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

QUESTION 1

Assignment Number: 3
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: November 14, 2017

---

1. We are given

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

For any $\boldsymbol{\theta}^{\mathrm{i}}$ we have

$$\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{MLE}}\right] \geq \mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i}}\right]$$

$$\log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{MLE}}\right] \geq \log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i}}\right] \tag{1}$$

Now from slide 44 of Lec-16 we have

$$\log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i}}\right] \geq \mathbb{E}_{\mathbf{Z}\sim\mathbb{P}\left[\mathbf{Z}\mid\mathbf{X},\boldsymbol{\theta}^{\mathrm{MLE}}\right]}\log\left[\frac{\mathbb{P}\left[X,Z \mid \theta^{i}\right]}{\mathbb{P}\left[Z \mid X,\theta^{\mathrm{MLE}}\right]}\right] \tag{2}$$

Also from slide 42 of Lec-16 we have

$$\mathbb{E}_{\mathbf{Z}\sim\mathbb{P}\left[\mathbf{Z}\mid\mathbf{X},\boldsymbol{\theta}^{\mathrm{MLE}}\right]}\log\left[\frac{\mathbb{P}\left[X,Z \mid \theta^{\mathrm{MLE}}\right]}{\mathbb{P}\left[Z \mid X,\theta^{\mathrm{MLE}}\right]}\right] = \log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{MLE}}\right] \tag{3}$$

Using (1), (2) and (3) we get

$$\mathbb{E}_{\mathbf{Z}\sim\mathbb{P}\left[\mathbf{Z}\mid\mathbf{X},\boldsymbol{\theta}^{\mathrm{MLE}}\right]}\log\left[\frac{\mathbb{P}\left[X,Z \mid \theta^{\mathrm{MLE}}\right]}{\mathbb{P}\left[Z \mid X,\theta^{\mathrm{MLE}}\right]}\right] \geq \mathbb{E}_{\mathbf{Z}\sim\mathbb{P}\left[\mathbf{Z}\mid\mathbf{X},\boldsymbol{\theta}^{\mathrm{MLE}}\right]}\log\left[\frac{\mathbb{P}\left[X,Z \mid \theta^{i}\right]}{\mathbb{P}\left[Z \mid X,\theta^{\mathrm{MLE}}\right]}\right]$$

Since this is true for all or any $\boldsymbol{\theta}^{\mathrm{i}}$ thus we get

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

2. We have from previous result as both are MLE solutions

$$\boldsymbol{\theta}^{1} \in \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\ Q_{\boldsymbol{\theta}^{1}}(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}^{2} \in \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\ Q_{\boldsymbol{\theta}^{2}}(\boldsymbol{\theta})$$

Since $\boldsymbol{\theta}^{1}$ is optimal MLE solution we must have reached it let's say after i iterations then $\boldsymbol{\theta}^{\mathrm{i}} = \boldsymbol{\theta}^{1}$. Now from slide 44 of Lec-16 we have

$$\log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i+1}}\right] \geq Q_{\boldsymbol{\theta}^{1}}(\boldsymbol{\theta}^{\mathrm{i+1}}) \geq \log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i}}\right]$$

Since $\boldsymbol{\theta}^{1}$ is MLE solution we have

$$\mathbb{P}\left[X \mid \boldsymbol{\theta}^{1}\right] \geq \mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i+1}}\right]$$

$$\log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{1}\right] \geq \log\mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i+1}}\right]$$

Thus we get

$$\log \mathbb{P}\left[X \mid \boldsymbol{\theta}^1\right] = Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{\mathrm{i}+1}) = \log \mathbb{P}\left[X \mid \boldsymbol{\theta}^{\mathrm{i}+1}\right]$$

Hence for any $\boldsymbol{\theta}^{\mathrm{i}+1}$ which satisfies the constraint will maximize $Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{\mathrm{i}+1})$ as that will be the globally maximum possible value. Hence for any $\boldsymbol{\theta}^{\mathrm{i}+1}$ which is MLE solution $Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta}^{\mathrm{i}+1})$ will be maximum. Thus we can say

$$\boldsymbol{\theta}^{\mathrm{i}+1} \in \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} \ \log \mathbb{P}\left[X \mid \boldsymbol{\theta}\right]$$
$$\boldsymbol{\theta}^{\mathrm{i}+1} \in \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\left[X \mid \boldsymbol{\theta}\right]$$

Thus we get

$$\operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\left[X \mid \boldsymbol{\theta}\right] = \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta})$$

Since $\boldsymbol{\theta}^2$ is also a MLE solution. Therefore we can say

$$\boldsymbol{\theta}^2 \in \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta})$$

Similarly along the same lines we can show

$$\operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\left[X \mid \boldsymbol{\theta}\right] = \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^2}(\boldsymbol{\theta})$$

and then finally

$$\boldsymbol{\theta}^1 \in \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} Q_{\boldsymbol{\theta}^2}(\boldsymbol{\theta})$$

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

QUESTION 2

Assignment Number: 3
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: November 14, 2017

---

1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a piecewise linear function with n partitions $\{\Omega_1, \ldots, \Omega_n\}$ of $\mathbb{R}^d$ with n linear models $\mathbf{w}^1, \ldots, \mathbf{w}^n$, then we have

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

To prove that $c \cdot f(\mathbf{x})$ is also a piecewise linear we construct a funtion for any scalar $c \in \mathbb{R}$

$g : \mathbb{R}^d \to \mathbb{R}$ with n partitions $\{\Omega_1, \ldots, \Omega_n\}$ of $\mathbb{R}^d$ with n linear models $c \cdot \mathbf{w}^1, \ldots, c \cdot \mathbf{w}^n$, thus

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

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

$$g(\mathbf{x}) = c \cdot f(\mathbf{x})$$

Thus $c \cdot f(\mathbf{x})$ is also piecewise linear for partitions $\{\Omega_1, \ldots, \Omega_n\}$ of $\mathbb{R}^d$ with $\mathbf{w}^1, \ldots, \mathbf{w}^n$ as n linear models.

2. Let $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ be a piecewise linear functions with n partitions $\left\{\Omega_1^f, \ldots, \Omega_n^f\right\}$ and m partitions $\{\Omega_1^g, \ldots, \Omega_n^g\}$ of $\mathbb{R}^d$ respectively and with $\mathbf{w}^{1,f}, \ldots, \mathbf{w}^{n,f}$ and $\mathbf{w}^{1,g}, \ldots, \mathbf{w}^{n,g}$ as n and m linear models respectively, then we have

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

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

Now consider nm distinct partitions $\left\{\Omega_{11}^h, \ldots, \Omega_{nm}^h\right\}$ of $\mathbb{R}^d$

$$\Omega_{ij}^h = \Omega_i^f \cap \Omega_j^g \quad \forall i \in [n], j \in [m]$$

Now consider nm linear models

$$\mathbf{w}^{ij,h} = \mathbf{w}^{i,f} + \mathbf{w}^{j,g} \quad \forall i \in [n], j \in [m]$$

Thus function $h : \mathbb{R}^d \to \mathbb{R}$ is a piecewise linear function with these nm defined partitions and linear models.

$$\begin{aligned}
h(\mathbf{x}) &= \sum_{i,j}^{n,m} \mathbb{I}\left\{\mathbf{x} \in \Omega_{ij}^h\right\} \cdot \left\langle \mathbf{w}^{ij,h}, \mathbf{x} \right\rangle \\
&= \sum_{i,j}^{n,m} \mathbb{I}\left\{\mathbf{x} \in \Omega_{ij}^h\right\} \cdot \left\langle \mathbf{w}^{i,f} + \mathbf{w}^{j,g}, \mathbf{x} \right\rangle \\
&= \sum_{i,j}^{n,m} \mathbb{I}\left\{\mathbf{x} \in \Omega_{ij}^h\right\} \cdot \left\langle \mathbf{w}^{i,f}, \mathbf{x} \right\rangle + \sum_{i,j}^{n,m} \mathbb{I}\left\{\mathbf{x} \in \Omega_{ij}^h\right\} \cdot \left\langle \mathbf{w}^{j,g}, \mathbf{x} \right\rangle
\end{aligned}$$

Now we have $\mathbf{x} \in \mathbb{R}^d$ and $\bigcup_{i=1}^n \Omega_i^f = \bigcup_{i=1}^m \Omega_i^g = \bigcup_{i=1,j=1}^{i=n,j=m} \Omega_{ij}^h = \mathbb{R}^d$, thus

$$\begin{aligned}
\sum_{i,j}^{n,m} \mathbb{I}\left\{\mathbf{x} \in \Omega_{ij}^h\right\} \cdot \left\langle \mathbf{w}^{i,f}, \mathbf{x} \right\rangle &= \sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \Omega_i^f\right\} \cdot \left\langle \mathbf{w}^{i,f}, \mathbf{x} \right\rangle \\
\sum_{i,j}^{n,m} \mathbb{I}\left\{\mathbf{x} \in \Omega_{ij}^h\right\} \cdot \left\langle \mathbf{w}^{j,g}, \mathbf{x} \right\rangle &= \sum_{j=1}^{m} \mathbb{I}\left\{\mathbf{x} \in \Omega_j^g\right\} \cdot \left\langle \mathbf{w}^{j,g}, \mathbf{x} \right\rangle
\end{aligned}$$

Thus we get

$$\begin{aligned}
h(\mathbf{x}) &= \sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \Omega_i^f\right\} \cdot \left\langle \mathbf{w}^{i,f}, \mathbf{x} \right\rangle + \sum_{j=1}^{m} \mathbb{I}\left\{\mathbf{x} \in \Omega_j^g\right\} \cdot \left\langle \mathbf{w}^{j,g}, \mathbf{x} \right\rangle \\
h(\mathbf{x}) &= f(\mathbf{x}) + g(\mathbf{x})
\end{aligned}$$

Hence sum of two piecewise linear functions is piecewise linear.

3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a piecewise linear function with n partitions $\{\Omega_1, \ldots, \Omega_n\}$ of $\mathbb{R}^d$ with n linear models $\mathbf{w}^1, \ldots, \mathbf{w}^n$, then we have

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

Now $g(\mathbf{x}) = f_{\text{ReLU}}(f(\mathbf{x})) = \max(f(\mathbf{x}), 0)$. Consider partitions $\left\{\Omega_1^{'}, \ldots, \Omega_n^{'}, \Omega_{n+1}^{'}\right\}$ and linear models $\mathbf{w}^{1'}, \ldots, \mathbf{w}^{n'}, \mathbf{w}^{(n+1)'}$ where $\Omega_{n+1}^{'}$ is the region of $\mathbb{R}^d$ for which $f(\mathbf{x}) < 0$ ie.

$$\begin{aligned}
\mathbf{x} \in \Omega_{n+1}^{'} &\Leftrightarrow f(\mathbf{x}) < 0 \\
\mathbf{w}^{(n+1)'} &= \mathbf{0}
\end{aligned}$$

And all other $\Omega_i^{'}$ and $\mathbf{w}^{i'}$ are defined as

$$\begin{aligned}
\Omega_i^{'} &= \Omega_i \setminus \Omega_i \cap \Omega_{n+1}^{'} \quad \forall i \in [n] \\
\mathbf{w}^{i'} &= \mathbf{w}^i
\end{aligned}$$

Now $\left\langle \mathbf{w}^{i'}, \mathbf{x} \right\rangle \geq 0$. Thus now define $g : \mathbb{R}^d \to \mathbb{R}$ is also a piecewise linear function with partitions $\left\{\Omega_1^{'}, \ldots, \Omega_n^{'}, \Omega_{n+1}^{'}\right\}$ and linear models $\mathbf{w}^{1'}, \ldots, \mathbf{w}^{n'}, \mathbf{w}^{(n+1)'}$ and

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

Hence proved that $g(\mathbf{x}) = f_{\text{ReLU}}(f(\mathbf{x}))$ is also piecewise linear.

4. We can prove that any neural network with a ReLU activation function computes a piecewise linear function by using induction. Lets say we have intermediate layers $\{L_1, \ldots, L_n\}$. We know that output of L_i layer acts as input for L_{i+1} layer. Let's assume that there is edge from all node of previous layer to all nodes in the next layer. Let n_i is the number of nodes in the layer L_i $\forall i \in [n]$. Let's say w_{ij}^l be the weight of edge connecting i^{th} node of layer l to j^{th} node of next layer. Finally let's say that output of i^{th} node of layer l is y_i^l.

   Induction: Output of i^{th} node of layer l is piecewise linear $\forall i \in [n_l]$ ie. y_i^l is piecewise linear $\forall i \in [n_l]$.

   <u>Base Case</u> : Output of L_1 ie. first hidden layer is piecewise linear. Consider we have n input vectors and w_{ij} be the weight vectors from inputs to first hidden layer $\forall i \in [n], j \in [n_1]$

$$y_j^1(\mathbf{x}) = f_{\text{ReLU}}(\sum_{i=1}^{n} w_{ij} \cdot \mathbf{x}^i) \quad \forall j \in [n_1]$$

   Clearly $\mathbf{x}^i$ is linear function and from Part-1 we see $w_{ij} \cdot \mathbf{x}^i$ is piecewise linear as w_{ij} is scaler. Also from Part-2 we can say $\sum_{i=1}^{n} w_{ij} \cdot \mathbf{x}^i$ is also piecewise linear and atlast from Part-3 we can deduce that $f_{\text{ReLU}}(\sum_{i=1}^{n} w_{ij} \cdot \mathbf{x}^i)$ is also piecewise linear. Hence base case $y_j^1(\mathbf{x})$ is also piecewise linear.

   Now we just have to show that output of all nodes of layer $(l+1)$ is piecewise linear. Now

$$y_j^{l+1}(\mathbf{x}) = f_{\text{ReLU}}(\sum_{i=1}^{n_l} w_{ij}^l \cdot y_i^l(\mathbf{x})) \quad \forall j \in [n_{l+1}]$$

   - By induction $y_i^l(\mathbf{x})$ is piecewise linear and w_{ij}^l is scaler and from the Part-1 we have $w_{ij}^l \cdot y_i^l(\mathbf{x})$ as piecewise linear.
   - Now we have $w_{ij}^l \cdot y_i^l(\mathbf{x})$ as piecewise linear and from Part-2 sum of piecewise linear is piecewise linear hence we can say that $\sum_{i=1}^{n_l} w_{ij}^l \cdot y_i^l(\mathbf{x})$ is also piecewise linear.
   - Clearly from Part-3 we have $f_{\text{ReLU}}(\sum_{i=1}^{n_l} w_{ij}^l \cdot y_i^l(\mathbf{x}))$ as piecewise linear.

   Hence we proved that output of $(l+1)^{th}$ layer is piecewise linear if output of l^{th} layer is piecewise linear.

   Hence any neural network with a ReLU activation function computes a piecewise linear function.

5. It corresponds to O(dD) pieces.

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

Assignment Number: 3
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: November 14, 2017

QUESTION 3

Algorithm for Kernel Perceptron

Algorithm 1: Kernel Perceptron Algorithm

Input: Online data points

Empty set $S \leftarrow \phi$, a constant $\beta \leftarrow 0$
while Data points are coming **do**
  Receive a point $P^t = \left(\mathbf{x}^t, y^t\right)$
  if S is empty **then**
  - $\alpha^t \leftarrow y^t$
  - $\beta \leftarrow \beta + y^t$
  - $S \leftarrow S \cup \left(\mathbf{x}^t, \alpha^t\right)$

  else
    if $sgn\left(\sum_{(\mathbf{x}^j, \alpha^t) \in S} \alpha^j K\left(\mathbf{x}^j, \mathbf{x}^t\right) + \beta\right) \neq y^t$ **then**
      if $\mathbf{x}^t$ is in set S **then**
      - $\alpha^t \leftarrow \alpha^t + y^t$
      - $\beta \leftarrow \beta + y^t$

      else
      - $\alpha^t \leftarrow y^t$
      - $\beta \leftarrow \beta + y^t$
      - $S \leftarrow S \cup \left(\mathbf{x}^t, \alpha^t\right)$

      end if
    end if
  end if
end while

Output: S, β

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

QUESTION 4

Assignment Number: 3
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: November 14, 2017

---

1. Consider $\varphi : \mathbb{R}^2 \to \mathbb{R}^7$ such that

$$\begin{aligned}
\varphi(\mathbf{z}) &= [\varphi_0(\mathbf{z}), \varphi_1(\mathbf{z}), \varphi_2(\mathbf{z})] \\
\varphi_0(\mathbf{z}) &= [1] \\
\varphi_1(\mathbf{z}) &= \left[\sqrt{2} \cdot \mathbf{z}\right] \\
\varphi_2(\mathbf{z}) &= [z_1 z_1, z_1 z_2, z_2 z_1, z_2 z_2]
\end{aligned}$$

s.t. $\varphi_0 : \mathbb{R}^2 \to \mathbb{R}$, $\varphi_1 : \mathbb{R}^2 \to \mathbb{R}^2$, $\varphi_2 : \mathbb{R}^2 \to \mathbb{R}^4$, $\mathbf{z} = (z_1, z_2) \in \mathbb{R}^2$. Now

$$\begin{aligned}
\left\langle \varphi(\mathbf{z}^1), \varphi(\mathbf{z}^2) \right\rangle &= \left[\varphi_0(\mathbf{z}^1), \varphi_1(\mathbf{z}^1), \varphi_2(\mathbf{z}^1)\right] \cdot \begin{bmatrix} \varphi_0(\mathbf{z}^2)^T \\ \varphi_1(\mathbf{z}^2)^T \\ \varphi_2(\mathbf{z}^2)^T \end{bmatrix} \\
&= \left[1, \sqrt{2} \cdot \mathbf{z}^1, z_1^1 z_1^1, z_1^1 z_2^1, z_2^1 z_1^1, z_2^1 z_2^1\right] \cdot \begin{bmatrix} 1 \\ (\sqrt{2} \cdot \mathbf{z}^2)^T \\ z_1^2 z_1^2 \\ z_1^2 z_2^2 \\ z_2^2 z_1^2 \\ z_2^2 z_2^2 \end{bmatrix} \\
&= 1 + 2 \cdot \left\langle \mathbf{z}^1, \mathbf{z}^2 \right\rangle + \sum_{i,j}^{2} z_i^1 z_j^1 z_i^2 z_j^2 \\
&= (\left\langle \mathbf{z}^1, \mathbf{z}^2 \right\rangle + 1)^2 \\
&= K(\mathbf{z}^1, \mathbf{z}^2)
\end{aligned}$$

Thus $\varphi_K = \varphi$ is feature map of K. Hence the kernel K is Mercer with $\varphi : \mathbb{R}^2 \to \mathcal{H}_K$ where $\mathcal{H}_K \equiv \mathbb{R}^7$. We chose $D = 7$. We can see that $D = 6$ will also work.

2. Here we assume $\mathbf{A} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \in \mathbb{R}^{2\times 2}$, $\mathbf{b} = [b_1, b_2] \in \mathbb{R}^2$ and $c \in \mathbb{R}$. Now for every quadratic function $f_{(A,\mathbf{b},c)}$ and for $\mathbf{z} \in \mathbb{R}^2$ we have

$$\begin{aligned}
f_{(A,\mathbf{b},c)}(\mathbf{z}) &= \langle \mathbf{z}, A\mathbf{z} \rangle + \langle \mathbf{b}, \mathbf{z} \rangle + c \\
&= \mathbf{z}^T A^T \mathbf{z} + \langle \mathbf{b}, \mathbf{z} \rangle + c \\
&= (a_1 z_1 z_1 + a_2 z_1 z_2 + a_3 z_2 z_1) + a_4 z_2 z_2 + \langle \mathbf{b}, \mathbf{z} \rangle + c \\
&= (a_1 z_1 z_1 + a_2 z_1 z_2 + a_3 z_2 z_1 + a_4 z_2 z_2) + (b_1 z_1 + b_2 z_2) + c
\end{aligned}$$

Also for $\mathbf{w} \in \mathcal{H}_K$, $\mathbf{w} = [w_1, \ldots, w_7]$

$$\begin{aligned}
\langle \mathbf{w}, \varphi_K(\mathbf{z}) \rangle &= \langle [w_1, \ldots, w_7], [\varphi_0(\mathbf{z}), \varphi_1(\mathbf{z}), \varphi_2(\mathbf{z})] \rangle \\
&= (w_4 z_1 z_1 + w_5 z_1 z_2 + w_6 z_2 z_1 + w_7 z_2 z_2) + \sqrt{2} \cdot (w_2 z_1 + w_3 z_2) + w_1
\end{aligned}$$

Now for any given quadratic function $f_{(A,\mathbf{b},c)}$ we construct $\mathbf{w} \in \mathbb{R}^7$, $\mathbf{w} = [w_1, \ldots, w_7]$ as

$$\begin{aligned}
w_1 &= c \\
w_2 &= \frac{b_1}{\sqrt{2}} \\
w_3 &= \frac{b_2}{\sqrt{2}} \\
w_4 &= a_1 \\
w_5 &= a_2 \\
w_6 &= a_3 \\
w_7 &= a_4
\end{aligned}$$

such that for all $\mathbf{z} \in \mathbb{R}^2$

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

3. Now for every $\mathbf{w} \in \mathcal{H}_K$ we can construct a triplet $(A, \mathbf{b}, c) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \times \mathbb{R}$ as

$$\begin{aligned}
c &= w_1 \\
b_1 &= \sqrt{2} \cdot w_2 \\
b_2 &= \sqrt{2} \cdot w_3 \\
a_1 &= w_4 \\
a_2 &= w_5 \\
a_3 &= w_6 \\
a_4 &= w_7
\end{aligned}$$

such that for all $\mathbf{z} \in \mathbb{R}^2$

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

4. Kernel Ridge Regression problem is

$$\min_{\mathbf{w} \in \mathcal{H}_K} \sum_{i=1}^{n} (y^i - \left\langle \mathbf{w}, \varphi_K(\mathbf{z}^i) \right\rangle)^2 + \lambda \left\| \mathbf{w} \right\|^2$$

Lets say it gives us $\mathbf{w}^{'} \in \mathcal{H}_K$ as output model. Then we have for any $\mathbf{w} \in \mathcal{H}_K$

$$\sum_{i=1}^{n} (y^i - \left\langle \mathbf{w}^{'}, \varphi_K(\mathbf{z}^i) \right\rangle)^2 + \lambda \left\| \mathbf{w}^{'} \right\|^2 \leq \min_{\mathbf{w} \in \mathcal{H}_K} \sum_{i=1}^{n} (y^i - \left\langle \mathbf{w}, \varphi_K(\mathbf{z}^i) \right\rangle)^2 + \lambda \left\| \mathbf{w} \right\|^2$$

Using part-3 we can construct a quadratic function $f_{(A,\mathbf{b},c)}$ over $\mathbb{R}^2$.
Let's say it is $\hat{f}$ for $\mathbf{w}^{'}$ and f for any $\mathbf{w}$

$$\begin{aligned}
\hat{f}_{(A,\mathbf{b},c)}(\mathbf{z}) &= \left\langle \mathbf{w}^{'}, \varphi_K(\mathbf{z}) \right\rangle \\
f_{(A,\mathbf{b},c)}(\mathbf{z}) &= \langle \mathbf{w}, \varphi_K(\mathbf{z}) \rangle
\end{aligned}$$

Thus we have

$$\sum_{i=1}^{n} (y^i - \hat{f}(\mathbf{z}^i))^2 + \lambda \left\| \mathbf{w}^{'} \right\|^2 \leq \min_{(A,\mathbf{b},c) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \times \mathbb{R}} \sum_{i=1}^{n} (y^i - f_{(A,\mathbf{b},c)}(\mathbf{z}^i))^2 + \lambda \left\| \mathbf{w} \right\|^2$$

Since $\lambda \to 0^+$

$$\lambda \left\| \mathbf{w}^{'} \right\|^2 \geq 0$$

Hence

$$\sum_{i=1}^{n} (y^i - \hat{f}(\mathbf{z}^i))^2 \leq \min_{(A,\mathbf{b},c) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \times \mathbb{R}} \sum_{i=1}^{n} (y^i - f_{(A,\mathbf{b},c)}(\mathbf{z}^i))^2 + \lambda \left\| \mathbf{w} \right\|^2$$

Since λ is finite and so is $\left\| \mathbf{w} \right\|^2$ thus for some finite ϵ we can say

$$\epsilon = \lambda \left\| \mathbf{w} \right\|^2$$

and also $\epsilon \to 0$ as $\lambda \to 0^+$. Hence we can conclude

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

We have with $C = WW^\top + \sigma^2 \cdot I_d$

$$\mathbb{P}\left[\mathbf{x}\right] = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, C)$$
$$\mathbb{P}\left[\mathbf{x}\right] = \frac{1}{\sqrt{(2\pi)^d|C|}} exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T C^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right)$$

Now

$$\mathbb{P}\left[X \mid \boldsymbol{\mu}, W, \sigma\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\mu}, W, \sigma\right]$$
$$\log \mathbb{P}\left[X \mid \boldsymbol{\mu}, W, \sigma\right] = \sum_{i=1}^{n} \log \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\mu}, W, \sigma\right]$$
$$= \sum_{i=1}^{n}\left(\log \frac{1}{\sqrt{(2\pi)^d|C|}} - \frac{1}{2}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)^T C^{-1}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)\right)$$
$$= \left(-\frac{nd}{2}\log 2\pi - \frac{n}{2}\log|C| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)^T C^{-1}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)\right)$$

So the complete expression for the data log-likelihood $\mathbb{P}\left[X \mid \boldsymbol{\mu}, W, \sigma\right]$ is

$$\log \mathbb{P}\left[X \mid \boldsymbol{\mu}, W, \sigma\right] = \left(-\frac{nd}{2}\log 2\pi - \frac{n}{2}\log|C| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)^T C^{-1}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)\right)$$

Now for $\boldsymbol{\mu}^{\text{MLE}}$

$$\boldsymbol{\mu}^{\text{MLE}} = \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\max}\ \mathbb{P}\left[X \mid \boldsymbol{\mu}, W, \sigma\right]$$
$$\boldsymbol{\mu}^{\text{MLE}} = \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\max} \log \mathbb{P}\left[X \mid \boldsymbol{\mu}, W, \sigma\right]$$
$$\boldsymbol{\mu}^{\text{MLE}} = \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\max} \sum_{i=1}^{n}\left(\log \frac{1}{\sqrt{(2\pi)^d|C|}} - \frac{1}{2}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)^T C^{-1}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)\right)$$
$$\boldsymbol{\mu}^{\text{MLE}} = \underset{\boldsymbol{\mu}\in\mathbb{R}^d}{\arg\max}\left(-\frac{n}{2}\log|C| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)^T C^{-1}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)\right)$$
$$L = -\frac{n}{2}\log|C| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)^T C^{-1}\left(\mathbf{x}^i-\boldsymbol{\mu}\right)(let)$$
$$\frac{\partial L}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\left(\frac{\partial\left(\sum_{i=1}^{n}\mathbf{x}^i - n\boldsymbol{\mu}\right)^T C^{-1}\left(\sum_{i=1}^{n}\mathbf{x}^i - n\boldsymbol{\mu}\right)}{\partial \boldsymbol{\mu}}\right)$$
$$\frac{\partial L}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\left(-2C^{-1}\left(\sum_{i=1}^{n}\mathbf{x}^i - n\boldsymbol{\mu}\right)\right)$$

Using first order optimality

$$\frac{\partial L}{\partial \boldsymbol{\mu}} = 0$$
$$\sum_{i=1}^{n} \mathbf{x}^i - n\boldsymbol{\mu} = 0$$
$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^i$$

Thus we get following expresion

$$\boldsymbol{\mu}^{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^i$$

Thus $\boldsymbol{\mu}^{\text{MLE}}$ is just the mean of data points.