Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION 1

Assignment Number: 3
Student Name: Nikhil Mittal
Roll Number: 17111056
Date: November 15, 2017

To show :

$$\theta^{MLE} \in \underset{\theta \in \Theta}{\arg\max}\; Q_{\theta^{MLE}}(\theta)$$

Given, MLE model

$$\boldsymbol{\theta}^{\text{MLE}} \in \Theta \text{ such that } \boldsymbol{\theta}^{\text{MLE}} \in \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\; \mathbb{P}\left[X \mid \boldsymbol{\theta}\right]$$

So,

$$\underset{\theta \in \Theta}{max}\; \mathbb{P}[X|\theta] = \mathbb{P}[X|\theta^{MLE}]$$

Taking log both sides :

$$\underset{\theta \in \Theta}{max}\; \log \mathbb{P}[X|\theta] = \log \mathbb{P}[X|\theta^{MLE}] \qquad -(1)$$

From Lecture 16,Slide 38

$$\log \mathbb{P}[X|\theta] \geq Q_{\theta^t}(\theta)\; \forall\; \theta^t \in \Theta$$

Since the above is true for any θ hence will be true for θ^{MLE} also.

$$\log \mathbb{P}[X|\theta] \geq Q_{\theta^{MLE}}(\theta)$$

Hence,

$$\max_{\theta \in \Theta}\; \log \mathbb{P}[X|\theta] \geq \max_{\theta \in \Theta} Q_{\theta^{MLE}}(\theta)$$

Using (1), we get

$$\log \mathbb{P}[X|\theta^{MLE}] \geq \max_{\theta \in \Theta} Q_{\theta^{MLE}}(\theta) \qquad -(2)$$

From Lecture16,Slide 42
We proved in class,

$$\log \mathbb{P}[X|\theta^0] = \; Q_{\theta^0}(\theta^0)$$

Hence this will be true for θ^{MLE} also, since $\theta^{MLE} \in \Theta$

$$\implies \log \mathbb{P}[X|\theta^{MLE}] = \; Q_{\theta^{MLE}}(\theta^{MLE})$$

Using above result and (2), we get

$$Q_{\theta^{MLE}}(\theta^{MLE}) \geq \max_{\theta \in \Theta} Q_{\theta^{MLE}}(\theta)$$

Hence,

$$\theta^{MLE} \in \underset{\theta \in \Theta}{\arg\max}\; Q_{\theta^{MLE}}(\theta)$$

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION 2

Assignment Number: 3
Student Name: Nikhil Mittal
Roll Number: 17111056
Date: November 15, 2017

An n-partition of a set $\mathcal{X}$ is a collection of n subsets $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ such that each $\mathcal{X}_i \subseteq \mathcal{X}$ and

- $\mathcal{X}_i \cap \mathcal{X}_j = \phi$ if $i \neq j$
- $\bigcup_{i=1}^{n} \mathcal{X}_i = \mathcal{X}$

1. To show: for a linear piecewise function $f(x)$, $c \cdot f(x)$ is also piecewise linear.

   A piecewise linear function $f : \mathbb{R}^d \to \mathbb{R}$ with $n > 0$ "pieces" is indexed by an n-partition $\{\Omega_1, \ldots, \Omega_n\}$ of $\mathbb{R}^d$ and n linear models $\mathbf{w}^1, \ldots, \mathbf{w}^n$ such that for any $\mathbf{x} \in \mathbb{R}^d$.

$$f(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x} \in \Omega_i\} \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle,$$

$$\begin{aligned} g(\mathbf{x}) &= c \cdot f(\mathbf{x}) && (1)\\ &= c \cdot \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x} \in \Omega_i\} \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle && (2)\\ &= \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x} \in \Omega_i\} \cdot c \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle && (3)\\ &= \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x} \in \Omega_i\} \cdot \left\langle c\mathbf{w}^i, \mathbf{x} \right\rangle && (4) \end{aligned}$$

   To define g from the definition of f, we only require to update all $\mathbf{w}^i$ to be $c\mathbf{w}^i$, $c\mathbf{w}^i$ is also a linear model. Hence this definition of g does not change the partition and the piecewise nature of the function.
   Hence $g(\mathbf{x}) = c \cdot f(\mathbf{x})$ is a piecewise function defined on same partition set with scaled model vectors.

2. To show, sum of 2 piecewise linear functions is also linear.

   Let,

$$f(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \Omega_i^f\right\} \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle,$$

   defined on an n-partition of $\mathbb{D}$ is a collection of n subsets $\left\{\Omega_1^f, \ldots, \Omega_n^f\right\}$ and set of model vectors $\left\{\mathbf{w}^1, \ldots \mathbf{w}^n\right\}$

   and

$$g(\mathbf{x}) = \sum_{i=1}^{m} \mathbb{I}\{\mathbf{x} \in \Omega_i^g\} \cdot \left\langle \mathbf{v}^i, \mathbf{x} \right\rangle,$$

defined on an n-partition of $\mathbb{D}$ is a collection of n subsets $\{\Omega_1^g, \ldots, \Omega_m^g\}$ and set of model vectors $\left\{\mathbf{v}^1, \ldots \mathbf{v}^m\right\}$

to define the sum of above function, $(f+g)(x) = f(x) + g(x)$, we define new partition set as: $\Omega = \{\Omega_{11}, \Omega_{12} \ldots \Omega_{1m} \ldots \Omega_{nm}\}$

where, $x \in \Omega_{ij}$ *iff* $x \in \Omega_i^f$ **and** $x \in \Omega_j^g$

Claim 3.1. Ω is a partition of $\mathbf{D}$

Proof. (a) To show : $\Omega_{ij} \cap \Omega_{pq} = \phi$ ***iff*** $i \neq p$ **OR** $j \neq m$

let, $\Omega_{ij} \cap \Omega_{pq} \neq \phi$ for some i, j, p, q

$$i.e., \mathbf{x} \in \Omega_{ij} \cap \Omega_{pq} \tag{5}$$

$$\implies \mathbf{x} \in \Omega_{ij} \tag{6}$$

$$\implies \mathbf{x} \in \Omega_i^f \text{ and } \mathbf{x} \in \Omega_j^g \tag{7}$$

$$\text{also, } \mathbf{x} \in \Omega_{pq} \tag{8}$$

$$\implies \mathbf{x} \in \Omega_p^f \text{ and } \mathbf{x} \in \Omega_q^g \tag{9}$$

now if either $i \neq p$ or $j \neq q$ we have at least one common element in intersection of 2 different partitions of either f or g, contradicting that Ω^f and Ω^g are partitions of $\mathbf{D}$.

(b) $\bigcup_{i=1,j=1}^{i=n,j=m} \Omega_{ij} = \mathbf{D}$

$$\forall \mathbf{x} \in \mathbf{D}, \exists i \mathbf{x} \in \Omega_i^f$$
$$\forall \mathbf{x} \in \mathbf{D}, \exists j \mathbf{x} \in \Omega_j^g$$
$$\implies \forall \mathbf{x} \in \mathbf{D} \exists i, j \mathbf{x} \in \Omega_{ij}$$

□

New partition is a valid partition for the "sum of functions" definition below:

$$f(\mathbf{x}) + g(\mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}\{\mathbf{x} \in \Omega_{ij}\} \left(\left\langle \mathbf{w}^i, \mathbf{x}\right\rangle + \left\langle \mathbf{v}^j, \mathbf{x}\right\rangle\right.$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}\{\mathbf{x} \in \Omega_{ij}\} \left(\left\langle \mathbf{w}^i + \mathbf{v}^j, \mathbf{x}\right\rangle\right)$$

3. To show: if $f(\mathbf{x})$ is peicewise linear then $g(\mathbf{x}) = f_{ReLU}(f(\mathbf{x}))$ is piecewise linear as well.

$$f_{ReLU}(f(\mathbf{x})) = \max(f(\mathbf{x}), 0)$$

$\forall \Omega_i \in \Omega$ partition Ω_i into

$$\Omega_i^0 = \left\{\mathbf{x}'s | f(\mathbf{x}) < 0\right\}$$

and

$$\Omega_i^1 = \left\{\mathbf{x}'s | f(\mathbf{x}) \geq 0\right\}$$

Note that some of new partitions might be empty, but that does not break the definition of partition given above.

And $\mathbf{w}$ is updated as $\mathbf{w}^{i0} = \mathbf{0}$ and $\mathbf{w}^{i1} = \mathbf{w}^{i}$

We have new n'-partition of the domain, with new n' model vectors.

With updated Ω and $\mathbf{w}$ f_{ReLU} is

$f_{ReLU}(f(x)) = \sum_{i=1}^{n'} \mathbb{I}\{\mathbf{x} \in \Omega_i\} (\langle \mathbf{w}^i, \mathbf{x} \rangle)$ which is piecewise linear with new Ω.

4. To show that neural nets with f_{ReLU} activation constructs piecewise linear functions

   Proof idea: From (1) and (2) we have shown that any linear combination of piecewise linear functions is also piecewise linear. Also in part (3) we proved, $f_{ReLU}(f(\mathbf{x}))$ is piecewise linear.

   Proof by Induction on number of layers

   Base case:

   Let NN be a neural net with only one layer of activation.

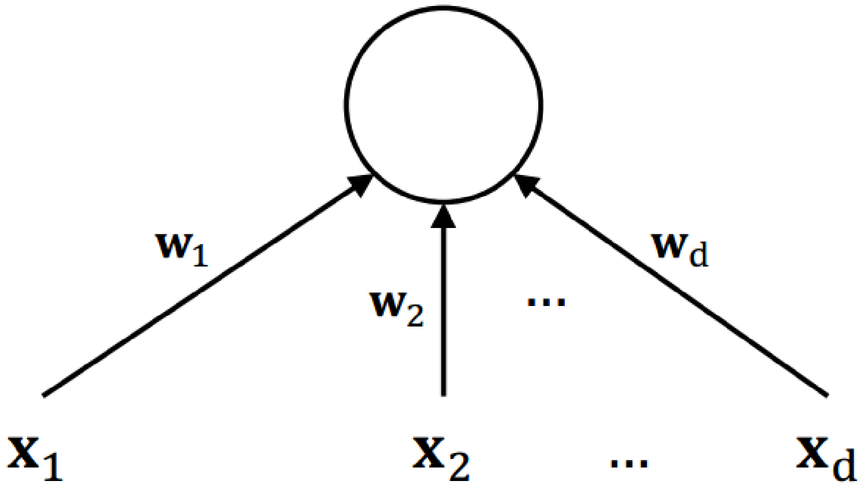

Figure 1:

   Evaluation, $\sum_{i=1}^{n_1} \mathbf{w}_i \cdot \mathbf{x}_i$ is a Piecewise Linear function.

   and the activation layer is $f_{ReLU}(g(\mathbf{x}))$.

   From part (3) we conclude output for above neural network is linearly piecewise.

   Therefore, $P(1)$ is *true*.

   Induction Hypothesis : Let's assume this claim is true for neural nets with m layers. And let z_i represent output of i^{th} component of m^{th} layer.

   Now adding a new layer $m+1$ above m^{th} layer can be seen as:

   Evaluation: $g = \sum_i w_i^m \cdot z_i$

   from (1) and (2), if $f(x)$ and $h(x)$ are piecewise linear, then so will be $g(x) = a \cdot f(x) + b \cdot h(x)$ for any scalars a, b.

   given the m^{th} layer outputs from each node as a piecewise linear function , Evaluation in $(m+1)^{st}$ layer is also piecewise linear.

   Next, (3) shows $f_{ReLU}(f(\mathbf{x}))$ is piecewise linear for any piecewise linear $f(\mathbf{x})$, *implies* output of $(m+1)^{st}$ layer is piecewise linear as well.

5. Given d input nodes and D hidden layer nodes of a single hidden layer.

   Each of the hidden layer node calculates f_{ReLU} which can produce twice the no. of peices than it's input function at maximum.

   This gives us $2Dd$ maximum possible pieces at the output of hidden layer. The output layer has a single node with f_{ReLU} as activation function, which can give maximum twice the no. of pieces.

   So, total no. of possible pieces computed by the network is **4Dd**.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 3
Student Name: Nikhil Mittal
Roll Number: 17111056
Date: November 15, 2017

QUESTION 3

Algorithm 1: kernelized perceptron

Input: Data $\mathbf{x}^1, \ldots, \mathbf{x}^n$ in online way

$\alpha \leftarrow 0$ //Initialize as 0 vector

Receive data point $z^t = (x^t, y^t)$
// Compute Activation i.e. $\langle w, \phi(x^t) \rangle$, no need to compute feature map. By Perceptron Representer Theorem $w = \sum_m \alpha_m \phi(x^m)$. So, $\langle w, \phi(x^t) \rangle = \langle \sum_m \alpha_m \phi(x^m), \phi(x^t) \rangle = \sum_m \alpha_m \langle \phi(x^m), \phi(x^t) \rangle$ which is $\sum_m \alpha_m K(x^m, x^t)$

Compute a $\leftarrow \sum_m \alpha_m K(x^m, x^t)$
if y^ta≤ 0 **then**
    $\alpha_t = \alpha_t + y^t$ // Make updates only when making a mistake
end if

By Perceptron Representer Theorem,

$$w = \sum_m \alpha_m \phi(x^m)$$

So calculating w not needed, its expression can be directly used to get value $< w, \phi(x) >$ at test time.

So for a test point x,

$$\langle w, \phi(x) \rangle \; = \; \left\langle \sum_m \alpha_m \phi(x_m), \phi(x) \right\rangle \; = \; \sum_m \alpha_m \langle \phi(x_m), \phi(x) \rangle$$

then,

$$\langle w, \phi(x) \rangle \; = \; \sum_m \alpha_m K(x^m, x)$$

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION 4

Assignment Number: 3
Student Name: Nikhil Mittal
Roll Number: 17111056
Date: November 15, 2017

3.4.1

$$K(z^1, z^2) = (\langle z^1, z^2 \rangle + 1)^2$$

where

$$z^1 = (x^1, y^1), z^2 = (x^2, y^2)$$

$$(\langle z^1, z^2 \rangle + 1)^2 = 1 + 2\langle z^1, z^2 \rangle + \langle z^1, z^2 \rangle^2$$

$$= 1 + 2(x^1x^2 + y^1y^2) + (x^1x^2 + y^1y^2)^2$$

$$= 1 + 2x^1x^2 + 2y^1y^2 + (x^1x^2)^2 + (y^1y^2)^2 + 2x^1x^2y^1y^2$$

$$= 1 + 2x^1x^2 + 2y^1y^2 + (x^1x^2)^2 + (y^1y^2)^2 + x^1x^2y^1y^2 + x^1x^2y^1y^2$$

$$= 1.1 + \sqrt{2}x^1.\sqrt{2}x^2 + \sqrt{2}y^1.\sqrt{2}y^2 + (x^1)^2.(x^2)^2 + (y^1)^2.(y^2)^2 + (x^1x^2).(y^1y^2) + (x^2x^1).(y^2y^1)$$

$$= (1, \sqrt{2}x^1, \sqrt{2}y^1, (x^1)^2, (y^1)^2, x^1y^1, y^1x^1)^T(1, \sqrt{2}x^2, \sqrt{2}y^2, (x^2)^2, (y^2)^2, x^2y^2, y^2x^2)$$

$$= \langle \phi(z^1), \phi(z^2) \rangle$$

where

$$\phi(z^1) = (1, \sqrt{2}x^1, \sqrt{2}y^1, (x^1)^2, (y^1)^2, x^1y^1, y^1x^1)$$

$\mathcal{H}_K = \mathrm{R}^7$
So, D is 7.

3.4.2
Let $w = (w_1, w_2, w_3, w_4, w_5, w_6, w_7)$

and

$$\phi(z^1) = (1, \sqrt{2}x^1, \sqrt{2}y^1, (x^1)^2, (y^1)^2, x^1y^1, y^1x^1)$$

$\langle w, \phi(z)\rangle = w_1 + \sqrt{2}xw_2 + \sqrt{2}yw_3 + x^2w_4 + xyw_5 + yxw_6 + y^2w_7$

Let

$$A = \begin{bmatrix} a & b' \\ c' & d \end{bmatrix} \, and \, b = \begin{bmatrix} e \\ f \end{bmatrix}$$

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

So,

$$\langle z, Az\rangle = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} a & b' \\ c' & d \end{bmatrix} * \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\langle b, z\rangle = ex + fy$$

$$f_{(A,b,c)}(z) = ax^2 + b'xy + c'yx + dy^2 + ex + fy + c$$

Comparing

$$\langle w, \phi(z)\rangle \; and \; f_{(A,b,c)}(z)$$

we get, (on comparing coefficients of x, y, x^2, y^2, xy and constant term)
So,

$$w_1 = c$$
$$\sqrt{2}xw_2 = ex$$
$$\sqrt{2}yw_3 = fy$$
$$x^2w_4 = ax^2$$
$$xyw_5 = b'xy$$
$$yxw_6 = c'yx$$
$$y^2w_7 = dy^2$$

Hence,

$$w_1 = c, w_2 = e/\sqrt{2}, w_3 = f/\sqrt{2}, w_4 = a, w_5 = b', w_6 = c', w_7 = d$$

$w = \left(c, \frac{e}{\sqrt{2}}, \frac{f}{\sqrt{2}}, a, b', c', d\right)$

3.4.3
Given, $w = (w_1, w_2, w_3, w_4, w_5, w_6, w_7)$

Let

$$A = \begin{bmatrix} a & b' \\ c' & d \end{bmatrix} \, and \; b = \begin{bmatrix} e \\ f \end{bmatrix}$$

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

So,

$$\langle z, Az \rangle = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} a & b' \\ c' & d \end{bmatrix} * \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\langle b, z \rangle = ex + fy$$

$$f_{(A,b,c)}(z) = ax^2 + b'xy + c'yx + dy^2 + ex + fy + c$$

Comparing

$$f_{(A,b,c)}(z) \; and \; \langle w, \phi(z) \rangle$$

we get, (on comparing coefficients of x, y, x$^2, y^2, xy$ and constant term)

$$c = w_1$$

$$ex = \sqrt{2} x w_2$$

$$fy = \sqrt{2} y w_3$$

$$ax^2 = x^2 w_4$$

$$b'xy = xyw_5$$

$$c'yx = yxw_6$$

$$dy^2 = y^2 w_7$$

Hence,

$$c = w_1, e = \sqrt{2} w_2, f = \sqrt{2} w_3, a = w_4, b' = w_5, c' = w_6, d = w_7$$

Therefore,

$$A = \begin{bmatrix} w_4 & w_5 \\ w_6 & w_7 \end{bmatrix} \, and \; b = \begin{bmatrix} \sqrt{2} w_2 \\ \sqrt{2} w_3 \end{bmatrix}$$

and

$$c = w_1$$

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION 5

Assignment Number: 3
Student Name: Nikhil Mittal
Roll Number: 17111056
Date: November 15, 2017

Given

$$\mathbb{P}[\mathbf{z}] = \mathcal{N}(\mathbf{0}, I_k) \in \mathbb{R}^k,$$

whereupon an affine transformation is applied to them and noise is added to produce the observed data point, i.e. for $W \in \mathbb{R}^{d \times k}, \boldsymbol{\mu} \in \mathbb{R}^d, \sigma \geq 0$

$$\mathbb{P}[\mathbf{x} \mid \mathbf{z}] = \mathcal{N}(\mathbf{x} \mid W\mathbf{z} + \mu, \sigma^2 \cdot I_d) \in \mathbb{R}^d.$$

Now using conjugacy properties of the Gaussian ([**BIS**] Chapter 12), we can show that

$$\mathbb{P}[\mathbf{x}] = \int_{\mathbf{z}} \mathbb{P}[\mathbf{x} \mid \mathbf{z}] \, \mathbb{P}[\mathbf{z}] \; d\mathbf{z} = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, C),$$

where $C = WW^\top + \sigma^2 \cdot I_d$. For a dataset $X = \left[\mathbf{x}^1, \ldots, \mathbf{x}^n\right]$.

Here, To find the mean and covariance of data X :

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[Wz + \mu + \epsilon]$$

By, linearity of expectaion,

$$= \mathbb{E}[Wz] + \mathbb{E}[\mu] + \mathbb{E}[\epsilon]$$

Given,

$$\mathbb{E}[z] = 0, \; \mathbb{E}[\epsilon] \; = \; 0 \; and \; \mathbb{E}[\mu] \; = \; \mu$$

So,

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[Wz] + \mathbb{E}[\mu] + \mathbb{E}[\epsilon] = \mu$$

Now, Covariance

$$\begin{aligned}
Cov[x] &= \mathbb{E}[(Wz + \epsilon)(Wz + \epsilon)^T] \\
&= \mathbb{E}[(Wz + \epsilon)(z^T W^T + \epsilon^T)] \\
&= \mathbb{E}[(Wzz^T W^T + \epsilon\epsilon^T)] \\
&= \mathbb{E}[(WW^T + \epsilon\epsilon^T)] \\
&= \mathbb{E}[WW^T] + \mathbb{E}[\epsilon\epsilon^T] \\
&= WW^T + \sigma^2 I_d
\end{aligned}$$

Let,

$$C = WW^T + \sigma^2 I_d$$

Likelihood Expression :

$$\mathbb{P}(x^i | \mu, W, \sigma^2) = \frac{1}{\sqrt{(2\pi)^D |C|}} \exp\left(-\frac{(x_i - \mu)^T C^{-1} (x_i - \mu)}{2}\right)$$

Taking log

$$\log \mathbb{P}(x^i|\mu, W, \sigma^2) = -\frac{D}{2}\log 2\pi - \frac{\log|C|}{2} - \frac{(x_i-\mu)^T C^{-1}(x_i-\mu)}{2}$$

$$\sum_i \log \mathbb{P}(x^i|\mu, W, \sigma^2) = \sum_i \left(-\frac{D}{2}\log 2\pi - \frac{\log|C|}{2} - \frac{(x_i-\mu)^T C^{-1}(x_i-\mu)}{2}\right)$$

$$\sum_i \log \mathbb{P}(x^i|\mu, W, \sigma^2) = -\frac{ND}{2}\log 2\pi - \frac{N\log|C|}{2} - \sum_i \frac{(x_i-\mu)^T C^{-1}(x_i-\mu)}{2}$$

$$\mathbb{P}(X|\mu, W, \sigma^2) = \prod_i \mathbb{P}(x^i|\mu, W, \sigma^2)$$

$$\log \mathbb{P}(X|\mu, W, \sigma^2) = \log \prod_i \mathbb{P}(x^i|\mu, W, \sigma^2)$$

$$\log \mathbb{P}(X|\mu, W, \sigma^2) = \sum_i \log \mathbb{P}(x^i|\mu, W, \sigma^2)$$

Complete Likelihood Expression :

$$\log \mathbb{P}(X|\mu, W, \sigma^2) = -\frac{ND}{2}\log 2\pi - \frac{N\log|C|}{2} - \sum_i \frac{(x_i-\mu)^T C^{-1}(x_i-\mu)}{2}$$

$$\log \mathbb{P}(X|\mu, W, \sigma^2) = -\frac{N}{2}\left(D\log 2\pi + \log|C| + \frac{1}{N}\sum_i \frac{(x_i-\mu)^T C^{-1}(x_i-\mu)}{2}\right)$$

Derivation for μ^{MLE} :

Taking derivative w.r.t μ

$$\frac{d}{d\mu}\log \mathbb{P}(X|\mu, W, \sigma^2) = \frac{d}{d\mu}\left(-\frac{N}{2}\left(D\log 2\pi + \log|C| + \frac{1}{N}\sum_i \frac{(x_i-\mu)^T C^{-1}(x_i-\mu)}{2}\right)\right)$$

$$\frac{d}{d\mu}\log \mathbb{P}(X|\mu, W, \sigma^2) = \frac{-1}{2}\frac{d}{d\mu}\left(\sum_i (x_i-\mu)^T C^{-1}(x_i-\mu)\right)$$

Setting this to Zero,

$$\frac{d}{d\mu}\left(\sum_i (x_i-\mu)^T C^{-1}(x_i-\mu)\right) = 0$$

$$\left(\sum_i (-2C^{-1})(x_i-\mu)\right) = 0$$

Hence,

$$\sum_i (x_i-\mu) = 0$$

$$\sum_i (x_i) - \sum_i (\mu) = 0$$

$$\sum_i (x_i) - n\mu = 0$$

So,

$$\mu = \frac{1}{n} \sum_i x_i$$

The Expression is :

$$\mu^{MLE} = \frac{1}{n} \sum_i x_i$$