Gary Chen
804361027

Homework 6

Problem 1

The Gaussian Discriminant Analysis (GDA) models the class conditional distribution as multivariate Gaussian, i.e, $P(x|y) \sim N(\mu_y, \Sigma)$. Suppose we want to enforce the Naive Bayes (NB) assumption, i.e. $P(x_i|y, x_j) = P(x_i|y), \forall j \neq i$, to GDA. Show that all off diagonal elements of Σ equals to 0: $\Sigma_{i,j} = 0, \forall i \neq j$ with the NB assumption.

$$\begin{aligned}\Sigma_{i,j} = \text{cov}(x_i, x_j) &= E[(x_i - \mu_i)(x_j - \mu_j)] \\ &= E[x_i x_j] - E[x_i \mu_j] - E[x_j \mu_i] + E[\mu_i \mu_j]\end{aligned}$$

Where $E[x_i x_j] = E[x_i]E[x_j]$ because x_i, x_j are conditionally independent

$$\begin{aligned}&= E[x_i]E[x_j] - \mu_j E[x_i] - \mu_i E[x_j] + \mu_i \mu_j \\ &= \mu_i \mu_j - \mu_i \mu_j - \mu_i \mu_j + \mu_i \mu_j \\ &= 0\end{aligned}$$

Problem 2

Consider the classification problem for two classes, C_0 and C_1. In the generative approach, we model the class-conditional distribution $P(x|C_0)$ and $P(x|C_1)$, as well as the class priors $P(C_0)$ and $P(C_1)$.The posterior probability for class C_0 can be written as

$$P(x|C_0) = \frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1) + P(x|C_1)P(C_1)} \tag{1}$$

a) **Show that $P(C_0|x) = \sigma(a)$ where $\sigma(a)$ is the sigmoid function defined by**

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

Find a in terms of $P(x|C_0)$, $P(x|C_1)$, $P(C_0)$ and $P(C_1)$.

$$
\begin{aligned}
P(x|C_0) &= \frac{1}{\frac{P(x|C_1)P(C_1)+P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}} \\
\frac{1}{1+\exp(-a)} &= \frac{1}{1+\frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}} \\
\implies \exp(-a) &= \frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)} \\
a &= \ln\left(\frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}\right) \\
&= \ln(P(C_0)) + \ln(P(x|C_0)) - \ln(P(C_1)) - \ln(P(x|C_1))
\end{aligned}
$$

b) **In GDA model, we have the class conditional distribution as follows**

$$
P(x|C_i) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right)
$$

Suppose we are able to find the maximum likelihood estimation of $\mu_0, \mu_1, \Sigma, P(C_0)$, and $P(C_1)$. Show that $a = w^Tx + b$ for some w and b. Find w and b in terms of $\mu_0, \mu_1, \Sigma, P(C_0)$, and $P(C_1)$. This shows that the decision boundary is linear.

From part a)

$$
\begin{aligned}
a =& \ln(P(C_0)) - \ln(P(C_1)) + \ln\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)\right) \\
&- \ln\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)\right) \\
=& \ln(P(C_0)) - \ln(P(C_1)) + \frac{1}{2}\left[(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right] \\
=& \frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 - x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}\mu_1) \\
&+ \ln\left(\frac{P(C_0)}{P(C_1)}\right) \\
=& \frac{1}{2}(x^T\Sigma^{-1}(\mu_1-\mu_0) + (\mu_1-\mu_0)^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 - \mu_1\Sigma^{-1}\mu_1) + \ln\left(\frac{P(C_0)}{P(C_1)}\right)
\end{aligned}
$$

Where the matrix Σ^{-1} is symmetric, $\Sigma^{-1} = \Sigma^{-T}$, and $a = w^Tx + b$

$$
w^Tx + b = (\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1\Sigma^{-1}\mu_1) + \ln\left(\frac{P(C_0)}{P(C_1)}\right)
$$

We find that $w = (\mu_1 - \mu_0)\Sigma^{-1}$ and $b = \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1\Sigma^{-1}\mu_1) + \ln(P(C_0)/P(C_1))$.

c) **Now let us consider two classes that have difference covariance matrix as follows**

$$P(x|C_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right)$$

Suppose we are able to find the maximum likelihood estimation of $\mu_0, \mu_1, \Sigma_0, \Sigma_1, P(C_0)$, and $P(C_1)$. Show that $a = x^TAx + w^Tx + b$ for some A, w and b. Find w and b in terms of $\mu_0, \mu_1, \Sigma_0, \Sigma_1, P(C_0)$, and $P(C_1)$. This shows that the decision boundary is quadratic.

$$\begin{aligned}
a =& \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \ln\left(\frac{1}{(2\pi)^{n/2}|\Sigma_0|^{1/2}}\right) - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) \\
&- \ln\left(\frac{1}{(2\pi)^{n/2}|\Sigma_1|^{1/2}}\right) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) \\
=& \frac{1}{2}[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)] + \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) \\
=& \frac{1}{2}[x^T\Sigma_1^{-1}x - \mu_1^T\Sigma_1^{-1}x - x^T\Sigma_1^{-1}\mu_1 + \mu_1\Sigma_1^{-1}\mu_1 - x^T\Sigma_0^{-1}x + \mu_0\Sigma_0^{-1}x + x^T\Sigma_0^{-1}\mu_0 - \mu_0^T\Sigma_0^{-1}\mu_0] \\
&+ \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) \\
=& \frac{1}{2}[x^T(\Sigma_1^{-1} - \Sigma_0^{-1})x - 2\mu_1^T\Sigma_1^{-1}x + 2\mu_0^T\Sigma_0^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0] \\
&+ \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) \\
=& \frac{1}{2}x^T(\Sigma_1^{-1} - \Sigma_0^{-1})x + (\mu_0^T\Sigma_0^{-1} - \mu_1^T\Sigma_1^{-1})x + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0) \\
&+ \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)
\end{aligned}$$

Where we find that $A = (\Sigma_1^{-1} - \Sigma_0^{-1})/2$, $w = (\mu_0^T\Sigma_0^{-1} - \mu_1^T\Sigma_1^{-1})^T$, and $b = \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0) + \ln(P(C_0)/P(C_1)) + \frac{1}{2}\ln(|\Sigma_1|/|\Sigma_0|)$

Problem 3

We are given a training set $\{(x^{(i)}, y^{(i)}); i = \{1, \ldots, m\}\}$ where $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \{0, 1\}$. We consider the Gaussian Discriminant Analysis (GDA) model, which models $P(x|y)$ using multivariate Gaussian. Writing out the model, we have:

$$P(y = 1) = \phi = 1 - P(y = 0)$$

$$P(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)$$
$$P(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

The log-likelihood of the data is given by:

$$L(\phi, \mu_0, \mu_1, \Sigma) = \ln\left(P(x^{(i)}, \ldots, x^{(m)}, y^{(i)}, \ldots, y^{(m)})\right) = \ln\left(\prod_{i=1}^{m} P(x^{(i)}, y^{(i)})P(y^{(i)})\right)$$

In this exercise, we want to maximize $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to ϕ, Σ.

a) **Write down the explicit expression for $P(x^{(i)}, \ldots, x^{(m)}, y^{(i)}, \ldots, y^{(m)})$ and $L(\phi, \mu_0, \mu_1, \Sigma)$**

$$\begin{aligned}
&P(x^{(i)}, \ldots, x^{(m)}, y^{(i)}, \ldots, y^{(m)}) \\
=&\prod_{i=1}^{m}\left[\frac{1-\phi}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right]^{1-y^{(i)}} \\
&\times\left[\frac{\phi}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right]^{y^{(i)}}
\end{aligned}$$

$$\begin{aligned}
&L(\phi, \mu_0, \mu_1, \Sigma) \\
=&\sum_{i=1}^{m}\left\{(1-y^{(i)})\left[\ln(1-\phi) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right]\right. \\
&\left.+\, y^{(i)}\left[\ln(\phi) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right]\right\}
\end{aligned}$$

b) **Find the maximum likelihood estimate for ϕ. How do you know such ϕ is the "best" but not the "worst"?**

$$
\begin{aligned}
\frac{\partial L}{\partial \phi} = 0 &= \sum_{i=1}^{m} \left\{ \frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi} \right\} \\
0 &= \sum_{i=1}^{m} \frac{y^{(i)} - \phi}{\phi(1-\phi)} \\
0 &= \sum_{i=1}^{m} (y^{(i)} - \phi) \\
m\phi &= \sum_{i=1}^{m} y^{(i)} \\
\phi &= \frac{\sum_{i=1}^{m} y^{(i)}}{m}
\end{aligned}
$$

To find if this is the maximum, we take the second derivative:

$$
\frac{\partial^2 L}{\partial \phi^2} = \sum_{i=1}^{m} \left\{ -\frac{y^{(i)}}{\phi^2} - \frac{1-y^{(i)}}{(1-\phi)^2} \right\}
$$

We can see that $\phi^2 \geq 0$, $(1-\phi)^2 \geq 0$, $y^{(i)} \geq 0$, and $(1-y^{(i)}) \geq 0$. This makes tit so that the full expression is ≤ 0, giving the maximum.

c) **Find the maximum likelihood estimate for μ_0. How do you know such μ_0 is the "best" but not the "worst"?**

$$
\begin{aligned}
\frac{\partial L}{\partial \mu_0} = 0 &= \frac{1}{2} \sum_{i=1}^{m} (1-y^{(i)}) \frac{\partial}{\partial \mu_0} \left[-x^{(i)^T} \Sigma^{-1} x^{(i)} + \mu_0^T \Sigma^{-1} x^{(i)} + x^{(i)^T} \Sigma^{-1} \mu_0 - \mu_0^T \Sigma^{-1} \mu_0 \right] \\
0 &= \frac{1}{2} \sum_{i=1}^{m} (1-y^{(i)}) \frac{\partial}{\partial \mu_0} [-x^{(i)^T} \Sigma^{-1} x^{(i)} + 2\mu_0^T \Sigma^{-1} x^{(i)} - \mu_0^T \Sigma^{-1} \mu_0] \\
0 &= \frac{1}{2} \sum_{i=1}^{m} (1-y^{(i)}) [2\Sigma^{-1} x^{(i)} - 2\Sigma^{-1} \mu_0] \\
\sum_{i=1}^{m} (1-y^{(i)}) \mu_0 &= \sum_{i=1}^{m} (1-y^{(i)}) x^{(i)} \\
\mu_0 &= \frac{\sum_{i=1}^{m} (1-y^{(i)}) x^{(i)}}{\sum_{i=1}^{m} (1-y^{(i)})}
\end{aligned}
$$

To find if this is the maximum, we find the second derivative:

$$\frac{\partial^2 L}{\partial \mu_0^2} = \frac{\partial^2}{\partial \mu_0^2} \Sigma^{-1} \sum_{i=1}^{m} (1 - y^{(i)})(x^{(i)} - \mu_0)$$
$$= -\Sigma^{-1} \sum_{i=1}^{m} (1 - y^{(i)})$$

Where see that $1-y^{(i)} \geq 0$ and Σ^{-1} is positive semi-definite, hence the second derivative is negative semi-definite, giving the maximum.

Problem 4

a) **Visualization. Is the data linearly separable?**

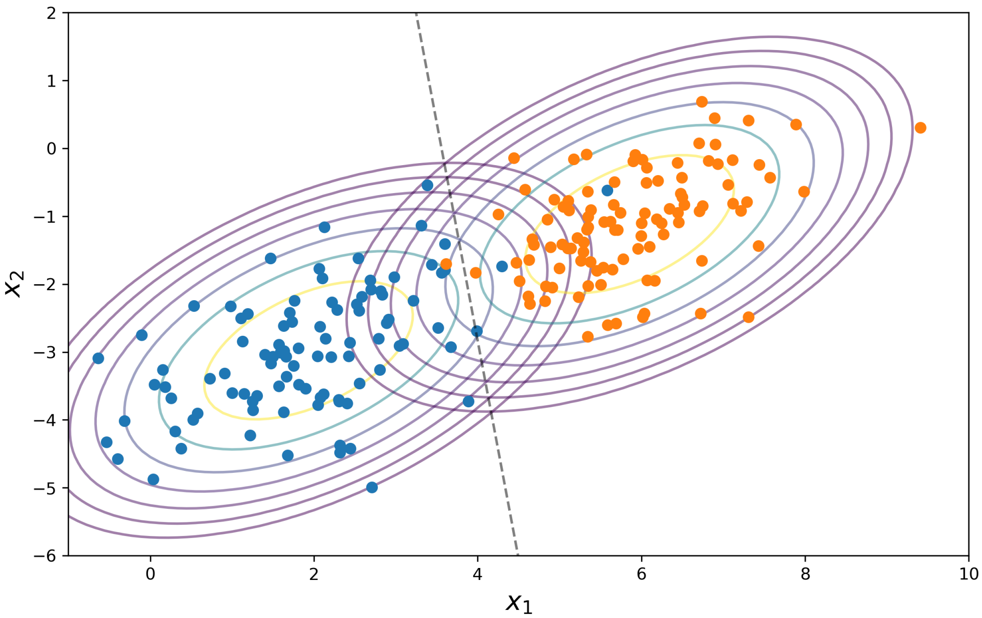

Figure 1: Labeled data with decision boundary and contours.

We can clearly see that the data is not linearly separable.

b) **Find the maximum likelihood estimate of the parameters $P(y = 0)$, μ_0, μ_1, and Σ given this data set.**

$$P(y = 0) = 0.485, \quad \mu_0 = \begin{bmatrix} 1.935 \\ -2.975 \end{bmatrix}, \quad \mu_1 = \begin{bmatrix} 5.856 \\ -1.118 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1.119 & 0.452 \\ 0.452 & 0.714 \end{bmatrix}$$

c) **Find the decision boundary parameterized by $w^T x + b = 0$. Report w, b and plot the decision boundary on the same plot.**

See Figure 1

$$w = \begin{bmatrix} -3.298 - 0.514 \end{bmatrix}, \qquad b = 11.736$$

d) **Visualize your results by plotting the contour of the two distributions $P(x|y = 0)$ and $P(x|y = 1)$. Your decision boundary should pass through points where the two distribution have equal probabilities. Explain why?**

See Figure 1

The decision boundary passes through the point where the classification of a point would be indeterminant, meaning, where the probabilities of being in any class are equal.

Problem 5

Suppose we have a data set $\{x_1, \ldots, x_N\}$ where $x_n \in \mathbb{R}^M$ and our goal is to partition the data set in to K clusters with μ_k representing the center of the k-th cluster. Recall that in K-means clustering we are attempting to minimize an objective function defined as follows:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|_2^2$$

where $r_{nk} \in \{0, 1\}$ and $r_{nk} = 1$ only if x_n is assigned to cluster k.

a) **What is the minimum value of the objective function when $K = N$ (the number of clusters equals to the number of samples)?**

When $K = N$, the centers of clusters are exactly each sample, thus, when $n = k = i$, we see that $x_i = \mu_i$. Then considering the objective function, $x_i - \mu_i = 0$ giving us $J = 0$.

b) **Adding a regularization term, the objective function now becomes:**

$$J = \sum_{k=1}^{K} \left(\lambda \mu_k + \sum_{n=1}^{N} r_{nk} \|x_n - \mu_k\|_2^2 \right)$$

Consider the optimization of μ_k with all r_{nk} known. Find the optimal μ_k for

$$\text{argmin}_{\mu_k} \lambda \|\mu_k\|_2^2 + \sum_{n=1}^{N} r_{nk} \|x_n - \mu_k\|_2^2$$

We find the derivative of J and set it to equal zero.

$$\begin{aligned}
\frac{\partial J}{\partial \mu_k} = 0 = 2\lambda\mu_k - 2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) \\
= \lambda\mu_k - \sum_{n=1}^{N} r_{nk}x_n + \mu_k\sum_{n=1}^{N} r_{nk} \\
\mu_k = \frac{\sum_{n=1}^{N} x_n r_{nk}}{\lambda + \sum_{n=1}^{N} r_{nk}}
\end{aligned}$$

To make sure this is a minimum, we take the second derivative:

$$\frac{\partial^2 J}{\partial \mu_k^2} = 2\lambda + \sum_{n=1}^{N} 2r_{nk} \geq 0$$

Where for positive λ, the objective function is convex.