Homework Assignment 5

Lecturer: Kyunghyun Cho

April 1, 2018

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

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

where

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

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

Let us assume that the covariance matrix Σ is a diagonal matrix, as below:

$$\Sigma = \left[\begin{array}{cccc} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{array}\right].$$

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.

2.

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

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

(b) We learned the definition of expectation:

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

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

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

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

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

(d) We learned the definition of variance:

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

Assuming X being a discrete random variable, show that

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

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

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

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)

4. Suppose that we have a data distribution $Y = f(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is a random vector, ε is an independent random variable with zero mean and fixed but unknown variance σ^2, and f is an unknown deterministic function that maps a vector into a scalar.

Now, we wish to approximate $f(\mathbf{x})$ with our own model $\hat{f}(\mathbf{x};\Theta)$ with some learnable parameters Θ.

(a) Show that considering all possible $\hat{f}$ and Θ, the minimum of L2 loss

$$\mathbb{E}_{\mathbf{X}}[(Y - \hat{f}(\mathbf{X};\Theta))^2]$$

is achieved when for all $\mathbf{x}$,

$$\hat{f}(\mathbf{x};\Theta) = f(\mathbf{x})$$

(Hint: find the minimum of L2 loss for a single example first.)

(b) If we train the same model varying initializations and examples from the underlying data distribution, we may end up with different Θ. So we can also consider Θ as a random variable if we fix $\hat{f}$.

Show that for a single unseen input vector $\mathbf{x}_0$ and a fixed $\hat{f}$, the expected squared error between the ground truth $f(\mathbf{x}_0)$ and the prediction $\hat{f}(\mathbf{x}_0;\Theta)$ can be decomposed into:

$$\mathbb{E}[(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0;\Theta))^2] = \left(\mathbb{E}[f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0;\Theta)]\right)^2 + \mathrm{Var}[\hat{f}(\mathbf{x}_0;\Theta)] + \sigma^2$$

(Side note: this is usually known as the *bias-variance decomposition*, closely related to *bias-variance tradeoff*, and other concepts such as underfitting and overfitting.)