Introduction to Machine Learning, Fall 2023
Homework 1

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

October 11, 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 $\mathbf{X}_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ which is a multivariate normal distribution, and samples X_i, 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) Let $X = (X_1, X_2, \cdots, X_n)^T$ be a random n dimensional vector.

let μ_i and σ_i^2 be the mean and variance of X_i,

and σ_{ij} be the covariance between X_i and X_j for $i \neq j$,

So the covariance matrix is a symmetric $n \times n$ matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} = E\left[(X-\mu)^T (X-\mu)\right], \quad \mu = (\mu_1, \mu_2, \cdots \mu_n)^T$$

For all $a \in R^n$

$$a^T \Sigma a = E\left[a^T (X-\mu)(X-\mu)^T a\right]$$
$$= E\left[(a^T (X-\mu))^2\right] \geq 0$$

So the covariance of X_i is a semi positive definite matrix.

(b) The PDF of the multivariate normal distribution

is $f(X_i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X_i-\mu)^T \Sigma^{-1}(X_i-\mu)\right)$

so the log-likelihood

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

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N} \Sigma^{-1}(X_i-\mu) = 0$$

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

(c) Since $\hat{\theta}$ is an unbiased estimator of θ, $E(\hat{\theta}) = \theta$

we need to prove $E(\hat{\theta}^2) \neq \theta^2$

$Var(\hat{\theta}^2) = (Var(\hat{\theta}))^2 > 0$

$E(\hat{\theta}^2) = Var(\hat{\theta}) + (E(\hat{\theta}))^2 = Var(\hat{\theta}) + \theta^2$

since $Var(\hat{\theta}) > 0$

$E(\hat{\theta}^2) > \theta^2$

so $\hat{\theta}^2$ 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) $\arg\max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - ax_i)^2\right)$

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

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

(c) We know that $\hat{y} = \hat{a}x_i + \hat{b}$

plug $(\bar{x}, \bar{y})$ into the equation

$\bar{y} = \hat{a}\bar{x} + \hat{b} = \hat{a}\bar{x} + \bar{y} - \hat{a}\bar{x} = \bar{y}$

so 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) It refers to the relationship between independent and dependent variables. A linear relationship means that a change in the independent variables is associated with a constant change in the dependent variables.

(b) when $\lambda > 0$, let v be a non zero vector

the $v^T(X^TX + \lambda I)v = v^TX^TXv + \lambda v^Tv = ||Xv||_2^2 + \lambda ||v||_2^2 > 0$

so $X^TX + \lambda I$ is invertible when $\lambda > 0$

define $f(\beta) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$

then $f'(\beta) = -2X^T(y - X\beta) + 2\lambda\beta$,

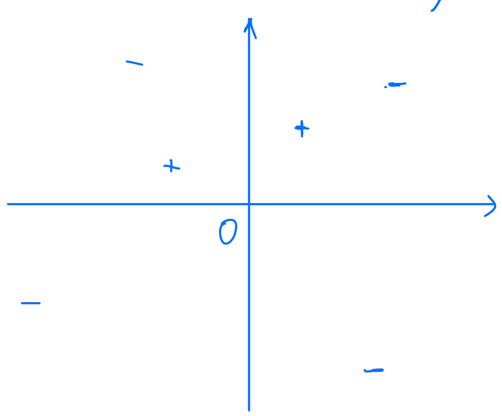
$f''(\beta) = 2X^TX + 2\lambda I > 0$

thus the optimal solution β_* is $f'(\beta) = 0$

$\beta(2X^TX + 2\lambda) - 2X^Ty = 0$

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

(C)



As the plot shows, it doesn't exist a single straight line to seperate the +1 and -1 labelled points. So the given data set is not linear separable.