HW5 Handwritten Assignment

Lecturer: Pei-Yuan Wu
TAs: Po-Yang Hsieh, Le-Rong Hsu

December 2024

As the final exam approaches, we will select some problems from old assignments and exams to help students review the course materials. Problems 1, 2, 3, and 5 must be completed, while only one among Problem 4 or Problem 6 has to be answered (which will be graded).

Problem 1 (Trace Optimization)(1%)

1. Let $\Sigma \in R^{m\times m}$ be a symmetric positive semi-definite matrix, $\mu \in R^m$. Please construct a set of points $x_1, ..., x_n \in R^m$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = \Sigma, \quad \frac{1}{n}\sum_{i=1}^{n}x_i = \mu$$

   Hint: n is not given by the problem. WLOG, you may assume $\mu = 0 \in \mathbb{R}^d$.

2. Let $1 \le k \le m$, solve the following optimization problem and justify with proof:

$$\begin{array}{ll} \text{minimize} & Trace(\Phi^T\Sigma\Phi) \\ \text{subject to} & \Phi^T\Phi = I_k \\ \text{variables} & \Phi \in R^{m\times k} \end{array}$$

   In other words, you need to find Φ and verify that your Φ minimize the trace.

Problem 2 (Gradient Boosting)(1%)

Consider the binary classification problem, where we are given training data set $\{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{1, -1\}$. Let $F = \{f \mid f : \mathbb{R}^d \to \{1, -1\}\}$ be the collection of classifiers. Given number of epochs $T \in \mathbb{N}$. Suppose that we want to find the function

$$g(x) = \sum_{t=1}^{N}\alpha_t f_t(x)$$

where $f_t \in F$ and $\alpha_t \in \mathbb{R}$ for all $t = 1, \ldots, T$, by which the aggregated classifier is given by

$$h(x) = \begin{cases} 1, & \text{if } g(x) > 0 \\ -1, & \text{if } g(x) \le 0. \end{cases}$$

Please apply gradient boosting to show how the functions f_t and the coefficients α_t are computed with an aim to minimize the following loss function

$$L(g) = \sum_{i=1}^{N}\log\left(1 + e^{-y_i g(x_i)}\right).$$

Problem 3 (EM algorithm for mixture of exponential model)(1%)

Given N samples $x_1, \ldots, x_N \in [0, \infty)$, we would like to cluster them into K clusters. Assume the samples are generated according to Exponential mixture models

$$X \sim \sum_{j=1}^{K} \pi_j Exp(\tau_j)$$

where $\pi_1 + \ldots + \pi_K = 1$, and $Exp(\tau)$ denotes the exponential distribution with probability density function

$$f_\tau(x) = \begin{cases} (1/\tau)e^{-x/\tau} & , x \geq 0 \\ 0 & , x < 0. \end{cases}$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \tau_k)\}_{k=1}^{K}$.

(a) Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{(\pi_k^{(t)}, \tau_k^{(t)})\right\}_{k=1}^{K}$ to $\theta^{(t+1)} = \left\{(\pi_k^{(t+1)}, \tau_k^{(t+1)})\right\}_{k=1}^{K}$ in the following form:

$$\tau_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} x_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}, \quad \pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{ik}^{(t)}$$

(b) What is the closed form expression of $\delta_{ik}^{(t)}$?

Problem 4 (Sparse SVM)(2%)

Given training data of N input-output pairs $\mathscr{D} = ((x_i, y_i))_{i=1}^{N}$, where $x_i \in \mathcal{X}$ and $y_i \in \{\pm 1\}$. One can give two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. Suppose instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the p-norm of the vector $\alpha = (\alpha_1, ..., \alpha_N)$ that defines the weight vector $\mathbf{w}$, for some $p \geq 1$. In this question we consider the case $p = 2$, which leads to the following optimization problem:

$$\begin{array}{ll} \text{minimize} & f(\alpha, b, \boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^{N} \alpha_i^2 + \sum_{i=1}^{N} C_i \xi_i \\ \text{subject to} & y_i \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b\right) \geq 1 - \xi_i,\ i \in \{1, ..., N\} \\ \text{variables} & b \in \mathbb{R}, \alpha_i \geq 0, \xi_i \geq 0,\ i \in \{1, ..., N\} \end{array}$$

which can be rewritten in the following primal problem:

$$\begin{array}{ll} \text{minimize} & f(\alpha, b, \boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^{N} \alpha_i^2 + \sum_{i=1}^{N} C_i \xi_i \\ \text{subject to} & \left.\begin{array}{l} g_{1,i}(\alpha, b, \boldsymbol{\xi}) = 1 - \xi_i - y_i \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b\right) \leq 0 \\ g_{2,i}(\alpha, b, \boldsymbol{\xi}) = -\alpha_i \leq 0 \\ g_{3,i}(\alpha, b, \boldsymbol{\xi}) = -\xi_i \leq 0 \end{array}\right\} i \in \{1, ..., N\} \\ \text{variables} & \alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N \end{array} \tag{1}$$

as well as its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\boldsymbol{\omega}, \beta, \boldsymbol{\gamma}) = \inf_{\alpha \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} L(\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma}) \\ \text{subject to} & \omega_i \geq 0, \beta_i \geq 0, \gamma_i \geq 0,\ i \in [\![1, N]\!] \\ \text{variables} & \boldsymbol{\omega} = (\omega_1, ..., \omega_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_N) \in \mathbb{R}^N \end{array} \tag{2}$$

1. Write down the Lagrangian function $L(\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma})$ in explicit form of $\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma}$.
2. Show that the duality gap between (1) and (2) is zero.
3. Derive $\theta(\boldsymbol{\omega}, \beta, \boldsymbol{\gamma})$ in explicit form of dual variables $\boldsymbol{\omega}, \beta, \boldsymbol{\gamma}$.

4. Show that the dual problem can be simplified as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} \omega_i - \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} \omega_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i\right)_+^2 \\ \text{subject to} & \sum_{i=1}^{N} \omega_i y_i = 0 \\ \text{variables} & 0 \le \omega_i \le C_i,\ i = 1, \ldots, N \end{array} \tag{3}$$

5. Suppose $(\bar{\alpha}, \bar{b}, \bar{\boldsymbol{\xi}})$ and $(\bar{\boldsymbol{\omega}}, \bar{\beta}, \bar{\boldsymbol{\gamma}})$ are the optimal solutions to problems (1) and (2) respectively. Denote $\bar{\mathbf{w}} = \sum_{j=1}^{N} \bar{\alpha}_j y_j \mathbf{x}_j$.

   (a) Prove that

$$\bar{\alpha}_i = \max\left(\sum_{j=1}^{N} \bar{\omega}_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i, 0\right) \quad \forall i = 1, \ldots, N \tag{4}$$

   (b) Prove that

$$\bar{b} = \arg\min_{b \in \mathbb{R}} \sum_{i=1}^{N} C_i \max\left(1 - y_i\left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + b\right), 0\right), \tag{5}$$

   (c) Prove that $\bar{\xi}_i = \max\left(1 - y_i\left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}\right), 0\right)$ for all $i = 1, \ldots, N$.

   (d) Prove that

$$\left.\begin{array}{ll} \bar{\omega}_i = C_i, & \text{if } y_i\left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}\right) < 1 \\ \bar{\omega}_i = 0, & \text{if } y_i\left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}\right) > 1 \\ 0 \le \bar{\omega}_i \le C_i, & \text{if } y_i\left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + \bar{b}\right) = 1 \end{array}\right\} \forall i = 1, \ldots, N$$

Problem 5 (Invited Talk)(1%)

Please share your thoughts about this talk.

Problem 6 (Bellman Optimality Equations)(2%)

In this problem, we aim to help students review and understand the proof of the Bellman optimality equations.

Theorem. *Let Π be the set of nonstationary and randomized policies. Define*

$$V^*(s) = \sup_{\pi \in \Pi} V^\pi(s)$$
$$Q^*(s,a) = \sup_{\pi \in \Pi} Q^\pi(s,a).$$

Then there exists a stationary and deterministic policy π such that for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$V^\pi(s) = V^*(s)$$
$$Q^\pi(s,a) = Q^*(s,a).$$

Remark. The notations are consistent with those in the lecture notes.

1. Verify that V^* and Q^* is bounded between 0 and $\frac{1}{1-\gamma}$. Hence, V^* and Q^* must be finite.

2. Show that given $(s_0, a_0, r_0, s_1) = (s, a, r, s')$, the optimal discounted value γV^*, from $t = 1$ onwards, does not depend on the initial conditions s, a, and r:

$$\sup_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, (s_0, a_0, r_0, s_1) = (s, a, r, s')\right] = \gamma V^*(s').$$

3. Let π^* be a policy such that

$$\forall\, s \in \mathcal{S}, \quad \pi^*(s) \in \arg\max_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')].$$

(a) Explain that π^* is deterministic.

(b) Now suppose that the transition of states and actions is deterministic. In order to show that π^* is an optimal policy, i.e. $V^*(s) = V^{\pi^*}(s)$, we have to show two inequalities: $V^* \geq V^{\pi^*}$ and $V^* \leq V^{\pi^*}$. The first one is trivial since $\pi^* \in \Pi$. Now, please show the other inequality $V^* \leq V^{\pi^*} < \infty$.

(c) Similarly, show that $Q^{\pi^*} = Q^*$ under the assumption that the transition is deterministic.