Homework 4 Solution: Mathematics for AI

---

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., $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0$, $\frac{\partial \mathcal{L}}{\partial b} = 0$, $\frac{\partial \mathcal{L}}{\partial \xi} = 0$.

Solution.

(a) **[5pt]** The Lagrangian is as follows:

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \gamma) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{n=1}^{N}\xi_n^2 - \sum_{n=1}^{N}\alpha_n(y_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n + b) - 1 + \xi_n) - \sum_{n=1}^{N}\gamma_n\xi_n, \tag{1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_N)$ and $\gamma = (\gamma_1, \ldots, \gamma_N)$.

[-1pt] minor mistake or removing ξ_n constraint w.o proper reasoning.

(b) **[10pt]** We first obtain following partial derivatives of $\mathcal{L}$ and equate them to zeros:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w}^\mathsf{T} - \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i^\mathsf{T} = 0 \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C\xi_i - \alpha_i - \gamma_i = 0 \tag{4}$$

By incorporating the Eq (2), (3), and (4) into Eq (1), then the dual function $\mathcal{D}(\alpha, \gamma)$ is as:

$$\mathcal{D}(\alpha, \gamma) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j x_i^\mathsf{T} x_j + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\frac{(\alpha_i + \gamma_i)^2}{2C}. \tag{5}$$

Hence, the dual problem is as:

$$\begin{aligned} \max_{\alpha,\gamma} \quad & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j x_i^\mathsf{T} x_j + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\frac{(\alpha_i + \gamma_i)^2}{2C} \\ \text{subject to} \quad & \sum_{i=1}^{N}\alpha_i y_i = 0, \alpha_i \geq 0, \gamma_i \geq 0. \end{aligned}$$

[-2pt] minor mistake or removing ξ_n constraint w.o proper reasoning.

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) $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\mathsf{T}\mathbf{x}' + c)^d$ ($c > 0$, $d \in \mathbb{N}$)

Solution.

Let ϕ_1 and ϕ_2 be the feature transformations of K_1 and K_2, respectively, i.e., $K_1(\mathbf{x}, \mathbf{x}') = \phi_1(\mathbf{x})^\mathsf{T} \phi_1(\mathbf{x}')$.

(a) **[3pt]** The function $K_3(\cdot)$ is as follows:

$$K_3(\mathbf{x}, \mathbf{x}') = \big(f(\mathbf{x})\phi_1(\mathbf{x})\big)^\mathsf{T} \big(f(\mathbf{x}')\phi_1(\mathbf{x}')\big).$$

Therefore, the $f(\mathbf{x})\phi_1(\mathbf{x})$ is a feature transformation of a given input $\mathbf{x}$ and K_3 is a valid kernel.

(b) **[3pt]** The function $K_3(\cdot)$ is as follows:

$$\begin{aligned} K_3(\mathbf{x}, \mathbf{x}') &= \phi_1(\mathbf{x})^\mathsf{T} \phi_1(\mathbf{x}') + \phi_2(\mathbf{x})^\mathsf{T} \phi_2(\mathbf{x}'). \\ &= \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{pmatrix}^\mathsf{T} \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{pmatrix} \end{aligned}$$

Therefore, the $\begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{pmatrix}$ is a feature transformation of a given input $\mathbf{x}$ and K_3 is a valid kernel.

(c) **[3pt]** Let $\phi_{1,i}$ and $\phi_{2,i}$ be the i-th element of ϕ_1 and ϕ_2, respectively. $K_3(\cdot)$ is as:

$$\begin{aligned} K_3(\mathbf{x}, \mathbf{x}') &= \phi_1(\mathbf{x})^\mathsf{T} \phi_1(\mathbf{x}') \cdot \phi_2(\mathbf{x})^\mathsf{T} \phi_2(\mathbf{x}') \\ &= \sum_{i=1}^{n} \phi_{1,i}(\mathbf{x}) \cdot \phi_{1,i}(\mathbf{x}') \cdot \sum_{j=1}^{m} \phi_{2,j}(x) \cdot \phi_{2,j}(x') \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \big(\phi_{1,i}(\mathbf{x}) \cdot \phi_{2,j}(\mathbf{x})\big) \cdot \big(\phi_{1,i}(\mathbf{x}') \cdot \phi_{2,j}(\mathbf{x}')\big) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{3,k}(\mathbf{x}) \cdot \phi_{3,k}(\mathbf{x}') \text{ where } \phi_{3,k} = \phi_{1,i} \cdot \phi_{2,j} \\ &= \phi_3(\mathbf{x})^\mathsf{T} \phi_3(\mathbf{x}'), \end{aligned}$$

where n and m is the number of element of ϕ_1 and ϕ_2, respectively. Therefore K_3 is a valid.

[-3pt] if one assume ϕ_1 and ϕ_2 has same dimension.

(d) **[6pt]**

(i) (Gaussian kernel) We will first prove that $K_3(\mathbf{x}, \mathbf{x}') := \exp\big(K_1(\mathbf{x}, \mathbf{x}')\big)$ is a valid kernel.

$$\exp\big(K_1(\mathbf{x}, \mathbf{x}')\big) = \sum_{n=0}^{\infty} \frac{1}{n!} K_1(\mathbf{x}, \mathbf{x}')^n,$$

is a valid kernel by 2-(b), i.e., the summation of kernels is a valid kernel, and 2-(c), i.e., the multiplication of kernels is a valid kernel.

The given original function is as:

$$e^{-c\|\mathbf{x}-\mathbf{x}'\|^2} = e^{-c\|\mathbf{x}\|^2} \cdot e^{-c\|\mathbf{x}'\|^2} \cdot e^{2c\mathbf{x}^\mathsf{T}\mathbf{x}'}.$$

Since, the function $e^{-c\|\mathbf{x}\|^2}$ is a scalar function and $e^{2c\mathbf{x}^\mathsf{T}\mathbf{x}'}$ is a valid kernel (by the above proof), one can utilizing 2-(a) to prove that $e^{-c\|\mathbf{x}-\mathbf{x}'\|^2}$ is a valid kernel.

(ii) As noticed, the problem contains an error, i.e., the kernel is not scalar, and all answers will be correct. Sorry again for the mistake.

(iii) (polynomial kernel) By the binomial expansion,

$$(\mathbf{x}^\mathsf{T}\mathbf{x}' + c)^d = \sum_{k=0}^{d} \binom{d}{k} (\mathbf{x}^\mathsf{T}\mathbf{x}')^{d-k} c^k.$$

For all $k \in \{0, \ldots, d\}$, $\binom{d}{k}(\mathbf{x}^\mathsf{T}\mathbf{x}')^{d-k}c^k$ is a valid kernel where the feature transformations is as: $\phi(\mathbf{x}) := \sqrt{\binom{d}{k}c^k}\mathbf{x}^{d-k}$. By utilizing 2-(b), i.e., the summation of kernels is a valid kernel, one can prove that $(\mathbf{x}^\mathsf{T}\mathbf{x}' + c)^d$ is a valid kernel.

[Condition]. If one *correctly* prove, i.e., provide a reasonable proof, that it is not a valid kernel by assuming $c \in \mathbb{R}$, will also get full points. Note that both are not a valid kernel for $c < 0$.

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

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

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 $s \in \mathbb{R}$,

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

(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).

Solution.

(a) **[5pt]**

$$\begin{aligned} M_{\bar{X}}(s) &= \mathbb{E}(e^{s\bar{X}}) = \mathbb{E}(e^{s\sum_{i=1}^{n} X_i}) \\ &= \mathbb{E}\left(\prod_{i=1}^{n} e^{sX_i}\right) \\ &= \prod_{i=1}^{n} \mathbb{E}(e^{sX_i}) \\ &= \prod_{i=1}^{n} M_{X_i}(s). \end{aligned}$$

(b) **[5pt]**

$$\begin{aligned} M_{X_i}(s) &= \mathbb{E}(e^{sX_i}) \\ &= p_i \cdot e^s + (1 - p_i) \cdot e^0 \text{ by the definition of expectation} \\ &= 1 + p_i(e^s - 1) \\ &\leq e^{p_i(e^s - 1)} \text{ since } 1 + a \leq e^a \text{ for all } a \in \mathbb{R}. \end{aligned}$$

(c) **[10pt]**

(i) We will first prove the first statement. For any $s > 0$,

$$\begin{aligned}
\text{Prob}(\bar{X} \geq (1+\delta)\mu) &= \text{Prob}(e^{s\bar{X}} \geq e^{(1+\delta)\mu}) && (6)\\
&\leq \mathbb{E}(e^{s\bar{X}})e^{-s(1+\delta)\mu} \quad \text{by Markov's inequality} && (7)\\
&= \prod_{i=1}^{n} M_{X_i}(s) \cdot e^{-s(1+\delta)\mu} \quad \text{by using 3-(a)} && (8)\\
&\leq e^{\mu(e^s-1)} \cdot e^{-s(1+\delta)\mu} \quad \text{by using 3-(b)} && (9)\\
&= e^{\mu(e^s-s-s\delta-1)}. && (10)
\end{aligned}$$

For a tight bound, we minimize the RHS w.r.t s by differentiating the term and set to zero:

$$\frac{\partial}{\partial s}(e^s - s - s\delta - 1) = e^s - 1 - \delta = 0,$$

which results in $s = \ln(1+\delta)$. By applying s into Eq. (10):

$$\text{Prob}(\bar{X}) \geq (1+\delta)\mu) \leq e^{\mu(\delta-(1+\delta)\ln(1+\delta))}. \tag{11}$$

For $\delta > 0$, one can obtain the following inequality (we will prove this proposition later),

$$\ln(1+\delta) \geq \frac{2\delta}{2+\delta}. \tag{12}$$

By utilizing Eq. (12) we get:

$$\mu(\delta - (1+\delta)\ln(1+\delta)) \leq -\frac{\delta^2}{2+\delta}\mu. \tag{13}$$

By applying Eq. (13) into Eq. (11), we result in:

$$\text{Prob}(\bar{X} \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{2+\delta}\mu}.$$

□

Proof of the Eq (12). Let

$$f(\delta) = \ln(1+\delta) - \frac{2\delta}{2+\delta}.$$

Since the gradient of the f is always positive for all $\delta > 0$ and $f(0) = 0$, $f(\delta)$ is always positive for all $\delta > 0$. □

(ii) We will prove the second statement. For any $s > 0$,

$$\begin{aligned}
\text{Prob}(\bar{X} \leq (1-\delta)\mu) &= \text{Prob}(e^{-s\bar{X}} \geq e^{-s(1-\delta)\mu}) && (14)\\
&\leq \mathbb{E}(e^{-s\bar{X}})e^{s(1-\delta)\mu} \quad \text{by Markov's inequality} && (15)\\
&= \prod_{i=1}^{n} M_{X_i}(-s) \cdot e^{s(1-\delta)\mu} \quad \text{by using 3-(a)} && (16)\\
&\leq e^{\mu(e^{-s}-1)} \cdot e^{s(1-\delta)\mu} \quad \text{by using 3-(b)} && (17)\\
&= e^{\mu(e^{-s}+s-s\delta-1)}. && (18)
\end{aligned}$$

For a tight bound, we minimize the RHS w.r.t s by differentiating the term and set to zero:

$$\frac{\partial}{\partial s}(e^{-s} + s - s\delta - 1) = -e^{-s} + 1 - \delta = 0,$$

which results in $s = -\ln(1-\delta)$. By applying s into Eq. (18):

$$\text{Prob}(\bar{X}) \leq (1-\delta)\mu) \leq e^{\mu(-\delta-(1-\delta)\ln(1-\delta))}. \tag{19}$$

For $0 < \delta < 1$, one can obtain the following inequality (we will prove this proposition later),

$$\ln(1-\delta) > \frac{\delta(\delta-2)}{2(1-\delta)}. \tag{20}$$

By utilizing Eq. (20) we get:

$$\mu(-\delta - (1-\delta)\ln(1-\delta)) \leq -\frac{\delta^2}{2}\mu. \tag{21}$$

By applying Eq. (21) into Eq. (19), we result in:

$$\text{Prob}(\bar{X} \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{2}\mu}.$$

□

Proof of the Eq (20). For $0 < \delta < 1$,

$$\ln(1-\delta) = -\sum_{n=1}^{\infty} \frac{\delta^n}{n}. \quad \text{by Taylor expanison}$$

Hence,

$$\begin{aligned}
(1-\delta)\ln(1-\delta) &= -\sum_{n=1}^{\infty} \frac{\delta^n}{n} + \sum_{n=1}^{\infty} \frac{\delta^{n+1}}{n}\\
&= -\Big(\delta + \frac{\delta^2}{2} + \sum_{n=3}^{\infty} \frac{\delta^n}{n}\Big) + \Big(\delta^2 + \sum_{n=2}^{\infty} \frac{\delta^{n+1}}{n}\Big)\\
&= \frac{\delta(\delta-2)}{2} + \sum_{n=3}^{\infty} \Big(\frac{1}{n^2-n}\Big)\delta^n\\
&> \frac{\delta(\delta-2)}{2}.
\end{aligned}$$

□

[-5pt] if did not minimize (10) and (18).
[-5pt] if only proved one statement.

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 $\mathbf{X}_1^\mathsf{T}\mathbf{X}_2$ converges to 0, i.e., orthogonal, as $d \to \infty$.

Solution.

[10pt] By the Strong Law of Large Numbers, for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}^d)$

$$\frac{1}{d}\|\mathbf{x}\|^2 = \frac{1}{d}\sum_i^d x_i^2 \to \mathbb{E}(x_i^2) = 1,$$

as $d \to \infty$, where x_i denotes the i-th element of $\mathbf{x}$.

Also using the fact that $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent, we have:

$$\frac{1}{d}\mathbf{x}_1^\mathsf{T}\mathbf{x}_2 = \frac{1}{d}\sum_{i=1}^d x_{1,i}x_{2,i} \to \mathbb{E}(x_{1,i}x_{2,i}) = \mathbb{E}(x_{1,i})\mathbb{E}(x_{2,i}) = 0,$$

as $d \to \infty$, where $x_{1,i}$ and $x_{2,i}$ denotes the i-th element of $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively.

Hence, as $d \to \infty$,

$$\mathbf{X}_1^\mathsf{T}\mathbf{X}_2 = \frac{\sqrt{d}}{\|\mathbf{x}_1\|} \cdot \frac{\sqrt{d}}{\|\mathbf{x}_2\|} \cdot \frac{1}{d}\mathbf{x}_1^\mathsf{T}\mathbf{x}_2 \to 0.$$

□