The Mathematics of Machine Learning
Homework Set 6

Due 21 April 2022 before 13:00
via Canvas

You are allowed to work on this homework in pairs. One person per pair submits the answers via Canvas. Make sure to put both names on the submission.

1 The Benefits of Averaging

Bagging combines bootstrapping with averaging of estimators. To understand the potential benefits of averaging, we will consider the idealized situation in which, instead of the bootstrap samples, we would have access to fresh data sets $T_1, \ldots, T_B$ of size N that were independently sampled from the true probability distribution P^*. We apply the same training procedure to each data set T_b to obtain an independent prediction function $\hat{f}_b$ for each $b = 1, \ldots, B$.

1. [2 pt] Consider regression with the squared loss. Let

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b$$

be the average of the estimators that were all trained on independent samples. Use the bias-variance decomposition to show that

$$\mathop{\mathbb{E}}_{T}[\mathrm{EPE}(\hat{f})] = \mathop{\mathbb{E}}_{X}[\mathrm{Var}(Y|X)] + \mathop{\mathbb{E}}_{X}\left[\big(\bar{f}(X) - f_{\mathrm{Bayes}}(X)\big)^2\right] + \frac{1}{B} \mathop{\mathbb{E}}_{T,X}\left[\big(\hat{f}_1(X) - \bar{f}(X)\big)^2\right],$$

where $\bar{f} = \mathbb{E}_T[\hat{f}_1] = \cdots = \mathbb{E}_T[\hat{f}_B]$ denotes the mean prediction function when we average over the draw of a data set of size N. Since the right-hand side is decreasing in B, the conclusion we can draw from this is that averaging improves the expected prediction error of estimators that are trained on independent samples.
Hint: For any independent random variables $A_1, \ldots, A_B$ and any constants $\alpha_1, \ldots, \alpha_B$, the variance satisfies

$$\mathrm{Var}(\alpha_1 A_1 + \ldots + \alpha_B A_B) = \alpha_1^2 \,\mathrm{Var}(A_1) + \ldots + \alpha_B^2 \,\mathrm{Var}(A_B).$$

2. [2 pt] Consider binary classification with 0/1-loss. Suppose that each $\hat{f}_b$ is a binary classifier with $\hat{f}_b(x) \in \{-1, +1\}$. Let

$$\hat{f}(x) := \operatorname{sign}\Big(\sum_{b=1}^{B} \hat{f}_b(x)\Big)$$

be the majority vote. Now let f_{Bayes} be the Bayes-optimal classifier and consider classification of a fixed data point x^*. Let

$$p := \Pr_{T_b}(\hat{f}_b(x^*) = f_{\text{Bayes}}(x^*))$$

be the probability that we sample a data set for which the trained classifier $\hat{f}_b$ classifies x^* the same way as $f_{\text{Bayes}}(x^*)$. Now we want to study how the majority vote $\hat{f}$ behaves as B becomes large. Show that:

(a) If $p > 1/2$, then $\Pr(\hat{f}(x^*) = f_{\text{Bayes}}(x^*)) \to 1$ as $B \to \infty$.

(b) If $p < 1/2$, then $\Pr(\hat{f}(x^*) = f_{\text{Bayes}}(x^*)) \to 0$ as $B \to \infty$.

The conclusion is that averaging helps as long as $\hat{f}_b$ is more likely to learn the optimal prediction on x^* than to learn the opposite prediction.
Hint 1: Note that $\operatorname{sign}\Big(\sum_{b=1}^{B} \hat{f}_b(x^*)\Big) \neq f_{\text{Bayes}}(x^*)$ *only if*
$n_1 := |\{b : \hat{f}_b(x^*) = f_{\text{Bayes}}(x^*)\}| \leq B/2$.
Hint 2: You may use Hoeffding's inequality, which states that, if $Z_1, \ldots, Z_B$ *are independent, identically distributed random variables with mean* μ *that take values in* $[0, 1]$*, then*

$$\Pr\Big(\big|\sum_{b=1}^{B} Z_b - B\mu\big| \geq \epsilon\Big) \leq 2e^{-\frac{2\epsilon^2}{B}}.$$

Apply this with $Z_b = \mathbf{1}[\hat{f}_b(x^*) = f_{\text{Bayes}}(x^*)]$ *and observe that* $\mu = \mathbb{E}[Z_b] = \Pr(\hat{f}_b(x^*) = f_{\text{Bayes}}(x^*)) = p$.