CS 4780/5780 Homework 6

Due: Thursday 11/08/18 11:55pm on Gradescope

Problem 1: Regularization Mitigates Overfitting

In this question, we are going to investigate how adding l2 regularization can help mitigate the effect of overfitting for ordinary least square regression. First, recall that in our notes for lecture 10, we mention that we can rewrite the objective function of l2-regularized least square regression (or ridge regression)

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 + \lambda ||\mathbf{w}||_2^2$$

as

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 \text{ subject to } ||\mathbf{w}||_2^2 \leq B^2$$

To simplify our analysis, we are going to focus on the second expression. In addition, we are going to assume the following:

(i) Each data point $(\mathbf{x}_i, y_i)$ is drawn identically and independently from the distribution $\mathcal{P}$, namely, the dataset $\mathcal{D} \sim \mathcal{P}^n$

(ii) For any $(\mathbf{x}, y)$ sampled from $\mathcal{P}$, we have $||\mathbf{x}||_2^2 = 1$

With the above assumption, we are going to prove that regularizing the classifier will reduce its *variance*:

(a) Notice that $\mathbf{w}(\mathcal{D})$ is a function of $\mathcal{D}$ and since $\mathcal{D}$ is random, we can consider $\mathbf{w}(\mathcal{D})$ to be a random variable. Define the expected classifier to be $\bar{\mathbf{w}} = \mathbb{E}_{\mathcal{D}}(\mathbf{w}(\mathcal{D}))$. Show that the squared distance from thee mean is bounded above as follows

$$||\mathbf{w}(\mathcal{D}) - \bar{\mathbf{w}}||_2^2 \leq 4B^2.$$

Hint: Remember that the regularizing constraint forces $\mathbf{w}$ (and the mean $\bar{\mathbf{w}}$) to be inside a ball around the origin with radius B. Use the triangular inequality (with the origin) to show that the maximum Euclidean distance between any two points can at most be $2B$.

(b) Define the model $h_{\mathcal{D}}(\mathbf{x}) = \mathbf{w}(\mathcal{D})^T \mathbf{x}$ and $\overline{h}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}(h_{\mathcal{D}}(\mathbf{x}))$. Show that the variance of the model is

$$\mathbb{E}_{\mathbf{x},\mathcal{D}}((h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x}))^2) \leq 4B^2$$

by first showing that

$$h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x}) = (\mathbf{w}(\mathcal{D}) - \overline{\mathbf{w}})^T \mathbf{x}$$

and then using the Cauchy-Schwarz inequality:

$$(a^T b)^2 \leq (a^T a)(b^T b)$$

to conclude the result.

Takeaway: By adding regularization, we essentially bound the variance of the model which reduces overfitting.

Problem 2: Bias and Variance in KNN

In this question, we are going to study the bias and variance of k-nearest neighbour regression. Suppose the data arises from a model $y_i = f(x_i) + \varepsilon_i$. All ε_i in different (x_i, y_i) are i.i.d. random variables with $E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$. Denote D as the training set. The expected prediction error at a single x is

$$\mathrm{EPE}_k(x) = E_{D,(x,y)}[(y - h_k(x))^2],$$

where $y = f(x) + \varepsilon$. (Note that ε is also i.i.d. as ε_i.) For simplicity, we assume that the values of x_i and x in the training sample are fixed in advance (nonrandom), while the value of y_i and y are random variables as defined. In the specific KNN regression model,

$$h_k(x) = \frac{1}{k}\sum_{l=1}^{k} y_{(l)} = \frac{1}{k}\sum_{l=1}^{k}(f(x_{(l)}) + \varepsilon_{(l)}),$$

where $x_{(l)}$ is the lth closest point to x in D. Try to decompose $\mathrm{EPE}_k(x)$ into three components: variance, noise and bias. Please also verify that the variance will drop off as k is increased. (Hint: Because training samples are nonrandom, $x_{(l)}$ $(l = 1, \cdots, k)$ are fixed too.)