Formalize Bootstrapping

One way to estimate the variance of a classifier would be to sample many data sets from the original data distribution, train a classifier on each one of them, and estimate their variance directly (as you have done in the demo in the previous module). However, this only works if you have access to the original data distribution, which you don't in most real-world settings.

We can, however, simulate drawing from the original data distribution $P(\mathbf{X}, Y)$ by drawing uniformly with replacement from the training data set $D \sim P^n$. Formally, let $Q\left(\mathbf{X}, Y | D\right)$ be a probability distribution that picks a training sample $(\mathbf{x}, y)$ from D uniformly at random.

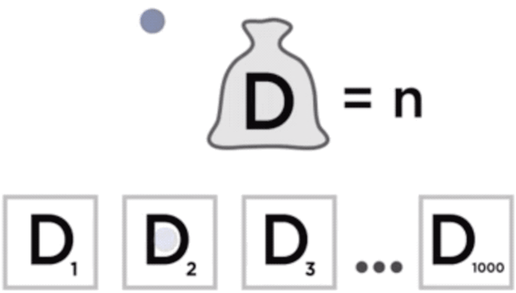
More formally, $Q\left((\mathbf{x}, y) \,|D\right) = \frac{1}{n} \quad \forall\, (\mathbf{x}, y) \in D$ with $n = |D|$.

☆ Key Points

Bootstrapping gives us the ability to estimate the variance of a classifier.

Sampling with replacement from the training sample results in the same distribution as the original data distribution. However, samples are no longer independent.

By training models on "bootstraps," we can calculate the mean prediction, and then calculate the variance of the model.



We sample the set $D_i \sim Q^n$; that is D_i of size n is picked **with replacement** from D as per the distribution Q. This way we obtain m data sets $D_1, \ldots, D_m$, each one with n elements sampled from D. Note, however, that these data sets are *not* all identical to D. The reason is that we may pick some samples multiple times in D_i, and others not at all. In fact, in expectation, the intersection of any D_i and D is only 63%, and the remaining samples of D_i are repetitions of each other. Because each D_i is picked with replacement from D, samples $(\mathbf{x}, y)$ in D_i are also drawn from the original data distribution P, but the samples are *not independent*.

In order to estimate the variance of our algorithm, we train one classifier h_{D_i} on each data set D_i. We can then compute the average classifier of these bootstraps: $\hat{h}_D(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} h_{D_i}(\mathbf{x})$.

The variance of our algorithm is thus estimated to be

$$\text{Variance} \approx \frac{1}{|D_v|}\sum_{(\mathbf{x},y)\in D_v} \frac{1}{m}\sum_{i=1}^{m}\left(h_{D_i}(\mathbf{x}) - \hat{h}_D(\mathbf{x})\right)^2$$

where D_v denotes the validation set.

Notice that $\hat{h}_D(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} h_{D_i}$ does not converge to the *true* average classifier $\bar{h}(\mathbf{x})$ as the number of subsample datasets, m, increases. We would only obtain $\bar{h}(\mathbf{x})$ if we had truly drawn i.i.d. different training data sets from the original data distribution P. The reason is that our data sets $D_1, \ldots, D_m$ are not independent of each other as they are all subsampled from the original data set D. However, in practice, bagging still reduces variance significantly.