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 $l(\mu, \mathbf{\Sigma})$ ① and MLE of μ ② [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). to prove the covariance of X_i is a semi positive definite matrix, we have to show that $a^T \Sigma_i a \geq 0$ where a can be any nonzero real vector

let Σ_i be the variance of X_i, so that

$\Sigma_i = E[(X_i-\mu_i)(X_i-\mu_i)^T]$, which $X=(X_1, \dots X_n)$ and $\mu=(E[X_1], E[X_2] \dots E[X_n])$

then $a^T\Sigma_i a = a^T\{E[(X_i-\mu_i)(X_i-\mu_i)^T]\}a$

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

let $Y_i = X_i - \mu_i$

then $a^T\Sigma_i a = E[a^T Y_i Y_i^T a]$

$= a^T E[Y_i Y_i^T] a$

$\because E[Y_i Y_i^T] \geq 0$, $\therefore a^T \Sigma_i a \geq 0$.

$\therefore \Sigma_i$ is a semi positive definite matrix.

b). ① 对数似然函数 $l(\mu, \Sigma) = \sum[\log(f(X_i;\mu,\Sigma))]$ 其中 $f(X_i;\mu,\Sigma)$ 表示多元正态分布的概率密度函数。

$f(X_i;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}$

$f(X_i;\mu,\Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)}$

$\log f(X_i;\mu,\Sigma) = -\frac{1}{2}d\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)$

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

② To find MLE of μ. $\frac{\nabla(\ell(\mu,\Sigma))}{\nabla\mu}=0$.

$$\frac{\nabla \sum_{i=1}^{n}\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)}{\nabla\mu}=0.$$

$$\frac{\nabla \sum_{i=1}^{n}\left[-\frac{1}{2}\left(x_i^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x_i + \mu^T\Sigma^{-1}\mu\right)\right]}{\nabla\mu}=0.$$

$\because \frac{\partial(A^T\mu)}{\partial\mu}=A.$ $\quad \frac{\partial(x^TAx)}{\partial x}=(A^T+A)x.$

$\therefore$ 原式 $=\sum_{i=1}^{n}\left(-\frac{1}{2}\left(-x_i\Sigma^{-1}-x_i\Sigma^{-1}+2\Sigma^{-1}\mu\right)\right).$

$=\sum_{i=1}^{n}\left(x_i\Sigma^{-1}-\Sigma^{-1}\mu\right).$ $\quad \because (\Sigma^{-1})^T=(\Sigma^{-1})$

$=\sum_{i=1}^{n}\Sigma^{-1}(x_i-\mu).=0.$

$\therefore \sum_{i=1}^{n}(x_i)=\sum_{i=1}^{n}\mu.=n\mu.$

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

(c). suppose $\hat{\theta}$ is an unbiased estimator of θ. then $E(\hat{\theta})=\theta$.

we need to prove that $(\hat{\theta})^2$ is not an unbiased estimator of θ, then.

we need to proce that $E[(\hat{\theta})^2] \neq \theta^2$

$\because \quad Var(X)=E(X^2)-(EX)^2$

$\therefore \quad Var(\hat{\theta}) = E[(\hat{\theta})^2] - [E(\hat{\theta})]^2$

$\therefore \quad E[(\hat{\theta})^2] = Var(\hat{\theta}) + [E(\hat{\theta})]^2$

$\because \quad Var(\hat{\theta}) > 0$. 且 $E(\hat{\theta})=\theta$

$\therefore \quad E[(\hat{\theta})^2] = Var(\hat{\theta}) + \theta^2 > \theta^2$

$\therefore \quad E[(\hat{\theta})^2] \neq \theta^2$.

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) $\because p(y_i|x_i, a, b) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2}$ and .

$L(a,b) = \prod_{i=1}^{n} p(y_i|x_i, a, b)$.

$\therefore L(a,b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2\right]$

$\therefore L(a,b) = \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp\left[-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2\right]$

$\therefore l(a,b) = \log(L(a,b)) = \sum_{i=1}^{n}\left[-\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2\sigma^2}(y_i - ax_i - b)^2\right]$

(b) To find the MLE problem for estimating a and b, we need to get the log-likelihood and $\frac{\partial l(a,b)}{\partial a}=0$ and $\frac{\partial (l(a,b))}{\partial b}=0$.

$$l(a,b)=\log(L(a,b))=\sum_{i=1}^{n}\left[-\frac{1}{2}\log 2\pi-\log\sigma-\frac{1}{2\sigma^2}(y_i-ax_i-b)^2\right]$$

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

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

$$\sum_{i=1}^{n}\cancel{\frac{1}{\sigma^2}}(y_ix_i-ax_i^2-bx_i)=0.$$

$$\sum_{i=1}^{n}ax_i^2=\sum_{i=1}^{n}y_ix_i-\sum_{i=1}^{n}bx_i \quad ①$$

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

$$=\sum_{i=1}^{n}\frac{1}{\sigma^2}(y_i-ax_i-b)=0.$$

$$\therefore \sum_{i=1}^{n}b=\sum_{i=1}^{n}y_i-ax_i.$$

$$nb=\sum_{i=1}^{n}(y_i-ax_i) \quad ②$$

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

now from ① ② we have.

$$\begin{cases} a\left(\sum_{i=1}^{n}x_i^2\right)=\sum_{i=1}^{n}y_ix_i-\sum_{i=1}^{n}bx_i \\ a\left(\sum_{i=1}^{n}x_i\right)=\sum_{i=1}^{n}y_i-nb. \end{cases}$$

$$a\sum_{i=1}^{n}(x_i^2)=\sum_{i=1}^{n}y_ix_i-\left(\sum_{i=1}^{n}x_i\right)\frac{1}{n}\left(\sum_{i=1}^{n}y_i-ax_i\right).$$

$$a\sum_{i=1}^{n}(x_i^2)-\frac{a}{n}\left[\sum_{i=1}^{n}(x_i)\right]^2=\sum_{i=1}^{n}y_ix_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)$$

$$\hat{a}=\frac{\sum_{i=1}^{n}y_ix_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sum_{i=1}^{n}(x_i)^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2} \qquad \hat{b}=\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\frac{\sum_{i=1}^{n}y_ix_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sum_{i=1}^{n}(x_i)^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2}$$

(c). $f(X) = \hat{a}X + \hat{b}$ 且 $\hat{a} = \dfrac{\sum_{i=1}^{n} y_i x_i - \hat{b}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$ 且.

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

$$\begin{aligned} f(\bar{x}) &= \hat{a}\bar{x} + \bar{b} \\ &= \hat{a}\,\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \bar{y} \end{aligned}$$

$\therefore$ $f(x)$ always passes $(\bar{x}, \bar{y})$.

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 refers to the relationship between the independent variables (predictors) and the dependent variable (response). It means that the model assumes a linear relationship between the independent variables and the expected value of the dependent variable.

(b). $L(\theta) = ||y - X\beta||_2^2 + \lambda||\beta||_2^2$

$$\frac{dL(\theta)}{d\beta} = \frac{d[(y-X\beta)^T(y-X\beta) + \lambda\beta^T\beta]}{d\beta}$$

$$= \frac{d[y^Ty - y^TX\beta - (X\beta)^Ty + \beta^TX^TX\beta + \lambda\beta^T\beta]}{d\beta}$$

$$= -X^Ty - X^Ty + 2X^TX\beta + 2\lambda\beta.$$

$$= 2X^T(X\beta - y) + 2\lambda\beta = 0.$$

$$\therefore \quad X^TX\beta - X^Ty + \lambda\beta = 0.$$

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

if $\lambda > 0$, then $X^TX + \lambda I$ is invertible. 不可逆的

we can prove it by contradiction: we suppose that it is vertible, there $\exists v$. v is a nonzero vector. s.t. $(X^TX + \lambda I)v = 0$.

$\therefore \quad X^TXv + \lambda v = 0.$

$v^TX^TXv + v^T\lambda v = 0.$

$(Xv)^TXv + \lambda v^Tv = 0.$

$\therefore \|Xv\|_2^2 + \lambda\|v\|_2^2 = 0$

$\because$ v is a nonzero vector and $\lambda > 0$ $\quad \therefore \lambda\|v\|_2^2 > 0$.

$\therefore \|Xv\|_2^2 + \lambda\|v\|_2^2 > 0$. our suppose is contradict.

$\therefore \forall v.\ (X^TX + \lambda I)v \neq 0.$

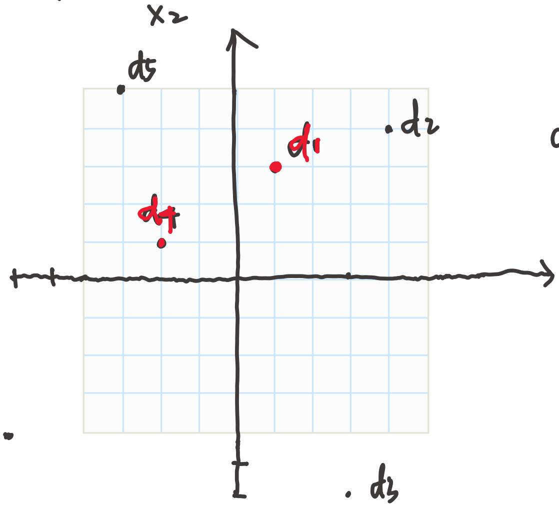
$X^TX + \lambda I$ is invertible. (可逆的).

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

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

(C). linear seperable means there is a hyperplane that can separate the data of different classes.

so we can suppose a linear function $h(x) = \theta_1x_1 + \theta_2x_2 + \theta_0$.



there isn't a linear which can separate d_1 d_4 and d_2. d_3. d_5. d_6, so the data set is not linear separable.