Formalize Bagging

☆ Key Points

The weak law of large numbers tells us that the sample mean of a random variable approaches the population mean as the number of samples goes to infinity.

We apply this to classifiers and average many classifiers trained on different data sets. If the data sets were truly sampled independently from the original data distribution, the variance term would reduce to 0.

As we do not have access to the underlying data distribution, we sample bootstraps instead. Although the resulting training sets are not independent, the resulting averaged classifier still has much reduced variance.

Remember the bias-variance decomposition

$$\underbrace{\mathbb{E}\left[(h_D(x)-y)^2\right]}_{\text{Generalization Error}} = \underbrace{\mathbb{E}\left[\left(h_D(x)-\bar{h}(x)\right)^2\right]}_{\text{Variance}} + \underbrace{\mathbb{E}\left[\left(\bar{h}_D(x)-\bar{y}(x)\right)^2\right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(\bar{y}(x)-y(x))^2\right]}_{\text{Noise}}$$

Our goal is to reduce the variance term: $\mathbb{E}\left[\left(h_D(x)-\bar{h}(x)\right)^2\right]$.

For this, we want $h_D \to \bar{h}$.

Weak Law of Large Numbers

The weak law of large numbers (roughly) says that for independent and identically distributed (i.i.d.) random variables x_i with mean $\bar{x}$, we have, $\frac{1}{m}\sum_{i=1}^{m} z_i \to \bar{z}$ as $m \to \infty$. In other words, as the number of samples approaches infinity, the mean of a sample approaches the true mean of the random variable.

Apply the law to classifiers: Assume we have m training sets $D_1, \ldots, D_m$ drawn from P^n. Train a classifier on each one, and at run time take the average result obtained across all classifiers:

$$\tilde{h} = \frac{1}{m}\sum_{i=1}^{m} h_{D_i}$$

We refer to such an average of multiple classifiers $\tilde{h}$ as an **ensemble** of classifiers.

Idea: If $\tilde{h} \to \bar{h}$, the variance component of the error will reduce to 0; that is, $\mathbb{E}\left[\left(\tilde{h}(x)-\bar{h}(x)\right)^2\right] \to 0$.

Problem: We don't have m data sets $D_1, \ldots, D_m$; we only have D. Recalling from the previous lectures, how can we use the one data set we have to simulate many different data

sets?

Solution: We use bootstrapping to obtain the data sets $D_1, \ldots, D_m$ by sampling them **with replacement** from the original data set D. Once again, we train a classifier on each one of these bootstraps $h_{D_1}, \ldots, h_{D_m}$ and compute their average $\hat{h}_D = \frac{1}{m} \sum_{i=1}^{m} h_{D_i}$. This average is the **bagged** classifier (**not the ensemble** classifier because D_i are not i.i.d. from P^n).

Notice that, although $\hat{h}_D$ gets closer to $\overline{h}$, we do not have full convergence as the weak law of large numbers suggests. That is, $\hat{h}_D = \frac{1}{m} \sum_{i=1}^{m} h_{D_i} \nrightarrow \overline{h}$. In other words, the bagged classifier does not converge to the true expected classifier $\overline{h}$.