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) $\mathbf{\Sigma} = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T]$
So $\forall y \in \mathbb{R}^n$,
$y^T\mathbf{\Sigma}y = E[y^T(\mathbf{X}-\mu)(\mathbf{X}-\mu)^Ty] = E[((\mathbf{X}-\mu)^Ty)^T((\mathbf{X}-\mu)^Ty)] = E(\|(\mathbf{X}-\mu)^Ty\|_2^2)$.
since $\|(\mathbf{X}-\mu)^Ty\|_2^2 \geq 0$, so $E(\|(\mathbf{X}-\mu)^Ty\|_2^2) \geq 0$
so $\forall y \in \mathbb{R}^n, y^T\mathbf{\Sigma}y \geq 0$
So $\mathbf{\Sigma}$ is a semi positive definite matrix.

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

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.
$\frac{\partial l(\mu,\mathbf{\Sigma})}{\partial \mu} = \frac{\partial b \cdot \log(\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}})}{\partial \mu} - \frac{\partial \sum\limits_{i=1}^{n} \dfrac{1}{2}(\mathbf{X}_i-\mu)^T\mathbf{\Sigma}(\mathbf{X}_i-\mu)}{\partial \mu}$
$= \sum\limits_{i=1}^{n} -\frac{\partial \dfrac{1}{2}(\mathbf{X}_i-\mu)^T\mathbf{\Sigma}(\mathbf{X}_i-\mu)}{\partial \mu}$

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

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

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

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

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

(c)

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)

(b)

(c)

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

(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)