STK2100- Machine Learning and Statistical Methods for Prediction and Classification

Spring 2025

Random Forests

Covers Sections 15.1-15.3 (except 15.3.2 and 15.3.3)

Ingrid Hobæk Haff
Department of Mathematics
University of Oslo

Random forests

- In Chapter 10, we encountered boosting tree models, that combine many small trees, in a way that keeps many of the advantages of classification and regression trees (CART), while giving much better predictive performance
- Another type of models that combine many trees to exploit the strengths of tree models, and also tend to give much better results than CART are random forests
- The main idea is to reduce the variance of the combination of trees by fitting each tree to a different bootstrap sample of the data, as well as randomly sampling a subset of the covariates during the splitting
- The resulting prediction is the average over the trees in the regression setting and a majority vote in the classification setting

Random forests for regression

- Let us first consider the regression setting, where the predictions from the trees are averaged
- An average over B independent identically distributed samples, each having variance σ^2, is $\frac{1}{B}\sigma^2$
- If the samples are identically distributed, but all share a positive pairwise correlation ρ (thus are not independent), the variance of the average is instead

$$\begin{aligned}\frac{1-\rho}{B}\sigma^2 + \rho\sigma^2 &= (1-\rho+B\rho)\frac{1}{B}\sigma^2 \\ &= (1+(B-1)\rho)\frac{1}{B}\sigma^2 > \frac{1}{B}\sigma^2, \quad \rho > 0\end{aligned}$$

- As B increases, the variance essentially becomes $\rho\sigma^2$, which does not decrease as B increases, and might not be much smaller than the variance σ^2 of a simple sample

Random forests for regression

- Hence, to reduce the variance of the average of B trees substantially compared to the variance of each tree, the correlation between the trees must not be too large
- This is achieved by random selection of the input variables during the tree-growing process
- More specifically, $m < p$ of the covariates are selected at random as candidates for splitting before each split in each tree
- Let $T(\mathbf{x}; \boldsymbol{\Theta}_1), \ldots, T(\mathbf{x}; \boldsymbol{\Theta}_B)$ be the trees that have been grown
- Then, the random forest predictor is given by

$$f(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T(\mathbf{x}; \boldsymbol{\Theta}_b)$$

- The procedure for growing trees is the same as for CART, except for the random selection of covariates before every split

Random forests for classification

- Random forests can also be constructed for classification
- Then, if $C(\mathbf{x}; \boldsymbol{\Theta}_b)$ is the classification from the tree grown on bootstrap sample b, the random forest classification is given by

$$C(\mathbf{x}) = \text{majority vote}_b C(\mathbf{x}; \boldsymbol{\Theta}_b)$$

- This means that one counts how many times each class was chosen by the B trees, and the class with the highest count is the random forest classification

Random forest algorithm

Random forest growing for regression or classification

for $b = 1, \ldots, B$ **do**
  Draw a bootstrap sample $\mathbf{z}_b^*$ of size N from the data
  Initialise tree T_b on the data $\mathbf{z}_b^*$ as empty
  while Nodesize n_{min} in T_b has not been reached **do**
    Select m covariates at random
    Choose the best split variable and split-point among these m covariates
    Add the split to T_b
  end while
end for
Return $T_1, \ldots, T_B$

Random forests vs. bagging and boosted trees

- Averaging over prediction models fitted to a large number of bootstrap samples in order to reduce the variance is called bootstrap aggregation, or bagging
- The main difference between random forests and bagging trees is the random selection of covariates prior to every tree split in the random forests, which reduces the correlation between trees
- Random forests improve over individual trees by reducing the variance compared to individual trees, but the bias is the same
- This is in contrast to boosted trees, that are grown adaptively to modified versions of the data in order to remove bias, and hence are not identically distributed

More details on random forests

- There are two parameters in random forests that must be chosen:
  - the number m of covariates used in each split
  - the number B of trees in the random forest
- The default values recommended by the inventors of random forests are:
  - $\lfloor p/3 \rfloor$ for regression
  - $\lfloor \sqrt{p} \rfloor$ for classification
- The optimal size of m depends on the application, and it should be treated as a tuning parameter
- Especially if p is large and only a small fraction of them are relevant for the response, one must not set m too small, as this may lead to a small probability of selecting relevant covariates in the splits

More details on random forests

- In principle, the maximum depth of the trees could also be tuning parameter, but studies have shown that little is gained by restricting the size of the individual trees (except from the minimum node size n_{min})
- The generalisation error of a random forest can be estimated by the out-of-bag (OOB) estimate

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \widehat{f}^{*b}(\mathbf{x}_i)),$$

  where $L(\cdot,\cdot)$ is the loss function that is used, C^{-i} is the set of indices of the bootstrap samples that do not contain observation i and $\widehat{f}^{*b}$ is the prediction from tree b in the random forest
- The OOB estimate is very similar to the LOOCV estimate, and can be used to tune m and B

Example: E-mail spam data

See R code