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 $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ form a random sample from a multivariate distribution:
   (a) Prove that the covariance of $\mathbf{X}_i$ is a semi positive definite matrix. [3 points]
   (b) Assuming $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ which is a multivariate normal distribution, and samples X, 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)

$$\mathbf{Cov}(X) = E[(X - E[X])^T (X - E[X])]$$

Let $\vec{a}$ is a vector. Then

$$\vec{a}^T \mathbf{Cov}(X)\vec{a} = \vec{a}^T E[(X-E[X])^T (X-E[X])]\vec{a} = E[\vec{a}^T (X-E[X])^T (X-E[X])\vec{a}] = E[m^2] = \mathbf{Var}(m) > 0$$

$$\text{where } m = (X - E[X])\vec{a}$$

$$\text{Therefore } \vec{a}^t \mathbf{Col}(X)\vec{a} > 0$$

Then, the covariance is a semi positive definite matrix.

(b)

$$l(\mu, \Sigma) = \sum_{i=1}^{N} \log Pr_\theta(X_i) = \sum_{i=1}^{N} \log \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_i-\mu)^T \Sigma^{-1}(X_i-\mu)}$$

$$= -\log(2\pi)^{\frac{N}{2}} - \log|\Sigma|^{\frac{1}{2}} - \sum_{i=1}^{N} \frac{1}{2}(X_i - \mu)^T \Sigma^{-1}(X_i - \mu)$$

$$\frac{\partial l(\mu, \Sigma)}{\partial \mu} = \frac{1}{\Sigma} \sum_{i=1}^{N} (X_i - \mu) = 0 \therefore \text{When } \mu = \frac{\sum_{i=1}^{n} X_i}{N}, \text{ the log-likelihood has its maximum value}$$

(c) Because $\hat{\theta}$ is unbiased, then we can say $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i = E[\hat{\theta}]$

Then, $\hat{\theta}^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} \hat{\theta}_i\right) \neq \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}^2 = e\hat{\theta}^2$

Then the $\hat{\theta}^2$ is not unbiased.

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)

$$l(a, b) = \log L(a, b) = \sum_{i=1}^{n} \log Pr_\theta(y_i|x_i, a, b) = -\frac{1}{n}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Then find the minimum of $l(a, b)$

(b) By the conclusion of 1(b), when $b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$, the MLE of a, b get the maximum value.

$$\frac{\partial l(a, b)}{\partial a} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0$$

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

$$\Leftrightarrow \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + \frac{a}{n}(\sum_{i=1}^{n} x_i)^2 = 0$$

$$\Leftrightarrow n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = a\left(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right)$$

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

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

(c) Because $b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} = \bar{y} - a\bar{x}$, therefore $\hat{f}(\bar{x}) = \hat{a}\bar{x} - \hat{b} = 0$, which means the linear regression function always 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:

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

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) The relation we supposed between y and x is linear, such as $y = ax + b$

(b)

$$\frac{\partial\, \text{argmin}_\beta}{\partial\beta} = X^TX\beta + \lambda\beta - X^TY = 0$$

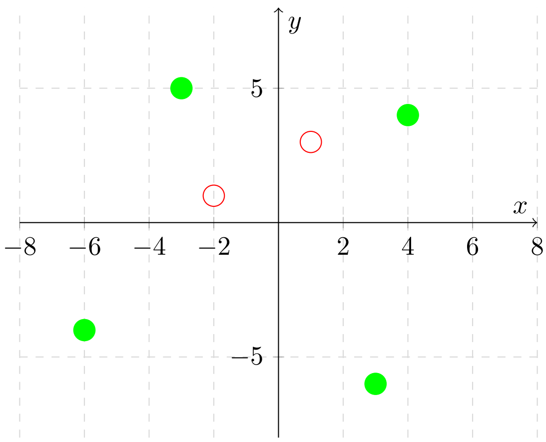
$$\text{By SVD decomposition, } X^TX = (U^TDV)^T(U^TDV) = (V^TDU)(U^TDV) = V^TD^2V$$

$$\therefore X^TX + \lambda I = V^TD^2V + \lambda IV^TV = V^T(D^2 + \lambda I)V$$

$$\because I > 0 \therefore D^2 + \lambda I \text{ is not a sigular matrix.} \therefore X^TX + \lambda I \text{ is invertible}$$

$$\therefore \beta = (X^TX + \lambda I)^{-1}X^TY$$

(c) We can draw the graph:



So we can easily know that we can not given a linear separable.