Prof. Yiping Lu
IEMS 402 Statistical Learning
November 3, 2024

Homework 3: Bias and Variance Trade-off

Question 1. Ensemble and Bias-Variance Trade-off As a reminder, the expected generalization error enjoys the following *bias-variance* trade-off:

$$\mathbb{E}\left[|h(x;\mathcal{D})-y|^2\right] = \underbrace{\mathbb{E}\left[(\mathbb{E}[h(x;\mathcal{D})|x]-y_*(x))^2\right]}_{\text{② bias}} + \underbrace{\mathbb{E}\left[(h(x;\mathcal{D})-\mathbb{E}[h(x;\mathcal{D})|x])^2\right]}_{\text{③ variance}} + \underbrace{\mathbb{E}\left[(y_*(x)-y)^2\right]}_{\text{Irreducible error}},$$

where x, y represent the sampled test data. $\mathcal{D}$ represents the sampled training dataset. $h(\cdot;\mathcal{D})$ is our prediction model learned using this dataset (i.e. $h(\cdot;\mathcal{D})$ is the learnt hypothesis given the training examples). $y_*(x)$ is the true model we want to learn and $\mathbb{E}[y|x] = y_*(x)$. In the following questions, we are interested in the generalization performance of ensemble.

1.0.1. **Weight Average or Prediction Average?** Does the ensemble of linear models using weight average or prediction average give the same expected generalization error? Does the ensemble of (nonlinear) neural networks using weight average or prediction average give the same expected generalization error?

1.0.2. **Bagging - Uncorrelated Models.** One way to construct an ensemble is through bootstrap aggregation, or bagging, that takes a dataset $\mathcal{D}$ and generates k new datasets, with replacement. In this question, we assume the generated dataset $\mathcal{D}_i$ has the same size as $\mathcal{D}$. Then train a model for each dataset, $\mathcal{D}_i$, resulting in k models. The ensemble model is following:

$$\bar{h}(x;\mathcal{D}) = \frac{1}{k}\sum_{i=1}^{k} h(x;\mathcal{D}_i)$$

For this part, we will make a very unrealistic assumption that the predictions of the ensemble members are uncorrelated. That is $\mathrm{Cov}(h(x;\mathcal{D}_j), h(x;\mathcal{D}_k)) = 0$.

1.0.2.1. *Bias with bagging.* Show that ensemble does not change the bias term in the generalization error.

$$\text{Show bias} = \mathbb{E}\left[\left(\mathbb{E}[\bar{h}(x;\mathcal{D})|x]-y_*(x)\right)^2\right] = \mathbb{E}\left[(\mathbb{E}[h(x;\mathcal{D})|x]-y_*(x))^2\right]$$

1.0.2.2. *Variance with bagging.* Assume the variance of a single predictor is σ^2, $\mathbb{E}\left[(h(x;\mathcal{D})-\mathbb{E}[h(x;\mathcal{D})|x])^2\right] = \sigma^2$. Derive the variance of ensemble in terms of σ^2 under uncorrelated predictions.

1.0.3. **Bagging - General Case.** In practice, there will be correlations among the k predictions of the ensemble members because the sampled training datasets would be very similar to each other. For simplicity, assume a non-zero pairwise correlation between the ensemble members, that is, ρ. The variance of the predictor for $h(x;\mathcal{D}_j)$ is σ_j^2.

$$\rho = \frac{\mathrm{Cov}(h(x;\mathcal{D}_j), h(x;\mathcal{D}_k))}{\sigma_j\sigma_k} \quad \forall j \neq k$$

1.0.3.1. *Bias under Correlation.* Does the correlation change the bias term in the generalization error? If so, derive the new expression in terms of ρ. Provide your justification.

1.0.3.2. *Variance under Correlation.* Assume the variance of a single predictor is σ^2. Derive the variance term of ensemble in terms of σ and ρ.

$$\text{Show variance} = \mathbb{E}\left[\left(\bar{h}(x;\mathcal{D})-\mathbb{E}[\bar{h}(x;\mathcal{D})|x]\right)^2\right] = \left(\rho + \frac{1-\rho}{k}\right)\sigma^2$$

1.0.3.3. *Intuitions on bagging.* Based on the derived variance, what happens to the variance when you increase the number of ensemble models k? What do $\rho = 0$, $\rho = 1$ represent and their consequences for the variance?

Question 2. Estimating the Derivatives via Kernel Smoothing Given a scalar $\beta > 1$, let p be a probability density function on $\mathbb{R}$ such that $p \in \Sigma(\beta)$ (i.e., β-th order Hölder class). We are interested in nonparametric estimation of the derivative p'.

Given a kernel function $K : \mathbb{R} \to \mathbb{R}$ supported on $[-1, 1]$ satisfying the conditions

$$\int_{\mathbb{R}} u^j K(u)\, du = \begin{cases} 1 & j = 1, \\ 0 & j = 0, 2, \cdots, \lfloor \beta \rfloor. \end{cases}$$

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p$. Given bandwidth $h > 0$, consider the kernel-based estimator

$$\hat{d}_n(x) := \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

For any x_0, and prove the MSE bound

$$\mathbb{E}\left[|\hat{d}_n(x_0) - p'(x_0)|^2\right] \leq n^{-\frac{2(\beta-1)}{1+2\beta}}.$$

Question 3. Estimating the Sobolev Ellipsoid via Spectral Methods Let $X_1, \ldots, X_n \sim P$ where $X_i \in [0, 1]$ and P has density p. Let $\varphi_1, \varphi_2, \ldots$ be an orthonormal basis for $L_2[0, 1]$. Hence $\int_0^1 \varphi_j^2(x)dx = 1$ for all j and $\int_0^1 \varphi_j(x)\varphi_k(x)dx = 0$ for $j \neq k$. Assume that the basis is uniformly bounded, i.e. $\sup_j \sup_{0 \leq x \leq 1} |\varphi_j(x)| \leq C < \infty$. We may expand p as $p(x) = \sum_{j=1}^{\infty} \beta_j \varphi_j(x)$ where $\beta_j = \int \varphi_j(x)p(x)dx$. Define

$$\hat{p}(x) = \sum_{j=1}^{k} \hat{\beta}_j \varphi_j(x)$$

where $\hat{\beta}_j = (1/n) \sum_{i=1}^{n} \varphi_j(X_i)$.

(a) Show that the risk is bounded by

$$\frac{ck}{n} + \sum_{j=k+1}^{\infty} \beta_j^2$$

for some constant $c > 0$.

(b) Define the Sobolev ellipsoid $E(m, L)$ of order m as the set of densities of the form $p(x) = \sum_{j=1}^{\infty} \beta_j \varphi_j(x)$ where $\sum_{j=1}^{\infty} \beta_j^2 j^{2m} < L^2$. Show that the risk for any density in $E(m, L)$ is bounded by $c[(k/n) + (1/k)^{2m}]$. Using this bound, find the optimal value of k and find the corresponding risk.

Northwestern University