Introduction to Machine Learning, Fall 2023

Homework 1

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

October 13, 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) As X_i's covariance matrix can be expressed as $E[(X_i - E(X_i))(X_i - E(X_i))^T]$

For arbitrary column vector a, we want to calculate

$$a \cdot E[(X_i - E(X_i))(X_i - E(X_i))^T] a^T$$

$$= E[a(X_i - E(X_i))(X_i - E(X_i))^T a^T]$$

Because $(X_i - E(X_i))^T a^T = (a(X_i - E(X_i)))^T$,

it can be rewritten as $E[(a(X_i - E(X_i))) \cdot (a(X_i - E(X_i)))^T]$,

which means that all elements in $[\,]$ are bigger or equal to 0.

Thus the expected value of it is bigger or equal to 0.

Thus the covariance of X_i is a semi positive definite matrix

(b) $$l(\mu, \Sigma) = -\frac{n}{2}\log(2\pi) - n\log\sqrt{\Sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\frac{d}{d\mu} l(\mu, \Sigma) = \frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{i=1}^{n}(X_i - \mu) = 0$$

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

(c) As $\hat{\theta}$ is unbiased.

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i$$

$$E(\hat{\theta}^2) = \frac{1}{n^2}\left(\sum_{i=1}^{n} \hat{\theta}_i\right)^2 \neq \frac{1}{n^2}\sum_{i=1}^{n} \theta_i^2 = E(\theta^2).$$

Thus $\hat{\theta}^2$ is biased.

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) Finding a and b to make $l(a,b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2\right)$ reach the maximum

(b) $\log l(a,b) = -\frac{n}{2}\log(2\pi) - n\cdot\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (ax_i + b))^2$

$$\frac{d}{da}\log l(a,b) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2x_i(-ax_i + y_i - b) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(-ax_i + y_i - b) \quad \cdots ①$$

$$\frac{d}{db}\log l(a,b) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(y_i - ax_i - b) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-ax_i + y_i - b) \quad \cdots ②$$

$$\begin{cases} ① = 0 \\ ② = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} x_i(-ax_i + y_i - b) = 0 \cdots ③ \\ \sum_{i=1}^{n}(-ax_i + y_i - b) = 0 \cdots ④ \end{cases} \Rightarrow \begin{cases} a = \dfrac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) - n\sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i)^2 - n\cdot\sum_{i=1}^{n} x_i^2} \\ b = -a\cdot\bar{x} + \bar{y} \\ = -\dfrac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) - n\sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i)^2 - n\cdot\sum_{i=1}^{n} x_i^2}\cdot\dfrac{\sum_{i=1}^{n} x_i}{n} + \dfrac{\sum_{i=1}^{n} y_i}{n} \end{cases}$$

③ $\Leftrightarrow -a\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i y_i - b\cdot\sum_{i=1}^{n} x_i = 0$

④ $\Leftrightarrow -a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i - n\cdot b = 0$.

$\Leftrightarrow -a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i = b$.

(c) From ④ in (b), we know that $-a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i = b \Leftrightarrow -a\cdot\bar{x} + \bar{y} = b$

$\Leftrightarrow \bar{y} = a\bar{x} + b$

which means that point $(\bar{x}, \bar{y})$ is always on $f(x)$

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:

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

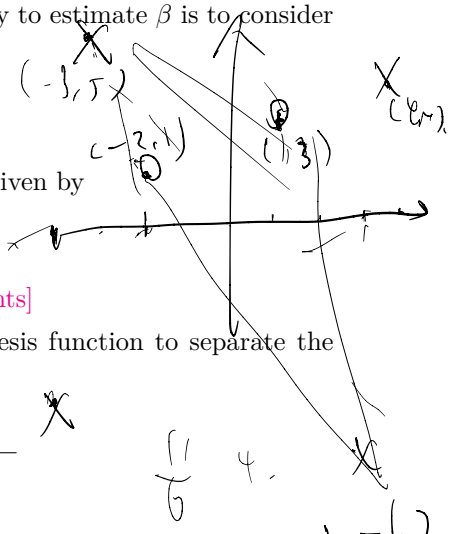
for some $\lambda > 0$. This is also known as Tikhonov regularization.
Show that the optimal solution β_* to the above optimization problem is given by

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

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) We suppose the regression function to be a linear function.

(b) Let $f(\beta) = ||y - X\beta||_2^2 + \lambda||\beta||_2^2$

$$\frac{df(\beta)}{d\beta} = 0 \iff -2X^T(y - X\beta) + 2\lambda I\beta = 0.$$

$$\iff (X^TX + \lambda I)\beta = X^Ty$$

As X^TX is symmetric, Let $X^TX = QSQ^T$

$$X^TX + \lambda I = Q(S + \lambda I)Q^T$$

Because all elements on the diagonal line of $S + \lambda I > 0$,

$S + \lambda I$ is invertible.

Thus $X^TX + \lambda I$ is invertible

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

(c) Not linear separable.

If linear seperable.

Suppose $Ax+By+C=0$.

$(-2,1)$, and $(1,3)$ should be at the same side of the line.

if $(3,-6)$ at another side, the point $(-3,5)$ can't at the same side as $(3,-6)$