Interactive lecture module 8 and 9

TMA4268 Statistical learning

Mette Langaas

14 March, 2019

Contents

8. Tree-based methods

Tree-based methods and solutions to RecEx

Topics in Module 8

- Method applicable both to regression and classification (K classes) and will give non-linear covariate effects and include interactions between covariates. Based on binary splits of each covariate at a time.
- Glossary: root, branches, internal nodes, terminal (leaf) nodes. Tree drawn upside down.
- A tree can also be seen as a division of the covariate space into non-overlapping regions.
- We build a tree from binary splits in one covariate at the time, chosen to improve some measure of error or impurity. The tree is created by not looking ahead - only at the current best split - thus a *greedy strategy.*
- Criterion to minimize
  - Regression: residual sums of squares
  - Classification: Gini or cross entropy impurity measure or deviance

---

- When to stop: decided stopping criterion - like minimal decrease in RSS or less than 10 observations in terminal node.
- Prediction in terminal nodes:
  - Regression: $\hat{y} = \frac{1}{N_j} \sum_{i:x_i \in R_j} y_i$
  - Classification: majority vote or fraction of each class in a node - and cut-off on probabiity.

- Grow full tree, and then prune back using pruning strategy: cost complexity pruning= cost function + penalty times number of terminal notes (hot handled in detail).

- From one tree to many trees= forest. Why? To improve prediction (but this will give worse interpretation).
- Bagging (bootstrap aggregation): draw B bootstrap samples and fit one full tree to each, used the average over all trees for prediction.
- Random forest: as bagging but only m (randomly) chosen covariates (out of the p) are available for selection at each possible split. Rule of thumb for m is $\sqrt{p}$ for classificaton and $p/3$ for regression.
- OOB: out-of-bag estimation can be used for model selection - no need for cross-validation.
- Variable importance plots: give the total amount of decrease in RSS or Gini index over splits of a predictor - averaged over all B trees. May also be calculated over randomization of OOB.
- Boosting: fit one tree with d splits, make residuals and fit a new tree, adjust residuals partly with new tree - repeat. Three tuning paramteers chosen by cross-validation.

Problems for interactive lecture

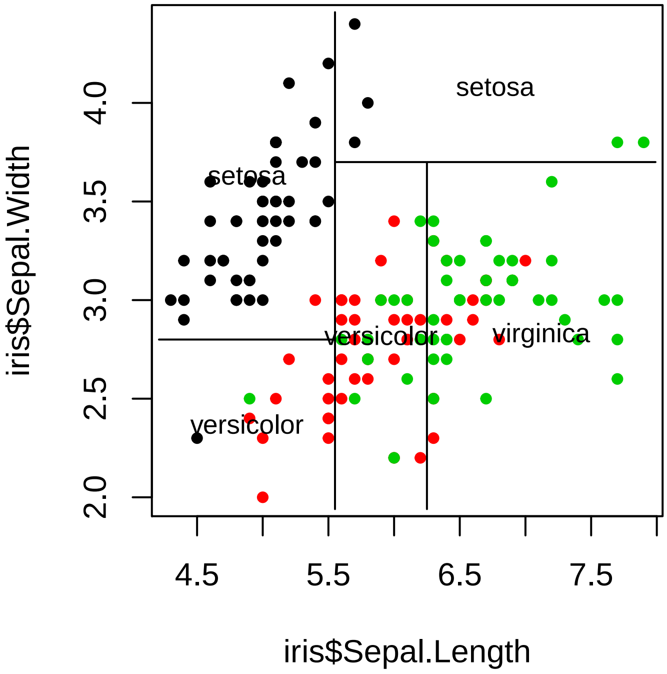
Problem 1: from regions to tree

We have a classification problem with covariates (predictors) `Sepal.Width` and `Sepal.Length` and reponse `Species` (three species)

The graph below gives a partition of the predictor space of variables `Sepal.Width` and `Sepal.Length`, where the observations are shown in different colors for the different species

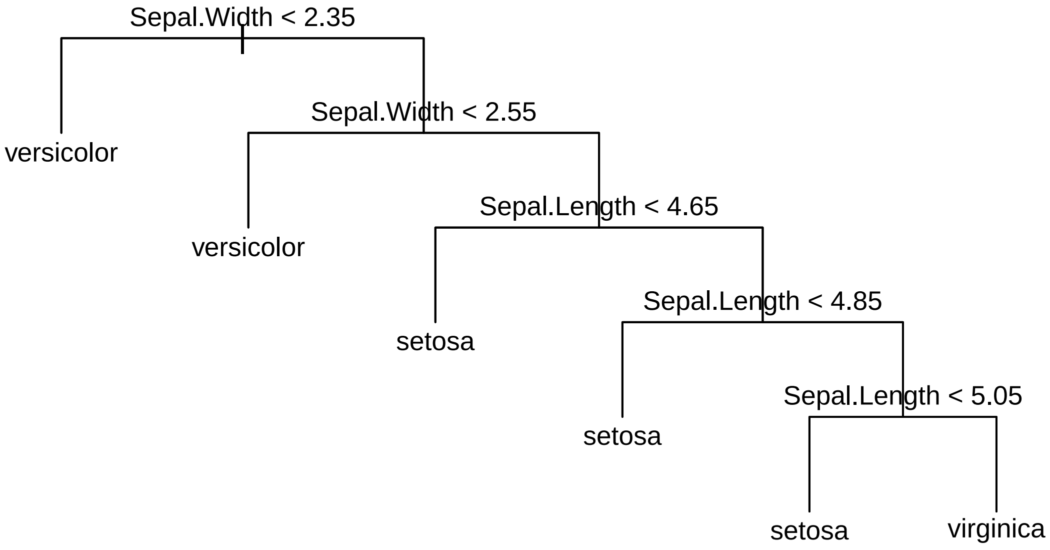
a) From regions to tree

Sketch the classification tree corresponding to the partition. Specify variables that are split on and an approximate value of the split point



b) From tree to regions

For the tree plot, draw the corresponding region plot.



Compulsory exercise 3 in 2018: Problem 1 on Classification with trees

https://www.math.ntnu.no/emner/TMA4268/2018v/Compulsory3.html#problem_1_-_classification_with_trees_%5B4_points%5D

Additonal questions/problems

- What does it mean that a method is *greedy*? Mention one greedy method that we have studied and explain why it is greedy.
- How do we choose that we perform a split in a tree? What is the natural cost function for regression? For classification we focus on node impurity - explain one possible cost function for node impurity.
- Image of tree, explain what you see. Predict the value for a new observation with numerical value given.
- Show full tree and pruned tree and results on test set: compare and argument for which of the models to choose.

- How do we choose the number of bootstrap samples B to be used in bagging and random forest? What about boosting?
- Why do we not have to use cross-validation to estimate error rates for bagging and random forest? What do we instead use, and how do we estimate error rates?
- (MA871 exam): What is boostrapping? We have looked at boostrapping for finding the standard error of an estimator and for bagging and random forest. What is the main idea behind bagging? What is the connection between bagging and random forests?
- For regression trees - how is a simple way to perform boosting?

Learning styles

ACT! project - how can knowing about your learning style help your in your study?

https://innsida.ntnu.no/forms/view.php?id=221738

Compulsory exercise 2

- with focus on the parts with Module 8 and 9!
- https://www.math.ntnu.no/emner/TMA4268/2019v/Compulsory2.html
- https://www.math.ntnu.no/emner/TMA4268/2019v/CompEx2mal.Rmd

9. Support vector machines

Support vector machines and solutions to RecEx.

Topics in Module 9

- SVM is a method for both classification and regression, but we have only studied two-class classification (classes are coded -1 and 1).
- Aim: find high dimensional hyperplan that separates two classes $f(\mathbf{x}) = \beta_0 + \mathbf{x}^T\beta = 0$. If $y_i f(\mathbf{x}_i) > 0$ observation $\mathbf{x}_i$ is correctly classified.
- Central: maximizing the distance (on both sides) from the class boundary to the closes observations= the margin M (maximal marginal classifier) - which is relaxed with slack variables (support vector classifiers), and to allow nonlinear functions of $\mathbf{x}$ by extending an inner product to kernels (support vector machine).
- Support vectors: observations that lie on the margin or on the wrong side of the margin.

- Kernels: generalization of an inner product to allow for non-linear boundaries and to speed up calculations due to inner products only involve support vectors. Most popular kernel is radial $K(x_i, x_i') = \exp(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2)$.
- Tuning parameters: cost and parameters in kernels - chosen by CV.
- Sad: not able to present details since then a course in optimization is needed.
- Nice connection to non-linar and ridge version of logistic regression - comparing hinge loss to logistic loss - but then without the computational advanges of the kernel method.

Compulsory exercise 3 in 2018: Problem 3:

https://www.math.ntnu.no/emner/TMA4268/2018v/Compulsory3.html#problem_2_-_nonlinear_class_boundaries_and_support_vector_machine_%5B2_points%5D

Additional questions/problems

- What is a support vector?
- What are differences between a maximal margin classifier and linear discriminant analysis classifier?
- What are the main differences between the maximal margin classifier and the support vector classifier? Explain the concept of a slack variable.

- What are important aspects of the support vector machine?

Team kahoot!