IITM-CS5691 : Pattern Recognition and Machine Learning — Release Date: August 30, 2023
Assignment 1 — Due Date : September 14, 2023, 23:59

Roll No: CS22Z121 — Name: Sandeep Kumar Suresh
Collaborators (if any):
References/sources : Duda and hart , Bishop Reference books ,Stack Exchange

Solution:

1.

(a) Given $\Sigma = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$fx(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(\frac{-1}{2}\frac{1}{ad-bc}[x_1 x_2]\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$$
$$= \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}\frac{1}{ad-bc}\left[dx_1^2 - cx_1x_2 - bx_1x_2 + ax_2^2\right]\right)$$
$$= \frac{1}{\sqrt{2\pi|\varepsilon|}} \exp\left(-\frac{1}{2}\cdot\frac{1}{\left(a-\frac{bc}{d}\right)}\left[x_1^2 - \frac{2bx_1x_2}{d} + \frac{ax_2^2}{d}\right]\right)$$
$$= \frac{1}{\sqrt{2\pi|\varepsilon|}} \exp\left(-\frac{1}{2}\frac{1}{\left(a-\frac{bc}{d}\right)}\left[x_1^2 - \frac{2b}{d}x_1x_2 + \left(\frac{bx_2}{d}\right)^2 - \left(\frac{bx_2}{d}\right)^2 + \frac{ax_2^2}{d}\right]\right)$$
$$= \frac{1}{\sqrt{2\pi|\varepsilon|}} \exp\left(\frac{-1}{2\left(a-\frac{bc}{d}\right)}\left[\left(x_1 - \frac{bx_2}{d}\right)^2 + \left(\frac{ad-b^2}{d}\right)x_2^2\right]\right)$$
$$= \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(\frac{-1}{2\left(a-\frac{bc}{d}\right)}\left(x_1 - \frac{bx_2}{d}\right)^2\right)\frac{1}{\sqrt{2\pi d}}\exp\left(\frac{-1}{2d}x_2^2\right)$$
$$= N\left(\frac{bx_2}{d}, a - \frac{bc}{d}\right) \quad N(0, d)$$

Similarly

(b)

$$g(x) = x_1^2 + x_2^2 + x_1x_2$$

Linear approximation around x

$$f(y) \approx f(x) + \nabla g(x)^\top (y - x)$$
$$\nabla f(x) = \begin{bmatrix} 2x_1 + x_2 \\ 2x_2 + x_1 \end{bmatrix}$$
$$\nabla f(v) = \begin{bmatrix} 811 \\ 13 \end{bmatrix}$$
$$f(y) = 3^2 + 5^2 + 5 \times 3 = 9 + 25 + 15 = 49$$
$$f(y) = 49 + \begin{bmatrix} 11 \\ 13 \end{bmatrix}^\top \begin{bmatrix} y_1 - 3 \\ y_2 - 5 \end{bmatrix}$$
$$f(y) = 49 + \begin{bmatrix} 11 & 13 \end{bmatrix}\begin{bmatrix} y_1 - 3 \\ y_2 - 5 \end{bmatrix}$$
$$f(y) = 49 + 11(y_1 - 3) + 13(y_2 - 5)$$

(c) The statement which are true are (i) and the vice versa is not true

Solution: 2 .
(a) Logarithm is a monotonically increasing function and θ that maximizes the log liklihood also maximise the liklihood. To argue that the stationary points obtained are the indeed the global maxima or minima , we need to show that

- Log-Liklihood is concave in μ

  Log likelihood function is given by

  $L\left(\mu, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$
  $\frac{dL}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right) = 0$

  Taking the double derivative

  $\frac{d^2L}{d\mu^2} = \frac{-n}{\sigma^2} < 0$

  Which implies it is a global maxima.

- Log Likelihood is concave in σ^2 maximum liklihood Estination of ε

  $\frac{dL}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 = 0$

  Taking the double derivative

  $\frac{d^2L}{d(\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\nu^2)^3}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 < 0$

The second derivative of the function will be less than zero , which implies that the likelihood function will be maximum.
To argue that this the global maximum , since there is only one term in the first derivative .

b) The mean of a of MLE is given is $\frac{1}{N}\sum_{i=1}^{N} x_i$

Bias of the mean

$$E[\bar{x}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x]$$
$$= \frac{1}{N} \times N \times E[x] = E[x] = \mu$$

Here the expected mean is equal to the true mean.

Let $\tilde{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \vec{x})^2$. We want to show $E\left[\sigma^2 \right| \neq \sigma^2$

$$E\left[\dot{\sigma}^2\right) = E\left[\tfrac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 \right| = \tfrac{1}{N}E\Big|\sum_{n=1}^{N}\left(x_n^2 - 2x_n\bar{x} + \bar{x}^2\right)\Big]$$
$$= \tfrac{1}{N}E\left[\sum_{n=1}^{N}x_n^2 - \sum_{n=1}^{N}2x_n\bar{x} + \sum_{n=1}^{N}\bar{x}^2\right]$$

Using the fact that $\sum_{n=1}^{N}x_n = N\bar{x}$ and $\sum_{n=1}^{N}\vec{x}^2 = Nx^2$,

$$\tfrac{1}{N}E\left[\sum_{n=1}^{N}x_n^2 - \sum_{n=1}^{N}2x_n\bar{x} + \sum_{n=1}^{N}\vec{x}^2\right] = \tfrac{1}{N}E\left[\sum_{n=1}^{N}x_n^2 - 2N\bar{x}^2 + N\bar{x}^2\right]$$
$$= \tfrac{1}{N}E\left[\sum_{n=1}^{N}x_n^3 - Nx^2\right] = \tfrac{1}{N}E\left[\sum_{n=1}^{N}x_n^2\right] - E\left[\bar{x}^2\right] = \tfrac{1}{N}\sum_{n=1}^{N}E\left[x_n^2\right] - E\left[\bar{x}^2\right]$$
$$= E\left[x_n^2\right] - E\left[\bar{x}^2\right]$$

From the def of variance $\sigma_x^2 = E\left[x^2\right] - E[x]^2$

$$E\left[x_n^2\right] - E\left[\tilde{x}^2\right] = \sigma_x^2 + E\left[x_n\right]^2 - \sigma_z^2 - E\left[x_n\right]^2 = \sigma_x^2 - \sigma_2^2 = \sigma_x^2 - \mathrm{Var}(\bar{x})$$
$$= \sigma_x^2 - \mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right) = \sigma_x^2 - \left(\frac{1}{N}\right)^2\mathrm{Var}\left(\sum_{n=1}^{N}x_n\right)$$

$$\sigma_x^2\left(\frac{1}{N}\right)^2\mathrm{Var}\left(\sum_{n=1}^{N}x_n\right) = \sigma_x^2 - \left(\frac{1}{N}\right)^2 N\sigma_x^2 = \sigma_x^2 - \frac{1}{N}\sigma_x^2 = \frac{N-1}{N}\sigma_x^2$$

Therefore the variance of the MLE is unbiased

Solution:
3.
(a) Let y_i be the class labels.
$p(y_1) = \frac{5}{14}$ $p(y_2) = \frac{4}{14}$ $p(y_3) = \frac{5}{14}$
$\mu_1 = -2.1$ $\mu_2 = 0.5$ $\mu_3 = 1.86$

$$P(x = 1 \mid y_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu 1)^2}{2}}$$
$$P(x = 2 \mid y_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu 2)^2}{2}}$$
$$P(x = 3 \mid y_4) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu 3)^2}{2}}$$

Let η_1, η_2 and η_3 be the posterior distribution

$$\eta_1 = \frac{e^{\frac{(x-\mu_1)^2}{2}}}{e^{-\frac{(x-\mu_1)^2}{2}} + e^{-\frac{(x-\mu_2)^2}{2}} + e^{-\frac{(x-\mu_3)^2}{2}}}$$
$$\eta_2 = \frac{e^{\frac{(x-\mu_1)^2}{2}}}{e^{-\frac{(x-\mu_1)^2}{2}} + e^{-\frac{(x-\mu_2)^2}{2}} + e^{-\frac{(x-\mu_3)^2}{2}}}$$
$$\eta_3 = \frac{e^{\frac{(x-\mu_1)^2}{2}}}{e^{-\frac{(x-\mu_1)^2}{2}} + e^{-\frac{(x-\mu_2)^2}{2}} + e^{-\frac{(x-\mu_3)^2}{2}}}$$

To find the bayesian decision boundary we need to equate
$\eta_1 = \eta_2$
$\eta_2 = \eta_3$

Given the Loss matrix

$$L = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$
$$h(x) = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} \eta_2 + 2\eta_3 \\ \eta_1 + \eta_3 \\ 2\eta_1 + \eta_2 \end{bmatrix}$$

$$\hat{y} = \arg\min P(y = c \mid x)$$

(b)

We need to minimise the expected loss for a vector x and need to assign x to a class c , such that the expected loss is minimised .

$$j = \arg\min_l \sum_R L_{kl} p\left(C_k \mid x\right)$$
$$j = \arg\min_l \sum_k L_{kl} p\left(c_k \mid x\right)$$

$$\text{choose} \begin{cases} \text{class } j, \text{ if } \min_i \sum_k L_{kl} P\left(c_k \mid x\right) < \psi \\ \text{reject, otherwise} \end{cases}$$

Solution: 4.

(a) Yes , Decision Boundaries can be discontinuous .

(b)

$$C_{\text{train}} = \begin{bmatrix} 100 & 10 \\ 30 & 120 \end{bmatrix} \quad C_{\text{lest}} = \begin{bmatrix} 90 & 45 \\ 30 & 8.5 \end{bmatrix}$$

Given the Confusion Matrix, we need to compute the posterior
Assuming the P(Positive) = P(Negative) = 0.5

$$P(\text{ data } \mid \text{ positive }) = \frac{TP}{TP+FN} = \frac{100}{100+20} = 0.83.$$

$$P(\text{ data } \mid \text{ negative }) = \frac{TN}{TN+FP} = \frac{120}{120+10} = 0.92.$$

$$\begin{aligned} P(\text{positive } \mid \text{data}) &\propto 0.83 \times 0.5 \\ &= 0.475 \end{aligned}$$

$$\begin{aligned} P(\text{negative } \mid \text{data}) &\propto 0.92 \times 0.5 \\ &= 0.46 \end{aligned}$$

Let $\eta_1 = 0.415 \quad \eta_2 = 0.46$

$$h_{i_{train}} = \begin{bmatrix} p & q \\ r & s \end{bmatrix} \begin{bmatrix} 0.415 \\ 0.46 \end{bmatrix} = \begin{bmatrix} 0.415p + 0.46q \\ 0.415x + 0.465 \end{bmatrix}$$

Similarly for C_{test}

$$\eta_1 = 0.37s \qquad \eta_2 = 0.32$$

$$h_{i_{test}} = \begin{bmatrix} p & q \\ h & s \end{bmatrix} \begin{bmatrix} 0.375 \\ 0.32 \end{bmatrix} = \begin{bmatrix} 0.375p + 0.32q \\ 0.375x + 0.32s \end{bmatrix}$$

The data belong to new class where hi is minimized , where the expected loss is minimized .

(c)

(i) Consider the prior probability

$P(ill) = 0.5$

$P(healthy) = 0.5$

Constructing a Likelihood Table for '+'

ouput	N	C	R
+ve (iil)	2/3	2/3	2/3
-ve (healthy)	1/3	1/3	1/3

for $(d7 : N =, C = +, R =)$ using a Naive Bayes classifier

$$P(ill|d_7) \propto P(ill) \cdot (1/3) \cdot (2/3) \cdot (1/3)$$

$$P(ill|d_7) \propto (1/2) \cdot (1/3) \cdot (2/3) \cdot (1/3) \propto 1/27$$

Similarly

$$P(healthy|d_7) \propto (1/2) \cdot (2/3) \cdot (1/3) \cdot (2/3) \propto 2/27$$

It is given that

$$P(healthy|d_7) > P(ill|d_7)$$

(ii) Naive Bayes Assumption is that each event is independent of the other and every event contributes to the outcome.The name bayes Formula is given by the following

$$P(C_k|x_1, x_2, \dots, x_n) \propto P(C_k) \cdot \prod_{i=1}^{n} P(x_i|C_k)$$

(iv) For the estimating the class conditionals I have used the Bernoulli Distribution.

Solution: 5 .

(a)

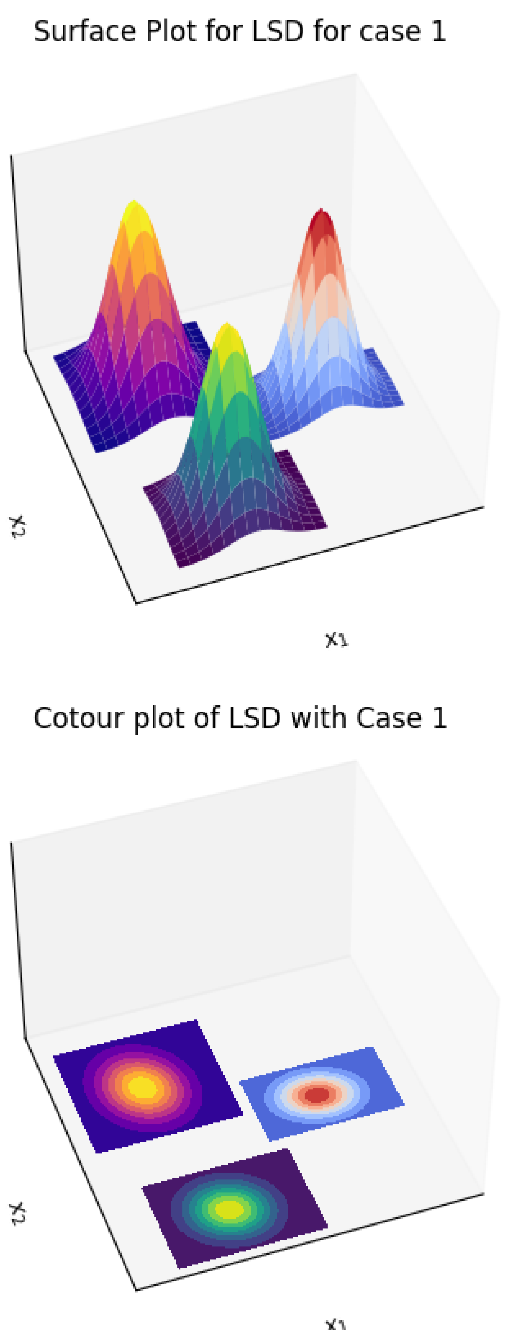

Figure 1: Linearly Seperable Data Case1

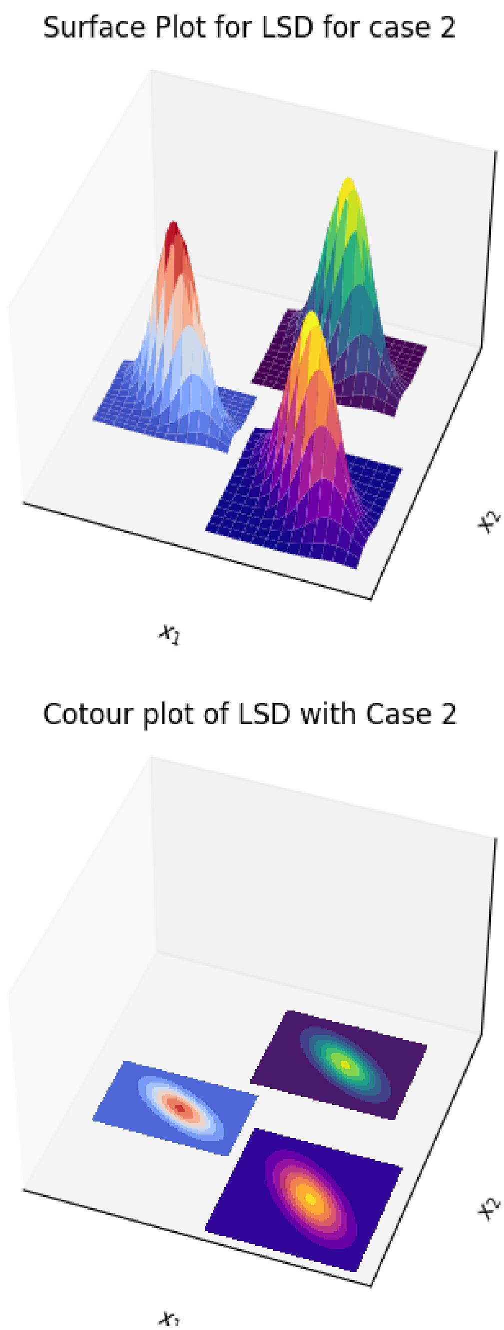

Figure 2: Linearly Seperable Data Case2

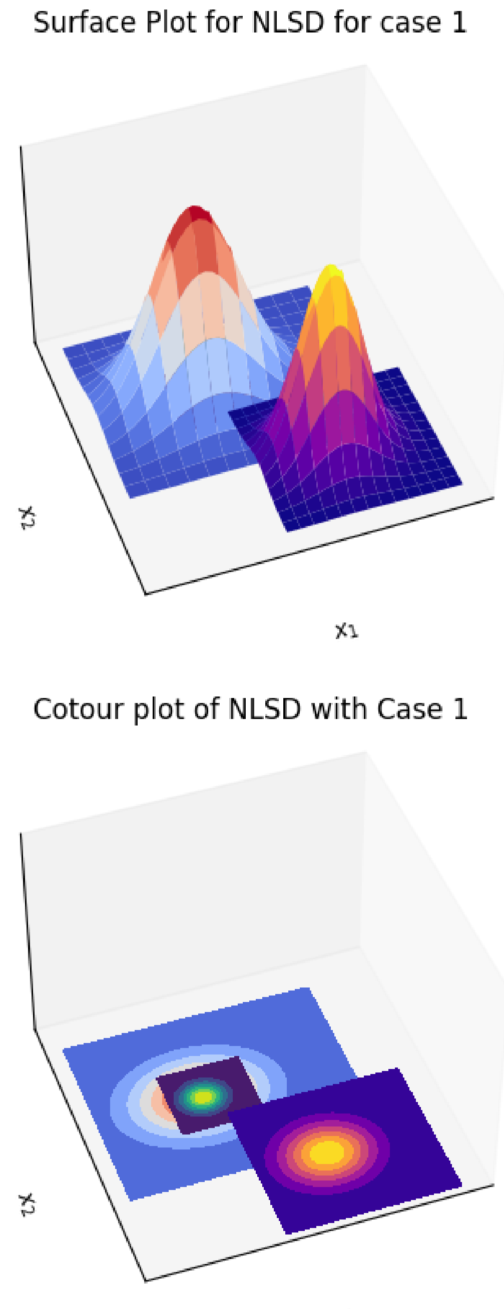

Figure 3: Non Linearly Seperable Data Case 1

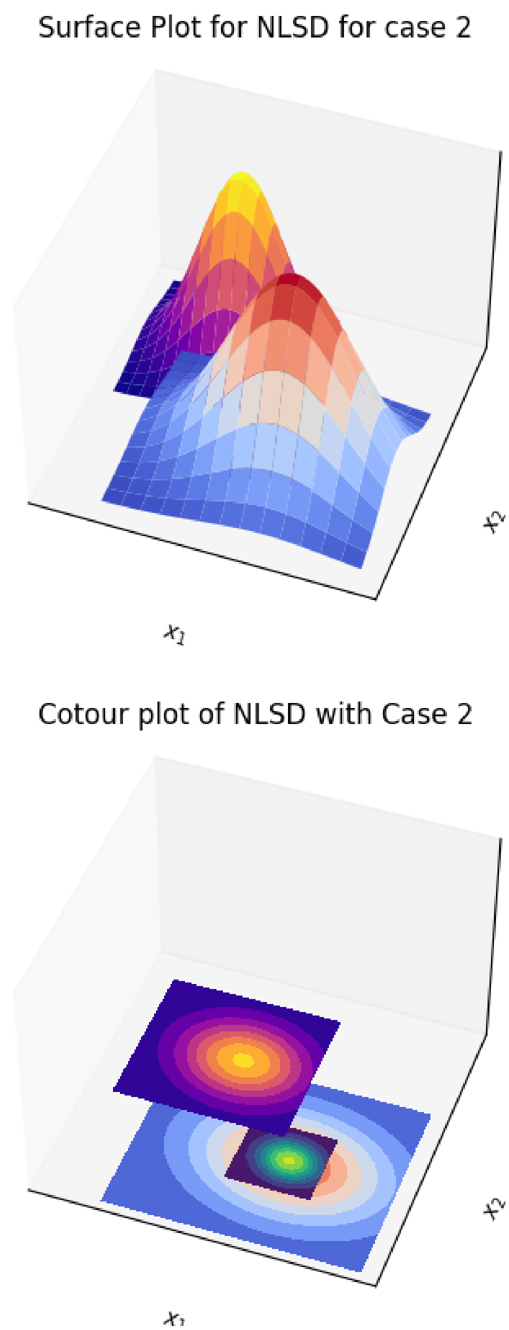

Figure 4: Non-Linearly Seperable Data Case2

(b)

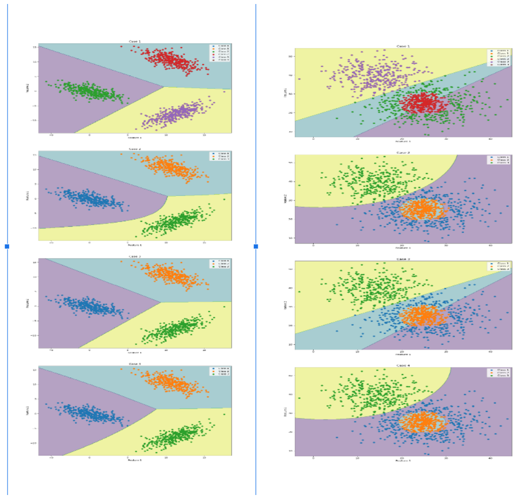

Figure 5: Decision Boundary for Various case . On the Left is Linearly Seperable Data and On the right is Non-Linearly Seperable Data. Top to Bottom represents Case 1 to Case 4

(d)

Case 2 can be used to modelled when we can assume that the overall covariance is similar for each classes.We can see that in Figure 5 , covariance matrix is unable to give a decision boundary for non linearly seperable data in case 2 . Therefore when the model becomes complex we cannot rely on the shared covariance matrix.
But if we look at the Case 2 of Linearly Seperable Data in Figure 5 , it makes no difference as the decision boundaries are clearly defined .

Therefore we can say that the Case 2 should be assumed when the model complexity is simpler and in general Case 1 should be used for better modelling of the data.