would give a higher misclassification rate than the second case but it could be still lower than that of the first case.

For the third case of $\Sigma_5 = \Sigma_6$ and both are diagonal matrices, the confusion matrix is given in Tab. III. The misclassification rate is $\frac{68}{333} = 20.42\%$, which is better than the first case. We can also observe here that there is a bias between the correct classification of 5 and 6 for the first two cases whereas this is almost absent in the third case.

TABLE I: Confusion matrix for Case 1.

Total N =333	Prediction of 5	Prediction of 6
Actual 5	106	49
Actual 6	27	151

TABLE II: Confusion matrix for Case 2.

Total N =333	Prediction of 5	Prediction of 6
Actual 5	134	21
Actual 6	27	151

TABLE III: Confusion matrix for Case 3.

Total N =333	Prediction of 5	Prediction of 6
Actual 5	132	23
Actual 6	45	133

Thus if we consider misclassification as the metric then Case 2 is the most appropriate classifier of the three examined. But if we consider the bias between the correct classification for both the classes as a metric then the Case 3 is more preferrable.

QUESTION 2

The problem involves two class classification with two attributes. We consider the likelihoods to be normal and estimate the parameters using the Maximum Likelihood Estimation.
The MLE estimate for Means(μ) is given by Eqn. 1.
Four cases for the covariance matrix are considered.

Case 1

We consider

$$\Sigma_0 = \Sigma_1 = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$$

The MLE estimate for a is given by Eqn. 4. Since we have assumed that $\Sigma_0 = \Sigma_1$ the summation must be taken over training data for both classes.

$$a = \frac{1}{2n} \sum_{i=1}^{n} (x_i - \mu)^t (x_i - \mu) \tag{4}$$

The isoprobability curves, descion boundary and the data points are shown in Fig. 1. We can observe that since the Σs are equal the probability distribution is similar, the descion boundary is linear and the means are different. Also since the covariance matrix is diagonal and the elements are equal we can observe that the isoprobabliltiy lines are circular.
The 3D probability distribution of the data points is shown in Fig. 2. The confusion matrix is shown in Tab. IV.

TABLE IV: Confusion matrix for Case 1.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	20	25
Actual C_1	27	18

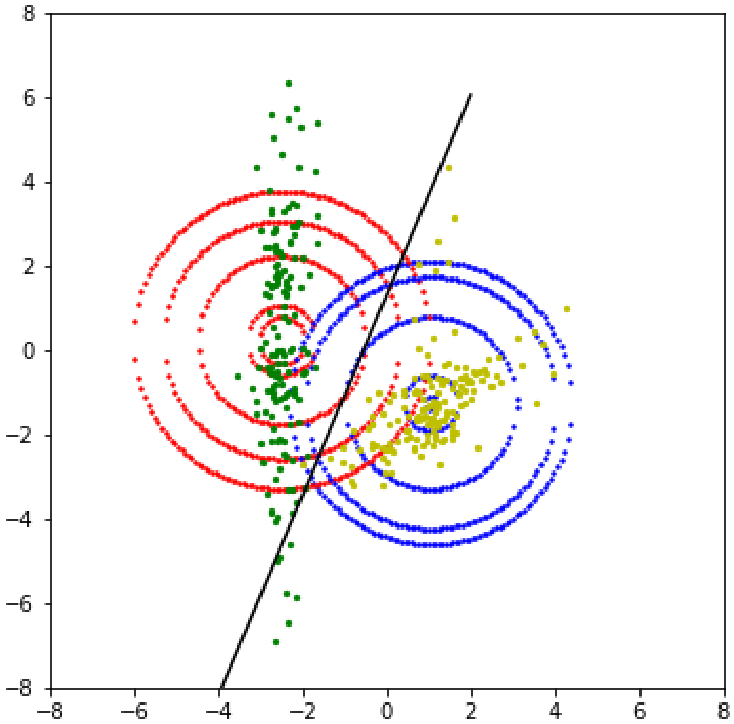

FIG. 1: Case 1: Equal and Diagonal Σs with equal diagonal elements.

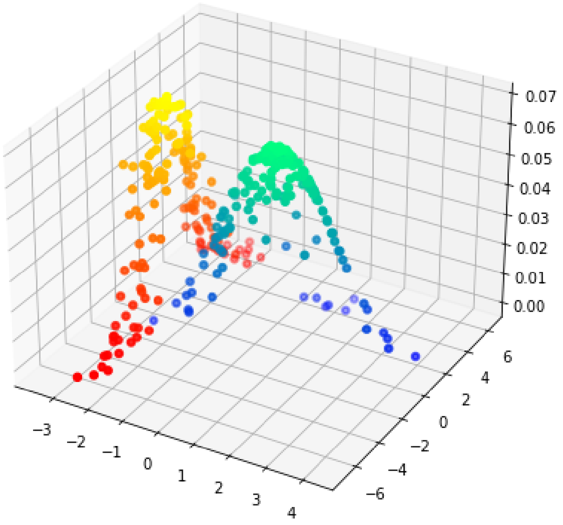

FIG. 2: Case 1: Probability Distribution.

Case 2

We consider

$$\Sigma_0 = \Sigma_1 = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

The MLE estimate for Σ is given by Eqn. 5. Since we have assumed that $\Sigma_0 = \Sigma_1$ the summation must be taken over training data for both classes.

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)_2^t(x_i - \mu)_2$$

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)_1^t(x_i - \mu)_1 \tag{5}$$

The isoprobability curves, descion boundary and the data points are shown in Fig. 3. We can observe that since the Σs are equal the probability distribution is similar, the descion boundary is linear and the means are different. Also since the covariance matrix is diagonal but the elements are unequal we can observe that the isoprobabliltiy contours are elliptical.

The 3D probability distribution of the data points is shown in Fig. 4. The confusion matrix is shown in Tab. V.

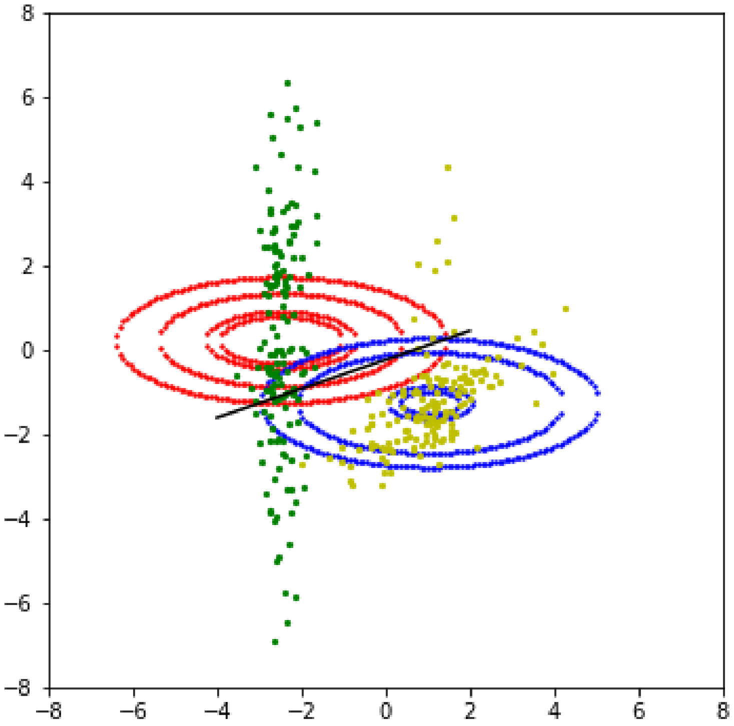

FIG. 3: Case 2: Equal and Diagonal Σs with unequal diagonal elements.

TABLE V: Confusion matrix for Case 2.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	15	30
Actual C_1	17	28

Case 3

We consider

$$\Sigma_0 = \Sigma_1 = arbitrary$$

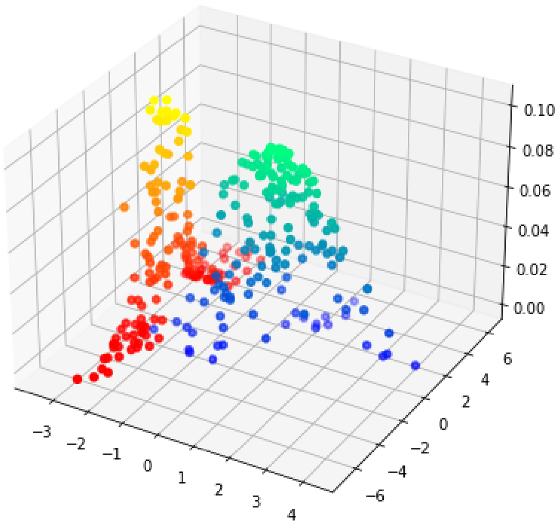

FIG. 4: Case 2: Probability Distribution.

The MLE estimate for Σ is same as that given by Eqn. 2. Since we have assumed that $\Sigma_0 = \Sigma_1$ the summation must be taken over training data for both classes. The isoprobability curves, descion boundary and the data points are shown in Fig. 5. We can observe that since the Σs are equal the probability distribution is similar, the descion boundary is linear and the means are different.

The 3D probability distribution of the data points is shown in Fig. 6. The confusion matrix is shown in Tab. VI.

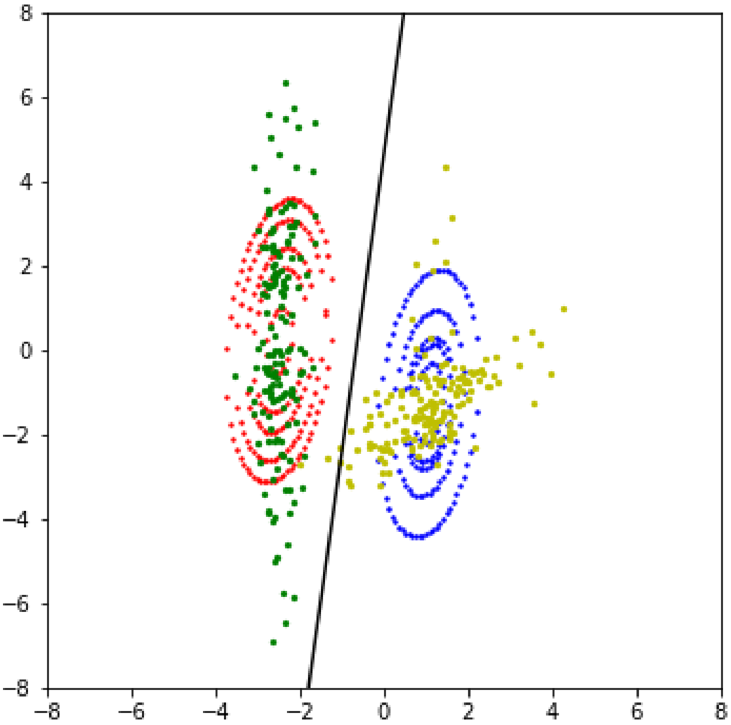

FIG. 5: Case 3: Equal Σs with arbitrary elements.

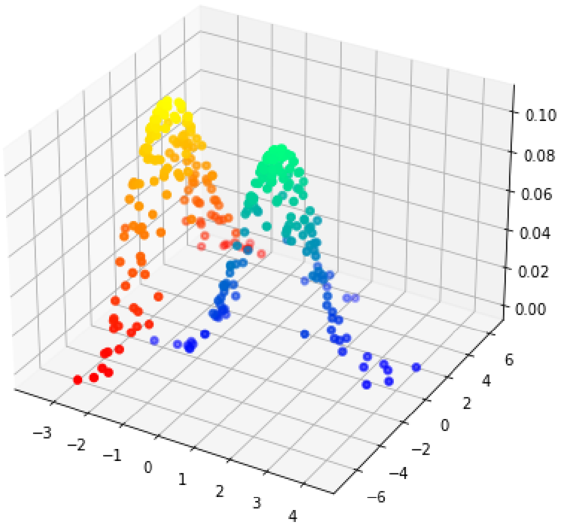

FIG. 6: Case 3: Probability Distribution.

TABLE VI: Confusion matrix for Case 3.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	23	22
Actual C_1	28	17

Case 4

We consider

$$\Sigma_0 \neq \Sigma_1 = arbitrary$$

The MLE estimate for Σ is same as that given by Eqn. 2. Since we have assumed that $\Sigma_0 \neq \Sigma_1$ the summation must be taken over training data for each classes seperately. The isoprobability curves, descion boundary and the data points are shown in Fig. 7. We can observe that since the Σs are unequal the probability distributions are not the same. Also the descion boundary unlike the three cases discussed is quadratic and not linear.

The 3D probability distribution of the data points is shown in Fig. 8. The confusion matrix is shown in Tab. VII. From the Confusion matrices we can observe that the misclassification rates for different cases are:

TABLE VII: Confusion matrix for Case 4.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	22	23
Actual C_1	28	17

1. 57.8% for Case 1.
2. 52.3% for Case 2.
3. 55.5% for Case 3.
4. 56.7% for Case 4.

From these data we can observe that the Best performance for the given testing dataset is in the Case 2 followed by Case 3,4 and 1 respectively. Actually case 4 considers the most general conditions but it shows more misclassification than that of Case 2 and 3. Thus for the given data the classifier generated in Case 2 is better than the rest. To generalize on a better classifier, since the misclassifiaction rates are close to each other, other sets of testing data can also be used to verify for any bias in the testing data and then the descion on a better classifier can be made.

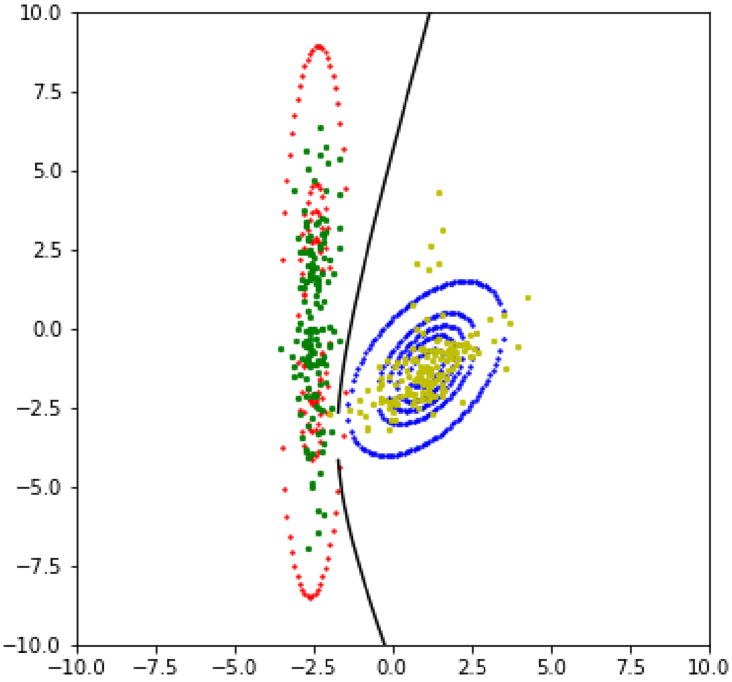

FIG. 7: Case 4: Unequal Σs with arbitrary elements.

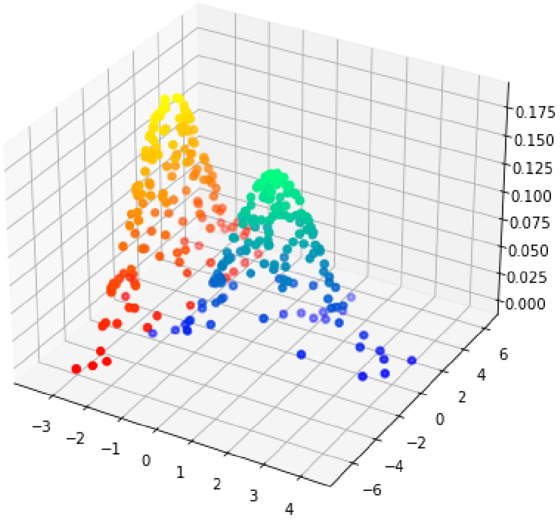

FIG. 8: Case 4: Probability Distribution.

QUESTION 3

The dataset corresponding to the problem is the 'Wage dataset' which contains the income survey information for a group of males from Atlantic region of the United States.
Out of the different attributes we take *Age, Education* and *Calender Year* into our consideration. The plots and polynomial curve fits for the different attributes versus Wage are given in the Figs. 9, 10 and 11.
Obervations that can be made from these plots are:

1. We can observe that there is a clear seperation, with a large majority of people remaining below 250 on the wage axis and only few people crossing the barrier.

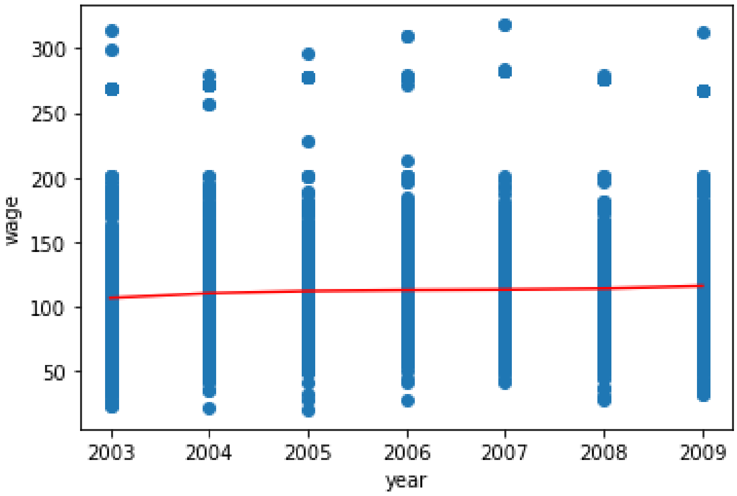

FIG. 9: Year vs Wage.

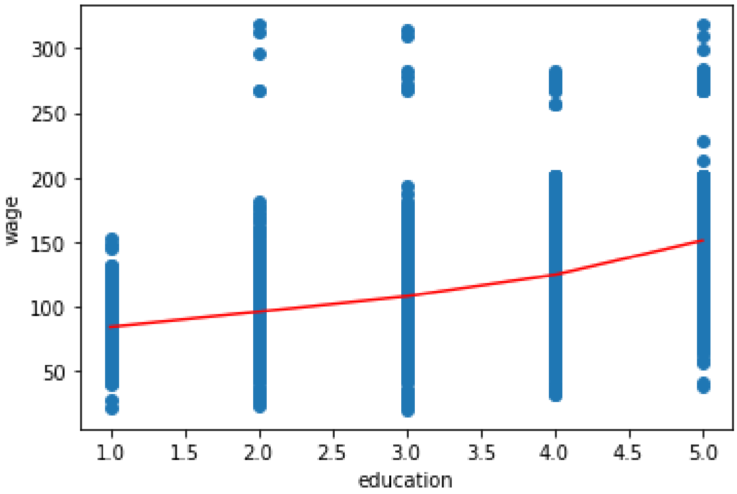

FIG. 10: Education vs Wage.

2. The people who have a wage higher than 250 are those who are in the age range of 30-70 and have higher points for education.

3. There is a gradual increase in the wage over the years whereas the increase is pronounced with increase in education.

4. The wage is almost a constant for the people of age between 30-70 and people falling out of this category have lesser wage.

The attributes certainly are a factor for the wage particularly education and age as they show more variations and trends than the calendar year. But from our experience we can tell that calendar year can also be considered as a factor due to inflation and its effect on wages. From our understanding we can conclude that the three attributes are almost independant. But largely we can also consider that as years go by more people get higher education and also as the years go ahead the inclusion of aged people in the working class can reduce. Hence to determine the degree to which these attributes affect and vary with respect to each other we use the Spearman's Correlation coefficient. The Spearman's correlation coefficient between two random variables x and y can be defined as in Eqn. 6:

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \tag{6}$$

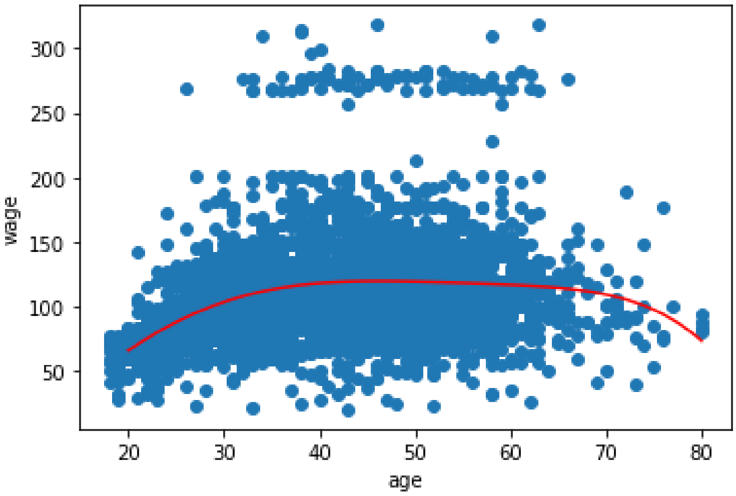

FIG. 11: Age vs Wage.

The correlation coefficients between different atrributes are tabulated in Tab. VIII. We can observe that as expected there is very low correlation between the three attributes and hence using only one or two of them for determining the wage will not lead to a good classifier. To emphasis this we create a classifier with each of the 3 attributes discussed as the attribute used to classify wage into 2 classes, above and below the mean wage. We also create a classifier that uses all the three attributes together to classify.

TABLE VIII: Correlation Coefficients.

$\rho_{education,calendaryear}$	0.01
$\rho_{education,Age}$	0.07
$\rho_{Calendaryear,Age}$	0.04

We assume the distributions to be normal and use MLE to obtain the estimates for mean and variance. We also divide the dataset into training and testing datasets to validate our results. We consider that C_0 corresponds to wage being lesser than mean wage and C_1 for wage higher than mean wage.

Case 1 : Using Age alone for wage prediction

The confusion matrix for this case is shown in Tab. IX.

TABLE IX: Confusion matrix for wage prediction using Age.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	49	0
Actual C_1	41	0

Case 2 : Using Education alone for wage prediction

The confusion matrix for this case is shown in Tab. X.

Case 1 : Using Calendar Year alone for wage prediction

The confusion matrix for this case is shown in Tab. XI.

TABLE X: Confusion matrix for wage prediction using Education.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	3	46
Actual C_1	12	29

TABLE XI: Confusion matrix for wage prediction using Calendar Year.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	49	0
Actual C_1	41	0

Case 4 : Using the three attributes for wage prediction

TABLE XII: Confusion matrix for wage prediction using three attributes.

Total N =90	Prediction of C_0	Prediction of C_1
Actual C_0	39	10
Actual C_1	19	22

The confusion matrix for this case is shown in Tab. XII. We can observe that the misclassification rate is the lowest for the last Case where all the three attributes were used. Also in cases where we have used calendar year and education as the attributes we have a very biased classification. Another thing to note is that here we have only done a two class classification and not predicted the wage. For more accurate predictions we have to increase the number of classes but even that will not be a better classifier than a classifier using three attributes. Thus we can't use any one of the attribute for determining the wage. Considering more and more uncorrelated attributes would improve the wage prediction.

APPENDIX

Maximum Likelihood Estimation

The expression for the normal denisty is

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(\frac{-1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)\right) \tag{7}$$

and that of the log likelihood can be given as

$$log\ \ likelihood = \sum\left(-\frac{d}{2}log2\pi - \frac{-1}{2}log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)\right) \tag{8}$$

Differentiating the log likelihood with respect to μ_i and equating it to zero we obtain

$$\sum \Sigma^{-1}(x-\mu_i) = 0 \tag{9}$$

Thus

$$\hat{\mu} = \frac{1}{n}\sum x_i \tag{10}$$

Assuming Σ to be arbitrary we otain the MLE estimate of Σ to be

$$\hat{\Sigma} = \frac{1}{n}\sum(x_i-\mu)(x_i-\mu)^t \tag{11}$$

If we assume Σ to be a diagonal matrix with diagonal elements 'a' and 'b' then differentiating the log likelihood with a and b we would obtain

$$\sum \left(\frac{-2b^2(x-\mu)_2^2}{4a^2b^2} - \frac{1}{2a} \right) = 0 \tag{12}$$

$$\sum \left(\frac{-2a^2(x-\mu)_1^2}{4a^2b^2} - \frac{1}{2b} \right) = 0 \tag{13}$$

Which would lead to

$$a = \frac{1}{n} \sum (x-\mu)_2^2 \tag{14}$$

$$b = \frac{1}{n} \sum (x-\mu)_1^2 \tag{15}$$

The subscript signifies the element number in the vector $(x-\mu)$.
Here we can observe that the first diagonal element sums over the second element of $(x-\mu)$ which is due to the fact that in Σ^{-1} the diagonal elements order is reversed. Along the same lines the Exp. 3 is obtained.
If we assume that 'a' and 'b' are equal the we would obtain

$$a = \frac{1}{2n} \sum (x-\mu)_2^2 + (x-\mu)_1^2 \tag{16}$$

which can also be represented as

$$a = \frac{1}{2n} \sum (x-\mu)^t (x-\mu) \tag{17}$$

Polynomail Regression [1]

When we have the given data as tuples of form (x, y) and we want to fit a polynomial to this data we employ the following method: Assuming

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

We minimize mean squared error to obtain the coefficients a_i.

$$E(a_0, a_1, \ldots a_n) = \sum \left(y_k - a_0 - a_1 x - a_2 x^2 - \cdots - a_n x^n \right)^2$$

Differentiating this with respect to the coefficients and equating them to zero we obtain:

$$\frac{\partial E}{\partial a_i} = -2 \sum x_k^i \left(y_k - a_0 - a_1 x - a_2 x^2 - \cdots - a_n x^n \right) = 0$$

This can be represented as a matrix equation of the form

$$\begin{bmatrix} n & \sum x_k & \sum x_k^2 & \ldots & \sum x_k^n \\ \sum x_k & \sum x_k^2 & \ldots & \sum x_k^n & \sum x_k^{n+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sum x_k^n & \sum x_k^{n+2} & \ldots & \sum x_k^{2n-1} & \sum x_k^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum y_k \\ \sum y_k x_k \\ \vdots \\ \sum y_k x_k^n \end{bmatrix} \tag{18}$$

and can be solved to obtain the polynomial coefficents.

[1] https://www.math.tamu.edu/~glahodny/Math442/Curve%20Fitting.pdf