Fall 2018 MATH895 HW1

Qinghong(Jackie) Xu

September 19, 2018

1 Ex2.4

Since inputs are drawn from a spherical multinormal distribution $\mathbf{x} \sim N(0, \mathbf{I}_p)$, with each component is drawn from a normal distribution, which is $x_i \sim N(0,1)$. Let $z = \mathbf{a^T x}$ be the projection of each of the training points on this direction. Then we have

$$z = \mathbf{a^T x} = \sum_{j=1}^{p} a_j x_j$$

Also we know that if $x_1, .., x_p$ are independent standard normal random variable, then their linear combination has normal distribution. Hence, since $\mathbf{a}$ is a unit vector, we have

$$\mathbf{E}(z) = \mathbf{E}(\sum_{j=1}^{p} a_j x_j) = \sum_{j=1}^{p} a_j \mathbf{E}(x_j) = 0$$

$$\mathbf{Var}(z) = \sum_{j=1}^{p} a_j^2 \mathbf{Var}(x_j) = \sum_{j=1}^{p} a_j^2 = 1$$

Therefore, $z \sim N(0,1)$ and $\mathbf{E}(z^2) = 1$. For any target point $x \sim N(0, \mathbf{I}_p)$, which is a spherically symmetric distribution, the expected squared distance is

$$\mathbf{E}(x^2) = \mathbf{E}(\sum_{j=1}^{p} z_j^2) = \sum_{j=1}^{p} \mathbf{E}(z_j^2) = p$$

Hence for $p = 10$, a randomly drawn test point is about $\sqrt{10} = 3.2$ standard deviations from the origin, while all the training points are on average one standard deviation along direction $\mathbf{a}$. So most prediction points see themselves as lying on the edge of the training set.

2 Ex2.6

Our goal is to minimize

$$RSS(\theta) = \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2$$

If there are observations with tied or identical values of x, then we can rewrite the least squares problem as

$$RSS(\theta) = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{n_i} (y_{ij} - f_\theta(x_i))^2$$

where $\bar{N}$ denotes the number of different x values and n_i is the number of different y values corresponding to x_i. Expand the square in the function gives:

$$RSS(\theta) = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{n_i} (y_{ij} - f_\theta(x_i))^2 = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{n_i} (y_{ij}^2 - 2y_{ij} f_\theta(x_i) + f_\theta^2(x_i))$$

Drop the term that doesn't depend on θ and we only need to minimize

$$RSS(\theta) = -2\sum_{i=1}^{\bar{N}} \sum_{j=1}^{n_i} y_{ij} f_\theta(x_i) + \sum_{i=1}^{\bar{N}} \sum_{j=1}^{n_i} f_\theta^2(x_i) = -2\sum_{i=1}^{\bar{N}} \sum_{j=1}^{n_i} y_{ij} f_\theta(x_i) + n_i \sum_{i=1}^{\bar{N}} f_\theta^2(x_i)$$

Denote $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, which is the mean of y values with the identical x_i. Then we have

$$RSS(\theta) = \sum_{i=1}^{\bar{N}} n_i(-2\bar{y}_i f_\theta(x_i) + f_\theta^2(x_i)) = \sum_{i=1}^{\bar{N}} n_i (f_\theta(x_i) - \bar{y}_i)^2 - \sum_{i=1}^{\bar{N}} n_i \bar{y}_i^2$$

Therefore we have a weighted reduced least squares problem, which is minimizing

$$RSS(\theta) = \sum_{i=1}^{\bar{N}} n_i (f_\theta(x_i) - \bar{y}_i)^2$$

And solving this problem is equivalent to solving the original least squares problem.

3 Ex2.7

3.1 (a)

For linear regression:

$$\hat{f}(x_0) = x_0^T \beta = x_0^T (A^T A)^{-1} A^T y = \sum_{i=1}^{N} x_0^T ((A^T A)^{-1} A^T)_i y_i$$

where

$$A = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix}$$

$x_1, ..., x_N$ are training inputs and $y_i, ..., y_N$ are training outputs, $((A^T A)^{-1} A^T)_i$ is the ith column of the matrix. Therefore, $l_i(x_0; \chi) = x_0^T ((A^T A)^{-1} A^T)_i$ in this case.
For k-nearest-neighbor regression:

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in N_k(x_0)} y_i$$

Therefore, $l_i(x_0; \mathcal{X}) = \begin{cases} \frac{1}{k}, & x_i \in N_k(x_0) \\ 0, & x_i \notin N_k(x_0) \end{cases}$

3.2 (b)Decompose the conditional mean-squared error

Since $f(x_0)$ is not random, we have

$$\begin{aligned}\mathbf{E}_{\mathcal{Y}|\mathcal{X}}(f(x_0)-\hat{f}(x_0))^2 &= \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(f(x_0))^2 - 2\mathbf{E}_{\mathcal{Y}|\mathcal{X}}(f(x_0)\hat{f}(x_0)) + \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))^2 \\ &= f^2(x_0) - 2f(x_0)\mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) + \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))^2 \\ &= (f(x_0)-\mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)))^2 + \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))^2 - (\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0))^2 \\ &= (\mathbf{bias}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0))^2 + \mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)\end{aligned}$$

3.3 (c)Decompose the unconditional mean-squared error

Follow the same process as in part(b), we have

$$\begin{aligned}\mathbf{E}_{\mathcal{Y},\mathcal{X}}(f(x_0)-\hat{f}(x_0))^2 &= \mathbf{E}_{\mathcal{Y},\mathcal{X}}(f(x_0))^2 - 2\mathbf{E}_{\mathcal{Y},\mathcal{X}}(f(x_0)\hat{f}(x_0)) + \mathbf{E}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0))^2 \\ &= f^2(x_0) - 2f(x_0)\mathbf{E}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) + \mathbf{E}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0))^2 \\ &= (f(x_0)-\mathbf{E}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)))^2 + \mathbf{E}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0))^2 - (\mathbf{E}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 \\ &= (\mathbf{bias}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 + \mathbf{Var}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0)\end{aligned}$$

3.4 (d)Some insights in the above two cases

There is another way to decompose the unconditional mean-squared error, which is using the law of the total expectations and the law of the total variance:

$$\begin{aligned}\mathbf{E}_{\mathcal{Y},\mathcal{X}}(f(x_0)-\hat{f}(x_0))^2 &= \mathbf{E}_{\mathcal{X}}\mathbf{E}_{\mathcal{Y}|\mathcal{X}}(f(x_0)-\hat{f}(x_0))^2 \\ &= \mathbf{E}_{\mathcal{X}}(\mathbf{bias}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0))^2 + \mathbf{E}_{\mathcal{X}}\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) \\ &= (f(x_0)-\mathbf{E}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 + \mathbf{Var}_{\mathcal{X}}\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) + \mathbf{E}_{\mathcal{X}}\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) \\ &= (f(x_0)-\mathbf{E}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 + \mathbf{Var}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0) \\ &= (\mathbf{bias}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 + \mathbf{Var}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0)\end{aligned}$$

Moreover, we want to explore some insight about the bias-variance trade-off of these two cases (b)&(c) under the class of estimator in this problem. So we plug in $\hat{f}(x_0)=\sum_{i=1}^N l_i(x;\mathcal{X})y_i$. We claim that $\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)=\sum_{i=1}^N l_i(x_0;\mathcal{X})f(x_i)$. This is because:

$$\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) = \mathbf{E}_{\mathcal{Y}|\mathcal{X}}\sum_{i=1}^N l_i(x_0;\mathcal{X})y_i = \sum_{i=1}^N \mathbf{E}_{\mathcal{Y}|\mathcal{X}} l_i(x_0;\mathcal{X})y_i$$

Since $l_i(x_0;\mathcal{X})$ is not dependent on $\mathcal{Y}$, and $\mathcal{X}$ is fixed, we have

$$\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) = \sum_{i=1}^N l_i(x_0;\mathcal{X})\mathbf{E}_{\mathcal{Y}|\mathcal{X}}y_i = \sum_{i=1}^N l_i(x_0;\mathcal{X})\mathbf{E}_{\mathcal{Y}|\mathcal{X}}(f(x_i)+\epsilon_i) = \sum_{i=1}^N l_i(x_0;\mathcal{X})f(x_i)$$

We also have $\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) = \sigma^2 \sum_{i=1}^N l_i^2(x_0;\mathcal{X})$. This is because

$$\begin{aligned}
\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) &= \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0) - \mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0))^2 \\
&= \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\sum_{i=1}^N l_i(x_0;\mathcal{X})y_i - \sum_{i=1}^N l_i(x_0;\mathcal{X})f(x_i))^2 \\
&= \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\sum_{i=1}^N l_i(x_0;\mathcal{X})(y_i - f(x_i)))^2 \\
&= \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\sum_{i=1}^N l_i(x_0;\mathcal{X})\epsilon_i)^2 \\
&= \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\sum_{i=1}^N l_i^2(x_0;\mathcal{X})\epsilon_i^2 + 2\sum_{i,j=1,i\neq j}^N \epsilon_i\epsilon_j l_i(x_0;\mathcal{X})l_j(x_0;\mathcal{X}))
\end{aligned}$$

And we know that the crossing term is 0 since ϵ_i, ϵ_j are independent random variables and $\epsilon \sim N(0,1)$. Therefore,

$$\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) = \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(\sum_{i=1}^N l_i^2(x_0;\mathcal{X})\epsilon_i^2) = \sum_{i=1}^N \mathbf{E}_{\mathcal{Y}|\mathcal{X}}(l_i^2(x_0;\mathcal{X})\epsilon_i^2) = \sum_{i=1}^N l_i^2(x_0;\mathcal{X})\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\epsilon_i^2 = \sigma^2\sum_{i=1}^N l_i^2(x_0;\mathcal{X})$$

So we have

$$(\mathbf{bias}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0))^2 = (f(x_0) - \sum_{i=1}^N l_i(x_0;\mathcal{X})f(x_i))^2$$

$$\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) = \sigma^2\sum_{i=1}^N l_i^2(x_0;\mathcal{X})$$

$$(\mathbf{bias}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 = (f(x_0) - \mathbf{E}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 = (f(x_0) - \mathbf{E}_{\mathcal{X}}\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0))^2 = (f(x_0) - \sum_{i=1}^N \mathbf{E}_{\mathcal{X}} l_i(x_0;\mathcal{X})f(x_i))^2$$

$$\mathbf{Var}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0) = \mathbf{Var}_{\mathcal{X}}\mathbf{E}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) + \mathbf{E}_{\mathcal{X}}\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) = \mathbf{Var}_{\mathcal{X}}\sum_{i=1}^N l_i(x_0;\mathcal{X})f(x_i) + \sigma^2\sum_{i=1}^N \mathbf{E}_{\mathcal{X}} l_i^2(x_0;\mathcal{X})$$

When applying k-NN method, we can analyze the bias-variance trade-off for part(b), in this case:

$$(\mathbf{bias}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0))^2 = (f(x_0) - \frac{1}{k}\sum_{i=1}^k f(x_i))^2$$

$$\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0) = \frac{\sigma^2}{k}$$

we can see the bias term will increase with the number of k. This is because when taking smaller number of points, where the values $f(x_i)$ are closer to $f(x_0)$, the average will be closer to $f(x_0)$, compared with taking larger number of points (since some may be far away from $f(x_0)$). While the variance term will decrease with the number of k increasing. However, for part(c), where $\mathcal{X}$ and $\mathcal{Y}$ are not random, it is hard to investigate the trade-off between the bias term and the variance term. Since it will be

$$(\mathbf{bias}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0))^2 = (f(x_0) - \frac{1}{k}\sum_{i=1}^k \mathbf{E}_{\mathcal{X}} f(x_i))^2$$

$$\mathbf{Var}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0) = \frac{1}{k^2}\sum_{i=1}^{k}\mathbf{Var}_{\mathcal{X}}f(x_i) + \frac{\sigma^2}{k}$$

So the bias term will depend on $\mathbf{E}_{\mathcal{X}}f(x_i)$ and the variance term will depend on $\mathbf{Var}_{\mathcal{X}}f(x_i)$. However, we can observe that $\mathbf{Var}_{\mathcal{Y},\mathcal{X}}\hat{f}(x_0)$ is always larger than $\mathbf{Var}_{\mathcal{Y}|\mathcal{X}}\hat{f}(x_0)$

4 Ex2.8

4.1 The results for linear regression method

the error for training data is: 0.575953923686%
the error for training data digit-2: 0.410396716826%
the error for training data digit-3: 0.759878419453%
the error for test data is: 4.12087912088%
the error for test data digit-2: 3.53535353535%
the error for test data digit-3: 5.42168674699%

4.2 The results for k-nearest neighbours method

when $k = 1$, the error for training data is 0.000000%
when $k = 3$, the error for training data is 0.503960%
when $k = 5$, the error for training data is 0.575954%
when $k = 7$, the error for training data is 0.647948%
when $k = 15$, the error for training data is 0.935925%
when $k = 1$, the error for test data is 2.472527%
when $k = 3$, the error for test data is 3.021978%
when $k = 5$, the error for test data is 3.021978%
when $k = 7$, the error for test data is 3.296703%
when $k = 15$, the error for test data is 3.846154%

4.3 Plot of the error ratio for each case