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

Assignment Number: 3
Student Name: Siddharth Agrawal
Roll Number: 150716
Date: November 14, 2017

QUESTION

1

1 Property 1

From the definition of MLE estimate:

$$\log \mathbb{P}[X|\boldsymbol{\theta}^{\mathrm{MLE}}] \geq \log \mathbb{P}\left[X|\theta\right] \quad \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{1}$$

Let

$$\mathcal{Q}_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i,\boldsymbol{\theta}^t\right]} \log \frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z}|\boldsymbol{\theta}\right]}{\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i, \boldsymbol{\theta}^t\right]}$$

From **Lec16 Slide 43**, we know that:

$$\mathcal{Q}_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) = \log \mathbb{P}\left[X|\boldsymbol{\theta}^t\right] \tag{2}$$

Using Equation 2 in Equation 1

$$\mathcal{Q}_{\boldsymbol{\theta}^{\mathrm{MLE}}}(\boldsymbol{\theta}^{\mathrm{MLE}}) \geq \log \mathbb{P}\left[X|\theta\right] \quad \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{3}$$

From **Lec16 Slide 44**, we know that:

$$\log \mathbb{P}\left[X|\theta\right] \geq \mathcal{Q}_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}) \quad \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{4}$$

Using Equation 4 in Equation 3

$$\mathcal{Q}_{\boldsymbol{\theta}^{\mathrm{MLE}}}(\boldsymbol{\theta}^{\mathrm{MLE}}) \geq \mathcal{Q}_{\boldsymbol{\theta}^{\mathrm{MLE}}}(\boldsymbol{\theta}) \quad \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{5}$$

Expanding Equation 5 using the definition of $\mathcal{Q}$:

$$\sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i,\boldsymbol{\theta}^{\mathrm{MLE}}\right]} \log \frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z}|\boldsymbol{\theta}^{\mathrm{MLE}}\right]}{\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i, \boldsymbol{\theta}^{\mathrm{MLE}}\right]} \geq \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i,\boldsymbol{\theta}^{\mathrm{MLE}}\right]} \log \frac{\mathbb{P}\left[\mathbf{x}^i, \mathbf{z}|\boldsymbol{\theta}\right]}{\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i, \boldsymbol{\theta}^{\mathrm{MLE}}\right]} \quad \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

$$\sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i,\boldsymbol{\theta}^{\mathrm{MLE}}\right]} \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z}|\boldsymbol{\theta}^{\mathrm{MLE}}\right] \geq \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}\left[\mathbf{z}|\mathbf{x}^i,\boldsymbol{\theta}^{\mathrm{MLE}}\right]} \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z}|\boldsymbol{\theta}\right] \quad \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

$$Q_{\boldsymbol{\theta}^{\mathrm{MLE}}}(\boldsymbol{\theta}^{\mathrm{MLE}}) \geq Q_{\boldsymbol{\theta}^{\mathrm{MLE}}}(\boldsymbol{\theta}) \quad \forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{6}$$

Maximizing the RHS in Equation 6 will maximize the LHS.
Hence,

$$\boldsymbol{\theta}^{MLE} \in \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, Q_{\boldsymbol{\theta}^{\mathrm{MLE}}}(\boldsymbol{\theta}) \tag{7}$$

2 Property 2

Since $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ are optimal MLE solutions, hence:

$$\mathbb{P}\left[X|\boldsymbol{\theta}^1\right] = \mathbb{P}\left[X|\boldsymbol{\theta}^2\right]$$

Let after t iterations of the EM algorithm, $\boldsymbol{\theta}^t = \boldsymbol{\theta}^1$.
Using the results derived in **Lecture 16, slides-43&44**, we can argue that

$$\mathbb{P}\left[X|\boldsymbol{\theta}^{t+1}\right] \geq \mathcal{Q}_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{t+1}) \geq \mathbb{P}\left[X|\boldsymbol{\theta}^1\right]$$

But, since $\boldsymbol{\theta}^1$ is an MLE solution, hence we have $\mathbb{P}\left[X|\boldsymbol{\theta}^1\right] \geq \mathbb{P}\left[X|\boldsymbol{\theta}^{t+1}\right]$ Hence, for all the subsequent $\boldsymbol{\theta}^{\tilde{t}}, \tilde{t} > t$, it will maximize $\mathcal{Q}_{\boldsymbol{\theta}^1}(\boldsymbol{\theta})$ and so will be the MLE estimate.

Remark 1.1. But how do I arrive at the result that this MLE estimate can be converted to $\boldsymbol{\theta}^2$?

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

QUESTION 2

Assignment Number: 3
Student Name: Siddharth Agrawal
Roll Number: 150716
Date: November 14, 2017

1 Scalar Multiplication

`Given:` a piecewise linear function $f : \mathbb{R}^d \to \mathbb{R}$, and a scalar c
`To Prove:` $g(\mathbf{x}) = c \cdot f(\mathbf{x})$ is also piecewise linear
`Proof:`

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

Take $\forall i \in [n]$:

$$\begin{aligned} \boldsymbol{\Omega}_i^g &= \boldsymbol{\Omega}_i \\ \mathbf{w}_g^i &= c \cdot \mathbf{w}^i \end{aligned}$$

Now rewrite $g(\mathbf{x})$ as:

$$g(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \boldsymbol{\Omega}_i^g\right\} \cdot \left\langle \mathbf{w}_g^i, \mathbf{x} \right\rangle$$

Hence, g is also a piecewise linear function.

$\square$

2 Addition

`Given:` 2 piecewise linear functions $f_1 : \mathbb{R}^d \to \mathbb{R}$, $f_2 : \mathbb{R}^d \to \mathbb{R}$
`To Prove:` $g(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$ is also piecewise linear
`Proof:`

$$f_1(\mathbf{x}) = \sum_{i=1}^{n_1} \mathbb{I}\left\{\mathbf{x} \in \mathbf{\Omega}_i^1\right\} \cdot \left\langle \mathbf{w}_1^i, \mathbf{x} \right\rangle \tag{8}$$

$$f_2(\mathbf{x}) = \sum_{i=1}^{n_2} \mathbb{I}\left\{\mathbf{x} \in \mathbf{\Omega}_i^2\right\} \cdot \left\langle \mathbf{w}_2^i, \mathbf{x} \right\rangle \tag{9}$$

Now,

$$\begin{aligned}
g(\mathbf{x}) &= f_1(\mathbf{x}) + f_2(\mathbf{x}) \\
&= \sum_{i=1}^{n_1} \mathbb{I}\left\{\mathbf{x} \in \mathbf{\Omega}_i^1\right\} \cdot \left\langle \mathbf{w}_1^i, \mathbf{x} \right\rangle + \sum_{i=1}^{n_2} \mathbb{I}\left\{\mathbf{x} \in \mathbf{\Omega}_i^2\right\} \cdot \left\langle \mathbf{w}_2^i, \mathbf{x} \right\rangle \\
&= \sum_{i=1,j=1}^{n_1,n_2} \mathbb{I}\left\{\mathbf{x} \in \mathbf{\Omega}_i^1 \cap \mathbf{\Omega}_j^2\right\} \cdot \left(\left\langle \mathbf{w}_1^i, \mathbf{x} \right\rangle + \left\langle \mathbf{w}_2^j, \mathbf{x} \right\rangle\right) \\
&= \sum_{i=1,j=1}^{n_1,n_2} \mathbb{I}\left\{\mathbf{x} \in \mathbf{\Omega}_i^1 \cap \mathbf{\Omega}_j^2\right\} \cdot \left\langle \mathbf{w}_1^i + \mathbf{w}_2^j, \mathbf{x} \right\rangle
\end{aligned}$$

Hence,

$$g(\mathbf{x}) = \sum_{i=1,j=1}^{n_1,n_2} \mathbb{I}\left\{\mathbf{x} \in \mathbf{\Omega}_i^1 \cap \mathbf{\Omega}_j^2\right\} \cdot \left\langle \mathbf{w}_1^i + \mathbf{w}_2^j, \mathbf{x} \right\rangle \tag{10}$$

Consider the partition:

$$\widetilde{\mathbf{\Omega}}^g = \bigcup_{i=1,j=1}^{n_1,n_2} \left(\mathbf{\Omega}_i^1 \cap \mathbf{\Omega}_j^2\right)$$

As, $\mathbf{\Omega}_i^1$ are disjoint $\forall i \in [n_1]$ and $\mathbf{\Omega}_j^2$ are disjoint $\forall j \in [n_2]$, hence $\mathbf{\Omega}_i^1 \cap \mathbf{\Omega}_j^2$ are disjoint $\forall\ i \in [n_1], j \in [n_2]$.
Hence g is indexed by $n_1 n_2$ partitions of $\mathbb{R}^d$ (say, $\left[\mathbf{\Omega}_1^g, \mathbf{\Omega}_2^g, ..., \mathbf{\Omega}_{n_1 n_2}^g\right]$).

The linear model corresponding to g is given by:-

$$\forall i \in [n_1 n_2] \quad \mathbf{w}_i^g = \mathbf{w}_j^1 + \mathbf{w}_k^2; \qquad \mathbf{\Omega}_i^g = \mathbf{\Omega}_j^1 \cap \mathbf{\Omega}_k^2$$

Hence, g is a piecewise linear function.

$\square$

3 ReLU Activation

`Given:` a piecewise linear function $f : \mathbb{R}^d \to \mathbb{R}$
`To Prove:` $g(\mathbf{x}) = f_{\text{ReLU}}\left(f(\mathbf{x})\right)$ is also piecewise linear
`Proof:`

$$\begin{aligned} g(\mathbf{x}) &= f_{\text{ReLU}}\left(f(\mathbf{x})\right) \\ &= f_{\text{ReLU}}\left(\sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \boldsymbol{\Omega}_i\right\} \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle\right) \\ &= \max\left(\sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \boldsymbol{\Omega}_i\right\} \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle, 0\right) \end{aligned}$$

Consider the set for which f becomes negative:

$$\boldsymbol{\Omega}_0^g = \left\{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^d, f(\mathbf{x}) < 0\right\}$$

Let its Corresponding linear model be:

$$\mathbf{w}_g^0 = [0, 0, ...d \text{ zeroes}]^\top$$

So,

$$\begin{aligned} g(\mathbf{x}) &= \sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \boldsymbol{\Omega}_i \backslash \boldsymbol{\Omega}_0^g\right\} \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle + \mathbb{I}\left\{\mathbf{x} \in \boldsymbol{\Omega}_0^g\right\} cdot \left\langle \mathbf{w}_g^0, \mathbf{x} \right\rangle \\ &= \sum_{i=0}^{n} \mathbb{I}\left\{\mathbf{x} \in \boldsymbol{\Omega}_i^g\right\} \cdot \left\langle \mathbf{w}_g^i, \mathbf{x} \right\rangle \end{aligned}$$

Where, $\forall i \in \{0, 1, 2, ..., n\}$

$$\begin{aligned} \boldsymbol{\Omega}_i^g &= \begin{cases} \left\{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^d, f(\mathbf{x}) < 0\right\} & i = 0 \\ \boldsymbol{\Omega}_i \backslash \boldsymbol{\Omega}_0^g & \text{Otherwise} \end{cases} \\ \mathbf{w}_g^i &= \begin{cases} [0, 0, ...d \text{ zeroes}]^\top & i = 0 \\ \mathbf{w}^i & \text{Otherwise} \end{cases} \end{aligned}$$

Hence, g is also a piecewise linear function.

$\square$

4 Neural Networks with ReLU activation function computes a piecewise linear function

We prove this result using induction on the layers of the neural network.

Base Case

The base case will consist of only 1 input and 1 output layers (no hidden layer).
Consider any i^{th} node of the output layer. For it, we can define the output as -

$$g^i(\mathbf{x}) = f_{\text{ReLU}}\left(\langle \mathbf{w}^i, \mathbf{x} \rangle\right)$$

. Using the result of part 3, since inner product is a piecewise linear function, hence g^i is also piecewise linear.

Induction Hypothesis

If k^{th} layer of the network computes piecewise linear function, then $(k+1)^{th}$ layer computes piecewise linear function.

Proof for Induction Hypothesis

Let for any i^{th} node in the $(k+1)^{th}$ layer, inputs come from all nodes in the k^{th} layer.
The outputs from k^{th} layer are mapped to any node i in the $(k+1)^{th}$ layer with weight $\mathbf{w}^i$. Let f^j be the output of the j^{th} node in this layer.
Hence, the output function g^i for the i^{th} node in $(k+1)^{th}$ layer is given by:

$$g^i = \sum_j \mathbf{w}^i_j \cdot f^j$$

Since scalar multiplication and addition on a piecewise linear function preserves its piecewise linearity (using results proved in part 1 & 2) and $\forall j, f^j$ are piecewise linear (induction hypothesis), hence g^i is piecewise linear.

$\square$

Hence, we have proved that any neural network with a ReLU activation function computes a piecewise linear function.

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

Assignment Number: 3
Student Name: Siddharth Agrawal
Roll Number: 150716
Date: November 14, 2017

QUESTION 3

For the below given algorithm, input points are being stored in a list instead of storing the whole model $\mathbf{w} \in \mathcal{H}_K$.
Also, it is assumed that the input points are in the form $\left(y^t, \mathbf{x}^t\right)$ with $y^t \in \mathbb{R}, \mathbf{x}^t \in \mathbb{R}^d$
Initial value is assumed as $\boldsymbol{\beta}^0$

Algorithm 1: Kernel Perceptron

Store an initial point $\boldsymbol{\beta}^0 \in \mathbb{R}^d$ in the list.
for each new data point $(y^t, \mathbf{x}^t)$ received **do**
  $\epsilon \leftarrow 0$
  for each point $\boldsymbol{\beta}^k$ in the list **do**
    $\epsilon \leftarrow \epsilon + K(\boldsymbol{\beta}^k, \mathbf{x}^t)$
  end for
  if $y^t \cdot \epsilon < 0$ **then**
    Append $\alpha_t y^t \mathbf{x}^t$ to the list (as the value for $\boldsymbol{\beta}^t$)
  end if
end for

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

Assignment Number: 3
Student Name: Siddharth Agrawal
Roll Number: 150716
Date: November 14, 2017

QUESTION

4

1 Proving that Kernel K is Mercer

The construction for the mapping $\varphi : \mathbb{R}^2 \to \mathcal{H}_K$ is:

$$
\begin{aligned}
&\varphi(\mathbf{z}) = [\varphi_0(\mathbf{z}), \varphi_1(\mathbf{z}), \varphi_2(\mathbf{z})] \in \mathbb{R}^{2^2+2+1} = \mathbb{R}^7 \\
&\text{where,} \\
&\varphi_0(\mathbf{z}) = 1 \in \mathbb{R}^1 \\
&\varphi_1(\mathbf{z}) = \sqrt{2} \cdot [\mathbf{z}_1, \mathbf{z}_2] \in \mathbb{R}^2 \\
&\varphi_2(\mathbf{z}) = [\mathbf{z}_1\mathbf{z}_1, \mathbf{z}_1\mathbf{z}_2, \mathbf{z}_2\mathbf{z}_1, \mathbf{z}_2\mathbf{z}_2] \in \mathbb{R}^4
\end{aligned}
$$

Remark 4.1. $D = 7. \therefore \mathcal{H}_K \equiv \mathbb{R}^7$

To show that K is a Mercer Kernel, it is sufficient to show that $K(\mathbf{z}^1, \mathbf{z}^2) = \left\langle \varphi(\mathbf{z}^1), \varphi(\mathbf{z}^2) \right\rangle$. Hence,

$$
\begin{aligned}
\left\langle \varphi(\mathbf{z}^1), \varphi(\mathbf{z}^2) \right\rangle &= \left\langle \varphi_0(\mathbf{z}^1), \varphi_0(\mathbf{z}^2) \right\rangle + \left\langle \varphi_1(\mathbf{z}^1), \varphi_1(\mathbf{z}^2) \right\rangle + \left\langle \varphi_2(\mathbf{z}^1), \varphi_2(\mathbf{z}^2) \right\rangle \\
&= 1 + 2 \cdot \left\langle \mathbf{z}^1, \mathbf{z}^2 \right\rangle + \sum_{i,j}^{2} \mathbf{z}_i^1 \mathbf{z}_j^1 \mathbf{z}_i^2 \mathbf{z}_j^2 \\
&= \left(\left\langle \mathbf{z}^1, \mathbf{z}^2 \right\rangle + 1 \right)^2 \\
&= K\left(\mathbf{z}^1, \mathbf{z}^2\right)
\end{aligned}
$$

2 Constructing $\mathrm{w} \in \boldsymbol{\mathcal{H}}_K$

We have to construct $\mathbf{w}$ such that $\forall\ \mathbf{z} \in \mathbb{R}^2$

$$\begin{aligned}
\langle \mathbf{w}, \varphi_K(\mathbf{z}) \rangle &= f_{(A,\mathbf{b},c)}(\mathbf{z}) \\
&= \langle \mathbf{z}, A\mathbf{z} \rangle + \langle \mathbf{b}, \mathbf{z} \rangle + c \\
&= A_{11}\mathbf{z}_1\mathbf{z}_1 + A_{12}\mathbf{z}_2\mathbf{z}_1 + A_{21}\mathbf{z}_1\mathbf{z}_2 + A_{22}\mathbf{z}_2\mathbf{z}_2 + \mathbf{b}_1\mathbf{z}_1 + \mathbf{b}_2\mathbf{z}_2 + c \\
&= c + \mathbf{b}_1\mathbf{z}_1 + \mathbf{b}_2\mathbf{z}_2 + A_{11}\mathbf{z}_1\mathbf{z}_1 + A_{12}\mathbf{z}_2\mathbf{z}_1 + A_{21}\mathbf{z}_1\mathbf{z}_2 + A_{22}\mathbf{z}_2\mathbf{z}_2 \\
&= c + \frac{1}{\sqrt{2}} \cdot \mathbf{b}_1\sqrt{2} \cdot \mathbf{z}_1 + \frac{1}{\sqrt{2}} \cdot \mathbf{b}_2\sqrt{2} \cdot \mathbf{z}_2 + A_{11}\mathbf{z}_1\mathbf{z}_1 + A_{12}\mathbf{z}_2\mathbf{z}_1 + A_{21}\mathbf{z}_1\mathbf{z}_2 + A_{22}\mathbf{z}_2\mathbf{z}_2
\end{aligned}$$

From the above derivation, we can clearly see that -

$$\mathbf{w} = \begin{bmatrix} c & \frac{1}{\sqrt{2}} \cdot \mathbf{b}_1 & \frac{1}{\sqrt{2}} \cdot \mathbf{b}_2 & A_{11} & A_{12} & A_{21} & A_{22} \end{bmatrix}^\top$$

3 Constructing a triplet $(A, \mathrm{b}, c) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2} \times \mathbb{R}$

We have to construct a triplet $(A, \mathbf{b}, c)$ such that $\forall\ \mathbf{z} \in \mathbb{R}^2$ given $\mathbf{w} = [\mathbf{w}_1, \ldots.., \mathbf{w}_7] \in \mathbb{R}^7$

$$
\begin{aligned}
f_{(A,\mathbf{b},c)}(\mathbf{z}) &= \langle \mathbf{w}, \varphi_K(\mathbf{z}) \rangle \\
&= \mathbf{w}_1 + \mathbf{w}_2 \cdot \sqrt{2} \cdot \mathbf{z}_1 + \mathbf{w}_3 \cdot \sqrt{2} \cdot \mathbf{z}_2 + \mathbf{w}_4\mathbf{z}_1\mathbf{z}_1 + \mathbf{w}_5\mathbf{z}_1\mathbf{z}_2 + \mathbf{w}_6\mathbf{z}_2\mathbf{z}_1 + \mathbf{w}_7\mathbf{z}_2\mathbf{z}_2 \\
&= \mathbf{w}_4\mathbf{z}_1\mathbf{z}_1 + \mathbf{w}_5\mathbf{z}_1\mathbf{z}_2 + \mathbf{w}_6\mathbf{z}_2\mathbf{z}_1 + \mathbf{w}_7\mathbf{z}_2\mathbf{z}_2 + \mathbf{w}_2 \cdot \sqrt{2} \cdot \mathbf{z}_1 + \mathbf{w}_3 \cdot \sqrt{2} \cdot \mathbf{z}_2 + \mathbf{w}_1 \\
&= \langle \mathbf{z}, A\mathbf{z} \rangle + \langle \mathbf{b}, \mathbf{z} \rangle + c
\end{aligned}
$$

where

$$
\begin{aligned}
A &= \begin{bmatrix} \mathbf{w}_4 & \mathbf{w}_5 \\ \mathbf{w}_6 & \mathbf{w}_7 \end{bmatrix} \\
\mathbf{b} &= [\mathbf{w}_2, \mathbf{w}_3] \\
c &= \mathbf{w}_1
\end{aligned}
$$

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

Assignment Number: 3
Student Name: Siddharth Agrawal
Roll Number: 150716
Date: November 14, 2017

QUESTION

5

1 Data log-likelihood expression

Since all the data points are independent, hence $\mathbb{P}\left[X|\boldsymbol{\mu}, W, \sigma\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i|\boldsymbol{\mu}, W, \sigma\right]$. Hence,

$$
\begin{aligned}
\mathbb{P}\left[X|\boldsymbol{\mu}, W, \sigma\right] &= \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i|\boldsymbol{\mu}, W, \sigma\right] \\
&= \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}^i|\boldsymbol{\mu}, C) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \left\|C\right\|}} \exp -\frac{1}{2}(\mathbf{x}^i - \boldsymbol{\mu})C^{-1}(\mathbf{x}^i - \boldsymbol{\mu})^\top \\
\implies \log \mathbb{P}\left[X|\boldsymbol{\mu}, W, \sigma\right] &= \sum_{i=1}^{n} -\frac{1}{2}(\mathbf{x}^i - \boldsymbol{\mu})C^{-1}(\mathbf{x}^i - \boldsymbol{\mu})^\top - \frac{1}{2}\log(2\pi \left\|C\right\|)
\end{aligned}
$$

2 Derivation for $\boldsymbol{\mu}^{\mathsf{MLE}}$

$$\begin{aligned}
\boldsymbol{\mu}^{\text{MLE}} &= \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\max}\, \mathbb{P}\left[X|\boldsymbol{\mu}, W, \sigma\right] \\
&= \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\max} \log \mathbb{P}\left[X|\boldsymbol{\mu}, W, \sigma\right] \\
&= \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \frac{1}{2}(\mathbf{x}^i - \boldsymbol{\mu})C^{-1}(\mathbf{x}^i - \boldsymbol{\mu})^\top + \frac{1}{2}\log(2\pi \left\|C\right\|) \\
&= \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} (\mathbf{x}^i - \boldsymbol{\mu})C^{-1}(\mathbf{x}^i - \boldsymbol{\mu})^\top
\end{aligned}$$

Now we Differentiate the RHS term w.r.t. $\boldsymbol{\mu}$ so as to get the MLE estimate.

$$\begin{aligned}
\frac{\partial \text{RHS}}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}} \left(\sum_{i=1}^{n} -\frac{1}{2}(\mathbf{x}^i - \boldsymbol{\mu})C^{-1}(\mathbf{x}^i - \boldsymbol{\mu})^\top \right) &= 0 \\
\sum_{i=1}^{n} -C^{-1}(\mathbf{x}^i - \boldsymbol{\mu}) &= 0 \\
\therefore \boldsymbol{\mu}^{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^i
\end{aligned}$$