EE 5184 Machine Learning, Fall 2021
Final Exam - Solution

Lecturer: Pei-Yuan Wu

January 6, 2022

This exam contains 6 problems and 120 pts in total. In this exam,

- For $x \in \mathbb{R}$, we denote $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$.
- $[\![m, n]\!]$ denotes the set of integers from m to n.

1. **(15%) Changing activation function in two-layered network**
Consider a two-layer network $\mathbf{f} : \mathbb{R}^D \to \mathbb{R}^K$ defined as

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_K(\mathbf{x}) \end{bmatrix}, \quad f_k(\mathbf{x}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x^{(i)} + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

where $\mathbf{x} = (x^{(1)}, ..., x^{(D)})$, and that h, σ denote the nonlinear activation functions for the hidden and output layers, respectively.

   (a) **(7%)** Please draw the neural network that computes function $\mathbf{f}(\mathbf{x})$. Please specify the nodes and the connections, as well as their correspondence to the parameters $w_{ji}^{(1)}$ and $w_{kj}^{(2)}$.

   (b) **(8%)** Suppose originally the activation function of the hidden layer is chosen as the sigmoid function $h(z) = \frac{1}{1+e^{-z}}$. Show that there exists an equivalent network, which computes exactly the same function $\mathbf{f}(\mathbf{x})$, but with hidden layer activation functions given by

$$h_{new}(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}.$$

   Hint: First find the relation between $h(z)$ and $\tanh(z)$, and then show that the parameters of the two networks differ by linear transformations.

2. **(15%) Linear regression interpretation**
Given training data of N input-output pairs $\mathcal{D} = ((x_i, y_i))_{i=1}^N$, where $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$. Consider the generative model $p_\theta(\xi|x) = \mathcal{N}(\xi; f_\theta(x), \sigma^2)$, where $\mathcal{N}(\xi; \mu, \sigma^2)$ denotes the probability density function of Gaussian distribution with mean μ and variance σ^2, namely

$$\mathcal{N}(\xi; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\xi-\mu)^2}{2\sigma^2}\right).$$

(a) **(7%)** Write down the error function $L_{ml}(\theta)$ to be minimized so as to maximize the log-likelihood function of generative model p_θ, assuming each of the N data points (x_i, y_i) is generated independently.

(b) **(8%)** Assume $\theta = (\theta^{(1)}, ..., \theta^{(L)})$, under Bayesian settings we assume θ follows some prior distribution where $\theta^{(1)}, ..., \theta^{(L)}$ are independent and Gaussian distributed with mean 0 and variance λ^2. Denote $\mathcal{D}_x = (x_i)_{i=1}^N$ and $\mathcal{D}_y = (y_i)_{i=1}^N$, we assume θ and $\mathcal{D}_x$ are independent. Show that the posterior of θ after observing the training data $\mathcal{D}$ is given by $p(\theta|\mathcal{D}) \propto \exp(-L_{Bayes}(\theta))$, where

$$L_{Bayes}(\theta) = \frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - f_\theta(x_i))^2 + \frac{1}{2\lambda^2}\sum_{i=1}^N \|\theta\|_2^2$$

with $\|\theta\|_p = \sqrt[p]{\sum_{l=1}^L \left(\theta^{(l)}\right)^p}$ denoting the p-norm of θ. Hence the regularization term can be interpreted as the prior we pose on the parameters θ.
Hint:

$$p(\theta|\mathcal{D}) = \frac{p(\theta, \mathcal{D}_y|\mathcal{D}_x)}{p(\mathcal{D}_y|\mathcal{D}_x)} = \frac{p(\theta|\mathcal{D}_x)p(\mathcal{D}_y|\theta, \mathcal{D}_x)}{\int p(\theta'|\mathcal{D}_x)p(\mathcal{D}_y|\theta', \mathcal{D}_x)d\theta'}$$

3. **(15%) Gradient boosting with class-dependent risk**
Let $\mathcal{X}$ be the input space, $\mathcal{F}$ be a collection of multiclass classifiers that map from $\mathcal{X}$ to $\{\pm 1\}$. Let $((x_i, y_i))_{i=1}^m$ be the training data set, where $x_i \in \mathcal{X}$ and $y_i \in \{\pm 1\}$. Given $T \in \mathbb{N}$, suppose we want to build classifer $h : \mathcal{X} \to \{\pm 1\}$ of the form

$$h(x) = \text{sign}(g_{T+1}(x)), \quad g_{T+1}(x) = \sum_{t=1}^T \alpha_t f_t(x),$$

where $f_t \in \mathcal{F}$ and $\alpha_t \in \mathbb{R}$ for all $t \in [\![1, T]\!]$. Please show how the functions f_t and coefficients α_t are chosen by gradient boosting with an aim to minimize the following loss function:

$$L(g_{T+1}) = C_1 \sum_{i:y_i=1} e^{-g_{T+1}(x_i)} + C_2 \sum_{i:y_i=-1} e^{g_{T+1}(x_i)}$$

where C_1 and C_2 are positive scalars indicating the panelty of misclassifying a postive (negative) sample as a negative (positive) one. Note that AdaBoost corresponds to the special case where $C_1 = C_2 = 1$.

4. **(20%) Gradient descent convergence**
In this question we will study a sufficient condition under which gradient descent is guaranteed to converge.

   (a) **(5%)** Let $f : [a, b] \to \mathbb{R}$ be a differentiable function, and that f' is γ-Lipschitz, namely

   $$|f'(x) - f'(y)| \leq \gamma |x - y|$$

   for all $x, y \in [a, b]$. Show that the following holds for all $x \in [a, b]$:

   $$|f(x) - f(a) - f'(a)(x - a)| \leq \gamma (x - a)^2/2.$$

   Hint: Recall by fundamental theorem of calculus that

   $$f(x) = f(a) + \int_a^x f'(t)dt.$$

   (b) **(10%)** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, and that ∇f is γ-Lipschitz, namely

   $$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \gamma \|\mathbf{x} - \mathbf{y}\|_2$$

   for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Recall that in gradient descent we start from some $\mathbf{x}_0 \in \mathbb{R}^n$ and iteratively apply the following updates:

   $$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta \nabla f(\mathbf{x}_{k-1}), \quad k = 1, 2, ...$$

   Show that

   $$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) - \eta(1 - \gamma\eta/2)\|\nabla f(\mathbf{x}_{k-1})\|_2^2.$$

   Hence the objective function decreases after gradient descent update.
   Hint: Consider $g(t) = f((1 - t)\mathbf{x}_{k-1} + t\mathbf{x}_k)$ and apply (a).

   (c) **(5%)** Continue (b), suppose we have the additional condition:

   $$\|\nabla f(\mathbf{x})\|^2 \geq \alpha(f(\mathbf{x}) - f(\mathbf{x}^*)) \tag{1}$$

   for all $\mathbf{x} \in \mathbb{R}^n$, where $f(\mathbf{x}^*) = \inf_{x \in \mathbb{R}^n} f(\mathbf{x})$ is the optimum. Show that

   $$f(\mathbf{x}) - f(\mathbf{x}^*) \leq (1 - \alpha\eta(1 - \gamma\eta/2))\,(f(\mathbf{x}) - f(\mathbf{x}^*)) \quad \forall 0 \leq \eta \leq 2/\gamma.$$

   Hence the objective function decreases exponentially. (1) is usually referred as the α**-PL (Polyak-Lojasiewicz)** condition.

5. **(20%) EM algorithm for mixture of Bernoulli model**
Consider the generative model parameterized by $\theta = (\pi_k, \boldsymbol{\mu}_k)_{k=1}^K$, where $\pi_1, ..., \pi_K \in [0,1]$ satisfies $\sum_{k=1}^K \pi_k = 1$, and that $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K \in [0,1]^D$, so that the probability of generating a D-dimensional binary vector $\mathbf{x} = (x^{(1)}, ..., x^{(D)}) \in \{0,1\}^D$ is

$$p(\mathbf{x}; \theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^D \mu_{kj}^{x^{(j)}} (1 - \mu_{kj})^{1 - x^{(j)}}$$

In other words, with given $\boldsymbol{\mu}_k$, the elements $x^{(1)}, ..., x^{(D)}$ are independent, where $x^{(j)}$ follows Bernoulli distribution of mean μ_{kj}. Suppose we observe training data of N binary vectors $\mathbf{x}_1, ..., \mathbf{x}_N \in \{0,1\}^D$, derive the E-step and M-step equations of the EM algorithm for optimizing the mixing coefficients π_k and the Bernoulli means μ_{kj} by maximum likelihood.

6. **(35%) Sparse SVM**
Given training data of N input-output pairs $\mathcal{D} = ((x_i, y_i))_{i=1}^N$, 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:

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

as well as its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\boldsymbol{\omega}, \beta, \boldsymbol{\gamma}) = \inf_{\alpha \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} L(\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma}) \\ \text{subject to} & \omega_i \geq 0, \beta_i \geq 0, \gamma_i \geq 0, \ i \in [\![1, N]\!] \\ \text{variables} & \boldsymbol{\omega} = (\omega_1, ..., \omega_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_N) \in \mathbb{R}^N \end{array} \tag{3}$$

(a) **(5%)** Write down the Lagrangian function $L(\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma})$ in explicit form of $\alpha, b, \boldsymbol{\xi}, \boldsymbol{\omega}, \beta, \boldsymbol{\gamma}$.

(b) **(5%)** Show that the duality gap between (2) and (3) is zero.

(c) **(5%)** Derive $\theta(\boldsymbol{\omega}, \beta, \boldsymbol{\gamma})$ in explicit form of dual variables $\boldsymbol{\omega}, \beta, \boldsymbol{\gamma}$.

(d) **(5%)** Show that the dual problem can be simplified as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} \omega_i - \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \omega_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i \right)_+^2 \\ \text{subject to} & \sum_{i=1}^{N} \omega_i y_i = 0 \\ \text{variables} & 0 \leq \omega_i \leq C_i, \ i = 1, ..., N \end{array} \tag{4}$$

(e) **(15%)** Suppose $(\bar{\alpha}, \bar{b}, \bar{\boldsymbol{\xi}})$ and $(\bar{\boldsymbol{\omega}}, \bar{\beta}, \bar{\boldsymbol{\gamma}})$ are the optimal solutions to problems (2) and (3) respectively. Denote $\bar{\mathbf{w}} = \sum_{j=1}^{N} \bar{\alpha}_j y_j \mathbf{x}_j$.

i. **(4%)** Prove that

$$\bar{\alpha}_i = \max \left(\sum_{j=1}^{N} \bar{\omega}_j y_j y_i \mathbf{x}_j \cdot \mathbf{x}_i, 0 \right) \ \forall i = 1, ..., N \tag{5}$$

ii. **(4%)** Prove that

$$\bar{b} = \arg\min_{b \in \mathbb{R}} \sum_{i=1}^{N} C_i \max \left(1 - y_i \left(\bar{\mathbf{w}} \cdot \mathbf{x}_i + b \right), 0 \right), \tag{6}$$

iii. **(4%)** 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 $i = [\![1, N]\!]$.

iv. **(3%)** Prove that

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