Contents

1 Ensemble

An **ensemble method** is an approach that combines many simple "building block" models in order to obtain a single and potentially very powerful model. These simple building block models are sometimes known as **weak learners**, since they may lead to mediocre predictions on their own.

1.1 Bagging

Bootstrap aggregation, or **bagging**, is a general-purpose procedure for reducing the variance (overfitting the specific split of training data) of a statistical learning method.

1.1.1 Regression

Recall that given a set of n independent observations $Z_1, ..., Z_n$, each with variance σ^2, the variance of the mean $\bar{Z}$ of the observations is given by $\frac{\sigma^2}{n}$. In other words, **averaging** a set of observations reduces variance. Hence a natural way to reduce the variance and increase the test set accuracy of a statistical learning method is to 1) take many training sets from the population, 2) build a separate prediction model using each training set, 3) and average the resulting predictions. We calculate $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ response functions using B separate training sets, and average them in order to obtain a single low-variance statistical learning model, given by

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$$

To obtain different random subsets of the training set, we can bootstrap, by taking repeated samples from the (single) training set. We generate B different bootstrapped training data sets. We then train our method on the b^{th} bootstrapped training set in order to get $\hat{f}^{*b}(x)$, and finally average all the predictions, to obtain

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

This is called bagging. The **aggregation function** in this case is a measure of central tendency— the mean. To apply bagging to decision trees, we simply construct B decision trees using B bootstrapped training sets, and average the resulting predictions. These trees are grown deep, and are not pruned. Hence each individual tree has high variance (likely overfitting the data) but low bias. Averaging these B trees reduces the variance.

1.1.2 Classification

Thus far, we have described the bagging procedure in the regression context, to predict a **quantitative** outcome Y. If Y is **qualitative**, we can use a voting classifier. For a given test observation, we can record the class predicted by each of the B trees, and take a majority vote: the overall prediction is the most commonly occurring majority class among the B predictions. This is also known as the the **hard voting** classifier. If all B classifiers are able to estimate class probabilities, then another approach is to predict the class with the highest class probability, averaged over all B individual classifiers. This is called **soft voting**.

1.1.3 Out-Of-Bag Estimation

There is a very straightforward way to estimate the test error of a bagged model without the need to perform cross-validation or the validation set approach. The key to bagging is that trees are repeatedly fit to bootstrapped subsets of the observations. Suppose that we obtain a bootstrap sample from a set of n observations. We will now derive the probability that a given observation is part of a bootstrap sample. Assuming that each selection from the set of n observations is independent, the probability that the j^{th} observation is not in the bootstrap sample is:

$$p_j(n) = \prod_{i=1}^{n} \pi_j = \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{1}{n}\right) = \left(1 - \frac{1}{n}\right)^n$$

Recall that the exponential function satisfies:

$$e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$$

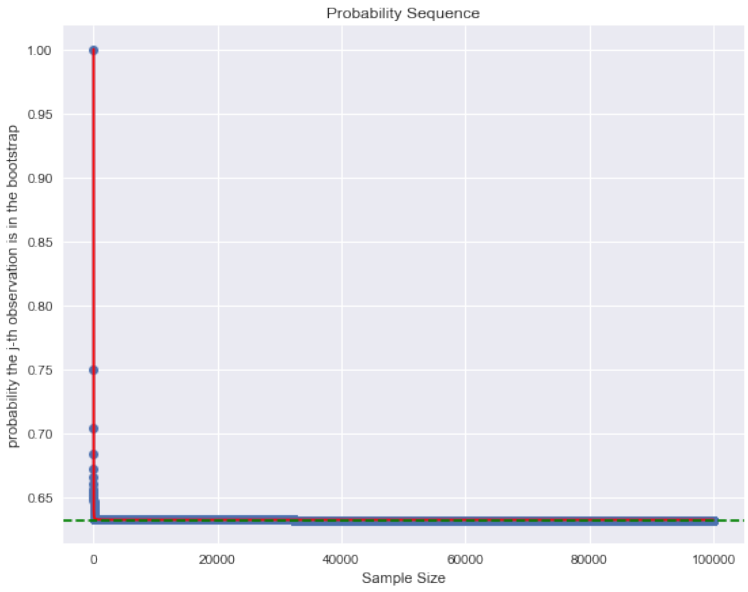
Therefore, letting $x = -1$

$$e^{-1} = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n$$

and the probability that the j^{th} observation is *not* in the bootstrap sample is

$$p_j := \lim_{n\to\infty} p_j(n) = e^{-1} \approx 0.36787944117$$

The probability that the j^{th} observation *is* in the bootstrap sample converges as $n \to \infty$:



This means that each bagged tree makes use of around two-thirds ($1 - 0.36787944117 = 0.6321205588$) of the observations. The remaining one-third of the observations not used to fit a given bagged tree are referred to as the **out-of-bag** (OOB) observations.

- We can predict the response for the i^{th} observation using each of the trees in which that observation was OOB. This will yield around $\frac{B}{3}$ predictions for the i^{th} observation.
- To obtain a single prediction for the i^{th} observation, we can then average these $\frac{B}{3}$ predicted responses (if regression is the goal) or can take a majority vote from $\frac{B}{3}$ votes (if classification is the goal). This leads to a single OOB prediction for the i^{th} observation.

An OOB prediction can be obtained in this way for each of the n observations, from which the overall OOB MSE (for a regression problem) or classification error (for a classification problem) can be computed. The resulting OOB error is a valid estimate of the test error for the bagged model, since the estimated response or classification for each observation is predicted using *only the trees that were not fit using that observation.* With B sufficiently large, OOB error is virtually equivalent to leave-one-out cross-validation error.