Bagging: Bootstrap Aggregating

By: Behzad Asadi

Generalization Error Decomposition

Consider the bias-variance-noise decomposition of the expected test error

$$E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x})-y)^2\right] = \underbrace{E_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x})-\bar{y}(x)\right)^2\right]}_{\text{Bias}^2} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(x)-y)^2\right]}_{\text{Noise}}$$

where $h_D(\mathbf{x})$ is the model fitted to the training dataset D, $\bar{h}(\mathbf{x})$ is the expected model calculated as

$$\bar{h}(\mathbf{x}) = E_D\left[h_D(\mathbf{x})\right] = \int_D h_D(\mathbf{x})p(D)d_D,$$

and $\bar{y}(\mathbf{x})$ is the expected label given $\mathbf{x}$, calculated as $\bar{y}(\mathbf{x}) = \int_y yp(y \mid \mathbf{x})d_y$.

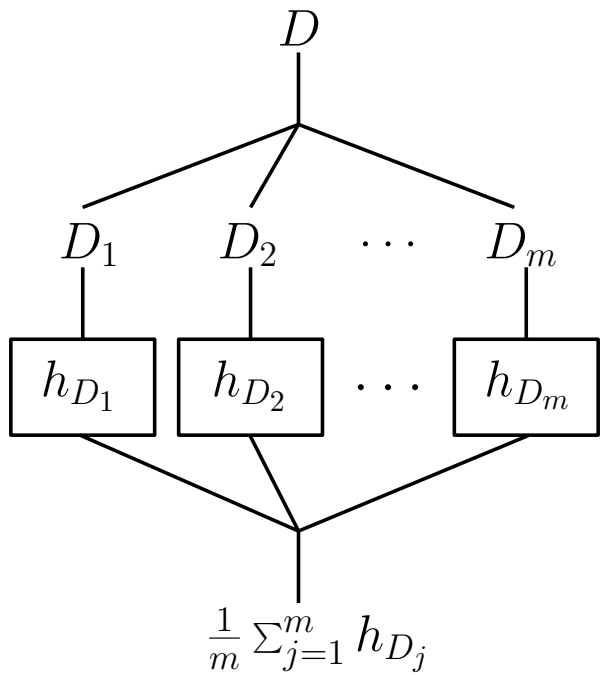
If we have m independent datasets, $D_1, D_2, \ldots, D_m$, and we fit a model to each of these datasets separately, and take the average, then by the weak law of large number

$$\hat{h}(x) = \frac{1}{m}\sum_{j=1}^{m} h_{D_j}(x) \to \bar{h}(x) \quad \text{as } m \to \infty.$$

Therefore the variance part of the error tends to zero. The problem is that we do not have access to m independent datasets; we have access to only one dataset D.

Bootstrapping

We create m datasets from the original dataset, D, by sampling with replacement. Although the m created datasets are not independent from each other, this approach can reduce the variance in practice.



The m models in bagging can be trained in parallel.

Bootstrapping

The datasets constructed via bootstrapping are not independent from each other, but it can be shown that they are drawn from the original distribution. Consider $Q(x_i)$ as the probability of observing x_i in the datasets constructed via bootstrapping. We here show that $Q(x_i)$ is the same as the original $p_i = P(x_i)$.

$$Q(x_i) = \sum_{k=1}^{n} \underbrace{\binom{n}{k} p_i^k (1-p_i)^{n-k}}_{\text{probability that the original dataset includes } k \text{ copies of } x_i} \underbrace{\frac{k}{n}}_{\text{probability of choosing one}}$$

$$= \frac{1}{n} \sum_{k=1}^{n} k \binom{n}{k} p_i^k (1-p_i)^{n-k}$$

$$= \frac{1}{n} n p_i$$

$$= p_i$$

Random Forest

Decision trees can be used as individual learners in the original bagging algorithm. The random forest algorithm differs from the original bagging by selecting only a random subset of features to train each individual tree. This random selection of features helps with reducing the correlation among trees.

Byproducts:

- **Out-of-bag (OOB) error:** One of the nice features of bagging is that, to estimate the test error, we do not need to split the training set into training and validation sets. We can use the whole training set for the training purpose. This is done by defining the set $S_i = \{k \mid (\mathbf{x}_i, y_i) \notin D_k\}$ corresponding to each sample in the training set. This set determines the models that have been trained without using this sample. The average of the models trained without using the training sample $(\mathbf{x}_i, y_i)$ is $\tilde{h}_i(\mathbf{x}) = \frac{1}{|S_i|} \sum_{k \in S_i} h_k(\mathbf{x})$. Therefore, the test error/loss can be estimated by

$$e_{OOB} = \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in D} \ell(\tilde{h}_i(\mathbf{x}_i), y_i)$$

- **Feature importance**

References

1. Kilian Weinberger, Lecture Notes, https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote18.html.
2. Kevin P. Murphy, Machine Learning: A Probabilistic Perspective.
3. Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning.