The Elements of Statistical Learning, WS 2018/19
Jilles Vreeken and Tobias Marschall
Michael Scherer, Fawaz Dabbaghie, Aryan Kamal
Center for Bioinformatics & Max Planck Institute for Informatics
CISPA Helmholtz Center i.G. & Cluster of Excellence MMCI

2019-01-17

Problem Set 7

Deadline: Thursday, January 31. 2019, 10:00 a.m.

Please read and follow the following requirements to generate a valid submission.
This problem set is worth 50 points. You may submit your solutions in groups of two students. The solutions to the theoretical problems should be submitted either digitally (in .pdf format) to `mscherer@mpi-inf.mpg.de` or as a hard copy before the lecture. **Label your hard copy submissions with your name(s).**
Solutions to programming problems and resulting plots need to be submitted in digital format (.pdf). For the programming problems you have to submit an executable version of your code (R script).

For digital submissions the subject line of your email should have the following format:

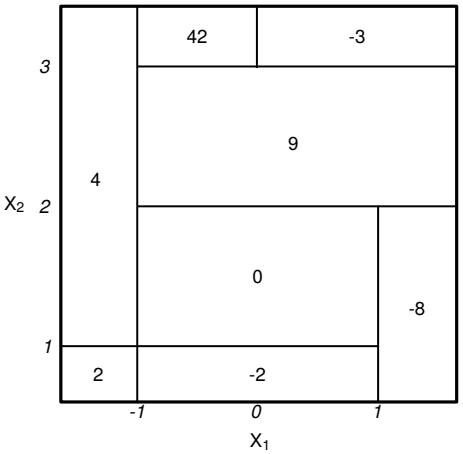
`[SL][problem set 7] lastname1,firstname1;lastname2,firstname2`

Please include the numbers of the problems you submitted solutions to (both digitally and analogously) in the email's body. **Please make sure that all the files are attached to the email.** The attached files should only include an executable version of your code as .R file and **one** .pdf file with all the other solutions.

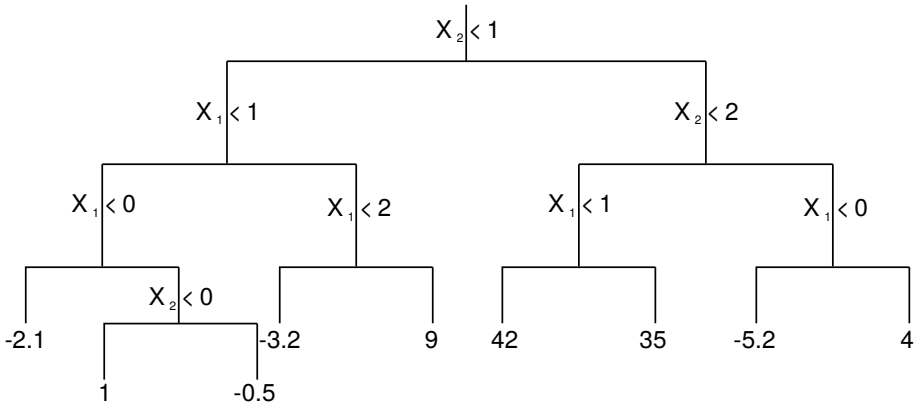
Problem 1 (T, 12 Points)

(modified version of Exercise 8.4.4 in ISLR)

(a) (4 P) Sketch the tree corresponding to the partition of the predictor space indicated in the figure below. The numbers inside the boxes indicate the mean of Y within each region.



(b) (4 P) Create a diagram similar to the one provided in a), using the tree illustrated below. You should divide up the predictor space into the correct regions, and indicate the mean for each region.



(c) (4 P) Create another tree representing the same partition of the predictor space as the one discussed in b), but with a different split at the root node.

Problem 2 (T, 10 Points)

Show that the criterion for the (soft margin) support vector classifier:

$$\begin{aligned} \underset{\beta_0,\ldots,\beta_p,\xi_1,\ldots,\xi_N}{\text{maximize}} \quad & M \\ \text{subject to} \quad & \|\beta\| = 1 \\ & \xi_i \geq 0; \\ & y_i f(x_i) \geq M(1-\xi_i) \quad \text{for } i = 1,\ldots,N \\ & \sum_{i=1}^{N} \xi_i \leq \text{constant} \end{aligned}$$

is equivalent to the formulation using the Hinge loss:

$$\underset{\beta_0,\beta}{\text{minimize}} \quad \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \lambda \|\beta\|^2 \ ,$$

where $y_i \in \{-1, 1\}$, $f(x) = x^T\beta + \beta_0$.

Hint: Show that both formulations are equivalent to

$$\begin{aligned} \underset{\beta_0,\beta}{\text{minimize}} \quad & \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & \xi_i \geq 0; \\ & y_i f(x_i) \geq 1 - \xi_i \quad \text{for } i = 1,\ldots,N, \end{aligned}$$

then $\lambda = \frac{1}{2}C > 0$.

Problem 3 (T, 8 Points)

(Exercise 10.7.1 in ISLR)
Show that the following equality holds:

$$\frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2 = 2\sum_{i \in C_k}\sum_{j=1}^{p}(x_{ij} - \bar{x}_{kj})^2,$$

where $\frac{1}{|C_k|}$ denotes the number of observations in cluster C_k, and $\bar{x}_{kj}$ the mean for feature j in cluster C_k. Argue on the basis of this identity, that the K-means clustering algorithm decreases the objective

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

at each iteration.

Problem 4 (P, 20 Points)

Go through **10.5 Lab 2: Clustering** (ISLR p.404–407) and **10.6 Lab 3: NCI60 Data Example** (ISLR p.407–413).

Fine needle aspiration cytology is an inexpensive technique used for diagnosing subcutaneous (just under the skin) lumps. It involves the extraction of cells from the liquid in the tumor mass. These cells are stained in dye and imaged under a microscope, resulting in micrographs. Your task is to identify subgroups among the tumors.

(a) (2P) The original dataset contains detailed information on the features and can be found at the UCL machine learning repository: UCL's Breast Cancer Wisconsin (Diagnostic) Data Set. Download the normalized dataset (wdbc.RData) from the course website. Use the first 400 as training, and the remaining as test observations.

(b) (6P) Apply Random Forest to predict cancer subtype (column: "Subtype") on the training set. Report train and test set misclassification error and comment on the importance of the individual predictors. (Useful package: `randomForest`).

(c) (6P) Apply SVM to predict cancer subtype. Use the radial-basis and the polynomial kernel and compare training and test error. Compare the results obtained with the Random Forest results. (Useful package: `e1071`)

(d) (6P) Use hierarchical clustering on the input features with the three linkage methods discussed in the lecture ("average","complete" and "single"). Visualize the tree representing the clustering and compare the resulting cluster to the subtypes annotated. How can you improve the overlap between annotated cancer subtypes and clustering results?