Homework 4: Mathematics for AI
Due date: November 25th, 11:59 pm.
Note: Submit your solution file to TA, Jihoon Tack (jihoontack@kaist.ac.kr), by email.
For questions, contact TA as well.

---

1. Lets consider a l_2 norm margin when optimizing the soft SVM (remark that we used l_1 norm in the lecture). For a classification task with a given dataset $\{\mathbf{x}_n, y_n\}_{n=1}^N \subseteq \mathcal{X} \times \mathcal{Y}$, where $\mathbf{x}_n \in \mathbb{R}^d$, and $\mathcal{Y} := \{-1, 1\}$, the optimization objective for soft SVM with l_2 norm margin is as:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{n=1}^{N} \xi_n^2 \\ \text{subject to} \quad & y_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n + b) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \text{for all } n, \end{aligned}$$

where $\xi = (\xi_n : n = 1, \ldots, N)$, ξ_n is the slack for the n-th sample $(\mathbf{x}_n, y_n)$, and C is the trade-off between width and slack.

(a) Compute that Lagrangian $\mathcal{L}$ of the given objective.

(b) Compute the dual of the given objective.

Hint: one should minimize $\mathcal{L}$ with respect to $\mathbf{w}, b$ and ξ, i.e., $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0$, $\frac{\partial \mathcal{L}}{\partial b} = 0$, $\frac{\partial \mathcal{L}}{\partial \xi} = 0$.

2. (Recap: Kernel function) Using some function (called as feature transformation) $\phi : \mathcal{X} \to \mathbb{R}^D$, the kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is as:

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\mathsf{T}\phi(\mathbf{x}').$$

Suppose K_1 and K_2 are kernel functions.

(a) Prove that for any scalar function $f : \mathcal{X} \to \mathbb{R}$, the function $K_3(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})f(\mathbf{x}')K_1(\mathbf{x}, \mathbf{x}')$ is a valid kernel.

(b) Prove that the sum $K_3(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + K_2(\mathbf{x}, \mathbf{x}')$ is a valid kernel.

(c) Prove that the product $K_3(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$ is a valid kernel.

(d) Prove that the following functions are valid kernels or not (one can use the properties (a-c)):

(i) $K(\mathbf{x}, \mathbf{x}') = e^{-c\|\mathbf{x}-\mathbf{x}'\|^2}$ (ii) $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^2$ (iii) $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\mathsf{T}\mathbf{x}' + c)^d$ ($c \in \mathbb{R}$, $d \in \mathbb{N}$)

3. Let $\bar{X} = \sum_{i=1}^{n} X_i$ where X_i is independent Bernoulli random variables (r.v):

$$X_i = \begin{cases} 0, & \text{probability with } 1-p \\ 1, & \text{probability with } p \end{cases}$$

One should prove the following statements:

$$\text{For all } \delta > 0,\ \text{Prob}(\bar{X} \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{2+\delta}\mu};$$
$$\text{For all } 0 < \delta < 1,\ \text{Prob}(\bar{X} \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2}{2}\mu};$$

where $\mu = \mathbb{E}(\bar{X}) = \sum_{i=1}^{n} p_i$. Follow the sub-problems to prove these statements.

(a) Prove the following statement where $M_X(s) := \mathbb{E}(e^{sX})$ (i.e., moment generating function):

$$M_{\bar{X}}(s) = \prod_{i=1}^{n} M_{X_i}(s).$$

(b) Prove the following statement: For a given Bernoulli r.v X_i and any $s \in \mathbb{R}$,

$$M_{X_i}(s) \leq e^{p(e^s - 1)}.$$

(c) Prove the original statements by using (a) and (b).

Hint: use Markov's inequality: for any $a > 0$, $\text{Prob}(x \geq a) \leq \frac{\mathbb{E}(x)}{a}$ (x is non-negative r.v).

4. Let $\mathbf{X}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$, $\mathbf{X}_2 = \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|}$ where $\mathbf{x}_1, \mathbf{x}_2 \sim \mathcal{N}(0, \mathbf{I}^d)$, are d-dimensional Gaussian r.v with mean 0, and standard deviation $\mathbf{I}^d$. Prove that $\mathbf{X}_1^\intercal \mathbf{X}_2$ converges to 0, i.e., orthogonal, as $d \to \infty$.