In the name of God

Sharif University of Technology
School of Electrical Engineering
Deep Learning - 25647
Dr. Fatemizadeh
Fall Semester 1400

Homework 1 - Section 1

Amirhossein Asadian - 96101187

Problem 1.

As wikipedia says: In Vapnik–Chervonenkis theory, the Vapnik–Chervonenkis (VC) dimension is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a set of functions that can be learned by a statistical binary classification algorithm. It is defined as the cardinality of the largest set of points that the algorithm can shatter, which means the algorithm can always learn a perfect classifier for any labeling of that many data points.

VC dimension of a classification model: A binary classification model f with some parameter vector θ is said to shatter a set of data points $(x_1, x_2, ..., x_n)$ if, for all assignments of labels to those points, there exists a θ such that the model f makes no errors when evaluating that set of data points.

The VC dimension of a model f is the maximum number of points that can be arranged so that f shatters them. More formally, it is the maximum cardinal D such that some data point set of cardinality D can be shattered by f.

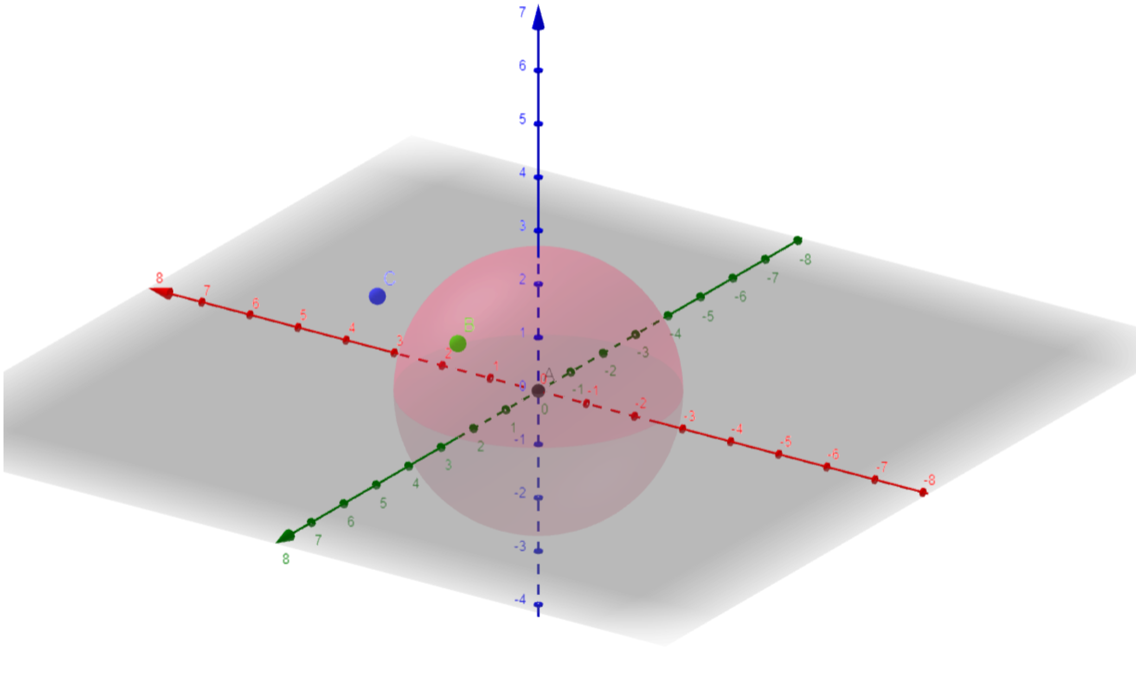

Figure 1: Green point: {(1, 1, 1): -1}, Blue point: {(2, 2, 2): +1}

(a) The VC dimension is 1.
First, we show that there exists 1 point, $x_1 \in \mathbb{R}^3$ that can be shattered. Pick (1, 1, 1). Consider the following sub-cases: $(r = \sqrt{(-\theta)})$
* The point is positive. Choose any origin-centered sphere that includes the point $(r < \sqrt{3})$
* The point is negative. Choose any origin-centered sphere that does not include the point $(r > \sqrt{3})$.
We now need to show that an origin-centered sphere cannot shatter 2 points. Pick (1, 1, 1) as +1 and (2, 2, 2) as -1. An origin-centered sphere cannot shatter these two points because $r_1 < r_2$. Each origin-centered sphere of radius r, divide the space into two parts, -1 and +1. Moreover, radius of class +1 is greater than radius of class -1. Therefore, we cannot shatter two points of different classes while the point of class -1 is closer to the origin (Figure 1).

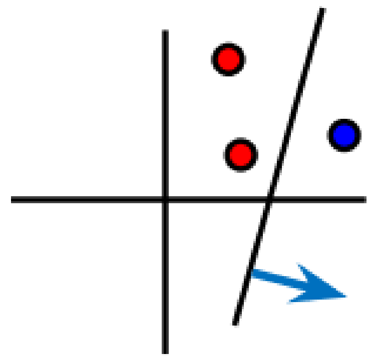

Figure 2: A two-dimensional line for splitting 3 points.

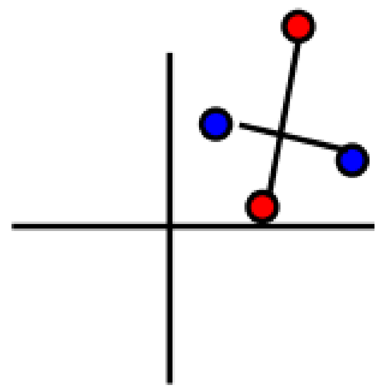

Figure 3: Two-dimensional lines for splitting 4 points.

(b) The VC dimension is 3.
First, we show that there exist 3 points, $x_1, x_2, x_3 \in \mathbb{R}^2$ that can be shattered.
* All of the points are positive or All of the points are negative: We can have a line which separates the points correctly (The points will be in one side of the line).
* Two points are positive and 1 point is negative or Two points are negative and 1 point is positive. Figure 2 shows that we can easily draw a line, which can separates the points correctly.
We now need to show that a two-dimensional line cannot shatter 4 points. Assume 4 points in 2 pairs. Any line through these points must split one pair (by crossing one of the lines) (Figure 3).

(c) We can show that the VC dimension of hyperplanes in m dimensions is m + 1.
Proof: Let $\mathcal{H}$ be the set of hyperplanes in m dimensions. First, we show that there exists a set S of m+1 points $\in \mathbb{R}^m$ shattered by $\mathcal{H}$.

Suppose $S = \{x_1, \ldots, x_m, x_{m+1}\}$, where x_i is a point $\in \mathbb{R}^m$, and our hyperplane is represented as $y = w^\top x + w_0$. Let $y_1, \ldots, y_m, y_{m+1}$ be any set of labels assigned to the m+1 points and construct the following linear system:

$$w^\top x_1 + w_0 = y_1 w^\top x_2 + w_0 = y_2 \ldots w^\top x_{m+1} + w_0 = y_{m+1}$$

.

Notice that the above linear system as having $m+1$ variables $w_0, \ldots, w_m$ and $m+1$ equations, hence it must have a solution as long as S satisfies the condition that $(1, x_1), \ldots, (1, x_m), (1, x_{m+1})$ are linearly independent.

Hence by choosing S s.t. the matrix $\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{m+1} \end{pmatrix}$ is full-rank, we can always solve for a m-dimensional hyperplane with bias term that separates the $m+1$ points in S. Since $y_1, \ldots, y_m, y_{m+1}$ can be any set of labels, the m-dimensional hyperplane shatters the $m+1$ points in S.

Secondly, we show that there exists no set S' of $m+2$ points can be shattered by $\mathcal{H}$.

Suppose to the contrary that $S' = \{x_1, \ldots, x_{m+1}, x_{m+2}\}$ can be shattered. This implies that there exist 2^{m+2} weight vectors $w^{(1)}, \ldots, w^{(2^{m+2})}$ such that the matrix of inner products denoted by $z_{i,j} = x_i^\top w^{(j)}$ has columns with all possible combination of signs (note here x_i contains the constant feature and $w^{(j)}$ contains the bias term). We use A to denote this matrix and

$$A = \begin{pmatrix} z_{1,1} & \ldots & z_{1,2^{m+2}} \\ \vdots & \ldots & \vdots \\ z_{m+2,1} & \ldots & x_{m+2,2^{m+2}} \end{pmatrix} s.t.\ sign(A) = \begin{pmatrix} - & - & \ldots & - & + \\ - & \cdot & \ldots & \cdot & + \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ - & + & \ldots & + & + \end{pmatrix}$$

Then the rows of A are linearly independent because there exists a column of A with the same signs which does not sum to zero hence there are no constants $c_1, c_2, \ldots, c_{m+2}$ such that $\sum_{i=1}^{m+2} c_i z_{i,:} = 0$. However, notice that row i of A can be written as $x_i^\top W$ where $W = [w^{(1)}, \ldots, w^{(2^{m+2})}]$. By linear algebra knowledge we know that $m+2$ vectors $\in \mathbb{R}^{m+1}$, are always linearly dependent (i.e. $x_1^\top, \ldots, x_{m+2}^\top$ are linearly dependent). Hence $x_1^\top W, \ldots, x_{m+2}^\top W$ should also be linearly dependent, which results in a contradiction. This contradiction proves that there are no m+2 points $\in \mathbb{R}^m$ which can be shattered by hyperplanes in m dimension. Thus, the VC dimension of $\mathcal{H}$ is $m+1$. [University of Pennsylvania, CIS250 wiki, 2018]

Problem 2.

The log likelihood of our model is:

$$log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \sum_{i=1}^{N} log\, p(y_i|\mathbf{x}_{\boldsymbol{i}}, \theta)$$

But since the noise ϵ is Gaussian, the likelihood is just:

$$\begin{aligned} log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \sum_{i=1}^{N} log\, N(y_i; \mathbf{x}_{\boldsymbol{i}}\mathbf{w}, \sigma^2) \\ &= \sum_{i=1}^{N} log\, \frac{1}{\sqrt{2\pi\sigma_e^2}} exp(-\frac{(y_i - \mathbf{x}_{\boldsymbol{i}}\mathbf{w})^2}{2\sigma_e^2}) \\ &= -\frac{N}{2}\, log\, 2\pi\sigma_e^2 - \sum_{i=1}^{N} \frac{(y_i - \mathbf{x}_{\boldsymbol{i}}\mathbf{w})^2}{2\sigma_e^2}) \end{aligned}$$

So:

$$\begin{aligned} \mathbf{w}_{MLE} &= \underset{w}{\text{argmax}} \; - \sum_{i=1}^{N} (y_i - \mathbf{x}_{\boldsymbol{i}}\mathbf{w})^2 \\ &= \underset{w}{\text{argmin}} \; \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{x}_{\boldsymbol{i}}\mathbf{w})^2 \\ &= \underset{w}{\text{argmin}} \; MSE_{train} \end{aligned}$$

That is, the parameters w chosen to maximise the likelihood are exactly those chosen to minimise the mean-squared error.

Problem 3.

We assume that the likelihood function ($\mathcal{L}$) is the Gaussian itself.

$$\begin{aligned}\mathcal{L} = p(\mathbf{X}|\theta) &= \mathcal{N}(\mathbf{X}|\theta)\\ &= \mathcal{N}(\mathbf{X}|\mu, \Sigma)\end{aligned}$$

Therefore, for MLE of a Gaussian model, we will need to find good estimates of both parameters: µ and Σ:

$$\begin{aligned}\mu_{MLE} &= \underset{\mu}{\operatorname{argmax}} \mathcal{N}(\mathbf{X}|\mu, \Sigma)\\ \Sigma_{MLE} &= \underset{\Sigma}{\operatorname{argmax}} \mathcal{N}(\mathbf{X}|\mu, \Sigma)\end{aligned}$$

For simplicity we will use log likelihood (the log() function is monotonically increasing). Now we want to get the best parameters $\theta = [\mu, \Sigma]$ for a dataset $\mathbf{X}$ evaluating on a Gaussian distribution.

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \log(\mathcal{N}(\mathbf{X}|\theta))$$

$$\begin{aligned}\mathcal{LL} = \log(\mathcal{N}(\mathbf{X}|\theta)) &= \sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\theta))\\ &= \sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\mu, \Sigma))\\ &= \sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\mu, \sigma^2))\\ &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp^{-\frac{1}{2}\left(\frac{(x_n-\mu)^2}{\sigma^2}\right)}\right)\\ &= \sum_{n=1}^{N} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp^{-\frac{1}{2}\left(\frac{(x_n-\mu)^2}{\sigma^2}\right)}\right)\right)\\ &= \sum_{n=1}^{N} \left(\log(1) - \log\left(\sqrt{2\pi\sigma^2}\right) + \log\left(\exp^{-\frac{1}{2}\left(\frac{(x_n-\mu)^2}{\sigma^2}\right)}\right)\right)\end{aligned}$$

$$= \sum_{n=1}^{N} \Big(\log(1) - \log\big(\sqrt{2\pi\sigma^2}\big) + \Big(-\frac{1}{2}\big(\frac{(x_n-\mu)^2}{\sigma^2}\big) \cdot \log(e) \Big) \Big)$$
$$= \sum_{n=1}^{N} \Big(-\log\big(\sqrt{2\pi\sigma^2}\big) + \Big(-\frac{1}{2}\big(\frac{(x_n-\mu)^2}{\sigma^2}\big) \Big) \Big)$$
$$= \sum_{n=1}^{N} \Big(-\frac{1}{2} \cdot \log(2\pi\sigma^2) - \frac{1}{2}\Big(\frac{(x_n-\mu)^2}{\sigma^2}\Big) \Big)$$
$$= -\frac{N}{2}\log(2\pi\sigma^2) + \sum_{n=1}^{N} -\frac{1}{2}\Big(\frac{(x_n-\mu)^2}{\sigma^2}\Big)$$
$$= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

So, now we're going to solve the problem for each variable one-by-one:

$$\underset{\mu}{\operatorname{argmax}} \mathcal{LL}(X|\mu,\sigma^2)$$
$$\underset{\sigma^2}{\operatorname{argmax}} \mathcal{LL}(X|\mu,\sigma^2)$$

MLE of μ:

$$\mathcal{LL} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$
$$\underset{\mu}{\operatorname{argmax}} \mathcal{LL}(X|\mu,\sigma^2) := \frac{\partial \mathcal{LL}}{\partial \mu} = 0$$

$$\frac{\partial \mathcal{LL}}{\partial \mu} = \frac{\partial}{\partial \mu}\Big(-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 \Big)$$
$$= \frac{\partial}{\partial \mu}\Big(-\frac{N}{2}\log(2\pi\sigma^2) \Big) + \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 \Big)$$
$$= 0 + \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 \Big)$$
$$= \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 \Big)$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \mu}\Big(\sum_{n=1}^{N} -\frac{1}{2\sigma^2}(x_n-\mu)^2\Big) \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}(x_n-\mu)^2\Big) \\
&= \sum_{n=1}^{N} \Big(\frac{\partial}{\partial \mu}\big(-\frac{1}{2\sigma^2}\big)\cdot(x_n-\mu)^2 + \big(-\frac{1}{2\sigma^2}\big)\cdot\frac{\partial}{\partial \mu}(x_n-\mu)^2\Big) \\
&= \sum_{n=1}^{N} \Big(0 + \big(-\frac{1}{2\sigma^2}\big)\cdot\frac{\partial}{\partial \mu}(x_n-\mu)^2\Big) \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} \frac{\partial}{\partial \mu}(x_n-\mu)^2 \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(x_n-\mu)\cdot -1 \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = 0 \\
0 &= \sum_{n=1}^{N}(x_n-\mu) \\
0 &= \sum_{n=1}^{N} x_n - \sum_{n=1}^{N}\mu \\
0 &= \sum_{n=1}^{N} x_n - N\cdot\mu \\
N\cdot\mu &= \sum_{n=1}^{N} x_n \\
\mu &= \frac{1}{N}\sum_{n=1}^{N} x_n
\end{aligned}
$$

MLE of σ^2:

$$
\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\Big(- \frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= \frac{\partial}{\partial \sigma^2}\Big(- \frac{N}{2}\log(2\pi\sigma^2)\Big) + \frac{\partial}{\partial \sigma^2}\Big(- \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= \frac{\partial}{\partial \sigma^2}\big(- \frac{N}{2}\big) \cdot \log(2\pi\sigma^2) + \big(- \frac{N}{2}\big) \cdot \frac{\partial}{\partial \sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial \sigma^2}\Big(- \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= 0 + \big(- \frac{N}{2}\big) \cdot \frac{\partial}{\partial \sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial \sigma^2}\Big(- \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= -\frac{N}{2} \cdot \frac{\partial}{\partial \sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial \sigma^2}\Big(- \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= -\frac{N}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 2\pi + \frac{\partial}{\partial \sigma^2}\Big(- \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= -\frac{N}{2} \cdot \frac{1}{\sigma^2} + \frac{\partial}{\partial \sigma^2}\Big(- \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= -\frac{N}{2\sigma^2} + \frac{\partial}{\partial \sigma^2}\Big(- \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(- \frac{1}{2\sigma^2}(x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(- \frac{1}{2\sigma^2}\big) \cdot (x_n - \mu)^2\big) + \big(- \frac{1}{2\sigma^2}\big) \cdot \frac{\partial}{\partial \sigma^2}(x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(- \frac{1}{2\sigma^2}\big) \cdot (x_n - \mu)^2\big) + 0\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(- \frac{1}{2\sigma^2}\big) \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(- \frac{1}{2} \cdot \sigma^{-2}\big) \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(- \frac{1}{2}\big) \cdot \sigma^{-2} + \big(- \frac{1}{2} \cdot \frac{\partial}{\partial \sigma^2}\sigma^{-2}\big) \cdot (x_n - \mu)^2\big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N} \Big(0 + \big(-\frac{1}{2} \cdot \frac{\partial}{\partial \sigma^2}\sigma^{-2}\big) \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N} \Big(-\frac{1}{2} \cdot \frac{\partial}{\partial \sigma^2}\sigma^{-2} \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N} \Big(-\frac{1}{2} \cdot \frac{\partial}{\partial \sigma^2}\big((\sigma^2)^{-1}\big) \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N} \Big(-\frac{1}{2} \cdot -1 \cdot (\sigma^2)^{-2} \cdot 1 \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N} \Big(\frac{1}{2} \cdot (\sigma^2)^{-2} \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N} \Big(\frac{1}{2\sigma^4} \cdot (x_n - \mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 \\
&= \frac{1}{2\sigma^2}\Big(-N + \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2\Big) \\
0 &= \frac{1}{2\sigma^2}\Big(-N + \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2\Big) \\
0 &= -N + \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \\
N\sigma^2 &= \sum_{n=1}^{N} (x_n - \mu)^2 \\
\sigma^2 &= \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2
\end{aligned}
$$

To conclude:

$$
\begin{aligned}
\mu_{MLE} &= \frac{1}{N} \sum_{n=1}^{N} x_n \\
\sigma^2_{MLE} &= \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2
\end{aligned}
$$

Problem 4.

A) Using the least squares method we want to find parameter values that minimizes the residual sum of squares (RSS):

$$
\begin{aligned}
RSS = \sum_{i=1}^{n} {r_i}^2 &= \sum_{i=1}^{n} (y_i - x_i W)^2 \\
&= ||y - XW||^2 \quad \textit{where} \quad X \in \mathbb{R}^{n \times (p+1)} \quad \textit{and} \quad y \in \mathbb{R}^n, W \in \mathbb{R}^{p+1} \\
&= (y - XW)^\top (y - XW)
\end{aligned}
$$

which leads to:

$$
\hat{W} = \underset{W}{\operatorname{argmin}} (y - XW)^\top (y - XW)
$$

Differentiate RSS this with respect to β:

$$
\begin{aligned}
RSS &= (y - XW)^\top (y - XW) \\
&= (y^\top - W^\top X^\top)(y - XW) \\
&= y^\top y - y^\top XW - W^\top X^\top y + W^\top X^\top XW
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial RSS}{\partial W} &= \frac{\partial (y^\top y - y^\top XW - W^\top X^\top y + W^\top X^\top XW)}{\partial W} \\
&= 0 - X^\top y - X^\top y + (X^\top X + (XX^\top)^\top)W \\
&= -2X^\top y + 2X^\top XW
\end{aligned}
$$

This first derivative should equal to 0. So,

$$
\begin{aligned}
-2X^\top y + 2X^\top XW &= 0 \\
X^\top XW &= X^\top y \\
W &= (X^\top X)^{-1} X^\top y
\end{aligned}
$$

B) L2 regularization:

$$
\begin{aligned}
\hat{W} &= \underset{W}{\operatorname{argmin}}(y - XW)^\top (y - XW) + \lambda W^\top W \\
\frac{\partial RSS}{\partial W} &= \frac{\partial (y^\top y - y^\top XW - W^\top X^\top y + W^\top X^\top XW + \lambda W^\top W)}{\partial W} \\
&= 0 - X^\top y - X^\top y + (X^\top X + (XX^\top)^\top)W + 2\lambda W \\
&= -2X^\top y + 2X^\top XW + 2\lambda W
\end{aligned}
$$

This first derivative should equal to 0. So,

$$
\begin{aligned}
-2X^\top y + 2X^\top XW + 2\lambda W &= 0 \\
(X^\top X + \lambda I)W &= X^\top y \\
W &= (X^\top X + \lambda I)^{-1} X^\top y
\end{aligned}
$$