

# Homework Assignment 5

Lecturer: Kyunghyun Cho

April 18, 2017

1. The probability density function of normal distribution is defined as

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right),$$

where

$$\begin{aligned} Z &= \int_{\mathbf{x} \in \mathbb{R}^d} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) d\mathbf{x} \\ &= (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2}, \end{aligned}$$

where  $|\boldsymbol{\Sigma}|$  is the determinant of the covariance matrix.

Let us assume that the covariance matrix  $\boldsymbol{\Sigma}$  is a diagonal matrix, as below:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}.$$

The probability density function simplifies to

$$f(\mathbf{x}) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left( -\frac{1}{2} \frac{1}{\sigma_i^2} (x_i - \mu_i)^2 \right).$$

Show that this is indeed true.

SOLUTION:

$$Z = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \quad |\boldsymbol{\Sigma}| = \prod_{i=1}^d \sigma_i^2$$

$$Z = (2\pi)^{-d/2} \left( \prod_{i=1}^d \sigma_i^2 \right)^{-1/2}$$

$$= (2\pi)^{-d/2} \prod_{i=1}^d \frac{1}{\sigma_i}$$

$$= \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i}$$

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

$$\begin{aligned}
&= \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\
&= \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}(x_i - \mu_i)\Sigma^{-1}(x_i - \mu_i)\right) \\
&= \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\frac{1}{\sigma_i^2}(x_i - \mu_i)^2\right)
\end{aligned}$$

2. (a) Show that the following equation, called Bayes' rule, is true.

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}.$$

SOLUTION: We know that  $p(X, Y) = p(X|Y)p(Y)$  and that  $p(Y|X) = \frac{p(X, Y)}{p(X)}$

Therefore:

$$p(Y|X) = \frac{p(X, Y)}{p(X)} = \frac{p(X|Y)p(Y)}{p(X)}$$

(b) We learned the definition of expectation:

$$\mathbb{E}[X] = \sum_{x \in \Omega} xp(x).$$

Assuming that  $X$  and  $Y$  are discrete random variables, show that

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

SOLUTION:

$$\begin{aligned}
\mathbb{E}[X + Y] &= \sum_{x, y \in \Omega} (x + y)p(x, y) = \sum_{x, y \in \Omega} xp(x, y) + \sum_{x, y \in \Omega} yp(x, y) \\
&= \sum_{x \in \Omega} x \sum_{y \in \Omega} p(x, y) + \sum_{y \in \Omega} y \sum_{x \in \Omega} p(x, y) \\
&= \sum_{x \in \Omega} xp(x) + \sum_{y \in \Omega} yp(y) \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

(c) Further assume that  $c \in \mathbb{R}$  is a scalar and is not a random variable, show that

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

$$\mathbb{E}[cX] = \sum_{x \in \Omega} c xp(x) = c \sum_{x \in \Omega} xp(x) = c\mathbb{E}[X]$$

(d) We learned the definition of variance:

$$\text{Var}(X) = \sum_{x \in \Omega} (x - \mathbb{E}[X])^2 p(x).$$

$\text{Var}(X) = \sum_{x \in \Omega} (x - \mathbb{E}[X])^2 p(x)$  Assuming  $X$  being a discrete random variable, show that

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

SOLUTION:

$$\begin{aligned} \text{Var}(X) &= \sum_{x \in \Omega} (x - \mathbb{E}[X])^2 p(x) \\ &= \sum_{x \in \Omega} (x^2 - 2x\mathbb{E}[X] + (\mathbb{E}[X])^2) p(x) \\ &= \sum_{x \in \Omega} x^2 p(x) - \sum_{x \in \Omega} 2x\mathbb{E}[X] p(x) + \mathbb{E}[X]^2 \\ &= \mathbb{E}[x^2] - 2\mathbb{E}[X] \sum_{x \in \Omega} x p(x) + \mathbb{E}[X]^2 \\ &= \mathbb{E}[x^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\ &= \mathbb{E}[x^2] - \mathbb{E}[X]^2 \end{aligned}$$

3. (a) An optimal linear regression machine (without any regularization term), that minimizes the empirical cost function given a training set

$$D_{\text{tra}} = \{(\mathbf{x}_1, \mathbf{y}_1^*), \dots, (\mathbf{x}_N, \mathbf{y}_N^*)\},$$

can be found directly without any gradient-based optimization algorithm. Assuming that the distance function is defined as

$$D(M^*(\mathbf{x}), M, \mathbf{x}) = \frac{1}{2} \|M^*(\mathbf{x}) - M(\mathbf{x})\|_2^2 = \frac{1}{2} \sum_{k=1}^q (y_k^* - y_k)^2,$$

derive the optimal weight matrix  $\mathbf{W}$ . (Hint: Moore–Penrose pseudoinverse)

- (b) (Extra Credit) Derive a probability density function of the predictive distribution of Bayesian linear regression. Follow the assumptions made during the lecture (see the lecture note.)