Ensemble Learning

Aldo Solari

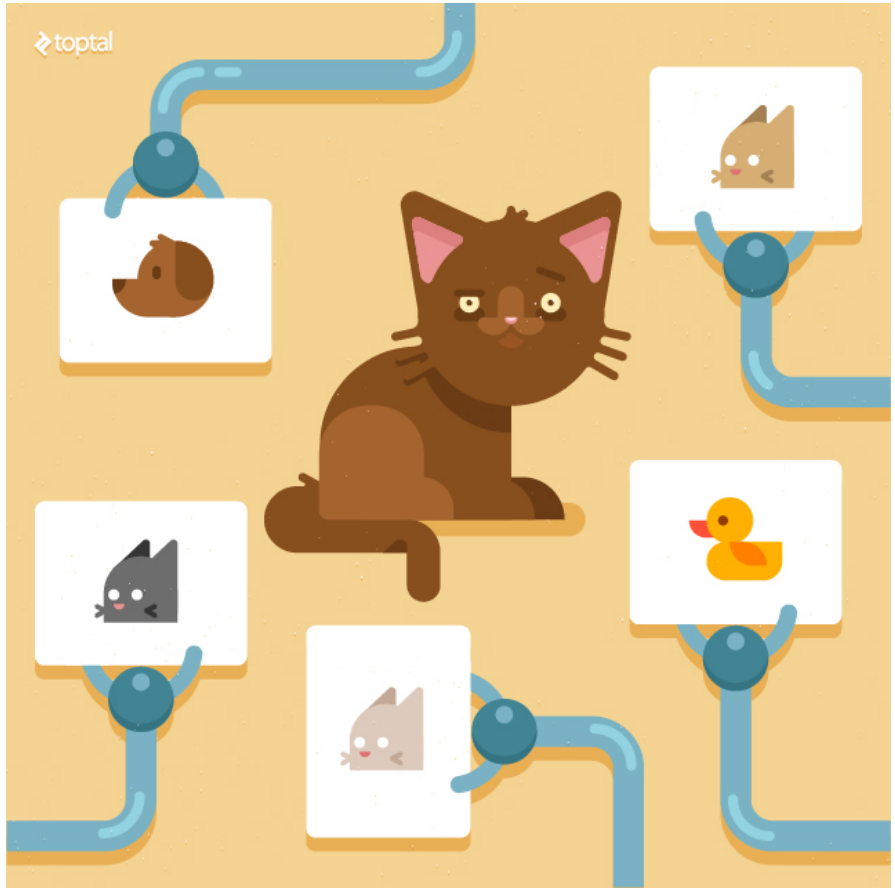
Ensembles

- Ensemble methods are techniques that create multiple models and then combine them to produce improved results
- These models, when used as inputs of ensemble methods, are called "base learners" or "weak learners"
- Ensemble learning is appealing because that it is able to boost weak learners which are slightly better than random guess to strong learners which can make very accurate predictions
- However, ensemble learning increases computation time and reduces interpretability

Majority voting

- Majority voting is a simple ensemble method used for classification
- Each classification model in the ensemble makes a prediction (vote) for each test observation and the final prediction is the one that receives more than half of the votes
- If none of the predictions get more than half of the votes, we may say that the ensemble method could not make a stable prediction for this observation

Majority voting



Majority voting

- Suppose that we have three independent classifiers
- Assume that each classifier has probability π of making the correct classification
- The number of correct classifications is

$$C \sim \text{Binomial}(3, \pi)$$

- The majority vote classifier makes the correct prediction when $C \geq 2$, that is, with probability

$$\Pr(C \geq 2) = 3\pi^2(1 - \pi) + \pi^3$$

- e.g. if $\pi = 70\%$, then the majority vote classifier makes the correct prediction with probability 78.4%

Ensemble of trees

- Classification and regression trees are simple and useful for interpretation
- However they are typically not competitive with other approaches in terms of prediction accuracy
- Ensemble methods such as bagging, random forests and boosting grow multiple trees which are then combined to yield a single prediction
- Combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some interpretation loss

Instability of trees

- The primary disadvantage of trees is that they are rather unstable (high variance)
- In other words, a small change in the data often results in a completely different tree
- One major reason for this instability is that if a split changes, all the splits under it change as well, thereby propagating the variability
- Leo Breiman's idea: use the instability!

Outline

Bagging

- Bootstrap aggregation, or bagging, is a general procedure for reducing the variance of a model and it is particularly useful in the case of regression or classification trees
- Recall that given a set of independent variables $Z_1, \ldots, Z_B$, each with variance σ^2, the variance of the average

$$\bar{Z} = \frac{1}{B} \sum_{b=1}^{B} Z_b$$

  is σ^2/B
- In other words, averaging a set of variables reduces variance. Of course, this is not practical because we generally do not have multiple training sets
- Instead, we can bootstrap, by taking repeated samples from the (single) training set

Bagging algorithm

- First, generate B different bootstrapped training sets

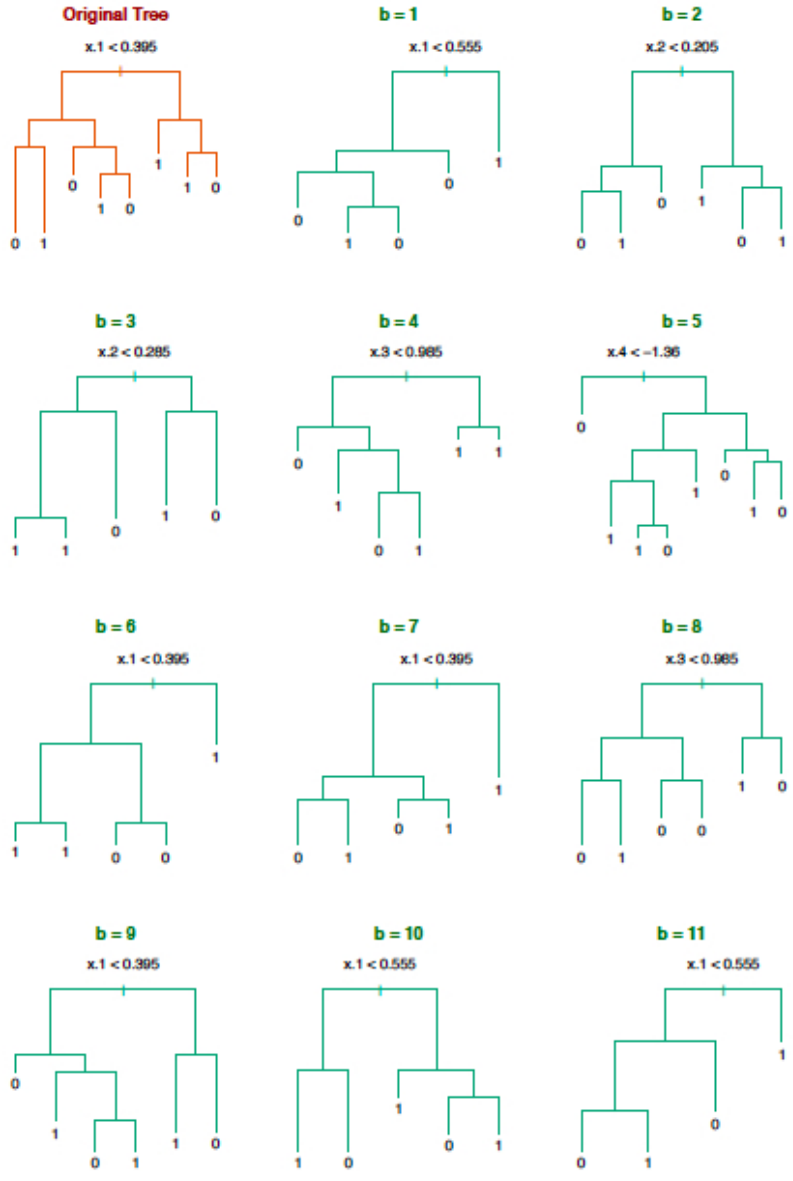
$$(x_1^{*b}, y_1^{*b}), (x_2^{*b}, y_2^{*b}), \ldots, (x_n^{*b}, y_n^{*b}) \quad b = 1, \ldots, B$$

  by sampling with replacement from the training set
- Then, fit a tree on the bth bootstrapped training set in order to get $\hat{f}^{*b}(x)$, the prediction at a point x
- Finally, average all the predictions to obtain

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

- For classification trees, we can record the class $\hat{C}^{*b}(x)$ predicted by each of the B trees, and take a majority vote

Many trees

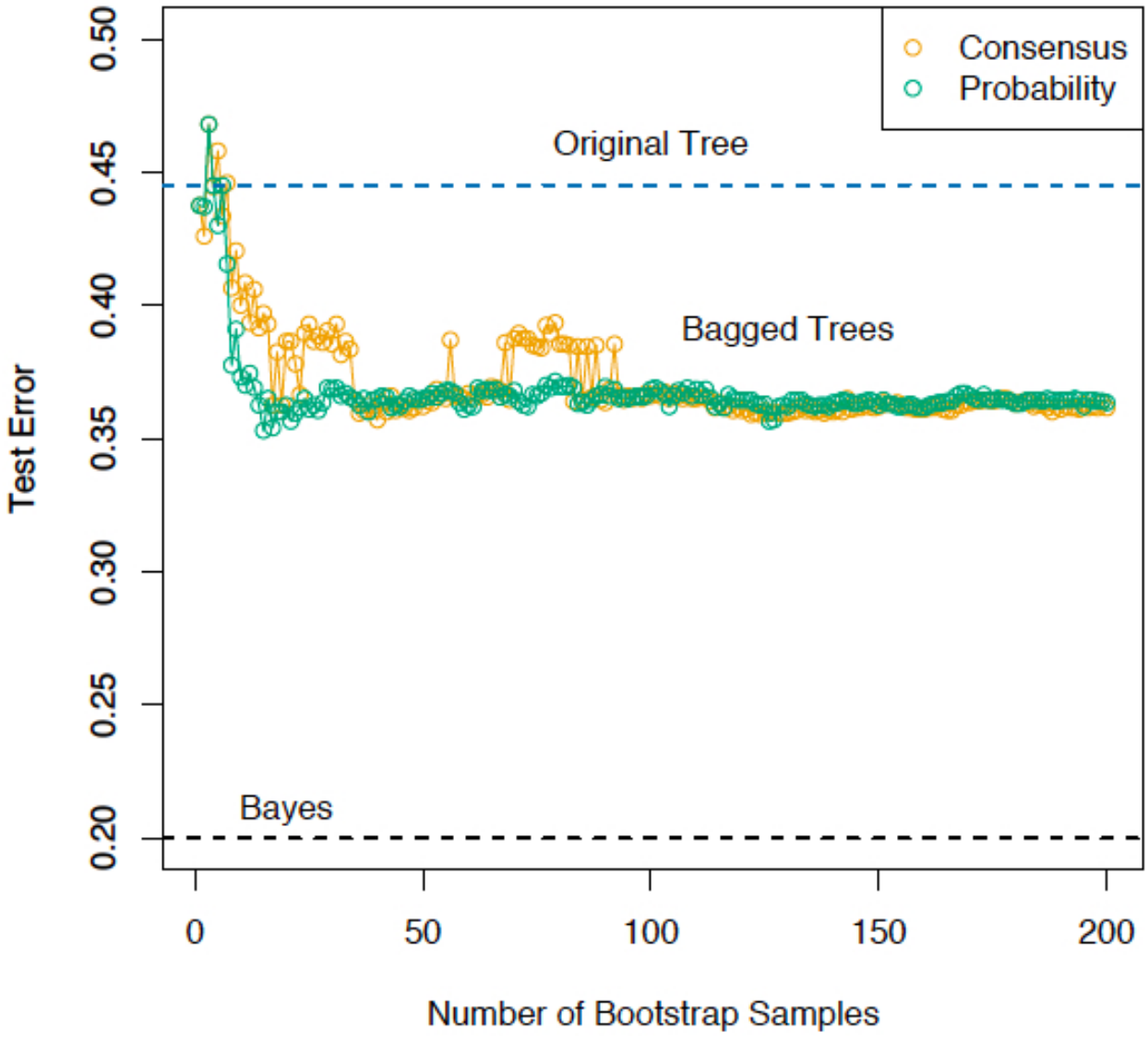


Source: Hastie et al. (2009) p. 284

`spam` data

- 4601 email messages sent to "George" at HP-Labs
- The goal is to build a customized spam filter for George: predict whether an e-mail message is spam (junk mail) or good
- For this problem not all errors are equal; we want to avoid filtering out good e-mail, while letting spam get through is not desirable but less serious in its consequences
- Recorded for each email message is the relative frequency of certain key words (e.g. `business`, `address`, `free`, `George`) and certain characters: (, [, !, $, #. Included as well are three different recordings of capitalized letters.

Bagged trees



Source: Hastie et al. (2009) p. 285

Out-of-bag observations

- In the bth bootstrapped training set

$$(x_1^{*b}, y_1^{*b}), (x_2^{*b}, y_2^{*b}), \ldots, (x_n^{*b}, y_n^{*b})$$

  the probability for one observation not to be drawn in any one of the n draws can be approximated by

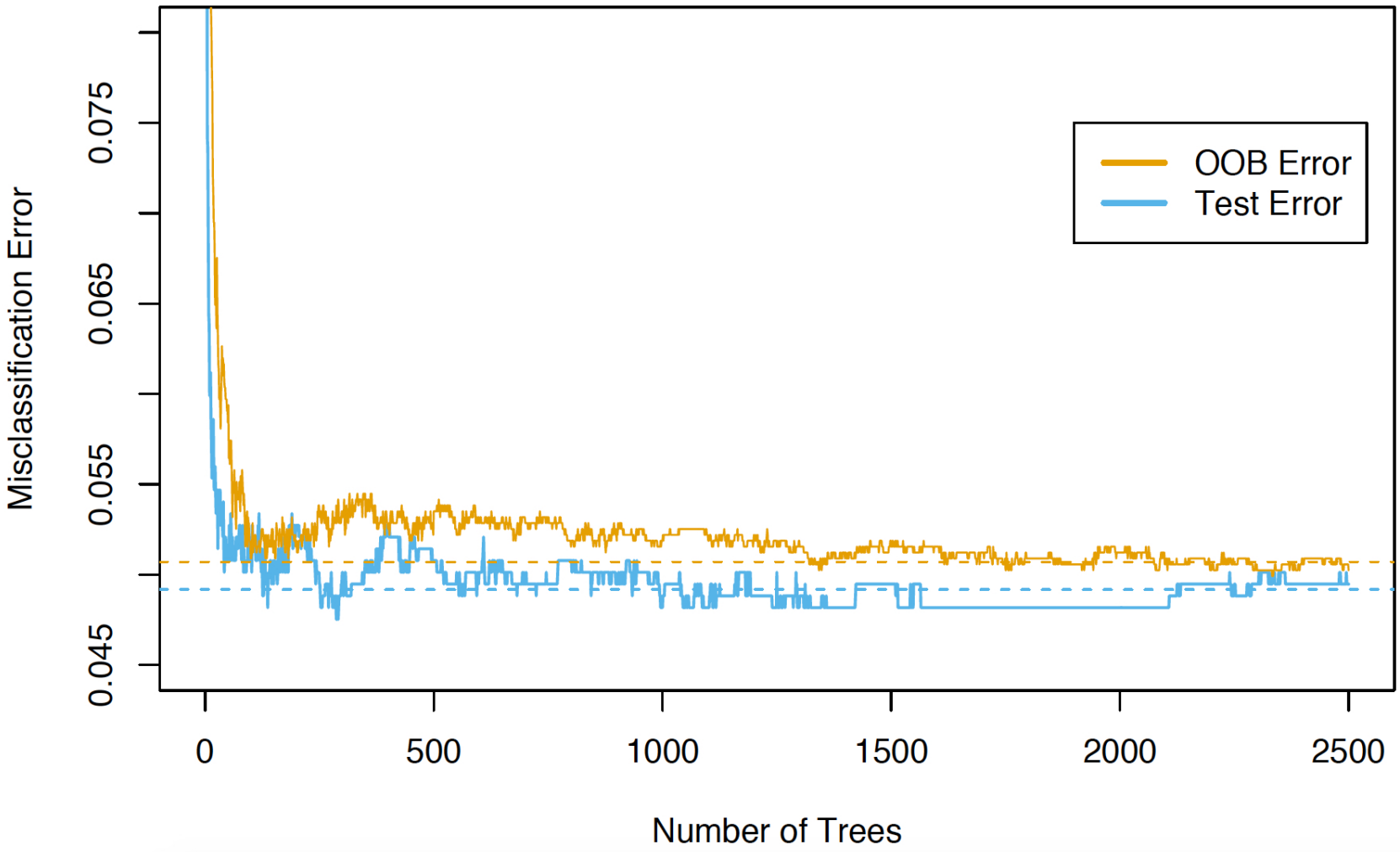
$$\lim_{n \to \infty} \left(1 - \frac{1}{n}\right)^n = \frac{1}{e} \approx 0.368$$

- $\approx 1/3$ of the n original observations are out-of-bag (OOB)

Out-of-bag error

- Each bagged tree makes use of $\approx 2/3$ of the original observations
- We can predict the response for the ith observation using each of the bagged trees in which that observation was OOB
- This yields $\approx B/3$ predictions for the ith observation, which we average/majority vote
- This estimate is essentially the LOOCV error for bagging, if B is large

`spam` data: OOB error



Source: Hastie et al. (2009) p. 592

Outline

Random forests

- Create even more variation in individual trees
- Bagging varies the rows of the training set (randomly draw observations)
- Random forests varies also the columns of the training set (randomly draw predictors)

Random forests algorithm

- Before each split, select $m \leq p$ of the predictors at random as candidates for splitting
- $m = p$ gives Bagging as a special case
- m is called a tuning parameter. Typically $m \approx \sqrt{p}$

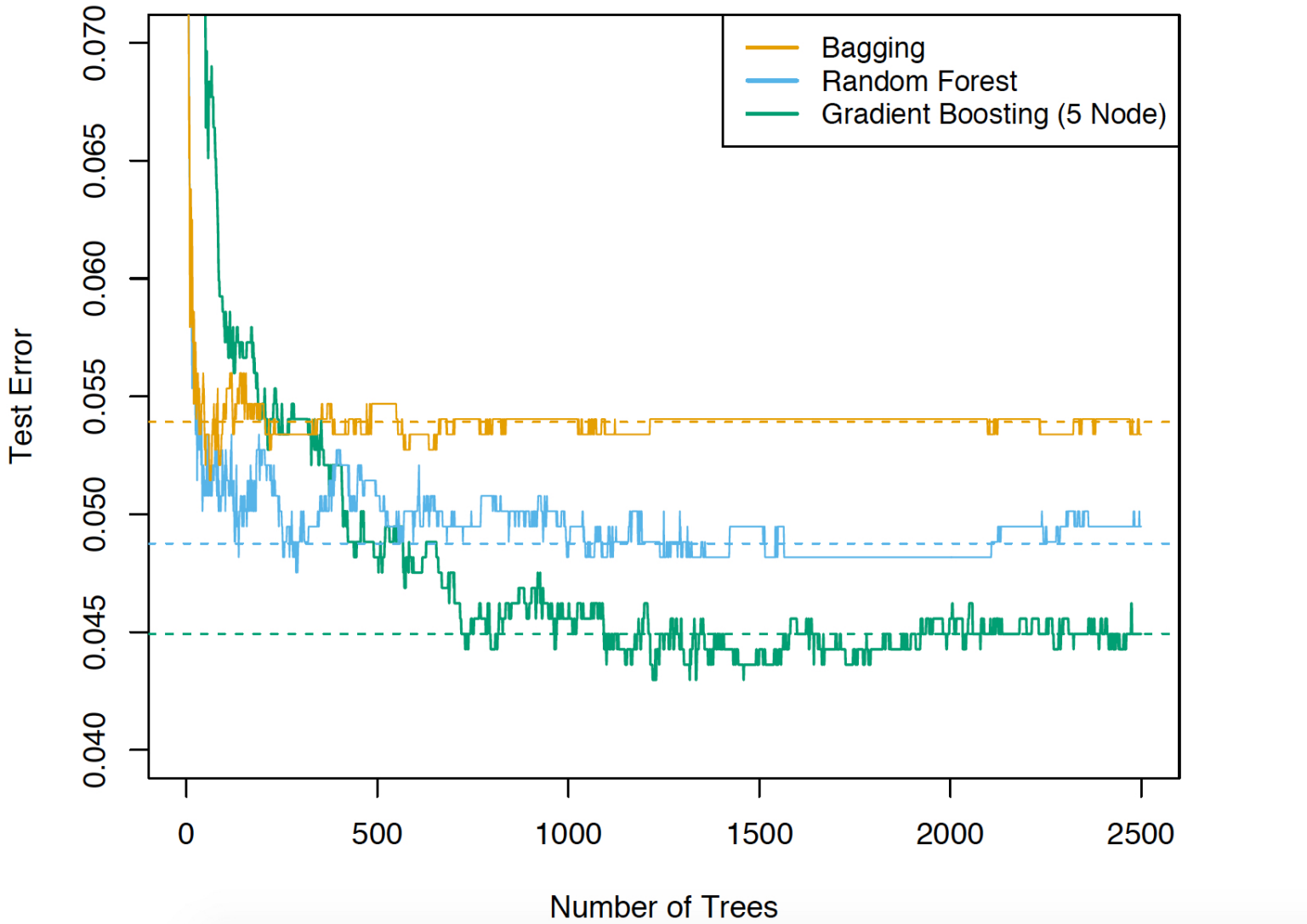
De-correlating trees

- Random sampling of the predictors decorrelates the trees. This reduces the variance when we average the trees
- Recall that given a set of variables $Z_1, \ldots, Z_B$ with pairwise correlation $\mathbb{C}\mathrm{orr}(Z_j, Z_l) = \rho$ and $\mathbb{V}\mathrm{ar}(Z_j) = \sigma^2$, then

$$\mathbb{V}\mathrm{ar}(\bar{Z}) = \rho\sigma^2 + \frac{(1-\rho)}{B}\sigma^2$$

- The idea in random forests is to improve the variance reduction of bagging by reducing the correlation ρ between the trees, without increasing the variance σ^2 too much

`spam` data: random forest

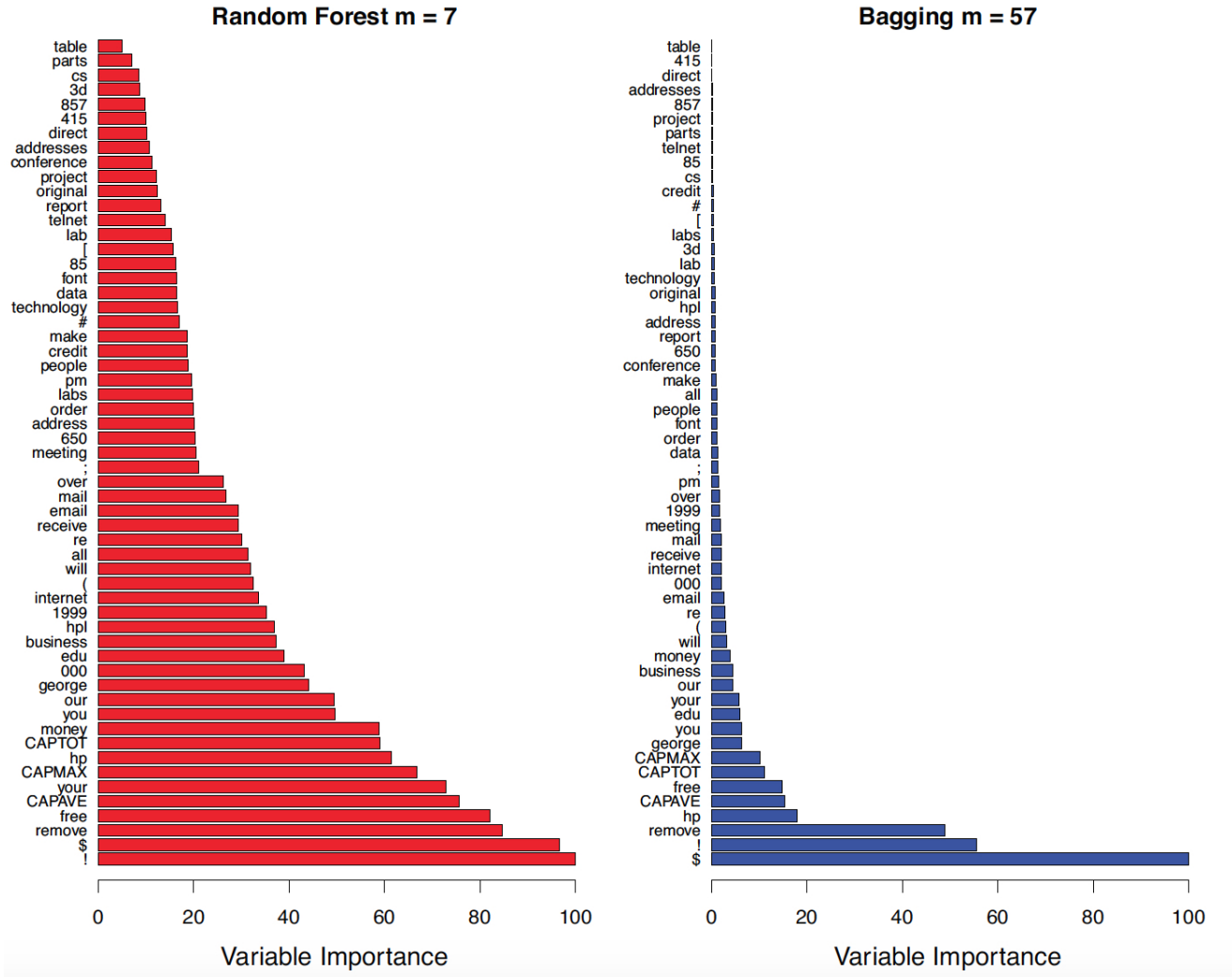


Source: Hastie et al. (2009) p. 589

Variable importance

- We can calculate the `importance` of a predictor X_j
- For each tree, record the accuracy on the OOB observations
- Do the same is done but with X_j values randomly permuted in the OOB observations
- Compute the decrease in each tree's accuracy
- If the average decrease over all the trees is large, then the predictor is considered important - its value makes a big difference in predicting the response
- If the average decrease is small, then the predictor doesn't make much difference to the response

spam data: variable importance



Source: Efron and Hastie (2016) p. 332

Bagging and random forest takeaways

- Bagging stabilizes decision trees and improves accuracy by reducing variance
- Random forests further improve decision tree performance by de-correlating the individual trees in the bagging ensemble
- Random forests' variable importance measures can help you determine which variables are contributing the most strongly to your model
- Because the trees in a random forest ensemble are unpruned and potentially quite deep, there's still a danger of overfitting