CS 182: Introduction to Machine Learning, Fall 2022
Homework 1

(Due on Monday, Oct. 10 at 11:59pm (CST))

Notice:

- Please submit your assignments via Gradescope. The entry code is G2V63D.
- Please make sure you select your answer to the corresponding question when submitting your assignments.
- Each person has a total of five days to be late without penalty for all the assignments. Each late delivery less than one day will be counted as one day.

1. [20 points] [*Probability Theory*]

   (a) Prove that the correlation matrix is positive semidefinite. [6 points]

   (b) Prove that if x_m and x_n are data points sampled from the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}[x_m x_n] = \mu^2 + I_{mn}\sigma^2,$$

   where $I_{mn} = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases}$. [7 points]

   (c) Prove

$$P(C_i \mid A, B) = \frac{P(C_i, B \mid A)}{\sum_{i=1}^{n} P(C_i, B \mid A)},$$

   where $\{C_i\}_{i=1}^{n}$ is a partition of the sample space.[7 points]

Solution:

(a) Let $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T}$ be a random n dimensional vector, with μ_i and σ_i^2 denoted as the mean and the variance of x_i, for $i = 1, \ldots, n$, and σ_{ij} as the covariance between x_i and x_j, for $i \neq j$. By definition, the covariance matrix is the symmetric $n \times n$ matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}],$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\mathsf{T}$. For all $\mathbf{a} \in \mathbb{R}^n$,

$$\begin{aligned} \mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{a} &= \mathbb{E}[\mathbf{a}^\mathsf{T}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{a}] \\ &= \mathbb{E}[(\mathbf{a}^\mathsf{T}(\mathbf{x} - \boldsymbol{\mu}))^2] \\ &\geq 0, \end{aligned}$$

so the covariance matrix is a positive semidefinite matrix.

(b) If $m = n$, then $x_m x_n = x_m^2$, and we have

$$\mathbb{E}[x_m^2] = \mu^2 + \sigma^2. \tag{1}$$

If $m \neq n$, as x_m and x_n are sampled from the same Gaussian distribution, they are implicitly independent. so

$$\mathbb{E}[x_m x_n] = \mathbb{E}[x_m]\mathbb{E}[x_n] = \mu^2. \tag{2}$$

Combine and , we have

$$\mathbb{E}[x_m x_n] = \mu^2 + I_{mn}\sigma^2.$$

(c)

$$\begin{aligned}
P(C_i \mid A, B) &= \frac{P(C_i, A, B)}{P(A, B)} \\
&= \frac{P(C_i, B \mid A)P(A)}{P(B \mid A)P(A)} \\
&= \frac{P(C_i, B \mid A)}{\sum_{i=1}^{n} P(C_i, B \mid A)}.
\end{aligned}$$

2. [20 points] [*Probability Theory, Bayesian Decision Theory*] Let $\{x_i\}_{i=1}^N$ be a set of random variables following Gaussian distribution with mean μ and variance σ^2, where μ is unknown.

   (a) Derive the maximum likelihood estimate μ_{ML}. [5 points]

   (b) Assume $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Derive the maximum a posteriori estimate μ_{MAP}. [10 points]

   (c) Show that the maximum a posteriori estimate tends to the maximum likelihood estimate ($\mu_{MAP} \to \mu_{ML}$) when $N \to \infty$. [5 points]

Solution:

(a) Denote the p.d.f. of the normal distribution as $p(x; \mu, \sigma^2)$. The likelihood function for $\{x_i\}_{i=1}^N$ is

$$\begin{aligned}\mathcal{L}(x_1, \ldots, x_N; \mu, \sigma^2) &= \prod_{i=1}^N p(x_i; \mu, \sigma^2) \\ &= \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp\left\{\sum_{i=1}^N -\frac{(x_i-\mu)^2}{2\sigma^2}\right\},\end{aligned}$$

and the log-likelihood function is

$$\log(\mathcal{L}) = -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Take the derivative of $\log(\mathcal{L})$ w.r.t. μ, we have

$$\frac{\partial \log(\mathcal{L})}{\partial \mu} = \sum_{i=1}^N \frac{(x_i-\mu)}{\sigma^2},$$

by setting which to zero we can get the MLE of μ as

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^N x_i.$$

(b) The MAP of μ can be obtained by maximizing

$$\begin{aligned}\log\left(\mathcal{L}(x_1, \ldots, x_N; \mu, \sigma^2) p(\mu; \mu_0, \sigma_0^2)\right) = &-\frac{N}{2}\log(2\pi) - N\log(\sigma) - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2} \\ &+ \log\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2},\end{aligned}$$

whose derivative w.r.t. μ is

$$\sum_{i=1}^N \frac{(x_i-\mu)}{\sigma^2} - \frac{(\mu-\mu_0)}{\sigma_0^2}.$$

By setting the above derivative to zero, the MAP of μ can be obtained as follows:

$$\mu_{MAP} = \frac{\frac{\sum_{i=1}^N x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}.$$

(c) $\lim\limits_{N\to\infty} \mu_{MAP} = \lim\limits_{N\to\infty} \frac{\sum_{i=1}^N x_i/\sigma^2 + \mu_0/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2} = \frac{\sum_{i=1}^N x_i/\sigma^2}{N/\sigma^2} = \mu_{ML}.$

3. [20 points] [*Introduction, Optimization Primer*] Given a set of data points $X = \{\mathbf{x}_i\}_{i=1}^N$, its convex hull is defined as

$$C(X) = \left\{ \mathbf{x} \mid \mathbf{x} = \sum_{i=1}^N \theta_i \mathbf{x}_i, \theta_i \geq 0, \sum_{i=1}^N \theta_i = 1 \right\}.$$

Similarly we have another data set $Y = \{\mathbf{y}_i\}_{i=1}^N$ and its corresponding convex hull $C(Y)$. Show that if convex hulls of two sets of points intersect, these two sets are not linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect. (Hint: Two sets of points are linearly separable, if there exists a vector $\mathbf{w}$ and a scalar b such that $\forall$ $\mathbf{x}_i$, $\mathbf{w}^\mathsf{T}\mathbf{x}_i + b > 0$ and $\forall$ $\mathbf{y}_i$, $\mathbf{w}^\mathsf{T}\mathbf{y}_i + b < 0$.)

Solution:

Define $C(Y) = \left\{ \mathbf{y} \mid \mathbf{y} = \sum_{i=1}^N \alpha_i \mathbf{y}_i, \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1 \right\}$.

- (a) If $C(X) \cap C(Y) \neq \emptyset$, then $\exists \mathbf{z} \in C(X) \cap C(Y)$, where $\mathbf{z} = \sum_{i=1}^N \theta_i \mathbf{x}_i = \sum_{i=1}^N \alpha_i \mathbf{y}_i$. Suppose that under this condition, X and Y are linearly separable, i.e., $\exists \mathbf{w}, b$ s.t., $\mathbf{w}^\mathsf{T}\mathbf{x}_i + b > 0$, $\forall \mathbf{x}_i$ and $\mathbf{w}^\mathsf{T}\mathbf{y}_i + b < 0$, $\forall \mathbf{y}_i$. Hence, we have

$$\begin{aligned} \mathbf{w}^\mathsf{T}\mathbf{z} + b &= \mathbf{w}^\mathsf{T}(\sum_{i=1}^N \theta_i \mathbf{x}_i) + 1 \cdot b \\ &= \sum_{i=1}^N \theta_i \mathbf{w}^\mathsf{T}\mathbf{x}_i + \sum_{i=1}^N \theta_i b \\ &= \sum_{i=1}^N \theta_i (\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \\ &> 0. \end{aligned} \tag{3}$$

  By similar deduction,

$$\mathbf{w}^\mathsf{T}\mathbf{z} + b = \sum_{i=1}^N \alpha_i (\mathbf{w}^\mathsf{T}\mathbf{y}_i + b) < 0. \tag{4}$$

  As (3) and (4) contradict each other, the hypothesis is wrong, i.e., X and Y must not be linearly separable.

- (b) If X and Y are linearly separable, then we can always find a hyperplane to separate them. By the definition of convex hull, $C(X)$ and $C(Y)$ can also be separated by the same hyperplane. Thus, $C(X) \cap C(Y) = \emptyset$.

4. [20 points] [*Linear Algebra*] Assume that there are n given training examples $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, where each input data point $\mathbf{x}_i$ has m real valued features. When $m > n$, the linear regression model is equivalent to solving an under-determined system of linear equations $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. One popular way to estimate $\boldsymbol{\beta}$ is to consider the so-called ridge regression:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

for some $\lambda > 0$. This is also known as Tikhonov regularization.

(a) Show that the optimal solution $\boldsymbol{\beta}_\star$ to the above optimization problem is given by:

$$\boldsymbol{\beta}_\star = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y},$$

given that the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is invertible. [10 points]

(b) Discuss the conditions on the matrix $\mathbf{X}$ and λ so that the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is guaranteed to be invertible. [10 points]

Solution:

(a) We define the cost function $f(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$, and we have $\nabla f(\boldsymbol{\beta}) = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}$, and $\nabla^2 f(\boldsymbol{\beta}) = 2\mathbf{X}^\top\mathbf{X} + 2\lambda\mathbf{I} \succ 0$. Thus the optimal solution $\boldsymbol{\beta}^\star$ holds when $\nabla f(\boldsymbol{\beta}) = 0 \Rightarrow -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star) + 2\lambda\boldsymbol{\beta}^\star = 0 \Rightarrow \boldsymbol{\beta}^\star = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$.

(b) When $\lambda > 0$, $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is guaranteed to be invertible. Let $\mathbf{v}$ be a non zero vector, then $\mathbf{v}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{v} = \mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v} + \lambda\mathbf{v}^\top\mathbf{v} = \|\mathbf{X}\mathbf{v}\|_2^2 + \lambda\|\mathbf{v}\|_2^2 > 0$. Thus we draw the conclusion.

5. [20 points] [*Optimization Primer*]

(a) Prove that if f is a convex function, then $\mathcal{C} = \{\mathbf{x} \mid f(\mathbf{x}) \leq 0\}$ is a convex set. [10 points]

(b) Prove that if x is a random variable and f is a convex function, then $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$. [10 points]

Solution:

(a) If $x, y \in \mathcal{C}$, then $f(x) \leq 0$ and $f(y) \leq 0$, so $f(\theta x + (1-\theta)y) \leq 0$ for $0 \leq \theta \leq 1$, and hence $\theta x + (1+\theta)y \in \mathcal{C}$. Set $\mathcal{C}$ is convex.

(b) Suppose we start with the inequality in the basic definition

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

for $0 \leq \theta \leq 1$. Using induction, this can fairly easily extended to convex combinations of more than one points,

$$f\left(\sum_{i=1}^{k} \theta_i x_i\right) \leq \sum_{i=1}^{k} \theta_i f(x_i)$$

for $\sum_{i=1}^{k} \theta_i = 1$ and $\theta_i \geq 0$. If x is a continuous random variable, the inequality can be written as

$$f\left(\int p(x)x dx\right) \leq \int p(x)f(x)dx$$

for $\int p(x)dx = 1$ and $p(x) \geq 0$, in which case the previous inequality can be written in terms of expectations,

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

If x is a discrete random variable, the proof is similar.