Homework Assignment 5

Lecturer: Kyunghyun Cho

April 15, 2018

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

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

where

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

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

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

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

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. Ans:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right),$$

We know Σ is a diagonal matrix, $det(\Sigma) = \sigma_1^2 * \sigma_2^2 * \ldots * \sigma_d^2$, therefore, $|\Sigma|^{1/2} = \prod_{i=1}^{d} \sigma_i$. And we also know $(2\pi)^{d/2} = \prod_{i=1}^{d} \sqrt{2\pi}$. So we have $\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i}$

We know that

$$\mathbf{x}-\mu = \begin{bmatrix} \mathbf{x}_1-\mu \\ \mathbf{x}_2-\mu \\ \vdots \\ \mathbf{x}_d-\mu \end{bmatrix}.$$

Therefore,

$$(\mathbf{x}-\mu)^\top = \begin{bmatrix} \mathbf{x}_1-\mu & \mathbf{x}_2-\mu & \cdots & \mathbf{x}_d-\mu \end{bmatrix}.$$

And

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_d^2} \end{bmatrix}.$$

Therefore,

$$(\mathbf{x}-\mu)^\top\Sigma^{-1} = \begin{bmatrix} \frac{\mathbf{x}_1-\mu}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{\mathbf{x}_2-\mu}{\sigma_2^2} & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{\mathbf{x}_d-\mu}{\sigma_d^2} \end{bmatrix}.$$

And $(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu) = \frac{\mathbf{x}_1-\mu}{\sigma_1^2} * \mathbf{x}_1-\mu + \frac{\mathbf{x}_2-\mu}{\sigma_2^2} * \mathbf{x}_2-\mu + \cdots + \frac{\mathbf{x}_d-\mu}{\sigma_d^2} * \mathbf{x}_d-\mu = \sum_{i=1}^d \frac{\mathbf{x}_i-\mu}{\sigma_i^2}$

We know sum in exponential function is the multiplication of individual exponential function, hence,

$\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^d \frac{\mathbf{x}_i-\mu}{\sigma_i^2}\right) = \prod_{i=1}^d \exp\left(-\frac{1}{2}\frac{\mathbf{x}_i-\mu}{\sigma_i^2}\right)$

Therefore, we have the solution,

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

2.

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

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

Ans: Suppose X and Y are random variables defined on a probability measure (Ω, A, P) where Ω is the outcome space, A is a σ-algebra and P is the corresponding probability measure. X maps to $(E, \mathscr{E})$ and Y maps to $(F, \mathscr{F})$. Suppose $y \in \mathscr{F}, x \in \mathscr{E}$,

$$p(Y|X) = p(Y \in y|X \in x) = p(\{\omega_y|Y(\omega_y) \in y\}|\{\omega_x|X(\omega_x) \in x\}) = \frac{P(\{\omega \in \omega_x \cap \omega_y\})}{P(\{\omega_x|X(\omega_x) \in x\})} = \frac{P(\omega \in \omega_x \cap \omega_y)}{P(X)}$$

Similarly,

$$P(X|Y) = \frac{P(\{\omega \in \omega_y \cap \omega_x\})}{P(Y)},$$
$$P(\{\omega \in \omega_y \cap \omega_x\}) = P(X|Y) * P(Y)$$

Therefore,

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

(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

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

Ans: Note instead of assuming $\Omega as outcome space for random variable outcome space$,I assume Ω is the outcome space for probability measure for the random variable.
We know, $\mathbb{E}[X] = \sum_{x \in \Omega} X(x)P(x)$. Similarly, $\mathbb{E}[Y] = \sum_{x \in \Omega} Y(x)P(x)$ and

$$\mathbb{E}[X+Y] = \sum_{x \in \Omega} (X+Y)(x)P(x) = \sum_{x \in \Omega} [X(x)+Y(x)]P(x) = \sum_{x \in \Omega} [X(x)P(x) + Y(x)P(x)]$$
$$= \sum_{x \in \Omega} X(x)P(x) + \sum_{x \in \Omega} Y(x)P(x) = \mathbb{E}[X] + \mathbb{E}[Y]$$

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

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

Similarly,

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

because c is a constant so, it can be out of the sum sign and stay the same.

(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

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

Ans:

$$\begin{aligned}\mathrm{Var}(X) = \sum_{x\in\Omega} (x - \mathbb{E}[X])^2 p(x) &= \sum_{x\in\Omega} [x^2 - 2\mathbb{E}[X]x + \mathbb{E}[X]^2]p(x) \\ &= \sum_{x\in\Omega} x^2 p(x) - 2\mathbb{E}[X] \sum_{x\in\Omega} x * p(x) + \mathbb{E}[X] \sum_{x\in\Omega} p(x) \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X] * \mathbb{E}[X] + \mathbb{E}[X]^2 * 1 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

Using linearity and scalar multiplication property, (note E[X] is a constant in the equation), and sum of probabilities over all outcomes is 1.

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)

Ans: To minimize D, we want to make $y_k^* - y_k$ as small as possible for every k. Now, the matrix representation of X with $x_i \in R^d$, is

$$X = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}.$$

$X \in R^{Nxd}$ And similarly $W \in R^{dx1}$, so the output is $y \in R^{Nx1}$ with each value corresponding to the result of $x_i * W$, the linear regression output.

Therefore, if we can make $X * W = y$, then $y_k^* - y_k = 0$. If X is invertible, then we directly derives $W = X^{-1}y$.

However, if X is not invertible, we can use "Moore Penrose Pseudo inverse" to generate X^+ such that it is the best approximating matrix to X for $||y - y_{truth}||_2^2$ and we have $W = X^+ * y$.

Reference: (http://buzzard.ups.edu/courses/2014spring/420projects/math420-UPS-spring-2014-macausland-pseudo-inverse.pdf) here provides many proof and theorems that I mentioned about Moore Penrose.

Now, let's show how we get W. One important condition for being a Moore Penrose Pseudo Inverse is $X * X^+ * X = X$ It is proved that $X+$ exists and is unique for any X matrix. There are many ways to compute the inverse $X+$, one way is using SVD decomposition. For any matrix, we can use SVD decomposition into $X = UDV^\top$ where U is a unitary matrix in size N x N and V D x D. D is a diagonal matrix with size N x D and D is in the following form:

$$D = \begin{bmatrix} S & \hat{0} \\ \hat{0} & \hat{0} \end{bmatrix}.$$

where $S \in R^{rxr}$ where r is the rank of D.
We can define D+ such that

$$D+ = \begin{bmatrix} S^{-1} & \hat{0} \\ \hat{0} & \hat{0} \end{bmatrix}$$

And we can show $X+ = VD+^\top U^\top$ Therefore, we have $W = VD+^\top U^\top y$ this is the optimized W weight matrix with minimized distance.

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.)
Ans: From the text, we want to approximate f(x) with our own model $\hat{f}(x)$. Now, we have $\hat{f}(x) = f(x)$, therefore, it has already achieved our target. When we replace $\hat{f}(x) = f(x)$ into the $\mathbb{E}_{\mathbf{X}}[(Y - \hat{f}(\mathbf{X};\Theta))^2]$, we have $\mathbb{E}_{\mathbf{X}}[(Y - f(x))^2] = \mathbb{E}_{\mathbf{X}}[(f(x) + \varepsilon - f(x))^2] = \mathbb{E}_{\mathbf{X}}[(\varepsilon)^2] = Var[\varepsilon] + \mathbb{E}^2_{\mathbf{X}}[(\varepsilon)]$, because ε is a random variable with zero mean, we have $\mathbb{E}_{\mathbf{X}}[(Y - \hat{f}(\mathbf{X};\Theta))^2] = Var[\varepsilon] = \sigma^2$ Because ε measures the noise of the distribution $f(x)$, we don't know it is corresponding variance, therefore, this is the part we cannot control and is irreducible. Therefore, minimum is achieved.

(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 $y_0 = f(\mathbf{x}_0) + \varepsilon$ and the prediction $\hat{f}(\mathbf{x}_0;\Theta)$ can be decomposed into:

$$\mathbb{E}[(y_0 - \hat{f}(\mathbf{x}_0;\Theta))^2] = \left(\mathbb{E}[y_0 - \hat{f}(\mathbf{x}_0;\Theta)]\right)^2 + \mathrm{Var}[\hat{f}(\mathbf{x}_0;\Theta)] + \sigma^2$$

(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.)

Ans: $\mathbb{E}[(y_0 - \hat{f}(\mathbf{x}_0;\Theta))^2] = \mathbb{E}[y_0^2] - 2\mathbb{E}[y_0 * \hat{f}] + \mathbb{E}[\hat{f}^2]$;
Since f(x0) is fixed, so we can consider it as a constant and Θ is independent with $f(x)$, so$\mathbb{E}[f(x_0) * \Theta] = \mathbb{E}[f(x_0)] * \mathbb{E}[\Theta]$.
We have
$\mathbb{E}[y_0^2] = \mathbb{E}[f(x_0)^2] + 2\mathbb{E}[f(x_0) * \Theta] + \mathbb{E}[\Theta^2] = f(x_0)^2 + 2\mathbb{E}[f(x_0)] * \mathbb{E}[\Theta] + E[\Theta^2]$
Also notice that, $f(x_0) = E[f(x_0)]$ and we can complete the square to get:
$\mathbb{E}[y_0^2] = \mathbb{E}[f(x_0)]^2 + 2\mathbb{E}[f(x_0)] * \mathbb{E}[\Theta] + \mathbb{E}[\Theta]^2 - \mathbb{E}[\Theta]^2 + E[\Theta^2]$
$= (\mathbb{E}[f(x_0)] + \mathbb{E}[\Theta])^2 - \mathbb{E}[\Theta]^2 + E[\Theta^2]$
$= (\mathbb{E}[f(x_0)] + \mathbb{E}[\Theta])^2 + \sigma^2$
$= \mathbb{E}[y_0]^2 + \sigma^2$.
Hence,

$\mathbb{E}[(y_0 - \hat{f}(\mathbf{x}_0;\Theta))^2] = \mathbb{E}[y_0]^2 - 2\mathbb{E}[y_0 * \hat{f}] + \mathbb{E}[\hat{f}^2] + \sigma^2$
$= \mathbb{E}[y_0]^2 - 2\mathbb{E}[y_0 * \hat{f}] + \mathbb{E}[\hat{f}]^2 - \mathbb{E}[\hat{f}]^2 + \mathbb{E}[\hat{f}^2] + \sigma^2,$ notice $-\mathbb{E}[\hat{f}]^2 + \mathbb{E}[\hat{f}^2] = Var[\hat{f}]$, we get
$\mathbb{E}[(y_0 - \hat{f}(\mathbf{x}_0;\Theta))^2] = \mathbb{E}[y_0]^2 - 2\mathbb{E}[y_0 * \hat{f}] + \mathbb{E}[\hat{f}]^2 + Var[\hat{f}] + \sigma^2 = \mathbb{E}[y_0 - \hat{f}]^2 + Var[\hat{f}] + \sigma^2$ by linearity.