IEMS 402 Statistical Learning - 2025 Winter
HW3

Tuorui Peng[1]

Exercise 1 Ensemble and Bias-Variance Trade-off

1.(a) Weight Average or Prediction Average?

Note that weight average & prediction average are both linear functions of the predictions. Thus, the two are equivalent if we are using linear models (do not expending the model space), while for non-linear models like nn, the two are not equivalent (new model space).

1.(b) Bagging - Uncorrelated Models

1.(b).(i) Bias with bagging

By total expectation law, we have

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{D}_i}\left[h(x;\mathcal{D}_i)|x\right]\right] = \mathbb{E}_{\mathcal{D}}\left[h(x;\mathcal{D}_i)|x\right]$$

Thus we have

$$\begin{aligned}
\text{bias} =& \mathbb{E}\left[\left(\mathbb{E}\left[\bar{h}(x;\mathcal{D})|x\right] - y_*(x)\right)^2\right] \\
=& \mathbb{E}\left[\left(\mathbb{E}\left[k^{-1}\sum_{i=1}^{k} h(x;\mathcal{D}_i)|x\right] - y_*(x)\right)^2\right] \\
=& \mathbb{E}\left[\left(\mathbb{E}\left[h(x;\mathcal{D})|x\right] - y_*(x)\right)^2\right]
\end{aligned}$$

1.(b).(ii) Variance with bagging

With uncorrelated assumption, we have

$$\begin{aligned}
\text{variance} =& \mathbb{E}\left[\left(\bar{h}(x;\mathcal{D}) - \mathbb{E}\left[\bar{h}(x;\mathcal{D})|x\right]\right)^2\right] \\
=& \mathbb{E}\left[\left(k^{-1}\sum_{i=1}^{k} h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)^2\right] \\
=& k^{-2}\sum_{i=1}^{k}\mathbb{E}\left[\left(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)^2\right] \\
&+ k^{-2}\sum_{i\neq j}\underbrace{\mathbb{E}\left[\left(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)\left(h(x;\mathcal{D}_j) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)\right]}_{0} \\
=& k^{-1}\sigma^2
\end{aligned}$$

[1] TuoruiPeng2028@u.northwestern.edu

1.(c) Bagging- General Case

1.(c).(i) Bias under Correlation

Note that in the derivation of bias, we only utilized the total expectation law, thus the bias term remains the same.

1.(c).(ii) Variance under Correlation

Turn to correlated case with $corr\big(h(x;\mathcal{D}_i), h(x;\mathcal{D}_j)\big) = \delta_{ij}(1-\rho)+\rho$, we have

$$\begin{aligned}
\text{variance} =& \mathbb{E}\left[\big(\bar{h}(x;\mathcal{D}) - \mathbb{E}\left[\bar{h}(x;\mathcal{D})|x\right]\big)^2\right] \\
=& \mathbb{E}\left[\big(k^{-1}\sum_{i=1}^{k} h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)^2\right] \\
=& k^{-2}\sum_{i=1}^{k}\mathbb{E}\left[\big(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)^2\right] \\
&+ k^{-2}\sum_{i\neq j}\mathbb{E}\left[\big(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)\big(h(x;\mathcal{D}_j) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)\right] \\
=& k^{-1}\sigma^2 + \frac{k-1}{k}\sigma^2\rho \\
=& \sigma^2\big(\rho + \frac{1-\rho}{k}\big)
\end{aligned}$$

1.(c).(iii) Intuitions on bagging

- Generall we should have a positive ρ due to the similarity between boostrapped models, thus the variance term decreases towards $\rho\sigma^2$ as k increases. i.e. more ensembles could reduce the variance, but we would still have variance induced by the intrinsic correlation.
- $\rho = 0$ case: variance term is σ^2/k, which could reduce to 0 as k increases.
- $\rho = 1$ case: variance term is σ^2, this case bagging does not help to reduce the variance.

Exercise 2 Central Limit Theorem for Kernel Density Estimator

2.(a)

Kernel density estimator is defined as

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})$$

1. For any $q > 1$, we have

$$\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h}) - p_h(x)\right|^q\right] \begin{cases} \geq & \frac{1}{2^q}\left[\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})\right|^q\right] - \mathbb{E}\left[|p_h(x)|^q\right]\right] \\ \leq & \frac{1}{2^q}\left[\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})\right|^q\right] + \mathbb{E}\left[|p_h(x)|^q\right]\right] \end{cases}$$

in which

$$\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})\right|^q\right] = \frac{1}{h^{q-1}}\int |K(\|v\|)|^q\, p(x+hv)\,\mathrm{d}v = \Theta(h^{-q})$$
$$\mathbb{E}\left[|p_h(x)|^q\right] = |\mathbb{E}\left[\hat{p}_h(x)\right]|^q = \Theta(n^{-q}h^{-q})$$

as a result, taking $q = 2+\delta$ we have

$$\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h}) - p_h(x)\right|^{2+\delta}\right] = \Theta(nh^{-1-\delta} + n^{-1-\delta}h^{-1-\delta}) = \Theta(nh^{-1-\delta})$$

2. Taking $q = 2$ we have

$$\big(\sum_{i=1}^{n}\sigma_i^2(x)\big)^{2+\delta} = \Theta((nh^{-1}))^{(2+\delta)/2}$$

3. Now we verify the condition for Lyaupnov CLT, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h}) - p_h(x)\right|^q\right]}{(\sum_{i=1}^{n}\sigma_i^2(x))^{(2+\delta/2)}} &= \lim_{n\to\infty}\frac{\Theta(nh^{-1-\delta})}{\Theta((nh^{-1}))^{(2+\delta)/2}}\\ &= \lim_{n\to\infty}\Theta(n^{-\delta/2}h^{-\delta/2}) = 0\end{aligned}$$

thus we have the Lyaupnov CLT for KDE as

$$\frac{\hat{p}_h(x) - p_h(x)}{s_n(x)} \xrightarrow{\mathrm{d}} N(0,1)$$

in which $s_n^2(x) = var(\hat{p}_h(x))$.

2.(b)

- Variance term:

$$s_n^2(x) = \Theta(n^{-1}h_n^{-1})$$

- Bias term:

$$\begin{aligned}|p_{h_n}(x) - p(x)| &= \left|\int K(\|v\|)\big(p(x+h_nv) - p_{x,\beta}(x+h_nv)\big)\,\mathrm{d}v\right|\\ &\sim Lh_n^{\beta}\int K(s)\,\|s\|^{\beta}\,\mathrm{d}s := ch_n^{\beta} \quad \text{generally speaking.}\end{aligned}$$

- Optimal h_n: choosing ℓ_2 risk as the criterion, we have

$$\text{error} \asymp \text{bias}^2 + \text{variance} = \Theta(h_n^{2\beta}) + \Theta(n^{-1}h_n^{-1})$$

which gives the optimal h_n as

$$h_n^* = \Theta(n^{-1/(2\beta+1)})$$

As a result, at $\beta = 2$ we have $h_n^* = \Theta(n^{-1/5})$ and

$$\frac{|p_{h_n}(x) - p(x)|}{s_n(x)} = \Theta(n^{1/2} h_n^{5/2}) = \Theta(1)$$

say the $\Theta(1)$ limit is $b(x)$, which gives

$$\frac{\hat{p}_h(x) - p(x)}{s_n(x)} = \frac{\hat{p}_h(x) - p_h(x)}{s_n(x)} + \frac{p_h(x) - p(x)}{s_n(x)} \xrightarrow{\mathrm{d}} N(0,1) + b(x) = N(b(x), 1)$$

Exercise 3 Estimating the Sobolev Ellipsoid via Spectral Methods

3.(a)

By the basis expansion, we have

$$\begin{aligned}
\text{risk} =& \mathbb{E}\left[\int_0^1 (\hat{p}(x) - p(x))^2 \,\mathrm{d}x\right] \\
=& \mathbb{E}\left[\int_0^1 \Big(\sum_{j=1}^{k} \left(\hat{\beta}_j - \beta_j\right) \phi_j(x) - \sum_{j=k+1}^{\infty} \beta_j(x)\Big)^2 \,\mathrm{d}x\right] \\
=& \sum_{j=1}^{k} \mathbb{E}\left[\left(\hat{\beta}_j - \beta_j\right)^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
=& \sum_{j=1}^{k} \mathbb{E}\left[\left(n^{-1} \sum_{i=1}^{n} \phi_j(X_i) - \beta_j\right)^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2
\end{aligned}$$

Now note that we have

$$\mathbb{E}\left[\phi_j(X_i)\right] = \int_0^1 \phi_j(x) p(x) \,\mathrm{d}x = \beta_j$$

thus we have

$$\begin{aligned}
\text{risk} =& \sum_{j=1}^{k} \mathbb{E}\left[\left(n^{-1} \sum_{i=1}^{n} \phi_j(X_i) - \beta_j\right)^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
=& \sum_{j=1}^{k} \left[\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left[\phi_j(X_i)^2\right] - 2\beta_j \frac{1}{n} \sum_{i=1}^{n} \beta_j + \beta_j^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
\leq& \sum_{j=1}^{k} \left[\frac{C^2}{n} - \beta_j^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
\leq& \frac{ck}{n} + \sum_{j=k+1}^{\infty} \beta_j^2.
\end{aligned}$$

3.(b)

We have for Sobolev ellipsoid $E(m, L)$ that

$$k^{2m} \sum_{j=k+1}^{\infty} \beta_j^2 \leq \sum_{j=1}^{k} \beta_j^2 j^{2m} + \sum_{j=k+1}^{\infty} \beta_j^2 j^{2m} < L^2 \Rightarrow \sum_{j=k+1}^{\infty} \beta_j^2 < \frac{L^2}{k^{2m}}$$

as a result we have

$$\text{risk} \leq \frac{ck}{n} + \frac{L^2}{k^{2m}} \lesssim \frac{k}{n} + \left(\frac{1}{k}\right)^{2m}$$

The optimal k is obtained at $k \asymp n^{1/(2m+1)}$, which gives the optimal risk $\lesssim n^{-2m/(2m+1)}$.