THE CHINESE UNIVERSITY OF HONGKONG, SHENZHEN
COMPUTER AND INFORMATION ENGINEERING

Homework #1

Student name: *Fanyi Meng (223015127)*

Course: *Advanced Machine Learning (AIR 6002)* – Professor: *Prof Tongxin Li*
Due date: *January 13th, 2024*

Question 1: Basics

(*a*) What is a hypothesis set?

(*b*) What is the hypothesis set of a linear model?

(*c*) What is overfitting?

(*d*) What are two ways to prevent overfitting?

(*e*) What are training data and test data, and how are they used differently? Why should you never change your model based on information from test data?

(*f*) What are the two assumptions we make about how our dataset is sampled?

(*g*) Consider the machine learning problem of deciding whether or not an email is spam. What could X, the input space, be? What could Y, the output space, be?

(*h*) What is the k-fold cross-validation procedure?

Answer.

(*a*) What is a hypothesis set?

(*b*) What is the hypothesis set of a linear model?

(*c*) What is overfitting?

(*d*) What are two ways to prevent overfitting?

(*e*) What are training data and test data, and how are they used differently? Why should you never change your model based on information from test data?

(*f*) What are the two assumptions we make about how our dataset is sampled?

(*g*) Consider the machine learning problem of deciding whether or not an email is spam. What could X, the input space, be? What could Y, the output space, be?

(*h*) What is the k-fold cross-validation procedure?

Question 2: Bias-Variance Tradeoff

(*a*) Derive the bias-variance decomposition for the squared error loss function. That is, show that for a model f_S trained on a dataset S to predict a target $y(x)$ for each x,

$$\mathbb{E}_S\left[E_{\text{out}}\left(f_S\right)\right] = \mathbb{E}_x[\text{Bias}(x) + \text{Var}(x)]$$

given the following definitions :

$$\begin{aligned} F(x) &= \mathbb{E}_S\left[f_S(x)\right] \\ E_{\text{out}}\left(f_S\right) &= \mathbb{E}_x\left[\left(f_S(x) - y(x)\right)^2\right] \\ \text{Bias}(x) &= (F(x) - y(x))^2 \\ \text{Var}(x) &= \mathbb{E}_S\left[\left(f_S(x) - F(x)\right)^2\right] \end{aligned}$$

(*b*) For each $N \in \{20, 25, 30, 35, \cdots, 100\}$:

i. Perform 5-fold cross-validation on the first N points in the dataset (setting aside the other points), computing both the training and validation error for each fold.
   - Use the mean squared error loss as the error function.
   - Use NumPy's `polyfit` method to perform the degree-d polynomial regression, and NumPy's `polyval` method to help compute the errors. (Refer to example code and NumPy documentation for details.)
   - When partitioning your data into folds, divide the data into K contiguous blocks (though in practice, you should randomize your partitions, for the purpose of this exercise, use contiguous blocks).

ii. Compute the average of the training and validation errors from the 5 folds.

iii. Create a learning curve by plotting both the average training and validation error as functions of N.

Answer. While this question leaves out the crucial element of the geographic origin of the swallow, according to Jonathan Corum, an unladen European swallow maintains a cruising airspeed velocity of **11 metres per second**, or **24 miles an hour**. The velocity of the corresponding African swallows requires further research as kinematic data is severely lacking for these species.

Question 3

Find the closed-form solutions of the following optimization problems ($\mathbf{W} \in \mathbb{R}^{K \times D}, N \gg D > K$):

(*a*) $\min\limits_{W,b} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{b}\|^2$

(*b*) $\min\limits_{W,b} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{b}\|^2 + \frac{\lambda}{2}\|\mathbf{W}\|_F^2$

Answer.

Question 4

Consider the following problem

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W}\Phi - \mathbf{Y}\|_F^2 + \frac{\lambda}{2}\|\mathbf{W}\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm; $Y \in \mathbb{R}^K \times N, \Phi = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_N)], \mathbf{x}_i \in \mathbb{R}^D$, $i = 1, 2, \ldots, N$ and ϕ is the feature map induced by a kernel function $k(\cdot, \cdot)$. Prove that for any x$\in \mathbb{R}^D$, we can make prediction as

$$\mathbf{y} = \mathbf{W}\phi(\mathbf{x}) = \mathbf{Y}\left(\mathbf{K} + \lambda\mathbf{I}\right)^{-1}\mathbf{k}(\mathbf{x}),$$

where $\mathbf{K} = \Phi^\top\Phi$ and $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}), k(\mathbf{x}_2, \mathbf{x}), \ldots, k(\mathbf{x}_N, \mathbf{x})]^\top$.

Answer.

Question 5

Compute the space and time complexities (in the form of big O, consider only the training stage) of the following algorithms:

(*a*) Ridge regression (Question 2(b)) with the closed-form solution

(*b*) N data points of $D-$dimension, choose d principal components)

(*c*) Neural network with architecture $D - H_1 - H_2 - K$ on a mini-batch of size B (consider only the forward process and neglect the computational costs of activation functions)

[Hint: the time complexity of $A \in \mathbb{R}^{m\times n} \times B \in \mathbb{R}^{n\times l}$ is $O(mnl)$; the time complexities of eigenvalue decomposition and inverse of an $n \times n$ matrix are both $O(n^3)$.]

Answer.

Question 6

Prove the convergence of the generic gradient boosting algorithm (AnyBoost). Specifically, suppose in the algorithm of AnyBoost (page 14 of Lecture 02), the gradient of the objective function $\mathcal{L}$ is L-Lipschitz continuous, i.e., there exists $L > 0$ such that

$$\|\nabla\mathcal{L}(H) - \nabla\mathcal{L}(H')\| \leq L\|H - H'\|$$

holds for any H and H'. Suppose in the algorithm, α is computed as

$$\alpha_{t+1} = -\frac{\langle\nabla\mathcal{L}(H_t), h_{t+1}\rangle}{L\|h_{t+1}\|^2}.$$

Then the ensemble model is updated as $H_{t+1} = H_t + \alpha_{t+1}h_{t+1}$. Prove that the algorithm either terminates at round T with $\langle\nabla\mathcal{L}(H_t), h_{t+1}\rangle$ or $\mathcal{L}(H_t)$ converges to a finite value, in which case

$$\lim_{t\to\infty}\langle\nabla\mathcal{L}(H_t), h_{t+1}\rangle = 0.$$

* [Hint: Using the following result: Suppose $\mathcal{L} : \mathcal{H} \to \mathbb{R}$ and $\|\nabla\mathcal{L}(F) - \nabla\mathcal{L}(G)\| \leq L\|F - G\|$ holds for any F and G in $\mathcal{H}$, then $\mathcal{L}(F + wG) - \mathcal{L}(F) \leq w\langle\nabla\mathcal{L}(F), G\rangle + \frac{Lw^2}{2}\|G\|^2$ holds for any $w > 0$.]

Answer.

Question 7:SGD

Linear regression learns a model of the form:

$$f(x_1, x_2, \cdots, x_d) = \left(\sum_{i=1}^{d} w_i x_i\right) + b$$

(*a*) We can make our algebra and coding simpler by writing $f(x_1, x_2, \cdots, x_d) = \mathbf{w}^T\mathbf{x}$ for vectors w and x. But at first glance, this formulation seems to be missing the bias term b from the equation above. How should we define x and w such that the model includes the bias term?

Linear regression learns a model bv minimizing the squared loss function L, which is the sum across all training data $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$ of the squared difference between actual and predicted output values:

$$L(f) = \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2$$

(*b*) We can make our algebra and coding simpler by writing $f(x_1, x_2, \cdots, x_d) = \mathbf{w}^T\mathbf{x}$ for vectors w and x. But at first glance, this formulation seems to be missing the bias term b from the equation above. How should we define x and w such that the model includes the bias term?

(c)-(f) Coding Part.

Answer.

Question 8

True or False? If False, then explain shortly.

(*a*) The inequality $G(\mathcal{F}, n) \leq n^2$ holds for any model class $\mathcal{F}$.

(*b*) The VC dimension of an axis-aligned rectangle in a 2D space is 4.

(*c*) The VC dimension of a circle in a 2D space is 4.

(*d*) The VC dimension of 1-nearest neighbor classifier in d-dimensional space is $d + 1$.

(*e*) Let d be the VC dimension of $\mathcal{F}$. Then the inequality $G(\mathcal{F}, n) \leq \left(\frac{en}{d}\right)^d$ always holds.

Answer.

Question 9

In LASSO, the model class is defined as $\mathcal{F} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_1 \leq \alpha\}$. Suppose $\mathbf{x} \in \mathbb{R}^d$, $y \in \{-1, +1\}$, the training data are $S = \{(\mathbf{x}_i, y_i)\}_i = 1^n$, and $\max_{1 \leq i \leq n} \|\mathbf{x}_i\|_\infty \leq \beta$, where $\|\cdot\|_\infty$ denotes the largest absolute element of a vector.

(a) Find an upper bound of the empirical Rademacher complexit, where σ_i are the Rademacher variables.

(b) Suppose the loss function is the absolute loss. Use the inequality (highlighted in blue) on page 30 and Lemma 5 on page 35 (i.e., (i.e.,$\mathcal{R}(\ell \circ \mathcal{F}) \leq \eta \mathcal{R}(\mathcal{F})$)) of Lecture 03 to derive a generalization error bound for LASSO.

* Hint: For question (a), please use the inequality $\langle \mathbf{a}, \mathbf{b} \rangle \leq \|\mathbf{a}\|_1 \|\mathbf{b}\|_\infty$ and the following lemma:

Lemma 1. Let $A \subseteq \mathbb{R}^n$ be a finite set of points with $r = \max_{\mathbf{x} \in A} \|\mathbf{x}\|_2$ and denote $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Then

$$\mathbb{E}_\sigma \left[\max_{\mathbf{x} \in A} \sum_{i=1}^{n} x_i \sigma_i \right] \leq r \sqrt{2 \log |A|}$$

where $|A|$ denotes the cardinality of set A and σ_i are the Rademacher variables.

Answer.