The Chinese University of Hong Kong, Shenzhen
School of Data Science

MDS5210 · Homework 2
Due: (23:59), March 26

Instructions:

- Homework problems must be carefully and clearly answered to receive full credit. Complete sentences that establish a clear logical progression are highly recommended.
- You must submit your assignment in Blackboard. Please upload a file or a zip file. The file name should be in the format **last name-first name-hw2**.
- The homework must be written in English.
- Late submission will not be graded.
- Each student **must not copy** homework solutions from another student or from any other source.

Problem 1 (20pts). Fundamental Knowledge for Generalization Theory

(1) When the random variables can have *negative* values, the Markov inequality may fail. Give an example.

(2) How will sample size and variance influence concentration? Explain it using Chebyshev's inequality.

(3) Suppose $X_1, X_2, \cdots, X_n$ are i.i.d sub-Gaussian random variables with bounded variance, then we can apply both Chebyshev's inequality and Hoeffding's inequality. Discuss which bound is tighter.

(4) Explain why i.i.d. sampling assumption is important for generalization theory.

(5) Suppose all training data points are from distribution $\mathcal{D}$ in an i.i.d. manner. If we increse the number of training data points, will the generalization error increase or decrease?

(6) Given a hypothesis class $\mathcal{H}$ with $\{-1, +1\}$-valued hypotheses. Consider the empirical Rademacher complexity defined in page 47 of lecture slides 4. When $|\mathcal{H}| = 1$, what is $\hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H})$? When $|\mathcal{H}| = 2^n$, what is $\hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H})$?

(7) By generalization theorey, explain how to make out-of-sample error small through .

Problem 2 (20pts).
Suppose $f \in \mathcal{H}$ maps sample space $\mathcal{X}$ to $\{0, 1\}$. The Rademacher complexity bound derived in our lecture is by applying McDiarmid's inequality to

$$h(\mathcal{S}) = \sup_{f \in \mathcal{H}} [\mathrm{Er}_{\mathrm{out}}(f) - \mathrm{Er}_{\mathrm{in}}(f)]$$

Try to obtain a generalization bound in terms of the empirical Rademacher complexity $\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H})$ by applying McDiarmid's inequality to

$$\Psi(S) = \sup_{f \in \mathcal{H}} [\mathrm{Er}_{\mathrm{out}}(f) - \mathrm{Er}_{\mathrm{in}}(f) - \widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H})]$$

Hint: You can follow the analysis in slides4 page 56-57 and use directly the result in slides6 page 58-60 that $\mathbb{E}[h(\mathcal{S})] \leq \mathcal{R}(\mathcal{H})$.

Problem 3 (40pts). Overfitting
Toy polynomial regression using ℓ_2-regularization and cross-validation. Suppose that we have the underlying model

$$y = x^2 + \varepsilon. \tag{1}$$

We collect $n = 10$ data points $\{(x_i, y_i)\}_{i=1}^n$; see the visualization in Figure 1. You can download the data from blackboard.

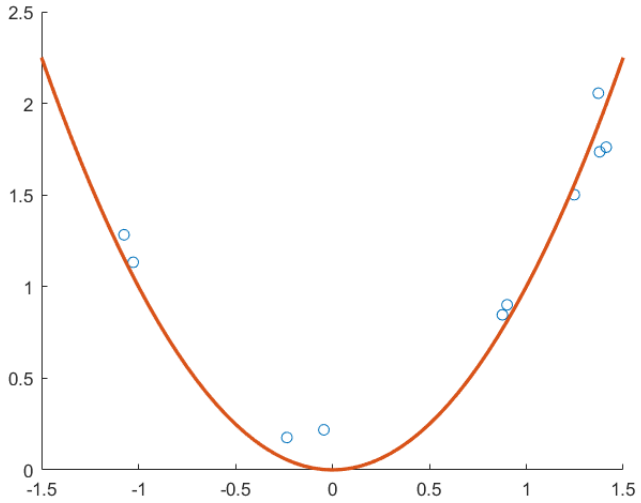

Figure 1: visualization of data

Now, suppose we are going to fit all the data using 8-th order polynomials:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_8 x^8. \tag{2}$$

(a1) **[2 points]** Denote the $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_8) \in \mathbb{R}^9$ as the parameter. We have the following linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}.$$

Specify $\boldsymbol{y}$ and $\boldsymbol{X}$ using the training data $\{(x_i, y_i)\}_{i=1}^n$.

(a2) **[5 points]** Furthermore, we can formulate the following least squares

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^9}{\operatorname{argmin}} \ \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2. \tag{3}$$

Calculate $\widehat{\boldsymbol{\theta}}$ defined in (3) and plot the fitted curve in Figure 1 (limit x-axis from -1.5 to 1.5 and limit y-axis from -0.5 to 2.5). You can download the code used to generate Figure 1 from blackboard.

(a3) **[3 points]** Using the test data set, calculate the test error $\|\boldsymbol{X}_{\text{test}}\widehat{\boldsymbol{\theta}} - \boldsymbol{y}_{\text{test}}\|_2$ of $\widehat{\boldsymbol{\theta}}$ defined in (3).

(b1) **[20 points]** Since we know that the underlying model in (1) is quadratic, while the fitting model in (2) is polynomial of order 8, we must have overfitting, which you can see from question (a2) and (a3). One way to prevent overfitting is regularization. Instead of using (3), we formulate the following ℓ_2-regularized least squares

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^9}{\operatorname{argmin}} \ \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2. \tag{4}$$

However, one difficulty to implementing (4) is determining the regularization parameter λ. A too large λ leads to underfitting, while a too small λ (e.g. $\lambda = 0$) results in overfitting. Suppose we set the set of candidates of λ as

$$[10^{-5}\ 10^{-4}\ 10^{-3}\ 10^{-2}\ 10^{-1}\ 0.3\ 0.5\ 0.8\ 1\ 2\ 5\ 10\ 15\ 20\ 50\ 100]$$

using **5-fold cross-validation** to select the regularization parameter λ, and plot the validation error versus the value of λ, where error is the y-axis and λ is the x-axis (set the x-axis to log-scale).

(b2) **[5 points]** Based on the result in (b1), set $\lambda = 0.01$, 0.1, 0.8, and 5 in (4) and solve for the corresponding $\widehat{\boldsymbol{\theta}}$, respectively. Plot the fitted curve using the former four choices of λ in Figure 1, you have to draw four figures separately (limit x-axis from -1.5 to 1.5 and limit y-axis from -0.5 to 2.5).

(b3) **[5 points]** Using the test data set, calculate the test error $\|\boldsymbol{X}_{\text{test}}\widehat{\boldsymbol{\theta}} - \boldsymbol{y}_{\text{test}}\|_2$ of each of $\widehat{\boldsymbol{\theta}}$ obtained in (b2).

Problem 4 (40pts). Multinomial and Ordinal Logistic Regression

In this problem we are going to use logistic regression to solve the Mobile Price Classification task. In particular, we are given a certain group of features for a phone, such as battery power, RAM, clock speed, etc. Our objective is to predict the price range of the phone, which can be viewed as a classification problem.

The data format is a csv file. The training data contains 1700 samples, and the test data contains 300 samples. The first 20 columns are features and the last column is the label. You may check the file for more detail.

Note: You are not allowed to use any pre-built packages for auto-differentiation.

(a) **[25 points] Multinomial Logistic Regression.** Each class $l = 1, \cdots, K$ is assigned a weight vector $\boldsymbol{\theta}^l$. The a-posterior of y_i is modeled by

$$\Pr[y_i = l|\boldsymbol{\Theta}, \boldsymbol{x}_i] = \frac{\exp(\langle \boldsymbol{\theta}^l, \boldsymbol{x}_i \rangle)}{\sum_{j=1}^{K} \exp(\langle \boldsymbol{\theta}^j, \boldsymbol{x}_i \rangle)}.$$

The problem can be formulated as

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d \times K}}{\operatorname{argmin}} \; -\frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{K} \mathbf{1}_{\{y_i = \ell\}} \log \left(\frac{\exp(\langle \boldsymbol{\theta}^\ell, \boldsymbol{x}_i \rangle)}{\sum_{j=1}^{K} \exp(\langle \boldsymbol{\theta}^j, \boldsymbol{x}_i \rangle)} \right), \tag{5}$$

where $\mathbf{1}_{\{y_i = \ell\}}$ is the indicator function defined as

$$\mathbf{1}_{\{y_i = \ell\}} = \begin{cases} 1, & y_i = \ell \\ 0, & \text{otherwise.} \end{cases}$$

The predicted label for sample $\boldsymbol{x}_i$ will be

$$\underset{j}{\operatorname{argmax}} \, e_j^T \hat{\boldsymbol{\Theta}}^T \boldsymbol{x}_i.$$

Learning Objective. You are asked to solve (5) by applying accelerated gradient descent (see Algorithm 1). The settings and tasks are

(1) Normalize the feature by $x \leftarrow \frac{x-\mu}{\sigma}$ before training.

(2) Set $\alpha = 0.01$, and maximum iteration T to 2000.

(3) Plot the training loss and test loss versus iteration in the same figure.

(4) Print out the training accuracy and test accuracy for the final iteration.

(5) Determine the most important feature which affects the price of the phone (Hint: compare the corresponding parameters' norm).

(b) **[15 points] Ordinal Logistic Regression.** We notice the price range have orders. To capture the order information, we consider parameterizing the cumulative probability as

$$\Pr[y_i \leq j|\boldsymbol{\theta}, \boldsymbol{x}_i] = h(\boldsymbol{\theta}^T \boldsymbol{x}_i - \boldsymbol{z}_j),$$

Algorithm 1 Accelerated Gradient Descent

Input: Observed data X, y and initialization parameter $\boldsymbol{\theta}_0$.
$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \alpha \nabla f(\boldsymbol{\theta}_0)$
for $t = 1$ to $T - 1$ **do**
  $y_t = \boldsymbol{\theta}_t + \frac{t-1}{t+2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})$
  $\boldsymbol{\theta}_{t+1} = y_t - \alpha \nabla f(y_t)$
end for
Return $\boldsymbol{\theta}_T$

where we use $h(t) := \frac{1}{1+e^{-t}}$ to represent the sigmoid function. Here, $\boldsymbol{\theta}$ and $\boldsymbol{z}_j$ are parameters to be learned. Note we have only one $\boldsymbol{\theta}$ for all class now. The probability of being class j can be calculated as

$$\begin{aligned}\Pr[y_i = j|\boldsymbol{\theta}, \boldsymbol{x}_i] &= \Pr[y_i \leq j|\boldsymbol{\theta}, \boldsymbol{x}_i] - \Pr[y_i \leq j-1|\boldsymbol{\theta}, \boldsymbol{x}_i] \\ &= h(\boldsymbol{\theta}^T \boldsymbol{x}_i - \boldsymbol{z}_j) - h(\boldsymbol{\theta}^T \boldsymbol{x}_i - \boldsymbol{z}_{j-1}).\end{aligned}$$

Therefore, the problem can be formulated as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} -\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{K} \mathbf{1}_{y_i = l} \log(h(\boldsymbol{\theta}^T \boldsymbol{x}_i - \boldsymbol{z}_j) - h(\boldsymbol{\theta}^T \boldsymbol{x}_i - \boldsymbol{z}_{j-1})). \tag{6}$$

You are asked to solve (6) by following the same instruction as the multinomial logistic regression. Compare the performance of two algorithms.