Homework 5 Learning Theory

Spring 2020
(Due: Friday, April 10, 2020, 11:59am EST)

Please submit your homework through **blackboard** by no later than 11:59am EST. To help us grade, please insert a front page that contains your name only. For subsequent pages, please insert page break between Q1 and Q2. Try writing in black ink and not pencil.

Objective

The objectives of this homework are:

(a) More concrete insights on VC dimensions

(b) Bias-variance study on a regularized learning problem

Exercise 1: VC Dimension

Recall from lecture that VC dimension of a hypothesis set is a number that characterizes the "complexity" of the set. We went through a sequence of reasoning leading to its theoretical roles in the generalization bounds. However, we are lacking in understanding of how it really behaves for concrete hypothesis sets; this is the motivation for this exercise. In part (a), we will practice computing the VC dimension on a few relatively simple hypothesis sets. In part (b), we shall show that, even though we might expect the number of tunable parameters of a hypothesis set should be approximately equal to the VC dimension of the set, it is *not* generally true.

(a) Compute the VC dimension of the following hypothesis sets.

(i) $\mathcal{H} = \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in [a,\infty), a \in \mathbb{R}\} \cup \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in (-\infty, a], a \in \mathbb{R}\}$. To clarify, the first subset is the positive ray, and the second subset is the negative ray. So the union is the set of all positive rays and negative rays.

(ii) $\mathcal{H} = \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in [a,b], a, b \in \mathbb{R}\} \cup \{h : \mathbb{R} \to \{-1,+1\} | h(x) = -1, \forall x \in [a,b], a, b \in \mathbb{R}\}$. So this is the union of the positive intervals and the negative intervals.

(iii) $\mathcal{H} = \{h : \mathbb{R}^2 \to \{-1,+1\} | h(\boldsymbol{x}) = +1, \forall \boldsymbol{x} \text{ where } \|\boldsymbol{x}\|_2 \leq b, \;\; b \in \mathbb{R}\}$. Note that this is a *concentric* circle.

(iv) $\mathcal{H} = \{h : \mathbb{R}^2 \to \{-1,+1\} | h(\boldsymbol{x}) = +1, \forall \boldsymbol{x} \text{ where } \|\boldsymbol{x} - \boldsymbol{a}\|_2 \leq b, \;\; \boldsymbol{a} \in \mathbb{R}^2, b \in \mathbb{R}\}$. Note that this is a circle with an arbitrary center $\boldsymbol{a}$.

(b) As we saw in lecture, the VC dimension of the perceptron hypothesis set corresponds to the number of parameters of the hypothesis set (recall the weights $\boldsymbol{w} \in \mathbb{R}^{d+1}$), and this observation is "usually" true for other hypothesis sets. But now, consider the following hypothesis set that has only one parameter $\alpha \in \mathbb{R}$:

$$\mathcal{H} = \{h_\alpha : \mathbb{R} \to \mathbb{R} | h_\alpha(x) = (-1)^{\lfloor \alpha x \rfloor}, \alpha \in \mathbb{R}\} \tag{1}$$

Prove that this hypothesis set has an infinite VC-dimension. What does this imply about the sample and model complexity of $\mathcal{H}$, based on what you learned in class? Do you think if this is a better hypothesis set than the perceptron hypothesis set? Why?

Remark: $\lfloor y \rfloor$ is the flooring operator which returns the closest integer smaller than or equal to y.

Hint: Recall that VC dimension requires you to know the growth function. The growth function is the worst case estimate of the number of dichotomies that can ever be created. So you need to construct a dataset containing $x_1, \ldots, x_N$ first. Move around these data points until you find the maximum number of dichotomies. The hint here is to consider $(x_1, x_2, ..., x_N) = (10^0, 10^1, ..., 10^{N-1})$. Say α is some number with at least $N-1$ decimal places. What do you notice about $\lfloor \alpha x_i \rfloor = \lfloor \alpha \times 10^i \rfloor$ for each i?

Exercise 2: Bias-Variance Trade-off

In this problem, we carry out a bias-variance study on a regularized learning problem. We introduce the setup of the problem first.

Often in a learning problem, the available training data is noisy. Assume a data space $\mathcal{X} = \mathbb{R}^d \times \{1\}$ (the $\{1\}$ accounts for the offset term, which will be useful to us when we look at linear classifiers) and target space $\mathcal{Y} = \mathbb{R}$. We denote the target function (ground truth) in the *noiseless* setting $f : \mathcal{X} \to \mathcal{Y}$. We assume a Gaussian noise, that is, the target function in the *noisy* setting is defined as $y : \mathbb{R}^d \to \mathbb{R}$ where $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In this exercise, we restrict $f(\boldsymbol{x}) = \boldsymbol{\theta}_f^T \boldsymbol{x}$, for some $\boldsymbol{\theta}_f \in \mathbb{R}^{d+1}$. We will give more details about why this assumption can be more realistic than it appears at the end of this exercise.

Suppose we wish to train a linear classifier on training data $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$, i.e. for each n, $y_n = \boldsymbol{\theta}_f^T \boldsymbol{x}_n + \epsilon_n$, and our hypothesis set is the set of linear classifiers $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y} | h(\boldsymbol{x}) = \boldsymbol{\theta}_h^T \boldsymbol{x}, \boldsymbol{\theta}_h \in \mathbb{R}^{d+1}\}$. To take into account the presence of noise in the available data and avoid overfitting, instead of directly using $E_{\text{in}}(h) = \frac{1}{N} \sum_{n=1}^N (\boldsymbol{\theta}_h^T \boldsymbol{x}_n - y_n)^2$ (convince yourself this equality is true!) as the training error measure, we choose the following *augmented* (regularized) measure:

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N} \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h \tag{2}$$

where $\lambda > 0$. The regularization term is often called the *weight decay*.

Let us study the bias and variance of this learning algorithm.

(a) Let us write the aggregated data matrix as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{x}_1^T & 1 \\ \boldsymbol{x}_2^T & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_N^T & 1 \end{bmatrix} \in \mathbb{R}^{N \times (d+1)} \tag{3}$$

Show that the optimal linear classifier $h^{(\mathcal{D})}(\boldsymbol{x}) = \boldsymbol{\theta}_{\mathcal{D}}^T \boldsymbol{x}$ with respect to error measure (2) has the following weight vector

$$\boldsymbol{\theta}_{\mathcal{D}} = (\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A}^T (\boldsymbol{A} \boldsymbol{\theta}_f + \boldsymbol{\epsilon}) \tag{4}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_N)$, and $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$.

(b) Use (a) to show that

$$\boldsymbol{\theta}_{\mathcal{D}} = \boldsymbol{\theta}_f - \lambda (\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\theta}_f + (\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A}^T \boldsymbol{\epsilon} \tag{5}$$

(c) Assume the data is normalized such that $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = \boldsymbol{I}$. We wish to show the following is true:

$$\begin{aligned} \text{bias} &\approx \frac{\lambda^2}{(\lambda + N)^2} \|\boldsymbol{\theta}_f\|_2^2 \\ \text{var} &\approx \frac{\sigma^2}{N} \mathbb{E}[\text{trace}(H^2(\lambda))] \end{aligned} \tag{6}$$

where $H(\lambda) = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T$, and N large. The following steps outline one way of proving the above results, feel free to use another approach if you prefer.

(i) Let us first compute the bias term. Recall that bias $= \mathbb{E}_{\boldsymbol{x}}[(\bar{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2]$, where in our case, $\bar{g}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}[h^{(\mathcal{D})}(\boldsymbol{x})]$, i.e. it is the expectation of the final hypothesis over all possible realizations of the dataset. Prove that

$$\bar{g}(\boldsymbol{x}) = \boldsymbol{\theta}_f^T\boldsymbol{x} - \lambda\boldsymbol{x}^T\mathbb{E}_{\mathcal{D}}[(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}]\boldsymbol{\theta}_f \tag{7}$$

Hint: Denote $\boldsymbol{y} = (y_1, ..., y_N)$, and consider the random variables $\boldsymbol{A}$ and $\boldsymbol{b}$. Note that $\mathbb{E}_{\mathcal{D}}[\cdots] = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\cdots] = \mathbb{E}_{\boldsymbol{A},\boldsymbol{y}}[\cdots] = \mathbb{E}_{\boldsymbol{A}}[\mathbb{E}_{\boldsymbol{y}|\boldsymbol{A}}[\cdots|\boldsymbol{A}]]$. Use independence of $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$.

(ii) Show that

$$(\bar{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 = \lambda^2\text{trace}(\boldsymbol{x}\boldsymbol{x}^T\mathbb{E}_{\boldsymbol{A}}[(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}]\boldsymbol{\theta}_f\boldsymbol{\theta}_f^T\mathbb{E}_{\boldsymbol{A}}[(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}]) \tag{8}$$

Hint: Make use of the identity that, for any positive semi-definite matrix $\boldsymbol{M}$, $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{x}] = \text{trace}(\boldsymbol{M}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^T])$. The identity follows from the facts that trace is invariant under cyclic permutations and linear.

(iii) Prove the expression for the bias term in (6), making use of the facts that $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^T] = \boldsymbol{I}$ and consequently $\boldsymbol{A}^T\boldsymbol{A} \approx N\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^T] = N\boldsymbol{I}$.

(iv) Now we compute the variance term. Recall it is defined by the expression var $= \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{\mathcal{D}}[(h^{(\mathcal{D})}(\boldsymbol{x}) - \bar{g}(\boldsymbol{x}))^2]]$. Let us compute the inner expression first, by showing that

$$\mathbb{E}_{\mathcal{D}}[(h^{(\mathcal{D})}(\boldsymbol{x}) - \bar{g}(\boldsymbol{x}))^2] = \sigma^2\mathbb{E}_{\boldsymbol{A}}[\text{trace}(\boldsymbol{x}\boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}] \tag{9}$$

Hint: The hints from before could still be useful here.

(v) Prove the variance expression in (6).

(d) What asymptotic properties can you observe about the bias and variance terms (from (c)) as functions of λ? For instance, what are their values when $\lambda = 0$? How do they behave when λ becomes larger and larger? Can you explain why? How about the behavior of these two terms with respect to N?

Remark: The assumption that the target function must be linear seems very restrictive, but with the help of nonlinear transformations of the data space, it can actually capture a much greater variety of target functions. For instance, suppose we have (potentially normalized) data lying in the space $\mathcal{Z} = [-1, 1]$, and we have a target function $f : \mathcal{Z} \to \mathcal{Y}$ that is a polynomial of degree Q_f. We can apply the nonlinear transformation

$$\mathcal{Z} \to \mathcal{X}, \boldsymbol{z} \mapsto \boldsymbol{x} = (1, L_0(\boldsymbol{z}), L_1(\boldsymbol{z}), ..., L_Q(\boldsymbol{z})) \tag{10}$$

where the L_i's are Legendre polynomials (a family of orthogonal polynomials useful for learning problems). Then if $Q \geq Q_f$, we can write the target function on the transformed space $\mathcal{X}$ in the form $f(\boldsymbol{x}) = \boldsymbol{\theta}_f^T\boldsymbol{x}$ for some $\boldsymbol{\theta}_f \in \mathbb{R}^{Q+1}$, and treat our learning problem on $\mathcal{X}$. You can read the lecture slides for details.

