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

Assignment Number: 3
Student Name: Gurpreet Singh
Roll Number: 150259
Date: November 15, 2017

QUESTION 1

Part 1

From the definition of $\boldsymbol{\theta}^{MLE}$, we have a MLE estimate $\boldsymbol{\theta}$ such that for any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\theta} \neq \boldsymbol{\theta}^{MLE}$

$$\mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^{MLE}\right] \quad > \quad \mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}\right] \tag{1}$$

Suppose that $\boldsymbol{\theta}^{MLE} \notin \arg\max_{\boldsymbol{\theta} \in \Theta} Q_{\boldsymbol{\theta}^{MLE}}(\boldsymbol{\theta})$.

From **Lecture 16, Slide 44**, we know that for any $\boldsymbol{\theta} \in \Theta$

$$log\left(\mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}\right]\right) \quad \geq \quad Q_{\boldsymbol{\theta}^{MLE}}(\boldsymbol{\theta})$$

From equation 1, we can write

$$log\left(\mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^{MLE}\right]\right) \quad > \quad Q_{\boldsymbol{\theta}^{1}}(\boldsymbol{\theta}^{2})$$

However, from **Lecture 16, Slide 42**, we also know

$$log\left(\mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^{MLE}\right]\right) \quad = \quad Q_{\boldsymbol{\theta}^{MLE}}(\boldsymbol{\theta}^{MLE})$$

Therefore, we can say that for any $\boldsymbol{\theta} \in \Theta$

$$Q_{\boldsymbol{\theta}^{MLE}}(\boldsymbol{\theta}^{MLE}) \quad = \quad Q_{\boldsymbol{\theta}^{MLE}}(\boldsymbol{\theta})$$

However, this is a contradiction, since we assumed that $\boldsymbol{\theta}^{MLE} \notin \arg\max_{\boldsymbol{\theta} \in \Theta} Q_{\boldsymbol{\theta}^{MLE}}(\boldsymbol{\theta})$. Hence, the self-consistency property needs to be satisfied.

Part 2

From the question, we know that $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ are the optimal MLE solutions. Therefore, we can say

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

Since $\boldsymbol{\theta}^1$ is one of the optimal solutions, say there is a chain of iterations such that we reach $\boldsymbol{\theta}^1$ through one of these chains in t^i iterations. Therefore, $\boldsymbol{\theta}^i = \boldsymbol{\theta}^1$. Again, from **Lecture 16, Slide 44**

$$\mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^{i+1}\right] \quad \geq \quad Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{i+1}) \quad \geq \quad \mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^1\right]$$

However, since $\boldsymbol{\theta}^1$ is a MLE solution, we have $\mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^{i+1}\right] \leq \mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^1\right]$. Hence,

$$\mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^1\right] \quad = \quad Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{i+1}) \quad = \quad \mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}^{i+1}\right]$$

Hence, for any $\boldsymbol{\theta}^{i+1}$ that satisies the above constraint will maximise $Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{i+1})$, as that will be the globally maximum possible value. Hence, for any $\boldsymbol{\theta}^{i+1}$ which is a MLE solution, $Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{i+1})$ will be maximum. Therefore, we can say

$$\begin{aligned} \left\{\boldsymbol{\theta}^{i+1}\right\} &= \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}\right] \\ \implies \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}\right] &= \arg\max_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}) \end{aligned}$$

Since, $\boldsymbol{\theta}^2$ is also a MLE solution, therefore we can say that $\boldsymbol{\theta}^2 \in \arg\max_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta})$
With a similar construction for $\boldsymbol{\theta}^2$, we can say

$$\begin{aligned} \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\left[X \,\middle|\, \boldsymbol{\theta}\right] &= \arg\max_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^2}(\boldsymbol{\theta}) \\ \boldsymbol{\theta}^1 &\in \arg\max_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^2}(\boldsymbol{\theta}) \end{aligned}$$

Hence, proved

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

Assignment Number: 3
Student Name: Gurpreet Singh
Roll Number: 150259
Date: November 15, 2017

QUESTION 2

In order to prove that a function f is piecewise linear, we need to prove that we can find a n-partition $\{\Omega_i\}_{i=1}^n$ and a set $\left\{\mathbf{w}^i\right\}_{i=1}^n$ such that

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

We use this to prove the parts in the question.

Part 1

Suppose we have a piecewise linear function f. Then we can write $g(\mathbf{x}) = c \cdot f(\mathbf{x})$ as

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

Hence, we have the set $\{\Omega_i^g\}_{i=1}^n$ and $\left\{\mathbf{w}_g^i\right\}_{i=1}^n$ which defines the piecewise linear function, where $\forall\, i \in [n]$, $\Omega_i^g = \Omega_i$ and $\mathbf{w}_g^i = c \cdot \mathbf{w}^i$.
Hence, g is also piecewise linear

Part 2

Suppose we have two piecewise linear functions f_1 and f_2. Then we can write $g(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$ as

$$\begin{aligned} g(\mathbf{x}) &= f_1(\mathbf{x}) + f_2(\mathbf{x}) \\ &= \sum_{i=1}^{n_1} \mathbb{I}\left[\mathbf{x} \in \Omega_i^1\right] \left\langle \mathbf{w}_1^i, \mathbf{x} \right\rangle + \sum_{i=1}^{n_2} \mathbb{I}\left[\mathbf{x} \in \Omega_i^2\right] \left\langle \mathbf{w}_2^i, \mathbf{x} \right\rangle \end{aligned}$$

Now, for g we need to consider all points in $\left\{\Omega_i^1\right\}_{i=1}^{n_1} \cup \left\{\Omega_i^2\right\}_{i=1}^{n_2}$

Since there can be instersections in these sets, we cannot directly add them. Hence we need to consider all the instersections. Therefore, we can write the complete set as

$$Q \quad = \quad \left\{\Omega_i^1 \cap \Omega_j^2\right\}_{i=1,j=1}^{n_1,n_2} \cup \left\{\Omega_i\right\}_{i=1}^{n_1} \cup \left\{\Omega_j\right\}_{j=1}^{n_2}$$

However, these are not disjoint. In order to make these sets disjoint, we can remove all intersections, since we are already adding them to the complete set. For this, we define two set quantities

$$\begin{aligned} Q_1 &= \sum_{i=1}^{n_1} \bigcup \Omega_i^1 \\ Q_2 &= \sum_{i=1}^{n_2} \bigcup \Omega_i^2 \end{aligned}$$

Then, we can transform the complete set as

$$Q = \left\{\Omega_i^1 \cap \Omega_j^2\right\}_{i=1,j=1}^{n_1,n_2} \cup \left\{\Omega_i \setminus Q_2\right\}_{i=1}^{n_1} \cup \left\{\Omega_j \setminus Q_1\right\}_{j=1}^{n_2}$$

Since, we are including all intersections, we can say that $Q = \left\{\Omega_i^1\right\}_{i=1}^{n_1} \cup \left\{\Omega_j^2\right\}_{j=1}^{n_2}$

Note I am using the concept that a n-partition is disjoint and not explicitly mentioning that. It is only becuase of that that the above term for Q is of this form

We can define the above set Q as our n-partion set for the function g. Since it is disjoint (from construction), it is a valid partion of sets. Now, we can define the mapping from Ω_i^g to $\mathbf{w}_g^i$.

$$\mathbf{w}_g^i = \begin{cases} \mathbf{w}_1^j & \Omega_i^g = \Omega_j^1 \setminus Q_2 \\ \mathbf{w}_2^j & \Omega_i^g = \Omega_j^2 \setminus Q_1 \\ \mathbf{w}_1^j + \mathbf{w}_2^k & \Omega_i^g = \Omega_j^1 \cap \Omega_k^2 \end{cases}$$

To prove that this is valid, and represents the same function g, we can see that for any point $\mathbf{x}$

$$g(x) = \begin{cases} \left\langle \mathbf{w}_1^j, \mathbf{x} \right\rangle & \text{if} \quad \mathbf{x} \in \Omega_j^1 \setminus Q_2 \\ \left\langle \mathbf{w}_2^j, \mathbf{x} \right\rangle & \text{if} \quad \mathbf{x} \in \Omega_j^2 \setminus Q_1 \\ \left\langle \mathbf{w}_1^j + \mathbf{w}_2^k, \mathbf{x} \right\rangle & \text{if} \quad \mathbf{x} \in \Omega_j^1 \cap \Omega_k^2 \end{cases}$$

This is clearly the same as the original defined function $g = f_1 + f_2$. Hence, we have found a representation which defines g and hence it is a piecewise linear function.

Hence, proved

Part 3

Suppose we have a piecewise linear function f and a function g such that $g = f_{ReLU}(f)$. Then we can write

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

We can also represent g as follows

$$g(\mathbf{x}) \quad = \quad \begin{cases} \langle \mathbf{w}^i, \mathbf{x} \rangle & \text{if} \quad \mathbf{x} \in \Omega_i \text{ and } \langle \mathbf{w}^i, \mathbf{x} \rangle > 0 \quad \forall\, i \in [n] \\ 0 & \text{else} \end{cases}$$

Therefore, we can define regions R^i as

$$R^i \quad = \quad \Omega_i \cap \left\{ \mathbf{x} \,\middle|\, \langle \mathbf{w}^i, \mathbf{x} \rangle > 0 \right\}$$

Then the function g can be written as

$$\begin{aligned} g(\mathbf{x}) \quad &= \quad \begin{cases} \langle \mathbf{w}^i, \mathbf{x} \rangle & \text{if} \quad \mathbf{x} \in \mathbb{R}^i \quad \forall\, i \in [n] \\ 0 & \text{else} \end{cases} \\ \implies g(\mathbf{x}) \quad &= \quad \sum_{i=1}^{n} \mathbb{I}\left[\mathbf{x} \in R^i\right] \langle \mathbf{w}^i, \mathbf{x} \rangle \end{aligned}$$

However, since $\sum_{i=1}^{n} \bigcup R^i$ does not cover the whole of $\mathbb{R}^d$, we introduce another set $\overline{R}$ such that $\overline{R} = \mathbb{R}^d \setminus \sum_{i=1}^{n} \bigcup R^i$ corresponding to $\overline{\mathbf{w}} = \mathbf{0}$. Therefore g becomes

$$g(\mathbf{x}) \quad = \quad \sum_{i=1}^{n} \mathbb{I}\left[\mathbf{x} \in R^i\right] \langle \mathbf{w}^i, \mathbf{x} \rangle + \mathbb{I}\left[\mathbf{x} \in \overline{R}\right] \langle \mathbf{0}, \mathbf{x} \rangle$$

From the above expression, we can say that g is a piecewise linear function.

Hence, Proved

Part 4

We prove this using induction. Let us say that for the r^{th} layer, this is true *i.e.* we can represent the value at every node for the r^{th} as a piecewise linear function. Then we prove that the value at every node in the $(r+1)^{th}$ layer can also be represented as a piecewise linear function.

Base Case

> Consider a neural network with only an input layer and an ouptut layer. This case is trivial, as we only need to consider the one layer.
>
> Since for the i^{th} node of the output layer, we can define the incoming input as $\langle \mathbf{w}^i, \mathbf{x} \rangle$, we can say that the value at output node i will be simply $f_{ReLU}\left(\langle \mathbf{w}^i, \mathbf{x} \rangle\right)$ which is, from part 3 can be claimed as a piecewise linear function

Hypothesis

> We assume that the condition is satisfied for the r^{th} layer *i.e.* for every node in the r^{th} layer, the value on that node is from a piecewise linear function

Induction

> Suppose that the input for the i^{th} node in the $(r+1)^{th}$ layer is coming from all the values in the previous layer
>
> For the previous layer, the value on the j^{th} node is given by a piecewise linear function f^j. Therefore, the input on the i^{th} node can be written as a weighted sum of the values of previous layer. Let these weights be defined by $\mathbf{w}^j$
> Since this will be not single value, we represent the input itself to be a function. And the value will be the activation function applied on the input. Hence the value on every node is actually a function, which will be computed on the value of the nodes in the input layer.
> Assume that the number of nodes in the r^{th} layer is n and $(r+1)^{th}$ layer is m

$$I^j \quad = \quad \sum_{i=1}^{n} w_i^j f^i$$

Since we proved in part 1, that multiplying a constant to a piecewise continuous function is also piecewise continuous, and in part 2 we proved that adding piecewise continuous functions also give piecewise continuous functions, we can say that $\forall\, j \in [m]$, I^j is piecewise continuous.

Also, in part 3 we proved that $f_{ReLU}(f)$ is piecewise continuous if f is piecewise continuous. Hence, $f_{ReLU}\left(I^j\right)$ is piecewise continuous.
Therefore, for all nodes, the neural network computes a piecewise continous function. Using the base case, hence, the neural network will compute a piecewise continuous function acting upon the input.

Bonus Question

Following the discussion from previous parts, we can say that every node in the hidden layer will have two sets or pieces, one consisting of points that ouptut something greater than zero, and the other being the complement of the first.

Also, as we showed in the case of addition of piecewise linear functions, the number of sets upon adding them is just analysing the number of subsets (of sets) an element belongs to. Since there might be two or more sets which are contained in another set, we can only comment on the upper bound of the number of pieces. Since the number of sets in the previous (hidden) layer is D.

We do not consider the rest complement sets, as they will be part of only those sets which the actual set is not of. We can also represent this as a subset of sets, and if a set does not belong to one set, we say it's complement belongs to that set. Hence, the number of subsets will simply be the powerset of these D subsets.

$$\text{Number of Pieces} \quad \leq \quad 2^D$$

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

Assignment Number: 3
Student Name: Gurpreet Singh
Roll Number: 150259
Date: November 15, 2017

QUESTION 3

For the following question, assume that $\mathbf{x} \in \mathbb{R}^d$

Algorithm 1: Kernel Perceptron — Storing all Points

1. Initialize a list structure, storing the head of the list
2. Add a single item $\mathbf{w}^0 \in \mathbb{R}^d$ as an initialization of $\mathbf{w}$ in the original space
3. While we recieve a point $t \longrightarrow \mathbf{z}^t = (y^t, \mathbf{x}^t)$
   (i) Initialize $e = 0$
   (ii) For all points $\mathbf{p}$ in the list, update $e \leftarrow e + K(\mathbf{p}, \mathbf{x}^t)$
   (iii) If $y^t \cdot e < 0$, add $\eta_t \cdot y^t \cdot \mathbf{x}^t$ to the head of the list

The list of points obtained are our support vectors, and can be used for predictions.

We can also have a subset of the points, instead of saving all points, we save only a subset of the points. We can use a simple array for this, instead of using a list, although there a tonne of ways to do this.

Algorithm 2: Kernel Perceptron — Storing a subset of Points

1. Initialize an array of k elements, all NULL (k is the number of support points required)
2. Add a single item $\mathbf{w}^0 \in \mathbb{R}^d$ as an initialization of $\mathbf{w}$ in the original space to the array
3. Initiliaze $head \leftarrow 1$ (Indexing starting from 0)
4. While we recieve a point $t \longrightarrow \mathbf{z}^t = (y^t, \mathbf{x}^t)$
   (i) Initialize $e = 0$
   (ii) For all points $\mathbf{p}$ in the array, update $e \leftarrow e + K(\mathbf{p}, \mathbf{x}^t)$
   (iii) If $y^t \cdot e < 0$,
   update array[$head$] $\leftarrow \eta_t \cdot y^t \cdot \mathbf{x}^t$
   update $head \leftarrow head + 1$

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

Assignment Number: 3
Student Name: Gurpreet Singh
Roll Number: 150259
Date: November 15, 2017

QUESTION 4

Part 1

In order to show that the quadratic kernel is Mercel, we need to find a mapping $\phi : \mathbb{R}^2 \to \mathcal{H}_K$, such that $K(\mathbf{z}^1, \mathbf{z}^2) = \langle \mathbf{z}^1, \mathbf{z}^2 \rangle$.

In order to do this, we need to change the representation of the quadratic kernel. We can alternatively represent it in the following form

$$
\begin{aligned}
\left(\langle \mathbf{z}^1, \mathbf{z}^2 \rangle + 1\right)^2 &= 1 + 2 \cdot \langle \mathbf{z}^1, \mathbf{z}^2 \rangle + \left(\langle \mathbf{z}^1, \mathbf{z}^2 \rangle\right)^2 \\
&= 1 + \left\langle \sqrt{2} \cdot \mathbf{z}^1, \sqrt{2} \cdot \mathbf{z}^2 \right\rangle + \sum_{i=1}^{d} \sum_{j=1}^{d} z_i^1 z_j^1 z_i^2 z_j^2 \\
&= 1 + \sum_{i=1}^{d} \left(\sqrt{2} \cdot z_i^1\right) \left(\sqrt{2} \cdot z_i^2\right) + \sum_{j=1}^{d} \sum_{k=1}^{d} \left(z_j^1 \cdot z_k^1\right) \left(z_j^2 \cdot z_k^2\right) \\
&= \begin{bmatrix} 1 \\ \sqrt{2} \cdot z_1^1 \\ \sqrt{2} \cdot z_2^1 \\ \vdots \\ \sqrt{2} \cdot z_d^1 \\ z_1^1 \cdot z_1^1 \\ z_1^1 \cdot z_2^1 \\ \vdots \\ z_1^1 \cdot z_d^1 \\ z_2^1 \cdot z_1^1 \\ \vdots \\ z_d^1 \cdot z_d^1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 \\ \sqrt{2} \cdot z_1^2 \\ \sqrt{2} \cdot z_2^2 \\ \vdots \\ \sqrt{2} \cdot z_d^2 \\ z_1^2 \cdot z_1^2 \\ z_1^2 \cdot z_2^2 \\ \vdots \\ z_1^2 \cdot z_d^2 \\ z_2^2 \cdot z_1^2 \\ \vdots \\ z_d^2 \cdot z_d^2 \end{bmatrix}
\end{aligned}
$$

The matrix representation is exactly the transformed Hilbert Space we require. Hence, we represent the Hilbert Space $\mathcal{H}_K$ in $d^2 + d + 1 = 7$ dimensions, such that $\phi : \mathbb{R}^2 \to \mathcal{H}_K$ is defined as

$$\begin{aligned}
\phi(\mathbf{z}) &= \begin{bmatrix} \phi_0(\mathbf{z}) & \phi_1(\mathbf{z}) & \phi_2(\mathbf{z}) \end{bmatrix} \\
&\quad \text{such that} \\
\phi(\mathbf{z}) &= [1] \quad \in \quad \mathbb{R}^1 \\
\phi(\mathbf{z}) &= \sqrt{2} \cdot \mathbf{z} \quad \in \quad \mathbb{R}^d \\
\phi(\mathbf{z}) &= \begin{bmatrix} z_1 z_1 & z_1 z_2 & z_2 z_1 & z_2 z_2 \end{bmatrix} \quad \in \quad \mathbb{R}^{d^2}
\end{aligned}$$

If $\mathcal{H}_K \equiv \mathbb{R}^D$ then $D = d^2 + d + 1$

Part 2

We have

$$f_{(A,\mathbf{b},c)}(\mathbf{x}) = \langle \mathbf{z}, A\mathbf{z}, + \rangle \langle \mathbf{b}, \mathbf{z}, + \rangle c$$

We can alternatively write this as

$$\begin{aligned}
f_{(A,\mathbf{b},c)}(\mathbf{x}) &= c + \langle \mathbf{b}, \mathbf{z} \rangle + \langle \mathbf{z}, A\mathbf{z} \rangle \\
&= c + \sum_{i=1}^{d} b_i z_i + \sum_{j=1}^{d} \sum_{k=1}^{d} A_{jk}\, z_j\, z_k \\
&= \begin{bmatrix} c \\ b_1/\sqrt{2} \\ b_2/\sqrt{2} \\ \vdots \\ b_d/\sqrt{2} \\ A_{11} \\ A_{12} \\ \vdots \\ A_{1d} \\ A_{21} \\ \vdots \\ A_{dd} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 \\ \sqrt{2} \cdot z_1^2 \\ \sqrt{2} \cdot z_2^2 \\ \vdots \\ \sqrt{2} \cdot z_d^2 \\ z_1^2 \cdot z_1^2 \\ z_1^2 \cdot z_2^2 \\ \vdots \\ z_1^2 \cdot z_d^2 \\ z_2^2 \cdot z_1^2 \\ \vdots \\ z_d^2 \cdot z_d^2 \end{bmatrix} \\
&= \langle \mathbf{w},\, \phi\,(\mathbf{z}) \rangle
\end{aligned}$$

Hence, we can clearly write the function $f_{(A,\mathbf{b},c)}(\mathbf{x})$ in the form $\langle \mathbf{w}, \mathbf{x} \rangle$ such that $\mathbf{w}$ has the form as given below

$$\mathbf{w} = \begin{bmatrix} c & b_1/\sqrt{2} & b_2/\sqrt{2} & A_{11} & A_{12} & A_{21} & A_{22} \end{bmatrix}^{\mathrm{T}}$$

Part 3

We can proceed similar to the previous part, but build build it backwards instead. Assume we have $\mathbf{w} \in \mathcal{H}_K$, then we can write $\langle \mathbf{w}, \phi(\mathbf{z}) \rangle$ as

$$\begin{aligned} \langle \mathbf{w}, \phi(\mathbf{z}) \rangle &= \sum_{i=1}^{D} w_i \cdot \phi(\mathbf{z})_i \\ &= w_1 + \sum_{i=1}^{d} w_{(1+i)} \cdot z_i + \sum_{j=1}^{d} \sum_{k=1}^{d} w_{(1+j \cdot d+k)} \, z_j z_k \end{aligned}$$

This is of the same form as a quadratic function. We can further transform this as follows. For the following discussion, we substitute d by 2. Firstly, assume

$$\begin{aligned} w_c &= w_1 \\ \mathbf{w}_b &= \begin{bmatrix} \sqrt{2}w_2 & \sqrt{2}w_3 \end{bmatrix}^{\mathrm{T}} \\ w_A &= \begin{bmatrix} w_4 & w_5 \\ w_6 & w_7 \end{bmatrix} \end{aligned}$$

Using these, we can write the dot product as

$$\langle \mathbf{w}, \phi(\mathbf{z}) \rangle = w_c + \langle \mathbf{w}_b, \mathbf{z} \rangle + \langle \mathbf{z}, w_A \mathbf{z} \rangle$$

Since, this is exactly of the form of a quadratic function. We can say that we have the parameters $(A, \mathbf{b}, c)$, where

$$\begin{aligned} A &= w_A \\ \mathbf{b} &= \mathbf{w}_b \\ c &= w_c \end{aligned}$$

such that

$$\langle \mathbf{w}, \mathbf{z} \rangle = \langle \mathbf{z}, A\mathbf{z} \rangle + \langle \mathbf{b}, \mathbf{z} \rangle + c$$

Part 4

We need to prove that if the function $\hat{f}$ offers the least squares error, then

$$\sum_{i=1}^{n}\left(y^i - \hat{f}(\mathbf{z}^i)\right)^2 \quad \leq \quad \min_{A,\mathbf{b},c} \quad \sum_{i=1}^{n}\left(y^i - f_{A,\mathbf{b},c}(\mathbf{z}^i)\right)^2 + \epsilon$$

From part 2, we know that every quadratic function can be represented in the form of the dot product of a parameter $\mathbf{w}$ in the Hilbert Space. Let $\hat{\mathbf{w}}$ represent $\hat{f}$. Therefore, we can transform the above equation as

$$\sum_{i=1}^{n}\left(y^i - \left\langle \hat{\mathbf{w}},\, \phi\left(\mathbf{z}^i\right)\right\rangle\right)^2 \quad \leq \quad \min_{\mathbf{w}\in\mathcal{H}_K} \quad \sum_{i=1}^{n}\left(y^i - \left\langle \mathbf{w},\, \phi\left(\mathbf{z}^i\right)\right\rangle\right)^2 + \epsilon \tag{2}$$

We can relate these to the concept of MLE and MAP solutions. From the case of linear regression, we already know the MLE and MAP solutions.

$$\begin{aligned}
\mathbf{w}^{MLE} &= \left(X^T X\right)^{-1} X^T \mathbf{y} \\
\mathbf{w}^{MAP} &= \left(X^T X + \lambda \mathbf{I}\right)^{-1} X^T \mathbf{y}
\end{aligned}$$

In the Hilbert Space, we can transform these as follows

$$\begin{aligned}
\mathbf{w}^{MLE} &= \left(\phi(X)^T\, \phi(X)\right)^{-1} \phi(X)^T\, \mathbf{y} \\
\mathbf{w}^{MAP} &= \left(\phi(X)^T\, \phi(X) + \lambda \mathbf{I}\right)^{-1} \phi(X)^T\, \mathbf{y}
\end{aligned}$$

Note In the Hilbert Space, $\mathbf{w}^{MAP} = \hat{\mathbf{w}}$ (comes directly from the definition of $\hat{f}$)

Also, from the definition of the MLE estimate, which is a minimum of the least squares error, we can say

$$\sum_{i=1}^{n}\left(y^i - \left\langle \mathbf{w}^{MLE},\, \phi\left(\mathbf{z}^i\right)\right\rangle\right)^2 \quad \leq \quad \min_{\mathbf{w}\in\mathcal{H}_K} \quad \sum_{i=1}^{n}\left(y^i - \left\langle \mathbf{w},\, \phi\left(\mathbf{z}^i\right)\right\rangle\right)^2$$

Therefore, if we prove that as $\lambda \longrightarrow 0^+$, $\hat{\mathbf{w}} \longrightarrow \mathbf{w}^{MLE}$, then we can say that $\epsilon \longrightarrow 0$, as the LHS in equation **??** will be a proper lower bound.

Note We can write $\phi(X)^T\, \phi(X)$ as the Gram's Matrix G
Consider the forms of $\mathbf{w}^{MLE}$ and $\hat{\mathbf{w}} = \mathbf{w}^{MAP}$, we can write

$$\begin{aligned}
\hat{\mathbf{w}} &= \left(G + \lambda \mathbf{I}\right)^{-1} \phi(X)^T\, \mathbf{y} \\
&= G^{-1}\left(\mathbf{I} - \frac{\lambda G^{-1}}{1 + \text{ trace}\left(G^{-1}\right)}\right) \phi(X)^T\, \mathbf{y} \\
&\longrightarrow G^{-1}\, \phi(X)^T\, \mathbf{y} \quad \text{as} \quad \lambda \longrightarrow 0 \\
&= \mathbf{w}^{MLE}
\end{aligned}$$

Hence, as $\lambda \longrightarrow 0$, $\hat{\mathbf{w}} \longrightarrow \mathbf{w}^{MLE}$, hence we can say $\epsilon \longrightarrow 0$

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

Assignment Number: 3
Student Name: Gurpreet Singh
Roll Number: 150259
Date: November 15, 2017

QUESTION 5

Part 1

We know that Gaussian-Gaussian has the conjugacy property, and hence, we can say that the marginal probability of $\mathbf{x}$ *i.e.* $\mathbb{P}[\mathbf{x}]$

Therefore, we can write the marginal probability of $\mathbf{x}$ as a gaussian distribution with mean $\boldsymbol{\mu}'$ and covariance matrix Λ.

$$\begin{aligned}\mathbb{P}\left[\mathbf{x} \,\middle|\, \boldsymbol{\mu}, W, \sigma\right] &= \int_{\mathbf{z}} \mathbb{P}\left[\mathbf{x} \,\middle|\, \mathbf{z}, \boldsymbol{\mu}, W, \sigma\right] \mathbb{P}\left[\mathbf{z} \,\middle|\, 0, \mathbf{I}\right] d\mathbf{z} \\ &= \mathcal{N}\left(\mathbf{x} \,\middle|\, \boldsymbol{\mu}', \Lambda\right)\end{aligned}$$

Also, we know that $\mathbf{x} = W\mathbf{z} + \boldsymbol{\mu} + \epsilon$, and from the properties of Gaussian, we know $\boldsymbol{\mu}' = \mathbb{E}[\mathbf{x}]$ and $\Lambda = cov(\mathbf{x})$. Hence

$$\begin{aligned}\boldsymbol{\mu}' &= \mathbb{E}[\mathbf{x}] \\ &= \mathbb{E}[W\mathbf{z} + \boldsymbol{\mu} + \epsilon] \\ &= W(\mathbb{E}[\mathbf{z}] = 0) + (\mathbb{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}) + (\mathbb{E}[\epsilon] = 0) \\ &= \boldsymbol{\mu}\end{aligned}$$

$$\begin{aligned}\Lambda' &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \\ &= \mathbb{E}\left[(W\mathbf{z} + \epsilon)(W\mathbf{z} + \epsilon)^T\right] \\ &= \mathbb{E}\left[W\mathbf{z}\mathbf{z}^T W^T + W\mathbf{z}\epsilon^T + \epsilon W\mathbf{z}^T + \epsilon\epsilon^T\right] \\ &= W\mathbb{E}\left[\mathbf{z}\mathbf{z}^T\right] W^T + 0 + 0 + \mathbb{E}\left[\epsilon\epsilon^T\right] \\ &= W\mathbf{I}\mathbf{I}^T W^T + \sigma^2\mathbf{I} \\ &= WW^T + \sigma^2\mathbf{I}\end{aligned}$$

Hence, we can write $\mathbb{P}[\mathbf{x}] = \mathcal{N}(\boldsymbol{\mu}, C)$ where $C = WW^T + \sigma^2\mathbf{I}$

Part 2

Note: I will use $C = WW^T + \sigma^2\mathbf{I}$ for the rest of the question
Since all the sample points are independent, we can write $\mathbb{P}[X] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i\right]$. Therefore

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

Part 3

We can write the MLE estimate of $\boldsymbol{\mu}$ as

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

We can simply differentiate the RHS term, and using differentials of matrices, we get

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

Therefore, the MLE estimate of $\boldsymbol{\mu}$ is, as expected, the emperical mean of all the data points. Hence we use this to centralize for the PCA algorithm.