Homework 2: Mathematics for AI
Due date: October 19th, 11:59 pm.
Note: Submit your solution file to TA, Seokin Seo (e-mail: siseo@ai.kaist.ac.kr), by email.
For questions, contact to TA as well.

---

1. Consider $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^\top, \mathbf{y} = [y_1, ..., y_N]^\top$ with $\mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R}$ for $i \in \{1, ..., N\}$. For $\theta_0 \in \mathbb{R}, \theta \in \mathbb{R}^D$, define $J_1(\mathbf{X}, \mathbf{y}, \theta, \theta_0) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\theta - \theta_0\mathbf{1}\|_2^2$. Then, denote $\hat{\theta} = \arg\min_\theta J_1(\mathbf{X}, \mathbf{y}, \theta, \theta_0)$, $\hat{\theta}_0 = \arg\min_{\theta_0} J_1(\mathbf{X}, \mathbf{y}, \theta, \theta_0)$.

(a) Let $\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i, \bar{y} = \frac{1}{N}\sum_i y_i$. Show that $\hat{\theta}_0 = \bar{y} - \bar{\mathbf{x}}^\top\theta$.

(b) Show that

$$\hat{\theta} = (\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top\mathbf{y}_c = \left[\sum_{i=1}^N(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\right]^{-1}\left[\sum_{i=1}^N(y_i - \bar{y})(\mathbf{x}_i - \bar{\mathbf{x}})\right]$$

where $\mathbf{X}_c$ is the centered input matrix containing $\mathbf{x}_i^c = \mathbf{x}_i - \bar{\mathbf{x}}$ along its rows, and $\mathbf{y}_c = \mathbf{y} - \bar{\mathbf{y}}$ is the centered output vector.

2. Consider the ridge regression problem, i.e.,

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^N(y_i - \theta_0 - \sum_{j=1}^D x_{ij}\theta_j)^2 + \lambda\sum_{j=1}^D\theta_j^2$$

Show that this problem is equivalent to the following optimization:

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^N(y_i - \theta_0 - \sum_{j=1}^D(x_{ij} - \bar{x}_j)\theta_j)^2 + \lambda\sum_{j=1}^D\theta_j^2$$

where $\bar{x}_j = \frac{1}{N}\sum_{i=1}^N x_{ij}$.

3. Assume $y_i \sim \mathcal{N}(\theta_0 + x_i^\top\theta, \sigma^2), i = 1, 2, ..., N$ and the parameters $\theta_j, j = 1, ..., D$ are each distributed as $\mathcal{N}(0, \tau^2)$, independently of each other. Assuming σ^2 and τ^2 are known, and θ_0 is not governed by a prior, show that the (minus) log-posterior density of θ is proportional to $\sum_{i=1}^N(y_i - \theta_0 - \sum_{j=1}^D x_{ij}\theta_j)^2 + \lambda\sum_{j=1}^D\theta_j^2 + C$ where $\lambda = \sigma^2/\tau^2$ and some constant C which is not dependent on θ.

4. Consider a set of random variables $(x_1, y_1), ..., (x_N, y_N)$ (training data) where each (x_i, y_i) is drawn from the same distribution $\mathcal{D}$ independently. Then, consider the least squares estimate $\hat{\theta} = \arg\min_\theta R_{\texttt{train}}(\theta)$ for the linear regression model, where $R_{\texttt{train}}(\theta) = \frac{1}{N}\sum_{i=1}^N(y_i - \theta^\top x_i)^2$. Now, consider another set of random variables $(\tilde{x}_1, \tilde{y}_1), ..., (\tilde{x}_M, \tilde{y}_M)$ (test data) where each $(\tilde{x}_i, \tilde{y}_i)$ is also drawn from the same distribution $\mathcal{D}$ independently. Prove that

$$\mathbb{E}[R_{\texttt{train}}(\hat{\theta})] \leq \mathbb{E}[R_{\texttt{test}}(\hat{\theta})],$$

where $R_{\texttt{test}}(\theta) = \frac{1}{M}\sum_{i=1}^M(\tilde{y}_i - \theta^\top\tilde{x}_i)^2$.