Introduction to Machine Learning, Fall 2023
Homework 1

(Due Thursday, Oct. 26 at 11:59pm (CST))

October 25, 2023

1. [10 points] [Math review] Suppose $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ are random samples from a random variable $\mathbf{X}$:

   (a) Prove that the covariance of $\mathbf{X}$ is a semi positive definite matrix. [3 points]

   (b) Assuming $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ which is a multivariate normal distribution, derive the the log-likelihood $l(\mu, \mathbf{\Sigma})$ and MLE of μ [4 points]

   (c) Suppose $\hat{\theta}$ is an unbiased estimator of θ and $\mathbf{Var}(\hat{\theta}) > 0$. Prove that $(\hat{\theta})^2$ is not an unbiased estimator of θ^2. [3 points]

(a) $\mathbf{\Sigma} = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]$
Suppose that the dimension of $\mathbf{X}$ is p.
So $\forall \mathbf{y} \in \mathbb{R}^p$,
$\mathbf{y}^T \mathbf{\Sigma} \mathbf{y} = \mathbf{y}^T E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]\mathbf{y} = E[\mathbf{y}^T(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T y]$
$= E[((\mathbf{X} - \mu)^T \mathbf{y})^T((\mathbf{X} - \mu)^T \mathbf{y})] = E(\|(\mathbf{X} - \mu)^T \mathbf{y}\|_2^2)$.
since $\|(\mathbf{X} - \mu)^T \mathbf{y}\|_2^2 \geq 0$, so $E(\|(\mathbf{X} - \mu)^T \mathbf{y}\|_2^2) \geq 0$
so $\forall \mathbf{y} \in \mathbb{R}^p, \mathbf{y}^T \mathbf{\Sigma} \mathbf{y} \geq 0$
So $\mathbf{\Sigma}$ is a semi positive definite matrix.
So above all, the covariance of $\mathbf{X}, \mathbf{\Sigma}$ is a semi positive definite matrix.

(b) From what we have learned in class, the PDF of the multivariate normal distribution $\mathbf{X}_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ is that $Pr(\mathbf{X}_i; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu))$, suppose that the dimension of $\mathbf{X}_i$ is p.
Since the sampling are independent, so the likelihood function is:
$Pr(\mathbf{X}_1, \cdots, \mathbf{X}_n; \mu, \mathbf{\Sigma}) = \prod_{i=1}^{n} Pr(\mathbf{X}_i; \mu, \mathbf{\Sigma})$
Then the log-likelihood function is:
$l(\mu, \mathbf{\Sigma}) = \sum_{i=1}^{n} \log(\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu)))$
$= n \cdot \log(\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}) - \sum_{i=1}^{n} \frac{1}{2}(\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu)$

And the MLE of μ is:
$\hat{\mu} = \underset{\mu}{\operatorname{argmax}}\, Pr(\mathbf{X}_1, \cdots, \mathbf{X}_n; \mu, \mathbf{\Sigma}) = \underset{\mu}{\operatorname{argmax}}\, l(\mu, \mathbf{\Sigma})$
Since $l(\mu, \mathbf{\Sigma})$ is a concave function, so we can get the optimal solution by setting the derivative of $l(\mu, \mathbf{\Sigma})$ to 0.
$\frac{\partial l(\mu, \mathbf{\Sigma})}{\partial \mu} = \frac{\partial n \cdot \log(\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}})}{\partial \mu} - \frac{\partial \sum_{i=1}^{n} \frac{1}{2}(\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu)}{\partial \mu}$
$= \sum_{i=1}^{n} -\frac{\partial \frac{1}{2}(\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu)}{\partial \mu}$

Since $\mathbf{\Sigma}$ is the covariance matrix, so $\mathbf{\Sigma}$ is a symmetric matrix, i.e.$\mathbf{\Sigma} = \mathbf{\Sigma}^T$.
So $(\mathbf{\Sigma}^{-1})^T = (\mathbf{\Sigma}^T)^{-1} = \mathbf{\Sigma}^{-1}$.
i.e. $\mathbf{\Sigma}$ is also a symmetric matrix.

So $\frac{\partial l(\mu, \mathbf{\Sigma})}{\partial \mu} = \sum_{i=1}^{n} -\frac{\partial \frac{1}{2}(\mathbf{X}_i - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}_i - \mu)}{\partial \mu}$

$= \sum_{i=1}^{n} -\frac{\partial \frac{1}{2}(\mathbf{X}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{X}_i-\mu)}{\partial(\mathbf{X}_i-\mu)}\frac{\partial(\mathbf{X}_i-\mu)}{\partial\mu}$
$= \sum_{i=1}^{n} -\frac{1}{2}(2\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu))(-1)$
$= \sum_{i=1}^{n} \mathbf{\Sigma}^{-1}(\mathbf{X}_i-\mu)$
$= \mathbf{\Sigma}^{-1}(\sum_{i=1}^{n} \mathbf{X}_i - n\mu)$

So $\frac{\partial l(\mu,\mathbf{\Sigma})}{\partial\mu} = 0 \Rightarrow \mathbf{\Sigma}^{-1}(\sum_{i=1}^{n} \mathbf{X}_i - n\mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$

So above all, the log-likelihood function is $l(\mu,\mathbf{\Sigma}) = n \cdot \log(\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}) - \sum_{i=1}^{n}\frac{1}{2}(\mathbf{X}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{X}_i-\mu)$

And the MLE of μ is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$

(c) Since $\hat{\theta}$ is the unbiased estimator of θ, so $E(\hat{\theta}) = \theta$.
And from the defination of varaiance, we could get that $Var(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2] = E[(\hat{\theta})^2] - [E(\hat{\theta})]^2$.
Since $Var(\hat{\theta}) > 0$, so $E[(\hat{\theta})^2] - [E(\hat{\theta})]^2 > 0$
i.e. $E[(\hat{\theta})^2] > [E(\hat{\theta})]^2 = (\hat{\theta})^2$
So $E[(\hat{\theta})^2] \neq (\hat{\theta})^2$
So $(\hat{\theta})^2$ is not an unbiased estimator of θ^2.

So above all, we have proved that if $\hat{\theta}$ is an unbiased estimator of θ and $Var(\hat{\theta}) > 0$, then $(\hat{\theta})^2$ is not an unbiased estimator of θ^2.

2. [10 points] Consider real-valued variables X and Y, in which Y is generated conditional on X according to

$$Y = aX + b + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Here ϵ is an independent variable, called a noise term, which is drawn from a Gaussian distribution with mean 0, and variance σ^2. This is a single variable linear regression model, where a is the only weight parameter and b denotes the intercept. The conditional probability of Y has a distribution $p(Y|X, a, b) \sim \mathcal{N}(aX + b, \sigma^2)$, so it can be written as:

$$p(Y|X, a, b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y - aX - b)^2\right).$$

(a) Assume we have a training dataset of n i.i.d. pairs (x_i, y_i), $i = 1, 2, ..., n$, and the likelihood function is defined by $L(a, b) = \prod_{i=1}^n p(y_i|x_i, a, b)$. Please write the Maximum Likelihood Estimation (MLE) problem for estimating a and b. [3 points]

(b) Estimate the optimal solution of a and b by solving the MLE problem in (a). [4 points]

(c) Based on the result in (b), argue that the learned linear model $f(X) = aX + b$, always passes through the point $(\bar{x}, \bar{y})$, 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. [3 points]

(a) the MLE of a and b is:

$\hat{a}, \hat{b} = \underset{a,b}{\operatorname{argmax}}\, L(a, b) = \underset{a,b}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|x_i, a, b) = \underset{a,b}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2)$

So above all, the MLE problem for estimating a and b is:

$\hat{a}, \hat{b} = \underset{a,b}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2)$

(b) Take log to the likelihood function, we could get:

$\hat{a}, \hat{b} = \underset{a,b}{\operatorname{argmax}} \sum_{i=1}^{n} \log(p(y_i|x_i, a, b)) = \underset{a,b}{\operatorname{argmax}} \sum_{i=1}^{n} \log(\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2))$

Since $\frac{1}{\sqrt{2\pi}\sigma}$ has nothing with a, b, and σ is just the variance of the noise term, so $\frac{1}{\sqrt{2\pi}\sigma}, -\dfrac{1}{2\sigma^2}$ are just constants.

so $\hat{a}, \hat{b} = \underset{a,b}{\operatorname{argmax}} \sum_{i=1}^{n} -(y_i - ax_i - b)^2 = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - ax_i - b)^2$

Since $\sum_{i=1}^{n} (y_i - ax_i - b)^2$ is a convex function both for a and b, so we just need to set the derivative of $\sum_{i=1}^{n} (y_i - ax_i - b)^2$ to 0 to get the optimal solution.

Let $f(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2, \bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i, \bar{y} = \dfrac{1}{n} \sum_{i=1}^{n} y_i$

So $\frac{\partial f}{\partial b} = \sum_{i=1}^{n} -2(y_i - ax_i - b) = 2nb - 2\sum_{i=1}^{n}(y_i - ax_i)$

$\dfrac{\partial f}{\partial b} = 0 \Rightarrow 2nb = 2\sum_{i=1}^{n}(y_i - ax_i) \Rightarrow b = \dfrac{1}{n}\sum_{i=1}^{n} y_i - \dfrac{1}{n} a \sum_{i=1}^{n} x_i$

$\Rightarrow b = \bar{y} - a\bar{x}$

Similarly, $\frac{\partial f}{\partial a} = \sum_{i=1}^{n} -2x_i(y_i - ax_i - b) = (-2)\sum_{i=1}^{n} x_i y_i - (-2)\sum_{i=1}^{n} a x_i^2 - (-2)\sum_{i=1}^{n} b x_i$

$\dfrac{\partial f}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0$

put $b = \bar{y} - a\bar{x}$ into the above equation, we could get:

$\Rightarrow \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - (\bar{y} - a\bar{x})\sum_{i=1}^{n} x_i = 0$

$$\Rightarrow a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

And put $a = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$ into $b = \bar{y} - a\bar{x}$, we could get:

$$b = \bar{y} - \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}\bar{x} = \frac{\sum_{i=1}^{n} x_i^2 \bar{y} - \sum_{i=1}^{n} x_i y_i \bar{x}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

So above all, the optimal solution of a and b is:

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

$$b = \bar{y} - a\bar{x} = \frac{\sum_{i=1}^{n} x_i^2 \bar{y} - \sum_{i=1}^{n} x_i y_i \bar{x}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

(c) From the analysis in (b), we could get that:

$$\frac{\partial f}{\partial b} = 2nb - 2\sum_{i=1}^{n}(y_i - ax_i) = 0 \Rightarrow b = \bar{y} - a\bar{x}$$

i.e. $b = \bar{y} - a\bar{x}$.

Put $(\bar{x}, \bar{y})$ into the linear model $f(X) = aX + b$, we could get:

$f(\bar{x}) = a\bar{x} + b = a\bar{x} + \bar{y} - a\bar{x} = \bar{y}$

So above all, the learned linear model $f(X) = aX + b$ always passes through the point $(\bar{x}, \bar{y})$.

3. [10 points] [Regression and Classification]

(a) When we talk about linear regression, what does 'linear' regard to? [2 points]

(b) Assume that there are n given training examples $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, where each input data point x_i has m real valued features. When $m > n$, the linear regression model is equivalent to solving an under-determined system of linear equations $\mathbf{y} = \mathbf{X}\beta$. One popular way to estimate β is to consider the so-called ridge regression:

$$\underset{\beta}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2$$

for some $\lambda > 0$. This is also known as Tikhonov regularization.
Show that the optimal solution β_* to the above optimization problem is given by

$$\beta_* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Hint: You need to prove that given $\lambda > 0$, $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is invertible. [5 points]

(c) Is the given data set linear separable? If yes, construct a linear hypothesis function to separate the given data set. If no, explain the reason. [3 points]

Data	(1,3)	(4,4)	(3,-6)	(-2,1)	(-3,5)	(-6,-4)
Label	+1	-1	-1	+1	-1	-1

(a) Linear is to the for all parameters of the regression variable β.

(b) As we have learned in linear algebra, we know that tha matrix X^TX must be similiar and diagonalizable. i.e. there must exist a matrix P and a diagonal matrix Λ such that $X^TX = P\Lambda P^{-1}$.
Also $\forall x \in \mathbb{R}^n$, we have $x^T(\mathbf{X}^T\mathbf{X})x = (\mathbf{X}x)^T(\mathbf{X}x) = \|\mathbf{X}x\|_2^2 \geq 0$.
So X^TX is positive semi-definite.
So all eigenvalues of X^TX are non-negative.
i.e. the diagonal matrix Λ's elements are all positive.
And since $\lambda > 0$, so λI's all elements are all also non-negative, and λI is also a diagonal matrix.
So $X^TX + \lambda I = P\Lambda P^{-1} + \lambda PIP^{-1} = P(\Lambda + \lambda I)P^{-1}$.
Since $\Lambda, \lambda I$ are all diagonal matrix, so $\Lambda + \lambda I$ is also a diagonal matrix.
And all elements in $\Lambda + \lambda I$ are all positive, this is because in $Lambda$, elements are non-nagetiva, in λI, all elements are positive. So $\Lambda + \lambda I$ is positive defined.
Since $X^TX + \lambda I = P(\Lambda + \lambda I)P^{-1}$, from the knowledge of similarity and diagonalizable, we could know that $X^TX + \lambda I$ is also positive defined.
So $X^TX + \lambda I$ is invertible.
And let $f(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta = \mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda\beta^T\beta$
Since $f(\beta)$ is convex, so we just need to set the derivative of $f(\beta)$ to 0 to get the optimal solution.
$\frac{\partial f(\beta)}{\partial\beta} = 2(-\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\beta + \lambda\beta)$
$\frac{\partial f(\beta)}{\partial\beta} = 0 \Rightarrow (\mathbf{X}^\mathbf{X} + \lambda I)\beta = \mathbf{X}^T\mathbf{y} \Rightarrow \beta = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$

Since we have proved that $X^TX + \lambda I$ is invertible, so $(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}$ exists. So $\beta* = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$

So above all, the optimal solution $\beta* = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$.

(c) Let the data be formed as $\mathbf{X}_i = (x_1, x_2)$.
And let the hypothesis function be $f(\mathbf{X}) = b_1x_1^2 + b_2x_2^2 + b_3$.
Let $b_1 = 1, b_2 = 1, b_3 = -25$, so the regression function is $f(\mathbf{X}) = x_1^2 + x_2^2 - 25$, and its a linear regression.
Make the separate line be $f(\mathbf{X}) = 0$, so the separate line is $x_1^2 + x_2^2 - 25 = 0$.
If $f(\mathbf{X}) \leq 0$, set the label to be $+1$, else set the label to be -1.
And we could get the result as below:

Data $\mathbf{X}$	(1,3)	(4,4)	(3,-6)	(-2,1)	(-3,5)	(-6,-4)
Function value $f(\mathbf{X})$	-15	7	20	-20	9	27
Label	+1	-1	-1	+1	-1	-1

So above all, we could find that we can construct a linear hypothesis function to separate the given data set.