Decision trees

- Aim: Separate the predictor variable space into areas R_m (subgroups) which have increasingly similar values on the response variable.

- Finding globally optimal partition: Computationally infeasible

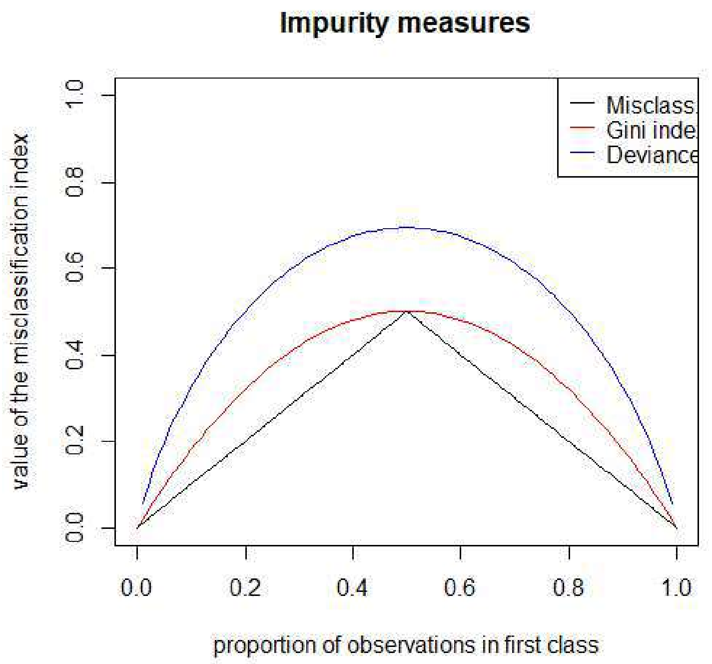
- Turn into feasible task:
  - R_m are *rectangular* and *non-overlapping*
  - Splits are recursive and involve one variable per split
  - Two-way splits only
  - Splits are found in 'greedy' fashion: split minimizing current loss is selected

- Double-edged sword: Resulting structure easy to visualize as a binary decision tree

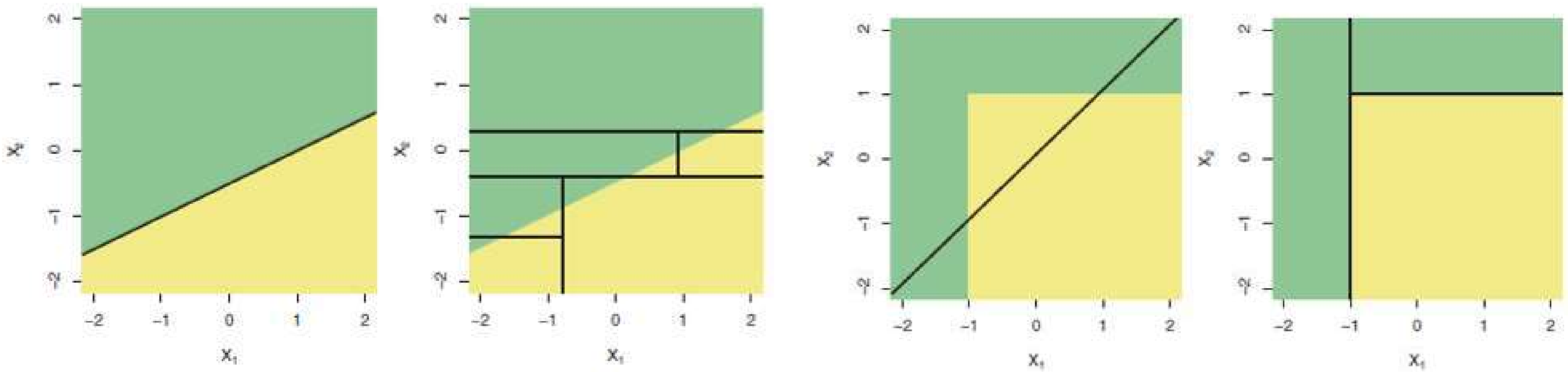
Loss functions (impurity measures)

To quantify the (dis)-similarity on the response variable, we can use several measures:

- Squared error loss: $\frac{1}{N_m}\sum_{x_i \in R_m}(y_i - \hat{y}_m)^2$
- Absolute error loss: $\frac{1}{N_m}\sum_{x_i \in R_m}|y_i - \hat{y}_m|$
- Misclassification error: $1 - \max_k \hat{p}_{mk}$
- Gini index: $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
- Cross-entropy or deviance: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$



Decision trees: (In)stability



- The larger the tree, the lower the bias and the higher the variance (instability)

- (Note: Amount of bias not only a function of the flexibility of the method, but also a function of (shape of) true associations in data)

Decision trees: (In)stability

- Thus, again we optimize (loss + penalty):

$$\sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

- Find α through k-fold CV (note: trees (quite) different in each fold)

Variable selection bias

- CART tends to prefer splitting variables with more possible cutpoints
- Solution: Separate variable and cutpoint selection
- E.g., conditional inference trees:

  1) Select splitting variable based on statistical association tests

  2) Select splitting value that minimizes loss (as usual)
- Uses conditional inference tests, developed by Strasser and Weber (1999)
- Tests provide natural stopping criterion for splitting