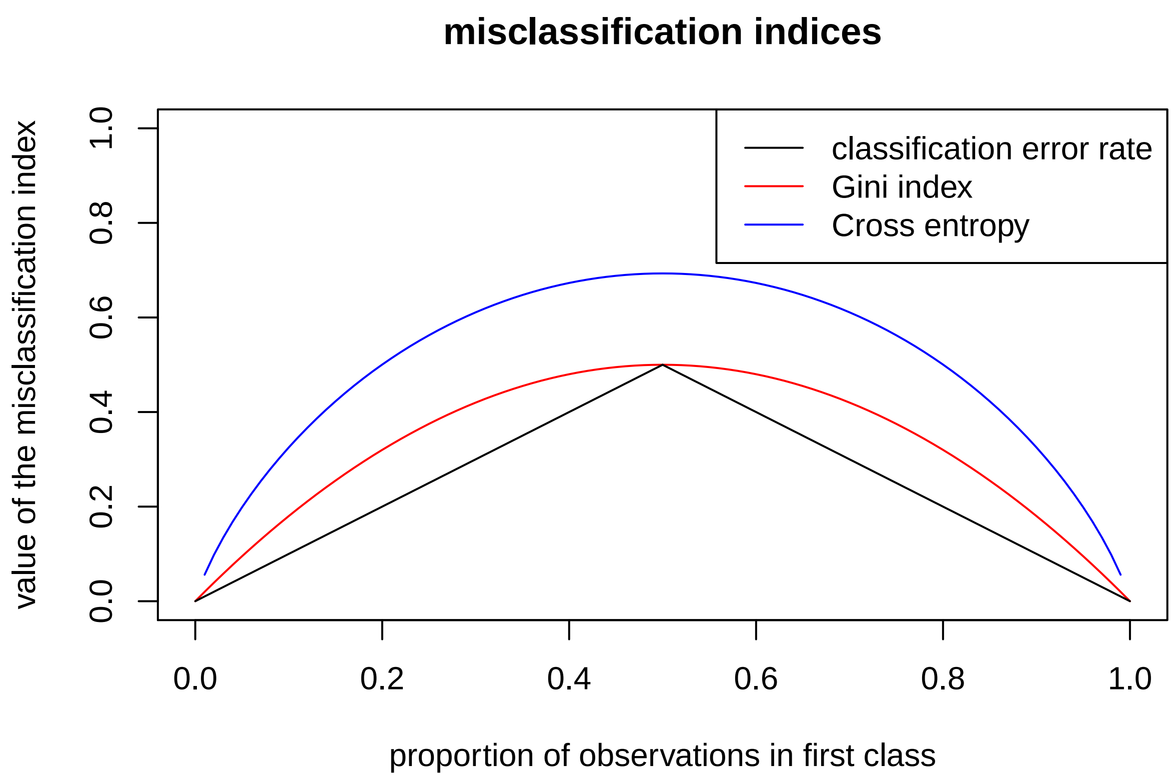
# Exercises IOPS course SLP: Trees and ensembles

## Exercise 1: Comparing misclassification indices




The plot above, show the values of the different misclassification indices, given the proportion of class-1 observations in a node, on the x-axis.

a) Based on the plot, do you expect each of the criteria to favor the same, or different potential splits?

b) Say, we have a mothernode with .7999 in class 1 (and .2111 in class 0). A given split would result in

- .565 going left, of which a proportion of 1.00 are class 1 observations
- .435 going right, of which a proportion of .54 are class 1 observations

Calculate the (average of the) classification error, Gini index and cross-entropy in the mothernode, and in the two daughternodes. Would the split improve purity according to the classification error, Gini index and cross-entropy?

Classification error in the mothernode is

```
1 - .7999
```

```
## [1] 0.2001
```

Classification in the daughter nodes will be

```
.565 * (1 - 1.0) + .435 * (1 - .54)
```

```
## [1] 0.2001
```

Gini index in the mothernode is

```
.7999 * .2001 + .2001 * .7999
```

```
## [1] 0.32012
```

Gini index in the daughternodes will be

```
.565 * (1.0 * 0.0 + 0.0 * 1.0) + .435 * (.54 * .46 + .46 * .54)
```

```
## [1] 0.216108
```

Cross-entropy in the mothernode is

```
- (.7999 * log(.7999) + .2001 * log(.20011))
```

```
## [1] 0.500531
```

Cross-entropy in the daughternodes will be (have to pick a very small value instead of zero, otherwise log is not defined)

```
- (.565 * (1.00 * log(1.00) + 0.00 * log(1e-50)) +
.435 * (.54 * log(.54) + .46 * log(.46)))
```

```
## [1] 0.3001255
```

Conclusion: According to the classification error rate, purity would not improve and no split would be made. According to the Gini index and cross-entropy, purity would improve and a split would be made.

## Exercise 2: Variable selection bias

a) Set the random seed and generate 200 observations from independent variables x1, x2 and e (you are free to choose the shape and parameters of the distribution yourself). Create two datasets consisting of x1, x2 and y: one where $y = e$ (the 'independent' dataset), and one where $y = x2 + e$ (the 'dependent' dataset).

b) Fit a regression tree using x1 and x2 to predict y, using each dataset. Which variable is most often selected for splitting? Is that what you would expect, given that both x1 and x2 were generated so as to be completely independent of y?

c) Prune the trees. Are there any splits left?

c) Use the `ctree()` function from the `partykit` package to fit a conditional inference tree to the independent and depndent same data. Also plot the resulting conditional inference tree.

d) Compare the results you obtained in part a, b and c.

## Exercise 3: Fitting trees and ensembles to the Carseats data

(Adaptation of exercise 8.8 ISLR)

In the lab session of chapter 8 (ISLR), a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

a) Split the data set into a training set and a test set.

b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of mtry, the number of variables considered at each split, on the error rate obtained.

f) Create a boosted ensemble to predict Sales. Compare the boosted ensemble with the bagged and random forest ensemble in terms of test MSE and (the effect of) important predictor variables. (Additional: Before creating the ensemble, use cross validation to determine the optimal parameter settings.)

## Exercise 4: Boston housing and OOB error estimates

(Adaptation of exercise 8.7 ISLR).

In the lab, a random forest was created for the Boston data using mtry=6 and using ntree=25 and ntree=500. For mtry values of $p$, $p/2$, and $\sqrt{p}$. Use ntree values of 1:750. Create a plot with the number of trees on the x-axis and the error rate on the y-axis. Plot both the OOB and test error.

Hints: Note that you only need to fit 3 ensembles, one for each value of mtry, because the fitted randomForest object contains a slot $mse, of which the i-th element ($1 \leq i \leq ntree$) is the OOB estimate of the MSE for all trees up to the i-th; and a slot $test$mse, of which the i-th element ($1 \leq i \leq ntree$) is the test MSE for the ensemble of trees up to the i-th.

To obtain both OOB and test eror, first separate the data in a test and training set and supply these to the `X.train`, `Y.train`, `xtest` and `ytest`arguments of the `randomForest()` function:

```
library(MASS)
set.seed(1)
train <- sample(1:nrow(Boston), nrow(Boston)/2)
X.train <- Boston[train, -14]
X.test <- Boston[-train, -14]
Y.train <- Boston[train, 14]
Y.test <- Boston[-train, 14]
```

a) Based on the plot, does the default setting of ntree=500 seem reasonable to you?

b) Based on the plot, would you prefer a random forest over a bagged ensemble?

c) Does the OOB error give a more realistic estimate of test error for bagged ensembles or for random forests? Can you explain this?

### Exercise 5: Fitting trees and ensembles for classification

Make exercise 8.8 a through e from ISLR.

**Additional question:**

Also fit a boosted tree ensemble. Compare the test MSE for the original tree, pruned tree, bagging, random forest and boosting.