CS 182: Introduction to Machine Learning, Fall 2021
Homework 1

(Due on Monday, Oct. 11 at 11:59pm (CST))

Notice:

- Please submit your assignments via Gradescope. The entry code is KYJ626.
- Please make sure you select your answer to the corresponding question when submitting your assignments.
- Each person has a total of five days to be late without penalty for all the homeworks. Each late delivery less than one day will be counted as one day.

1. [20 points]

(a) Given a set of observation pairs $\{(x_i, y_i)\}_{i=1}^N$, where $x_i, y_i \in \mathbb{R}$, $i = 1, 2, ..., N$. By assuming the linear model is a reasonable approximation, we consider to fit the model via the least squares method. Thus, our goal is to estimate the coefficients $\hat{\omega}_0$ and $\hat{\omega}_1$ to minimize the residual sum of squares (RSS),

$$[\hat{\omega}_0,\ \hat{\omega}_1] = \underset{\omega_0,\ \omega_1}{\operatorname{argmin}} \sum_{i=1}^{N} [y_i - (\omega_1 x_i + \omega_0)]^2. \tag{1}$$

Please show that

$$\begin{cases} \hat{\omega}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}, \\ \hat{\omega}_0 = \bar{y} - \hat{\omega}_1 \bar{x}, \end{cases} \tag{2}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$ denote the sample means. [10 points]

- **Solution:** Firstly, we compute ω_0.

$$\begin{aligned} &\tfrac{\partial}{\partial \omega_0} \textstyle\sum_{i=1}^N (y_1 - \omega_0 - \omega_1 x_i)^2 = \sum_{i=1}^N -2(y_i - \omega_0 - \omega_1 x_i) = 0 \\ &\Rightarrow \textstyle\sum_{i=1}^N (y_i - \omega_1 x_i) = \sum_{i=1}^N \omega_0 = N\omega_0 \\ &\Rightarrow \omega_0 = \tfrac{1}{N}\textstyle\sum_{i=1}^N (y_i - \omega_1 x_i) = \tfrac{1}{N}\sum_{i=1}^N y_i - \omega_1 \tfrac{1}{N}\sum_{i=1}^N x_i = \bar{y} - \omega_1 \bar{x} \\ &\Rightarrow \hat{\omega}_0 = \bar{y} - \hat{\omega_1}\bar{x} \end{aligned}$$

Plug ω_0 into $\sum_{i=1}^N (y_1 - \omega_0 - \omega_1 x_i)^2$ and differentiate with ω_1 ,

$$\begin{aligned} &\tfrac{\partial}{\partial \omega_1} \textstyle\sum_{i=1}^N (y_1 - \omega_0 - \omega_1 x_i)^2 = \tfrac{\partial}{\partial \omega_1}\sum_{i=1}^N (y_1 - \bar{y} + \omega_1\bar{x} - \omega_1 x_i)^2 = \sum_{i=1}^N 2[y_i - \bar{y} + \omega_1(\bar{x} - x_i)](\bar{x} - x_i) = 0 \\ &\Rightarrow \textstyle\sum_{i=1}^N (y_i - \bar{y})(\bar{x} - x_i) = -\omega_1 \sum_{i=1}^N (\bar{x} - x_i)^2 \\ &\Rightarrow \hat{\omega_1} = \tfrac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} \end{aligned}$$

In conclusion,

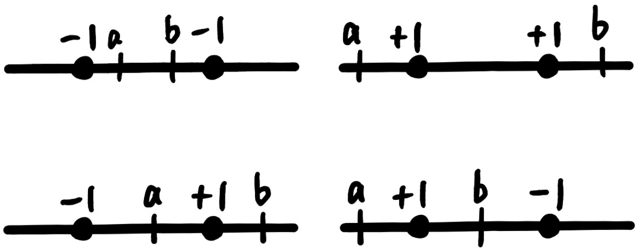
$$\hat{\omega_1} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}, \quad \hat{\omega_0} = \bar{y} - \hat{\omega_1}\bar{x}$$

(b) Assume now we want to classify the examples $\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}$, $i = 1, \cdots, N$ and the hypothesis class is

$$\mathcal{H}(x) = \begin{cases} 1 & a \le x \le b, \quad a, b \in \mathbb{R}, a < b \\ 0 & otherwise. \end{cases} \tag{3}$$

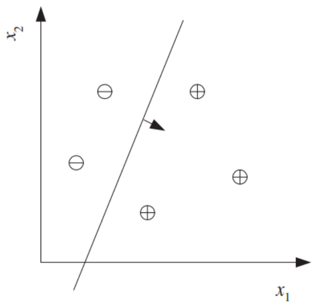
What is the VC dimension of $\mathcal{H}$ and why? (You need to show that if the VC dimension is k, k points can be shattered but $k+1$ points cannot. See P40 of Lecture 01.) [5 points]

- **Solution:** 2. We first show that 2 points can be shattered.

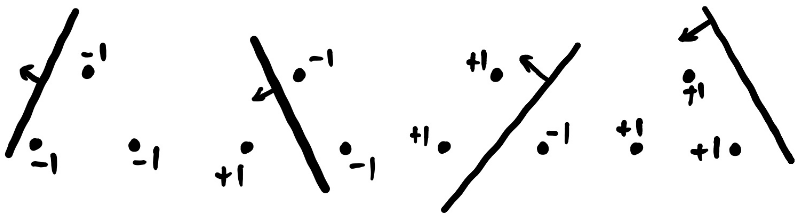


Then we show that 3 points cannot be shattered.

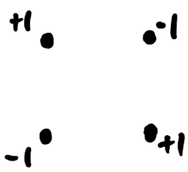
(c) Assume now the examples $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^2$, $i = 1, \cdots, N$ and the hypothesis class is the set of lines. What is the VC dimension of $\mathcal{H}$ and why? [5 points]



- **Solution:** 3. We first show that 3 points can be shattered.



Then we show that 4 points cannot be shattered.



2. [20 points] Suppose we have a two-class recognition problem with ω_1 and ω_2. The $p(x|\omega_i)$ follows normal distribution such that

$$p(x|\omega_i) \sim \mathcal{N}(\mu_i, \sigma^2) \tag{4}$$

and $p(\omega_i)$ is known. Suppose we have $\mu_2 > \mu_1$.

(a) Write the discriminant functions $g_i(x)$ and the classification rule. [10 points]

(b) Derive the boundary of the decision regions. [10 points]

Solution:

(a) The discriminant functions are

$$g_i(x) = \ln(p(x|\omega_i)p(\omega_i))\; i = 1, 2.$$

Define $g(x) = g_1(x) - g_2(x)$. The classification rule is

$$\text{choose} \begin{cases} \omega_1, & g(x) > 0, \\ \omega_2, & \text{otherwise.} \end{cases}$$

(b) Since

$$g(x) = -\frac{1}{2\sigma^2}(-2(\mu_1 - \mu_2)x + (\mu_1^2 - \mu_2^2)) + \ln\frac{p(\omega_1)}{p(\omega_2)},$$

and $\mu_2 > \mu_1$, when $x = \frac{\sigma^2(\ln\frac{p(\omega_1)}{p(\omega_2)})}{\mu_2 - \mu_1} + \frac{\mu_1 + \mu_2}{2}$, we have $g(x) = 0$. The boundary of the decision regions is

$$x = \frac{\sigma^2(\ln\frac{p(\omega_1)}{p(\omega_2)})}{\mu_2 - \mu_1} + \frac{\mu_1 + \mu_2}{2}.$$

3. [20 points] Given a set of observations $\{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}$, $i = 1, 2, ..., N$. Assume $\{x_i\}_{i=1}^N \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$, where σ, θ_0 and σ_0 are known constants.

(a) Derive the MLE of θ. [6 points]

Solution:

For normal distribution we have

$$P(x \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}.$$

The likelihood function is

$$\mathcal{L}(x_1, \ldots, x_N \mid \theta, \sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} e^{\sum_{i=1}^N -\frac{(x_i-\theta)^2}{2\sigma^2}},$$

and the log-likelihood function is

$$\log(\mathcal{L}) = -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \sum_{i=1}^N \frac{(x_i - \theta)^2}{2\sigma^2}.$$

The derivative of $\log(\mathcal{L})$ w.r.t. θ is

$$\frac{\partial \log(\mathcal{L})}{\partial \theta} = \sum_{i=1}^N \frac{(x_i - \theta)}{\sigma^2},$$

by setting which to zero we can get the MLE of θ as

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^N x_i.$$

(b) Derive the MAP of θ. [7 points]

Solution:

The MAP of θ can be obtained by maximizing

$$\begin{aligned}
\log\left(\mathcal{L}(x_1, \ldots, x_N \mid \theta, \sigma^2) P(\theta \mid \theta_0, \sigma_0^2)\right) &= -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \sum_{i=1}^N \frac{(x_i-\theta)^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}}\right) \\
&= -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \sum_{i=1}^N \frac{(x_i-\theta)^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right) - \frac{(\theta-\theta_0)^2}{2\sigma_0^2},
\end{aligned}$$

whose derivative w.r.t. θ is

$$\sum_{i=1}^N \frac{(x_i - \theta)}{\sigma^2} - \frac{(\theta - \theta_0)}{\sigma_0^2}.$$

By setting the above derivative to zero, the MAP of θ can be obtained as follows:

$$\theta_{\text{MAP}} = \frac{\frac{\sum_{i=1}^N x_i}{\sigma^2} + \frac{\theta_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}.$$

(c) Derive the Bayes' estimator of θ. [7 points]

Solution:

Observe that

$$\begin{aligned}
P(\theta \mid x_1, \ldots, x_N) &= \mathcal{L}(x_1, \ldots, x_N \mid \theta, \sigma^2) P(\theta \mid \theta_0, \sigma_0^2). \\
&= \frac{1}{(2\pi)^{\frac{N+1}{2}} \sigma^N \sigma_0} e^{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2} - \sum_{i=1}^N \frac{(x_i-\theta)^2}{2\sigma^2}} \\
&= \frac{1}{(2\pi)^{\frac{N+1}{2}} \sigma^N \sigma_0} e^{-\frac{\sigma^2(\theta-\theta_0)^2 + \sigma_0^2 \sum_{i=1}^N (x_i-\theta)^2}{2\sigma_0^2\sigma^2}} \\
&= \frac{1}{(2\pi)^{\frac{N+1}{2}} \sigma^N \sigma_0} e^{-\frac{\sigma^2\theta^2 + \sigma^2\theta_0 - 2\sigma^2\theta\theta_0 + \sigma_0^2 N\theta^2 + \sigma_0^2\sum_{i=1}^N x_i^2 - \sigma_0^2 2\sum_{i=1}^N x_i\theta}{2\sigma_0^2\sigma^2}} \\
&= ke^{-\frac{\left(\theta - \frac{\sigma^2\theta_0 + \sigma_0^2\sum_{i=1}^N x_i}{\sigma^2 + \sigma_0^2 N}\right)^2}{2\sigma_1^2}},
\end{aligned}$$

where k and σ_1 are constants.

The Bayes' estimator of θ is given by

$$\theta_{\text{Bayes'}} = \mathbb{E}\left[\theta \mid x_1, \ldots, x_N\right] = \frac{\sigma^2\theta_0 + \sigma_0^2 \sum_{i=1}^N x_i}{\sigma^2 + \sigma_0^2 N}.$$

4. [20 points] Given a set of observation pairs $\{(x_i, y_i)\}_{i=1}^N$, where $x_i, y_i \in \mathbb{R}$, $i = 1, 2, ..., N$. By assuming the polynomial model is a reasonable approximation, we consider to fit the model via the least squares estimate. Consider the polynomial regression function of order k:

$$g(x_i|\omega_0, \cdots, \omega_k) = \sum_{j=0}^{k} \omega_j x_i^j. \tag{5}$$

Define $\boldsymbol{\omega} = [\omega_0, \cdots, \omega_k]^T$. Show that the least squares estimate of $\boldsymbol{\omega}$ (assuming that $\mathbf{A}^T\mathbf{A}$ is invertible) is

$$\hat{\boldsymbol{\omega}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}, \tag{6}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^k \end{bmatrix}, \tag{7}$$

$$\hat{\boldsymbol{\omega}} = [\hat{\omega}_0, \cdots, \hat{\omega}_k]^T \tag{8}$$

and

$$\mathbf{y} = [y_1, \cdots, y_N]^T. \tag{9}$$

Solution: We can rewrite (5) as matrix form

$$g(x_i \mid \omega_0, \cdots, \omega_k) = \mathbf{A}^{(i)}\omega,$$

where $\mathbf{A}^{(i)}$ stands for the i-th row of $\mathbf{A}$. Then the error function is

$$\mathcal{E} = \frac{1}{2}\sum_{i=1}^{N}(y_i - g(x_i \mid \omega_0, \cdots, \omega_k))^2 = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\omega\|_2^2.$$

Letting the gradient equal to $\mathbf{0}$, we can get the result

$$\nabla\mathcal{E} = \mathbf{A}^T(\mathbf{y} - \mathbf{A}\omega) = 0 \Rightarrow \hat{\omega} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}.$$

5. [20 points] Given a set of observations $\{x_i\}_{i=1}^N$ that are drawn i.i.d. from a Poisson distribution

$$P(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

with parameter $\lambda > 0$.

(a) Derive the MLE of λ and determine whether it is unbiased or not. [10 points]

Solution:

The likelihood function is computed as

$$\mathcal{L}(x_1, \ldots, x_N \mid \lambda) = \prod_{i=1}^N \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-N\lambda} \frac{\lambda^{\sum_{i=1}^N x_i}}{\prod_{i=1}^N x_i!},$$

and the log-likelihood function is given by

$$\log(\mathcal{L}) = -N\lambda + \log(\lambda) \sum_{i=1}^N x_i - \sum_{i=1}^N \log(x_i!).$$

Then the detivative of $\log(\mathcal{L})$ w.r.t. λ is

$$\frac{\partial \log(\mathcal{L})}{\partial \lambda} = -N + \frac{1}{\lambda} \sum_{i=1}^N x_i,$$

by setting which to zero we can get the MLE of λ:

$$\hat{\lambda} = \frac{1}{N} \sum_{i=1}^N x_i.$$

Since $\{x_i\}_{i=1}^N$ are drawn from a Poisson distribution, we have

$$\mathbb{E}[x_i] = \lambda,$$

hence the mean of $\hat{\lambda}$ is

$$\mathbb{E}\left[\hat{\lambda}\right] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i\right] = \lambda,$$

which means it is unbiased.

(b) Derive the MLE of $\eta = e^{-2\lambda}$ and determine whether it is unbiased or not. [10 points]

Solution:

We can get $\lambda = -\frac{1}{2}\log(\eta)$, hence the log-likelihood function is given by

$$\log(\mathcal{L}) = \frac{N}{2}\log(\eta) + \log(-\frac{1}{2}\log(\eta)) \sum_{i=1}^N x_i - \sum_{i=1}^N \log(x_i!).$$

The derivative of $\log(\mathcal{L})$ w.r.t. η is

$$\frac{\partial \log(\mathcal{L})}{\partial \lambda} = \frac{N}{2\eta} + \frac{1}{\eta \log(\eta)} \sum_{i=1}^N x_i,$$

by setting which to zero we can get the MLE of η:

$$\hat{\eta} = e^{-\frac{2}{N}\sum_{i=1}^N x_i}.$$

The mean of $\hat{\eta}$ is

$$
\begin{aligned}
\mathbb{E}\left[\hat{\eta}\right] &= \mathbb{E}\left[e^{-\frac{2}{N}\sum_{i=1}^{N} x_i}\right] \\
&= \left(\mathbb{E}\left[e^{-\frac{2}{N} x_i}\right]\right)^N \\
&= \left(\sum_{z\geq 0} e^{-\frac{2}{N} z} \frac{\lambda^z e^{-\lambda}}{z!}\right)^N \\
&= e^{-N\lambda}\left(\sum_{z\geq 0} \frac{(\lambda e^{-\frac{2}{N}})^z}{z!}\right)^N \\
&= e^{-N\lambda}\left(\sum_{z\geq 0} \frac{(\lambda e^{-\frac{2}{N}})^z}{z!}\right)^N.
\end{aligned}
$$

Using Taylor series for exponential functions we can get

$$
\mathbb{E}\left[\hat{\eta}\right] = e^{-N\lambda} e^{N\lambda e^{-\frac{2}{N}}} = e^{N\lambda\left(e^{-\frac{2}{N}}-1\right)}.
$$

Therefore, the MLE of η is biased and the bias is

$$
e^{N\lambda\left(e^{-\frac{2}{N}}-1\right)} - e^{-2\lambda}.
$$