Introduction to Machine Learning, Fall 2023
Homework 1
(Due Thursday, Oct. 26 at 11:59pm (CST))

November 2, 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]

Answer:

(a) Let $\mu = \mathbf{E}(\mathbf{X})$

$$\Sigma = \mathbf{E}\left[\mathbf{X} - \mu\right]\left[\mathbf{X} - \mu\right]^T$$

For any $\mathbf{u} \in \mathbb{R}^n$

$$\mathbf{u}^T R \mathbf{u} = \mathbf{u}^T \mathbf{E}\left[\mathbf{X} - \mu\right]\left[\mathbf{X} - \mu\right]^T \mathbf{u} = \mathbf{E}\left[\mathbf{u}^T(\mathbf{X} - \mu)\right]\left[\mathbf{u}^T(\mathbf{X} - \mu)\right]^T = \mathbf{E}\left[\mathbf{u}^T(\mathbf{X} - \mu)\right]^2 \geq 0$$

(b) Let p be the dimension of $\mathbf{X}$. Let $A = \sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^T = X^T X - n\overline{X}^T\overline{X}$
Notice that Σ is positive definite, we have

$$(\mathbf{X}_i - \mu)^T \Sigma^{-1} (\mathbf{X}_i - \mu) = \mathrm{tr}((\mathbf{X}_i - \mu)^T \Sigma^{-1} (\mathbf{X}_i - \mu)) = \mathrm{tr}(\Sigma^{-1} (\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^T)$$

Hence we have

$$\begin{aligned}
l(\mu, \Sigma) &= ln(L(\mu, \Sigma)) \\
&= \prod_{i=1}^n \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\tfrac{1}{2}(\mathbf{X}_i - \mu)^T \Sigma^{-1} (\mathbf{X}_i - \mu)\right] \\
&= -\frac{np}{2}\ln 2\pi - \frac{n}{2} ln|\Sigma| - \frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^n (\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^T\right] \\
&= C - \frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}A + n\Sigma^{-1}(\overline{X} - \mu)(\overline{X} - \mu)^T\right] \\
&= C - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}A\right) - \frac{n}{2}(\overline{X} - \mu)\Sigma^{-1}(\overline{X} - \mu)^T \\
&\leq C - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}A\right)
\end{aligned}$$

Notice that $l(\mu, \Sigma) = C - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}A\right)$ if and only if $(\overline{X} - \mu)\Sigma^{-1}(\overline{X} - \mu)^T = 0 \iff \overline{X} = \mu$, hence, $\max_\mu l(\mu, \Sigma) = l(\overline{X}, \Sigma)$, $\hat{\mu} = \overline{X}$ (3 points for log-likelihood and 1 points for MLE)

(c)

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

Given that $\mathbf{Var}(\hat{\theta}) > 0$

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

Hence, $(\hat{\theta})^2$ 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]

Answer:

(a)

[HINT:] You will lose 1 point if neither argmin/argmax nor "maximize" / "minimize" is appeared in your answer.

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

ii. $\arg\max_a \prod_i \exp\left(-\frac{1}{2\sigma^2}(y_i - ax_i)^2\right)$

iii. $\arg\min_a \frac{1}{2}\sum_i (Y_i - aX_i)^2$

All of the above answers are correct.

(b)

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \\ \hat{b} &= \bar{y} - \hat{a}\bar{x}, \end{aligned} \tag{1}$$

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.

(c) We can plug $(\bar{x}, \bar{y})$ into the equation $\hat{y} = \hat{a}x_i + \hat{b}$, and we find $\bar{y} = \hat{a}\bar{x} + \bar{y} - \hat{a}\bar{x} = \bar{y}$ satisfies. So the least squares line 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

Answer:

(a) linear is regard to the parameters θs. The hypnosis can be written in the form of $h(\mathbf{x}, \theta) = \theta^T\phi(\mathbf{x})$ (notice that $\phi(\cdot)$ dose not necessary to be a linear function. i.e.

$$f(x) = \theta \begin{bmatrix} x, \\ x^2 \end{bmatrix}$$

is a linear regression.)

(b) Let $L(\beta) = ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2$

$$L'(\beta) = -2(\mathbf{y} - \mathbf{X}\beta)\mathbf{X} + 2\lambda\beta = -2X^Ty + 2X^TX\beta + 2\lambda\beta$$

To minimize $L(\beta)$, $L'(\beta_*) = 0$, $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta_* = \mathbf{X}^T\mathbf{y}$.
Notice that for all $\mathbf{u} \neq \mathbf{0}$

$$\mathbf{u}^T(X^TX + \lambda\mathbf{I})\mathbf{u} = \mathbf{u}^TX^TX\mathbf{u} + \lambda\mathbf{u}^T\mathbf{u} >= \lambda\mathbf{u}^T\mathbf{u} > 0$$

$X^TX + \lambda\mathbf{I}$ is positive definite, and so it is invertible. Hence, $\beta_* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ (2points for solution and 2 points for invertible)

(c) It is not linear separable because there is no hyper-plane which can separate the data, in another word, there is no $\mathbf{w}$ and b such that for all x, y, $y = \text{sign}(w^Tx + b)$. To proof this, notice that the convex set of positive labels and the convex set of negative labels are intersect, thus it is not separable. (Student may plot a image to explain the reason which is acceptable.)