Introduction to Machine Learing: Homework 1

Due on Oct 10th, 2022 at 11:59pm

Professor Ziping Zhao

Bingnan Li
2020533092

1 (a) Prove that the correlation matrix is positive semidefinite:

Proof:

Assume that $\boldsymbol{A}$ is the $m \times m$ correlation matrix with corresponding coordinates a_{ij}, then for any $m \times 1$ column vector $\boldsymbol{y}$ with corresponding coordinates y_i, we have

$$\begin{aligned}
\boldsymbol{A} &= \mathbb{E}\left[\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)^T\right] \\
\boldsymbol{y^T A y} &= \boldsymbol{y^T}\mathbb{E}\left[\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)^T\right]\boldsymbol{y} \\
&= \mathbb{E}\left[\boldsymbol{y^T}\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)^T\boldsymbol{y}\right] \\
&= \mathbb{E}\left[\left|\left|\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)^T\boldsymbol{y}\right|\right|_2^2\right] \\
&\geq 0
\end{aligned}$$

Therefore, the correlation matrix is positive semidefinite.

(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 $I_{mn} = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases}$

Proof:

Given the properties of Gaussian distribution, we have

$$\mathbb{E}[x] = \mu \quad \mathrm{Var}[x] = \sigma^2$$

Additionally,

$$\begin{aligned}
\mathrm{Cov}(x_m, x_n) &= \mathbb{E}[(x_m - \mu)(x_n - \mu)] \\
&= \int_{\mathbb{R}} \mathbb{E}[(x_m - \mu)(x_n - \mu)|x_m = x]p(x)dx \\
&= \begin{cases} \displaystyle\int_{\mathbb{R}} (x-\mu)^2 p(x)dx, & m = n \\ \displaystyle\int_{\mathbb{R}} (x-\mu)\mathbb{E}[x_n - \mu]p(x)dx, & m \neq n \end{cases} \\
&= \begin{cases} \mathrm{Var}(x) = \sigma^2, & m = n \\ \displaystyle\int_{\mathbb{R}} (x-\mu)(\mu-\mu)p(x)dx = 0, & m \neq n \end{cases}
\end{aligned}$$

Moreover, we know that

$$\mathrm{Cov}(x_m, x_n) = \mathbb{E}[x_m x_n] - \mu^2 = \begin{cases} \sigma^2, m = n \\ 0, m \neq n \end{cases}$$

Thus,

$$\mathbb{E}[x_m x_n] = \mathrm{Cov}(x_m, x_n) + \mu^2 = \begin{cases} \mu^2 + \sigma^2, m = n \\ \mu^2, m \neq n \end{cases}$$

(c) Prove

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

Proof:

Define the sample space S, then $A, B \subset S$ and $\{C_i\}_{i=1}^{n}$ is a partition of S. Thus, we know that for any subset of S, say D, then we have

$$P(D) = \sum_{i=1}^{n} P(C_i, D)$$

Based on that, we have

$$\begin{aligned} P(C_i|A,B) &= \frac{P(C_i,A,B)}{P(A,B)} = \frac{P(C_i,B|A)P(A)}{P(B|A)P(A)} \\ &= \frac{P(C_i,B|A)}{P(B|A)} \end{aligned}$$

Since B is a subset of S, then we have

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

then we get

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

2 (a) Derive the maximum likelihood estimate μ_{ML}.

Solution:

Since $x_i \sim \mathcal{N}(\mu, \sigma^2)$, we have the likelihood function

$$L(\mu|\mathcal{X}) = P(\mathcal{X}|\mu) = \prod_{i=1}^{N} P(x_i|\mu) = \left(\sqrt{2\pi}\sigma\right)^{-N} e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

and the log likelihood function

$$\mathcal{L}(\mu|\mathcal{X}) = -N\log\sqrt{2\pi}\sigma - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

Then

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu} &= \frac{1}{\sigma}\sum_{i=1}^{N}\frac{x_i-\mu}{\sigma} = 0 \\ \Rightarrow \mu_{ML} &= \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \end{aligned}$$

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

Solution:

By definition, we have

$$\mu_{MAP} = \arg\max_{\mu} P(\mu|\mathcal{X}) = \arg\max_{\mu} P(\mathcal{X}|\mu)P(\mu) = \arg\max_{\mu} \left(\log P(\mathcal{X}|\mu) + \log P(\mu)\right)$$

Let $\mathcal{L} = \log P(\mathcal{X}|\mu) + \log P(\mu)$, then

$$\mathcal{L} = -N \log \sqrt{2\pi}\sigma - \frac{1}{2}\sum_{i=1} N\left(\frac{x_i - \mu}{\sigma}\right)^2 - \log\sqrt{2\pi}\sigma_0 - \frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i - N\mu\right) - \frac{\mu - \mu_0}{\sigma_0^2} = 0$$

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

(c) Show that the maximum a posteriori estimate tends to the maximum likelihood estimate when $N \to \infty$.
Solution:

$$\begin{aligned}\lim_{N\to\infty} \frac{\sigma_0^2 \sum_{i=1}^{N} x_i + \sigma^2 \mu_0}{\sigma_0^2 N + \sigma^2} &= \lim_{N\to\infty} \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\frac{1}{N}\sum_{i=1}^{N} x_t + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 \\ &= \lim_{N\to\infty} \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 \\ &= \mu_{ML}\end{aligned}$$

3 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.
Proof:

First, two sets are linearly separable, then their convex hulls do not intersect:

Define $C(Y) = \{y | y = \sum_{i=1}^{m} \beta_i y_i, \beta_i \geq 0, \sum_{i=1}^{m} \beta_i = 1\}$. Given that X and Y are linearly separable, then assume that there is a hyperplane that can linearly separate X and Y, say $\exists \boldsymbol{w}$ and a scalar b s.t. $\forall \boldsymbol{\xi_i} \in C(X), \boldsymbol{\zeta_i} \in C(Y)$, we have $\boldsymbol{w^T \xi_i} + b > 0$ and $\boldsymbol{w^T \zeta_i} + b < 0$.

Based on that, we have

$$\begin{aligned}\boldsymbol{w^T \xi_i} + b &= \boldsymbol{w^T} \sum_{i=1}^{n} \theta_i \boldsymbol{x_i} + b \\ &= \sum_{i=1}^{n} \boldsymbol{w^T} \theta_i \boldsymbol{x_i} + \sum_{i=1}^{n} \theta_i b \\ &= \sum_{i=1}^{n} \theta_i (\boldsymbol{w^T x_i} + b)\end{aligned}$$

Similarly, for ζ_i, we have

$$\boldsymbol{w^T \zeta_i} + b = \sum_{i=1}^{m} \beta_i (\boldsymbol{w^T y_i} + b)$$

Then assume that $C(X) \cap C(Y) \neq \varnothing$, let $\boldsymbol{z}$ be a point that both in $C(X)$ and $C(Y)$, we have

$$\boldsymbol{w^T z} + b = \sum_{i=1}^{n} \theta_i (\boldsymbol{w^T x_i} + b) > 0$$

and

$$\boldsymbol{w^T z} + b = \sum_{i=1}^{m} \beta_i (\boldsymbol{w^T y_i} + b) < 0$$

which is a contradiction and proof done.

Then given that "if convex hulls of two sets of points intersect, these two sets are not linearly separable" is the Inverse Negative Proposition of the above one, thus this one is also true.

4 (a) Show that the optimal solution $\boldsymbol{\beta_\star}$ is given by:

$$\boldsymbol{\beta_\star} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Proof:

Let $\boldsymbol{F} = ||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2$, then

$$\begin{aligned}\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\beta}} &= \frac{\partial||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2}{\partial \boldsymbol{\beta}} \\ &= \frac{\partial(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})}{\partial \boldsymbol{\beta}} + \lambda\frac{\partial \boldsymbol{\beta}^T\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \\ &= -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X\beta}) + 2\lambda\boldsymbol{\beta} = 0 \\ &\Rightarrow (\lambda\boldsymbol{I} + \boldsymbol{X^TX})\boldsymbol{\beta} = \boldsymbol{X^Ty}\end{aligned}$$

Given that $\boldsymbol{X^TX} + \lambda\boldsymbol{I}$ is invertible, we have

$$\boldsymbol{\beta_\star} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

(b) Discuss the conditions on the matrix $\boldsymbol{X}$ and λ so that the matrix $\boldsymbol{X^TX} + \lambda\boldsymbol{I}$ is guaranteed to be invertible.

Solution:

Let $\boldsymbol{A} = \left(\boldsymbol{X^TX}+\right)\lambda\boldsymbol{I}$, then we know that $\boldsymbol{A}$ is a symmetric matrix. Next, for any non-zero vector $\boldsymbol{v}$, we have

$$\begin{aligned}\boldsymbol{v^TAv} &= \boldsymbol{v^TX^TXv} + \lambda\boldsymbol{v^TIv} \\ &= ||\boldsymbol{Xv}||_2^2 + \lambda||\boldsymbol{v}||_2^2\end{aligned}$$

since $\boldsymbol{v}$ is a non-zero vector, then $||\boldsymbol{v}||_2^2 \neq 0$. Moreover, $\lambda > 0$, then

$$\boldsymbol{v^TAv} \geq \lambda||\boldsymbol{v}||_2^2 > 0$$

Thus, $\boldsymbol{A}$ is a positive definite matrix and $\boldsymbol{A}$ is invertible.

5 (a) Prove that if f is a convex function, then $\mathcal{C} = \{x|f(x) \leq 0\}$ is a convex set.

Proof:

For $\forall \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathcal{C}$, we can determine a line l between $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, and $\exists \boldsymbol{x} \in l$, we can denote $\boldsymbol{x}$ as

$$\boldsymbol{x} = \boldsymbol{x_1} + t(\boldsymbol{x_2} - \boldsymbol{x_1}) = (1-t)\boldsymbol{x_1} + t\boldsymbol{x_2}$$

Then, we have

$$\begin{aligned}f(\boldsymbol{x}) &= f((1-t)\boldsymbol{x_1} + t\boldsymbol{x_2}) \\ &\leq (1-t)f(\boldsymbol{x_1}) + tf(\boldsymbol{x_2})\end{aligned}$$

since $f(\boldsymbol{x_1}), f(\boldsymbol{x_2}) \leq 0$ and $t \in [0,1]$, we know that

$$f(\boldsymbol{x}) \leq 0$$

which means for any points that lay in the line between $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, we know that $\boldsymbol{x} \in \mathcal{C}$. Thus, $\mathcal{C}$ is a convex set.

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

If x is a discrete random variable, then by the property of convex function, we know that for x that only has 2 possible values, we have

$$\begin{aligned} f(\mathbb{E}[x]) &= f(p_1x_1 + p_2x_2) \\ &\leq p_1f(x_1) + p_2f(x_2) = \mathbb{E}[f(x)] \end{aligned}$$

Then assume that inequality holds for x that has k possible values, then for x that has $k+1$ possible values, we have

$$\begin{aligned} f(\mathbb{E}[x]) &= f(\sum_{i=1}^{k} p_ix_i + p_{k+1}x_{k+1}) \\ &= f\left(\sum_{i=1}^{k} p_i \sum_{i=1}^{k} \frac{p_ix_i}{\sum_{i=1}^{k} p_i} + p_{k+1}x_{k+1}\right) \\ &\leq \sum_{i=1}^{k} p_i f\left(\sum_{i=1}^{k} \frac{p_ix_i}{\sum_{i=1}^{k} p_i}\right) + p_{k+1}f(x_{k+1}) \\ &\leq \sum_{i=1}^{k} p_i \sum_{i=1}^{k} \frac{p_ix_i}{\sum_{i=1}^{k} p_i} f(x_i) + p_{k+1}f(x_{k+1}) \\ &= \sum_{i=1}^{k+1} p_if(x_i) = \mathbb{E}[f(x)] \end{aligned}$$

If x is a continuous random variable(domain is $\mathbb{D}$), and $g(x)$ is the corresponding pdf, then

$$\mathbb{E}[x] = \int_{\mathbb{D}} xg(x)dx = \lim_{||T||\to 0} \sum_{i=1}^{n} \xi_i g(\xi_i)\Delta x_i$$

where T is a partition of domain such that $\mathbb{D}$ is divided into n intervals, then denote $||T|| = max(||intervals||)$.

Based on that, we have

$$\begin{aligned} f(\mathbb{E}[x]) &= f\left(\lim_{||T||\to 0} \sum_{i=1}^{n} \xi_i g(\xi_i)\Delta x_i\right) \\ &= \lim_{||T||\to 0} f\left(\sum_{i=1}^{n} \xi_i g(\xi_i)\Delta x_i\right) \end{aligned}$$

By definition, we have $\sum_{i=1}^{n} f(\xi_i)\Delta x_i = 1$. Thus, by the discrete form of above proof, we have

$$f(\mathbb{E}[x]) \leq \lim_{||T||\to 0} \sum_{i=1}^{n} g(\xi_i)\Delta x_i f(\xi_i) = \int_{\mathbb{D}} f(x)g(x)dx = \mathbb{E}[f(x)]$$