Math 342W/642/742W

Recitation – Day #20 (4.29.25)

I. Model Averaging/Bagging

(i) What does the meta-algorithm called "*Model Averaging*" entail?

Consider models $g_1, g_2, \ldots, g_M$ and let $g_{\text{avg}} = \dfrac{g_1 + g_2 + \cdots + g_M}{M}$.

(ii) What is the MSE of g_{avg}?

$$\begin{aligned}
\text{MSE} &= \sigma^2 + \mathbb{E}_X\left[\text{Bias}[g_{\text{avg}}]^2\right] + \mathbb{E}_X\left[\text{Var}[g_{\text{avg}}]\right] \\
&= \sigma^2 + \mathbb{E}_X\left[\left(f - \frac{g_1 + \cdots + g_M}{M}\right)^2\right] + \mathbb{E}_X\left[\text{Var}\left[\frac{g_1 + \cdots + g_M}{M}\right]\right] \\
&= \sigma^2 + \mathbb{E}_X\left[\frac{1}{M^2}\left((f - g_1) + \cdots + (f - g_M)\right)^2\right] + \frac{1}{M^2}\mathbb{E}_X\left[\text{Var}\left[g_1 + \cdots + g_M\right]\right] \\
&\overset{\star}{=} \sigma^2 + \mathbb{E}_X\left[\left(\text{Bias}[g_1]\right)^2\right] + \frac{1}{M}\mathbb{E}_X\left[\text{Var}\left[g_1\right]\right] \\
&= \sigma^2 + \mathbb{E}_X\left[\left(\text{Bias}[g_1]\right)^2\right] \quad (M \to \infty) \\
&= \sigma^2 \quad (\textit{Use low bias models such as trees})
\end{aligned}$$

$\star$ Assume (1) same bias for each model, (2) Models are independent, (3) variances are the same.

(iii) Why is the result of the MSE found in part (ii) impossible?

It is impossible because models $g_1, \ldots, g_M$ are really *dependent* since they are trained with the same training dataset $\mathbb{D}$.

(iv) How do we make the models $g_1, \ldots, g_M$ "*more*" independent? What is this technique called?

Neil Breiman (1994) came up with the meta-algorithm called "*bootstrap aggregation*", a.k.a. "*bagging*." We make the models $g_1, \ldots, g_M$ more independent to each other by

- $\mathbb{D}_1$ = sample with replacment from $\mathbb{D}$ and $g_1 = \mathcal{A}(\mathbb{D}_1, \mathcal{H})$
- $\mathbb{D}_2$ = sample with replacment from $\mathbb{D}$ and $g_2 = \mathcal{A}(\mathbb{D}_2, \mathcal{H})$

  $\vdots \qquad \vdots \qquad \vdots \qquad \vdots$
- $\mathbb{D}_M$ = sample with replacment from $\mathbb{D}$ and $g_M = \mathcal{A}(\mathbb{D}_M, \mathcal{H})$

We then get $g_{\text{bag}} = \dfrac{g_1 + \cdots + g_M}{M}$.

(v) What bonus feature do we get with bagging?

From bagging we get a validation set, a.k.a. "*out-of-bag*" (OOB) for free. The oos for g_1 for example is,

$$\mathbb{D} = \underbrace{\mathbb{D}_1}_{\approx \frac{2}{3}n} \cup \underbrace{(\mathbb{D} \setminus \mathbb{D}_1)}_{\approx \frac{1}{3}n}$$

Bootstrap Validation: Average the oob errors over all M models.

Note: Theoretically, this is similar to $K \approx 2$.

II. Math 241 Review

Let $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ where the X_i's are dependent random variables.

(i) Find $\text{Var}[\bar{X}]$.

$$\text{Var}[\bar{X}] = \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left(\text{Var}[X_1] + \ldots + \text{Var}[X_n] + \sum_{i \neq j}\text{Cov}[X_i, X_j]\right)$$

(ii) Recall definition of $\text{Cov}[X_i, X_j]$.

$$\text{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \cdot \mathbb{E}[X_j]$$

(iii) Find the expression for the correlation coefficient ρ.

$$\rho = \text{Corr}[X_i, X_j] = \frac{\text{Cov}[X_i, X_j]}{\text{SD}[X_i] \cdot \text{SD}[X_j]} \in [-1, 1] \implies \text{Cov}[X_i, X_j] = \rho \cdot \text{SD}[X_i] \cdot \text{SD}[X_j]$$

Assume σ^2 is the same for all X_i, ρ is the same for all X_i, X_j where $i \neq j$.

(iv) Complete the expression for $\text{Var}[\bar{X}]$ in part (i).

$$\text{Var}[\bar{X}] = \frac{1}{n^2}\left(n\sigma^2 + (n^2 - n)\rho\sigma^2\right) = \frac{1}{n}\left(\sigma^2(1-\rho) + n\sigma^2\rho\right)$$

III. More on Bagging

(i) Assuming that $\text{Var}[g_i] = \sigma^2$ and $\text{Corr}[g_i, g_j] = \rho$, find the MSE for bagging.

Assuming $\text{Var}[g_i] = \sigma^2$ and $\text{Corr}[g_i, g_j] = \rho$

$$\begin{aligned}
\text{MSE} &= \mathbb{E}_X\left[\text{Bias}[g]^2\right] + \mathbb{E}_X\left[\text{Var}[g]\right] + \sigma^2 \\
&= \mathbb{E}_X\left[\text{Bias}[g_1]^2\right] + \mathbb{E}_X\left[\rho\text{Var}[g_1] + \frac{1-\rho}{M} \cdot \text{Var}[g_1]\right] + \sigma^2 \\
&= \mathbb{E}_X\left[\text{Bias}[g_1]^2\right] + \rho\,\mathbb{E}_X\left[\text{Var}[g_1]\right] + \frac{1-\rho}{M}\,\mathbb{E}_X\left[\text{Var}[g_1]\right] + \sigma^2 \\
&= \mathbb{E}_X\left[\text{Bias}[g_1]^2\right] + \rho\,\mathbb{E}_X\left[\text{Var}[g_1]\right] + \sigma^2 \qquad\qquad M \to \infty
\end{aligned}$$

This derivation shows how bagging works in reducing the variance component of MSE.

(ii) How can we decorrelate the trees even more leading to minimizing the variance of the MSE even further?

Breiman (2001) invented "Random Forests" as a generalization of bagged trees by changing the $\mathcal{A}$ of CART in the following way. Let $m_{\text{try}} < p_{\text{raw}}$, # of features in $\mathbb{D}$. At each split, choose a subset of the p_{raw} features of size m_{try} where typically $m_{\text{try}} \approx \sqrt{p_{\text{raw}}}$. Then, find optimal local split.