CS4487: Home Assignment №3

Department of Computer Science
City University of Hong Kong

Due on October 30, 2019, 7pm

Exercise 1

[5 points]. In Section 5.4 of the lecture note, we have given a detailed derivation of the dual form of SVM with soft margin. With simpler arguments, derive the dual form of SVM with hard margin

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^T w \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m. \end{aligned}$$

Exercise 2

[5 points]. Considering the two-way partitioning problem

$$\begin{aligned} \min_{x} \quad & \frac{1}{2} x^T W x \\ \text{s.t.} \quad & x_i^2 = 1, \quad i = 1, \ldots, n, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the parameter vector to be optimized and $W \in \mathbb{R}^{n \times n}$ is a fixed weight matrix. One interpretation of this problem is to partition $\{1, 2, \ldots, n\}$ in two sets; W_{ij} is cost of assigning i, j to the same set; $-W_{ij}$ is cost of assigning to different sets.

a) [1 point]. Is it a convex optimization problem? Give your reasons.

b) [3 points]. Derive the Lagrange dual function of the above problem. Hint: recall the definition of a positive definite matrix[1].

[1] `https://en.wikipedia.org/wiki/Definiteness_of_a_matrix`

ZHANG Dehang

15199998

Exercise 1

[5 points]. In Section 5.4 of the lecture note, we have given a detailed derivation of the dual form of SVM with soft margin. With simpler arguments, derive the dual form of SVM with hard margin

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^T w \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m. \end{aligned}$$

Lagrangian:

$$L(\vec{w}, \vec{b}, \vec{\alpha}) = \frac{1}{2}\|\vec{w}\|_2^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} (\vec{w}^T x^{(i)} + b) - 1 \right]$$

$$= \frac{1}{2} w^T w - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i$$

$$\nabla_{\vec{w}} L = \vec{w} - \sum_{i=1}^{m} \alpha_i y^{(i)} \vec{x}^{(i)} = 0$$

$$\nabla_{\vec{b}} L = - \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Lagrange dual function:

$$g(\alpha) = \inf_{\vec{w},\vec{b}} (L(\vec{w}, \vec{b}, \vec{\alpha}))$$

$$= \frac{1}{2} \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right) - \sum_{i=1}^{m} \alpha_i y^{(i)} \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x^{(i)} + \sum_{i=1}^{m} \alpha_i$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}.$$

Dual form:

$$\max_{\vec{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T (x^{(j)})$$

subject to $\nabla_{\vec{b}} L = - \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$.

$\alpha_i \geq 0$

Exercise 2

[5 points]. Considering the two-way partitioning problem

$$\begin{aligned} \min_{x} \quad & \frac{1}{2} x^T W x = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_i x_j \\ \text{s.t.} \quad & x_i^2 = 1, \quad i = 1, \ldots, n, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the parameter vector to be optimized and $W \in \mathbb{R}^{n \times n}$ is a fixed weight matrix. One interpretation of this problem is to partition $\{1, 2, \ldots, n\}$ in two sets; W_{ij} is cost of assigning i, j to the same set; $-W_{ij}$ is cost of assigning to different sets.

a) [1 point]. Is it a convex optimization problem? Give your reasons.

(a) The objective (cost function):

$\min_{\vec{x}} \frac{1}{2} x^T W x = \min_x \frac{1}{2} \sum_{i,j=1}^{n} x_i x_j w_{ij}$.

where $x_i^2 = 1$, $i = 1, 2, \cdots, n$.

$\frac{1}{2} x^T W x$ is nonconvex, $x_i^2 - 1 = 0$ is convex functions for $\vec{x}$, since $\frac{1}{2}(\theta \vec{x} + (1-\theta)\vec{y})^T W (\theta \vec{x} + (1-\theta)\vec{y})$ for $i = 1, \cdots, n$.

$$= \frac{1}{2} \left(\theta^2 x^T W x + \theta(1-\theta) x^T W y + \theta(1-\theta) y^T W x + (1-\theta)^2 y^T W y \right)$$

$$= \frac{1}{2} \left[\theta^2 x^T W x + 2\theta(1-\theta) x^T W y + (1-\theta)^2 y^T W y \right]$$

$$\leq \frac{1}{2} \theta x^T W x + \frac{1}{2} (1-\theta) y^T W y$$

proof:

$$\theta x^T W x + (1-\theta) y^T W y \geq \theta^2 x^T W x + 2\theta(1-\theta) x^T W y + (1-\theta)^2 y^T W y$$

$$\theta(1-\theta) x^T W x + \theta(1-\theta) y^T W y - 2\theta(1-\theta) x^T W y \geq 0.$$

since $0 < \theta < 1$, $\theta(1-\theta) > 0$

$$x^T W x + y^T W y - 2 x^T W y \geq 0.$$

$$(x-y)^T W (x-y) \geq 0.$$

W might be a negative metrix.

for eg. $\begin{bmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & -1 \end{bmatrix}$, therefore we cannot prove the inequality.

(b).

b) [3 points]. Derive the Lagrange dual function of the above problem. Hint: recall the definition of a positive definite matrix[1].

[1] https://en.wikipedia.org/wiki/Definiteness_of_a_matrix

Lagrangian: $x^T \alpha^T x$

$$L(x, \alpha) = \frac{1}{2} x^T W x + \sum_{i=1}^{n} \alpha_i (x_i^2 - 1)$$

$$= \frac{1}{2} x^T W x + \sum_{i=1}^{n} \alpha_i x_i^2 - \sum_{i=1}^{n} \alpha_i$$

$$= \frac{1}{2} x^T W x + x^T A x - \sum_{i=1}^{n} tr(A)$$ where A is a diagonal metrix $\begin{bmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_n \end{bmatrix}$

$\nabla_x L = 2x^T (\frac{1}{2} W + A)$

Lagrange dual function.

$$g(\alpha) = \inf_{\vec{x}} (L(\vec{x}, \alpha))$$

$$= \inf_x x^T (\tfrac{1}{2} W + A) x - tr(A) = \begin{cases} -tr(A), & \text{when } \frac{1}{2} W + A \geq 0. \\ -\infty & \text{otherwise.} \end{cases}$$

c) [1 point]. Try to give a nontrivial lower bound (*i.e.*, other than $-\infty$) of the optimal value p^* of the above problem. Hint: use the lower bound property.

$p^* = \frac{1}{2}(X^*)^T W X^*$

$\geq L(X^*, a) \geq g$

$= -tr(A)$ $\left(\inf \frac{1}{2}W + A \geq 0.\right)$

Therefore, the lower bound of $\frac{1}{2}X^T W X$ is $-tr(A)$. where A is the diagnal metrix $\begin{bmatrix} \alpha_1 & & \\ & \alpha_2 & \\ & & \ddots \\ & & & \alpha_n \end{bmatrix}$ where α_i is a lagrange multiplier.

Exercise 3

[5 points]. Prove the following matrix identity

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}, \qquad (1)$$

where $P \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times n}$, and $R \in \mathbb{R}^{m \times m}$. P and R are invertible. Note that if $m \ll n$, it will be much cheaper to evaluate the right-hand side than the left-hand side. Hint: right multiplying both sides by $(BPB^T + R)$. Using Eq. (1) to prove a special case

$$(I + AB)^{-1} A = A(I + BA)^{-1},$$

where $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$.

Proof: L.H.S $= \left(P^{-1} + B^T R^{-1} B\right)^{-1} B^T R^{-1}$

$= \left[P + B^T R (B^T)^{-1}\right] B^T R^{-1}$

$= P B^T R^{-1} + B^T R (B^T)^{-1} B^T R^{-1}$

$= P B^T R^{-1} + B^{-1}$

$= P B^T R^{-1} + P B^T (P B^T)^{-1} B^{-1}$

$= P B^T \left(R^{-1} + (P B^T)^{-1} B^{-1}\right)$

$= P B^T \left(B P B^T + R\right)^{-1}$ = R.H.S ※

Exercise 4 - Optional

For a pair of random variables (X, Y), sampled from the joint distribution $P(X, Y) = P(X)P(Y|X)$, the mutual information between X and Y is defined as

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)}$$
$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

Prove that for fixed $P(Y|X)$, the mutual information $I(X;Y)$ is a concave function of $P(X)$. Hint: recall (in Home Assignment I) the relationship between mutual information and entropy which is defined as

$$H(X) = -\sum_{x \in X} p(X = x) \log p(X = x).$$

$\sum P(y|x) P(x)$

$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(y|x)}{P(y)}$

$\sim (P)$

$= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(y|x) - \sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(y)$

$= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(y|x) - \sum_{y \in Y} P(y) \log P(y)$

$= \sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \log P(y|x) - \sum_{y \in Y} P(y) \log P(y).$

$= H(Y) - H(Y|X)$

$A = \theta \left[I(X;Y) \sim (P)\right] + (1-\theta)\left[I(X;Y) \sim (q)\right]$

$= \sum_{x \in X} \left[\theta P(x) + (1-\theta) q(x)\right] \sum_{y \in Y} P(y|x) \log P(y|x)$

$- \theta \sum_{y \in Y} \left(P(y|x) P(x)\right) \log \left(P(y|x) P(x)\right)$

$- (1-\theta) \sum_{y \in Y} \left(P(y|x) q(x)\right) \log \left(P(y|x) q(x)\right)$

$B = I(X;Y) \sim \left(\theta P + (1-\theta) q\right)$

$= \sum_{x \in X} \left[\theta P(x) + (1-\theta) q(x)\right] \sum_{y \in Y} P(y|x) \log P(y|x)$

$- \sum_{y \in Y} P(y|x) \left(\theta P(x) + (1-\theta) q(x)\right) \log \left(P(y|x)\left(\theta P(x) + (1-\theta) q(x)\right)\right)$

$B - A = \sum_{x \in X} \Big[\theta P(y|x) P(x) \log \left(P(y|x) P(x)\right) + (1-\theta) P(y|x) q(x) \log \left(P(y|x) q(x)\right)$

$- P(y|x) \left(\theta P(x) + (1-\theta) q(x)\right) \log \left(P(y|x)\left(\theta P(x) + (1-\theta) q(x)\right)\right) \Big]$

I is concave w.r.t. $P(x)$. iff. $B - A \geq 0$.

For each $y \in Y$. if the $B_y - A_y \geq 0$ holds, we can prove it.

For each $y \in Y$, denotes $P(y|x) = k$, $P(x) = p$, $q(x) = q$.

$\theta k p \log(kp) + (1-\theta) k q \log(kq)$

$- k\left(\theta p + (1-\theta) q\right) \log\left(k\left(\theta p + (1-\theta) q\right)\right) \geq 0$

$k > 0$

$\Leftrightarrow \theta \left[p \log p + p \log k\right] + (1-\theta)\left[q \log q + q \log k\right]$

$- \left(\theta p + (1-\theta) q\right) \log\left(\theta p + (1-\theta) q\right)$

$- \theta p \log k - (1-\theta) q \log k \geq 0$

$\Leftrightarrow \theta(p \log p) + (1-\theta)(q \log q) \geq \left[\theta p + (1-\theta) q\right] \log\left(\theta p + (1-\theta) q\right).$

$\Leftrightarrow f(x) = x \log x$ is convex. $(x > 0)$

$f''(x) = \frac{\partial}{\partial x}(1 + \log x) = \frac{1}{x} > 0.$ ※

c) [1 point]. Try to give a nontrivial lower bound (*i.e.*, other than $-\infty$) of the optimal value $p^\star$ of the above problem. Hint: use the lower bound property.

Exercise 3

[5 points]. Prove the following matrix identity

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}, \tag{1}$$

where $P \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times n}$, and $R \in \mathbb{R}^{m \times m}$. P and R are invertible. Note that if $m \ll n$, it will be much cheaper to evaluate the right-hand side than the left-hand side. Hint: right multiplying both sides by $(BPB^T + R)$. Using Eq. (1) to prove a special case

$$(I + AB)^{-1} A = A(I + BA)^{-1},$$

where $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$.

Exercise 4 - Optional

For a pair of random variables (X, Y), sampled from the joint distribution $P(X, Y) = P(X)P(Y|X)$, the mutual information between X and Y is defined as

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log \frac{p(X = x, Y = y)}{p(X = x) p(Y = y)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x) p(y)}. \end{aligned}$$

Prove that for fixed $P(Y|X)$, the mutual information $I(X;Y)$ is a concave function of $P(X)$. Hint: recall (in Home Assignment I) the relationship between mutual information and entropy which is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(X = x) \log p(X = x).$$