EE 5184 Machine Learning Final Exam

Date: 2020/01/03

The paper is double-sided, 5 pages, consisting of 5 questions. Total 100 points.

Problem 1: (30 pts) Multiple Selection (多選題有倒扣，最多倒扣至本大題零分)

Please answer the following multiple selection questions. Wrong selections will result in inverted scores. ***No derivation required.***

(1) Suppose a SVM classifier is trained from data set $\{(\boldsymbol{x}_i,\ y_i)\}_{i=1}^{N}$, where $y_i \in \{\pm 1\}$ denotes the labels, and the classifier classifies $\boldsymbol{x}$ as positive label if $\mathrm{f}(\mathbf{x}) = \boldsymbol{w}^T\boldsymbol{x} + b \geq 0$.
The primal problem for solving $\boldsymbol{w}$ is given by

Minimize $\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$

Subject to $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \forall i = 1, \dots, \mathrm{N}$

Variables $\boldsymbol{w} \in \mathbb{R}^{\boldsymbol{d}},\ b \in \mathbb{R}, \xi_1, \dots, \xi_N \geq 0$

The dual problem for solving α_i's in $\boldsymbol{w} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i$ is given by

Maximize $\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i{}^T\boldsymbol{x}_j)$

Subject to $\sum_{i=1}^{N}\alpha_i y_i = 0$

Variables $0 \leq \alpha_i \leq C$

Upon achieving optimal in both primal and dual problems,

(A) If $\alpha_i = C$ then $\xi_i > 0$.
(B) If $\alpha_i > 0$ then $\xi_i > 0$.
(C) If $\alpha_i = 0$ then $\xi_i = 0$.
(D) If $\xi_i > 0$ then $\alpha_i > 0$.
(E) If $\xi_i > 0$ then $\alpha_i = C$.

(2) Select all that belong to supervised learning algorithms.
(A) Deep auto-encoder
(B) Hierarchical Agglomerative Clustering
(C) K-means
(D) Linear regression
(E) Logistic regression
(F) Locally Linear Embedding (LLE)
(G) Principle Component Analysis (PCA)
(H) Random forest
(I) Support Vector Machine (SVM)
(J) t-Distributed Stochastic Neighbor Embedding (t-SNE)

(3) Suppose you are using a kernel SVM to 2 class classification problem, where the data points are distributed on the x-y plane (i.e., data points are 2 dimensional). Suppose we choose kernel function as $\mathrm{k}\big((\mathrm{x},\mathrm{y}),(\mathrm{x}',\mathrm{y}')\big) = (xx' + yy')^2$, which of the following decision boundaries, as described by equation $\mathrm{f}(\mathrm{x},\mathrm{y}) = 0$, are possible?
(A) $\mathrm{f}(\mathrm{x},\mathrm{y}) = (x-1)^2 + 3(y+2)^2 - 2$.
(B) $\mathrm{f}(\mathrm{x},\mathrm{y}) = 2\mathrm{x} + 5\mathrm{y} - 4$.
(C) $\mathrm{f}(\mathrm{x},\mathrm{y}) = x^2 + 4\mathrm{xy} + y^2 - 7$.
(D) $\mathrm{f}(\mathrm{x},\mathrm{y}) = \mathrm{y} - \max(\mathrm{x}, 0) + 6$.
(E) $\mathrm{f}(\mathrm{x},\mathrm{y}) = |\mathrm{x}| - 3$.

(4) Suppose you are using a kernel SVM to 2 class classification problem, where the data points are distributed on the x-y plane (i.e., data points are 2 dimensional). Suppose we choose kernel function as $k((x,y),(x',y')) = (1 + xx' + yy')^2$, which of the following decision boundaries, as described by equation $f(x,y) = 0$, are possible?

(A) $f(x,y) = (x-1)^2 + 3(y+2)^2 - 2$.
(B) $f(x,y) = 2x + 5y$.
(C) $f(x,y) = x^2 + 4xy + y^2 - 7$.
(D) $f(x,y) = y - \max(x,0) + 6$.
(E) $f(x,y) = |x| - 3$.

(5) Given training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ and their corresponding labels $y_1, \ldots, y_N \in \{\pm 1\}$, a linear classifier $h(\mathbf{x}) = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$ is often determined with parameters $(\boldsymbol{w}, b)$ minimizing some loss function

$$L_{tot}(\mathbf{w}, b) = \sum_{i=1}^{N} \ell\left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\right) + \lambda L_{reg}(\mathbf{w})$$

where $\ell(\cdot)$ describes the fitting error, and $L_{reg}(\cdot)$ is the regularization term.

(A) In SVM, the fitting error takes the form $\ell(z) = \max(z, 0)$.
(B) In SVM, the regularization term takes the form $L_{reg}(\mathbf{w}) = \|\mathbf{w}\|_1$ (L1-norm).
(C) In logistic regression, the fitting error takes the form $\ell(z) = \log(1 + e^{-z})$.
(D) In logistic regression, the fitting error takes the form $\ell(z) = 1/(1 + e^{z})$
(E) In AdaBoost, the fitting error takes the form $\ell(z) = e^{-z}$.

(6) Following (1), one may rewrite the SVM primal formulation as:

$$\underset{\boldsymbol{w}\in\mathbb{R}^d,\, b\in\mathbb{R}}{\text{minimize}} \; L(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \max(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b), 0)$$

(A) Upon solving such optimization problem, gradient descent (with learning rate decreasing towards zero) always converges to global optimal, regardless of initialization.
(B) If $(\boldsymbol{w}_1, b_1)$ and $(\boldsymbol{w}_2, b_2)$ are two local optimal solutions, then $\boldsymbol{w}_1 = \boldsymbol{w}_2$ and $b_1 = b_2$.
(C) If $(\boldsymbol{w}_1, b_1)$ and $(\boldsymbol{w}_2, b_2)$ are two local optimal solutions, then $L(\boldsymbol{w}_1, b_1) = L(\boldsymbol{w}_2, b_2)$.
(D) $L(\boldsymbol{w}, b)$ is a convex function.
(E) If $(\bar{\boldsymbol{w}}, \bar{b})$ is a global optimal solution, then $L(\bar{\boldsymbol{w}}, \bar{b}) \leq NC$.

(7) Consider applying Expectation Maximization (EM) algorithm for maximum likelihood estimation of Gaussian Mixture Model parameters θ. Let $\theta^{(0)}$ be the initial parameters, and let $\theta^{(1)}, \theta^{(2)}, \ldots$ be the subsequent parameters in each epoch. Let $f(\theta)$ be the log-likelihood function.
<u>Hint: An upper-bounded non-decreasing sequence always converges.</u>

(A) The likelihood function is always non-decreasing regardless of initialization.
(B) EM algorithm always converges to the same parameters, regardless of initialization. That is, $\theta^{(t)}$ converges (elementwise) to some fixed θ^* regardless of $\theta^{(0)}$.
(C) EM algorithm always converges to the same log-likelihood, regardless of initialization. That is, $f(\theta^{(t)})$ converges to some fixed number $r \in \mathbb{R}$ regardless of $\theta^{(0)}$.
(D) The likelihood function of EM algorithm always converges, regardless of initialization. That is, $\lim_{t\to\infty} f(\theta^{(t)})$ exists regardless of $\theta^{(0)}$.
(E) EM algorithm always converges to a global optimal $\bar{\theta}$ that yields the maximum likelihood function.

(8) Which of the following statement(s) are true?

(A) In the training of a fully-connected neural network classifier (with ReLU activation function) where the cross-entropy loss is to be minimized through gradient descent, the loss is always non-increasing in each epoch regardless of initialization.

(B) A 1,000-layer fully-connected neural network with linear activation function is equivalent to a single layer neural network with linear activation function.

(C) A 1,000-layer fully-connected neural network with ReLU activation function is equivalent to a piecewise linear function.

(D) A 1,000-layer fully-connected neural network with sigmoid activation function is differentiable.

(E) The output of a softmax layer is always within the range $[0,1]$.

(9) Which of the following activation functions can be realized by maxout network?

(A) Identity function

(B) ReLU

(C) Leaky-ReLU

(D) Quadratic function

(E) Sigmoid function

(10) In the setting of variational auto-encoder, given a collection of generative models p_θ (parameterized by θ) and dataset $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \in \mathcal{X}$, one aims to find θ that maximizes the log-likelihood function $\log p_\theta(X)$. Introduce latent variables $Z = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_N\} \in \mathcal{Z}$, and for arbitrary probability distribution q_ϕ (parameterized by ϕ) on $\mathcal{X}$ and $\mathcal{Z}$, define

$$L(p_\theta, q_\phi, X) = \int_{\mathcal{Z}} q_\phi(Z|X) \log \frac{p_\theta(Z,X)}{q_\phi(Z|X)} dZ$$

$$R(p_\theta, q_\phi, X) = -\int_{\mathcal{Z}} q_\phi(Z|X)\, log \frac{p_\theta(Z|X)}{q_\phi(Z|X)} dZ$$

Which of the following statements are true?

(A) $R(p_\theta, q_\phi, X)$ is always non-negative.

(B) $R(p_\theta, q_\phi, X)$ is always non-positive.

(C) $\log p_\theta(X) = L(p_\theta, q_\phi, X) + R(p_\theta, q_\phi, X)$

(D) Fix θ and adjust ϕ, then the maximum of $L(p_\theta, q_\phi, X)$ is achieved when $q_\phi(Z|X) = p_\theta(Z|X)$ (Assume such ϕ exists).

(E) Assume $q_\phi(Z|X) = p_\theta(Z|X)$, and suppose $L(p_{\theta'}, q_\phi, X) > L(p_\theta, q_\phi, X)$, then $\log p_{\theta'}(X) > \log p_\theta(X)$.

Problem 2: (10 pts) Principle Component Analysis (PCA)

Given m samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^2$. Suppose

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i = \boldsymbol{0}, \qquad \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T = \begin{bmatrix} 66 & 12 \\ 12 & 59 \end{bmatrix}$$

(1) **(3 pts)** Find $\frac{1}{N}\sum_{i=1}^{m} \|\boldsymbol{x}_i\|_2^2$.

(2) **(4 pts)** Find the first principle axis after performing PCA on this data set.

(3) **(3 pts)** Denote $\boldsymbol{u}_i$ as the projection of $\boldsymbol{x}_i$ to the first principle axis. Find $\frac{1}{N}\sum_{i=1}^{m} \|\boldsymbol{u}_i\|_2^2$.

Problem 3: (20 pts) Concentric disks are PAC-learnable

Let $\mathcal{X} = \mathbb{R}^2$ be the input space and consider the set of concepts of the form $c = \{(x, y): x^2 + y^2 \leq r^2\}$ for some real number r. Show that this class can be (ϵ, δ)-PAC-learned from training data of size $m \geq (1/\epsilon)\log(1/\delta)$.

Problem 4: (20 pts) Expectation Maximization and Exponential Mixture Models

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

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

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

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

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

(1) **(16 pts)** 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:

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

(2) **(4 pts)** What is the closed form expression of $\delta_{ik}^{(t)}$?

Problem 5: (20 pts) Support Vector Machine with Quadratic Hinge Loss

Given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ and their corresponding labels $y_1, \dots, y_N \in \{\pm 1\}$, consider soft-margin SVM with quadratic hinge loss (referred as *quadSVM* in the following context):

$$\begin{aligned}
&\text{minimize} && \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i^2 \\
&\text{subject to} && y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i,\ \ \mathrm{i} = 1, \dots, \mathrm{N} \\
&\text{variables} && \xi_i \geq 0,\ \ \mathrm{i} = 1, \dots, \mathrm{N}
\end{aligned}$$

with $\mathrm{C} > 0$. We may rewrite *quadSVM* in the standard **primal formulation/problem**

$$\begin{aligned}
&\text{minimize} && \mathrm{f}(\mathbf{w}, \mathrm{b}, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i^2 \\
&\text{subject to} && g_i(\mathbf{w}, \mathrm{b}, \boldsymbol{\xi}) = 1 - \xi_i - \left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\right) \leq 0,\ \ \mathrm{i} = 1, \dots, \mathrm{N} \\
&\text{variables} && \xi_i \geq 0,\ \ \mathrm{i} = 1, \dots, \mathrm{N}
\end{aligned}$$

(1) **(15 pts)** Show that the **dual formulation/problem** of *quadSVM* can be written as

$$\begin{aligned}
&\text{maximize} && \theta(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j - \frac{1}{4C}\sum_{i=1}^{N} \alpha_i^2 \\
&\text{subject to} && \sum_{i=1}^{N} \alpha_i y_i = 0 \\
&\text{variables} && \alpha_i \geq 0,\ \ \mathrm{i} = 1, \dots, \mathrm{N}
\end{aligned}$$

(2) **(5 pts)** Let $(\overline{\boldsymbol{w}}, \bar{b}, \bar{\boldsymbol{\xi}})$ be a primal optimal solution, $\bar{\boldsymbol{\alpha}}$ be a dual optimal solution. Which of the following statements are true?

(A) $\mathrm{f}(\overline{\boldsymbol{w}}, \bar{b}, \bar{\boldsymbol{\xi}}) = \theta(\bar{\boldsymbol{\alpha}})$

(B) $\bar{\xi}_i = max\left(1 - y_i\left(\overline{\boldsymbol{w}}^T\boldsymbol{x}_i + \bar{b}\right), 0\right)$

(C) If $y_i\left(\overline{\boldsymbol{w}}^T\boldsymbol{x}_i + \bar{b}\right) > 1$, then $\bar{\alpha}_i = 0$.

(D) $0 \leq \bar{\alpha}_i \leq \mathrm{C}$ for all $\mathrm{i} = 1, \dots, \mathrm{N}$.

(E) There exists $\gamma > 0$ such that $\bar{\xi}_i = \gamma\bar{\alpha}_i$ for all $\mathrm{i} = 1, \dots, \mathrm{N}$.