Exercise 6

Philipp Hanslovsky, Robert Walecki

June 10, 2012

6.2.2

The gini Δ as well as class distributions for labels 1 and 2 are shown in figure 1. The y-value of the corresponding label does not have a meaning. It just makes both labels more distinct. The x-value of the largest gini Δ is where both classes have the least overlap.

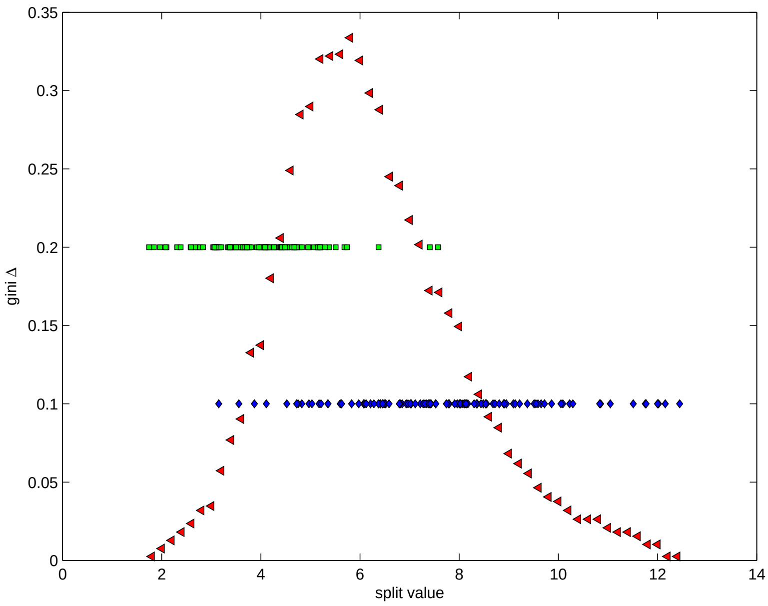


Figure 1: gini Δ as a function of the split value. The class distributions are shown for label 1 (blue) and label 2 (green).

6.3.1

The covariance is a symmetric bilinear form. Therefore the following rules apply:

$$Cov(X, Y) = Cov(Y, X) \tag{1}$$
$$Cov(aX + b, Y) = a * Cov(X, Y) \tag{2}$$
$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \tag{3}$$

3 can be generalized to:

$$Cov(\sum_i X_i, Y) = \sum_i Cov(X_i, Y) \tag{4}$$

Furthermore the covariance is a generalization of the variance:

$$Var(X) = Cov(X, X) \tag{5}$$

The relation between covariance and correlation is given by:

$$\rho_{ij} = \frac{cov(X_i, X_j}{\sigma_i \sigma_j} \overset{i.d.}{=} \frac{cov(X_i, X_j)}{\sigma^2} \tag{6}$$

With this preconditions we get:

$$\begin{aligned}
Var(\frac{1}{B}\sum_{i=1}^{B} X_i) &\overset{5}{=} Cov(\frac{1}{B}\sum_{i=1}^{B} X_i, \frac{1}{B}\sum_{i=1}^{B} X_i) && (7)\\
&\overset{2,3}{=} \frac{1}{B^2}\sum_{i,j=1}^{B}(X_i, X_j) && (8)\\
&\overset{5}{=} \frac{1}{B^2} sum_{i=1}^{B} Var(X_i) + \frac{2}{B^2}\sum_{i=1}^{B-1}\sum_{j=i+1}^{B} cov(X_i, X_j) && (9)\\
&\overset{6,i.d.}{=} \frac{1}{B^2} B\sigma^2 + \frac{2\rho\sigma^2}{B^2}\sum_{i=1}^{B-1}\sum_{j=i+1}^{B} 1 && (10)\\
&= \frac{\sigma^2}{B} + \frac{2\rho\sigma^2}{B^2}\sum_{i=1}^{B-1} B - i && (11)\\
&= \frac{\sigma^2}{B} + \frac{2\rho\sigma^2}{B^2}\left(B\,(B-1) - \frac{B\,(B-1)}{2}\right) && (12)\\
&= \frac{\sigma^2}{B} + \frac{\rho\sigma^2}{B}(B-1) && (13)\\
&= \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{B} && (14)
\end{aligned}$$

6.3.2

The probability p_{oob} for an observation to be out of bag is given by the ratio of the number N_{all} of all possible bags and the number N_{-1} of all possible bags that do not contain given observation. Let N be the number of all observations and k the number of elements contained in a bag. Then the total number of possible bags is given by the number of possible configurations $\frac{N!}{(N-k)!}$ divided by the number of corresponding permutations $k!$ (i.e. choosing observations 1, 2 and 4 is the same as choosing 1, 4 and 2) or the binomial coefficient.

$$\begin{aligned}
N_{all} &= \frac{N!}{k!(N-k)!} && (15)\\
&= \binom{N}{k} && (16)
\end{aligned}$$

The number of possible bags not containing a certain observation can be obtained in a similar fashion. Let x_i be an observation fixed to be not in a bag. Then the number of possible bags for that configuration is given by the number of possible configurations for the remaining $N-1$ samples $\frac{(N-1)!}{(N-k-1)!}$ divided by the number of corresponding permutations $k!$.

$$N_{-1} = \frac{(N-1)!}{k!(N-1-k)!} \tag{17}$$

$$= \binom{N-1}{k} \tag{18}$$

Therefore the resulting probability p_{oob} is given by:

$$p_{oob} = \frac{N_{-1}}{N_{all}} \tag{19}$$

$$= \frac{(N-1)!}{k!(N-1-k)!}\frac{k!(N-k)!}{N!} \tag{20}$$

$$p_{oob} = \frac{N-k}{N} = 1 - \frac{k}{N} \tag{21}$$

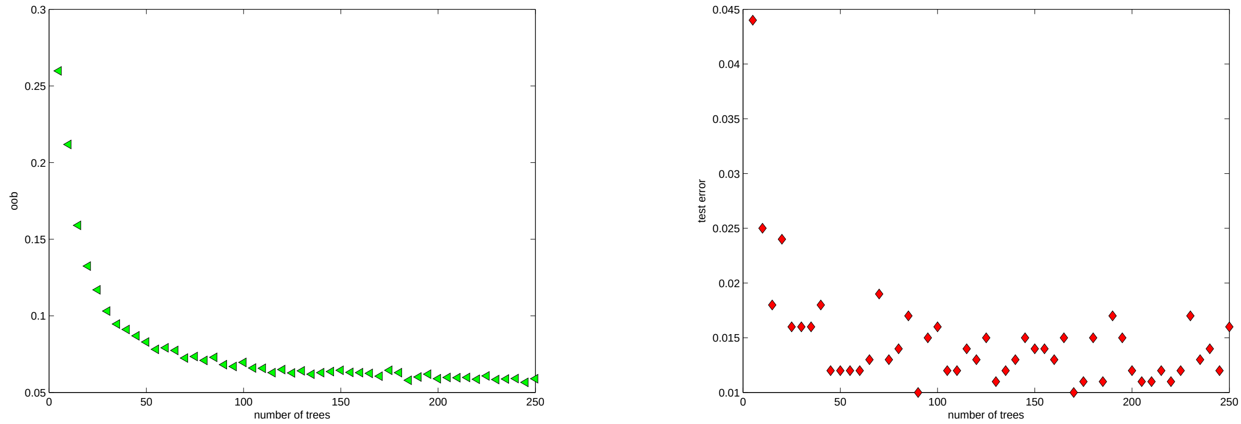
That probability holds for a single bootstrap sample. Given M trees based on M bootstrap samples, the probability for an observation to be out-of-bag in at least one tree can be calculated with the help of the probability that the observation is contained in every bootstrap sample. $(1-p_{oob})^M$ is the probability for an observation to be contained in all bootstrap samples.

$$p = 1 - (1 - p_{oob})^M \tag{22}$$

$$p = 1 - (\frac{k}{N})^M \tag{23}$$

6.3.3

The oob error decreases with an increasing number of trees M. The test error is decreasing as well, however it's fluctuating, so it might be constant for $M > M_{min}$. Therefore choosing a good k might save a lot of computation time.



(a) oob error

(b) test error

Figure 2: oob and test error as functions of the number of trees