20th March, 2019

STK-2100

Mandatory assignment 2 of 2

Submission deadline

Thursday 4th April 2019, 14:30 at Devilry (devilry.ifi.uio.no).

Instructions

You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with LaTeX). The assignment must be submitted as a single PDF file. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. Students who fail the assignment, but have made a genuine effort at solving the exercises, are given a second attempt at revising their answers. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (e-mail: studieinfo@math.uio.no) well before the deadline.

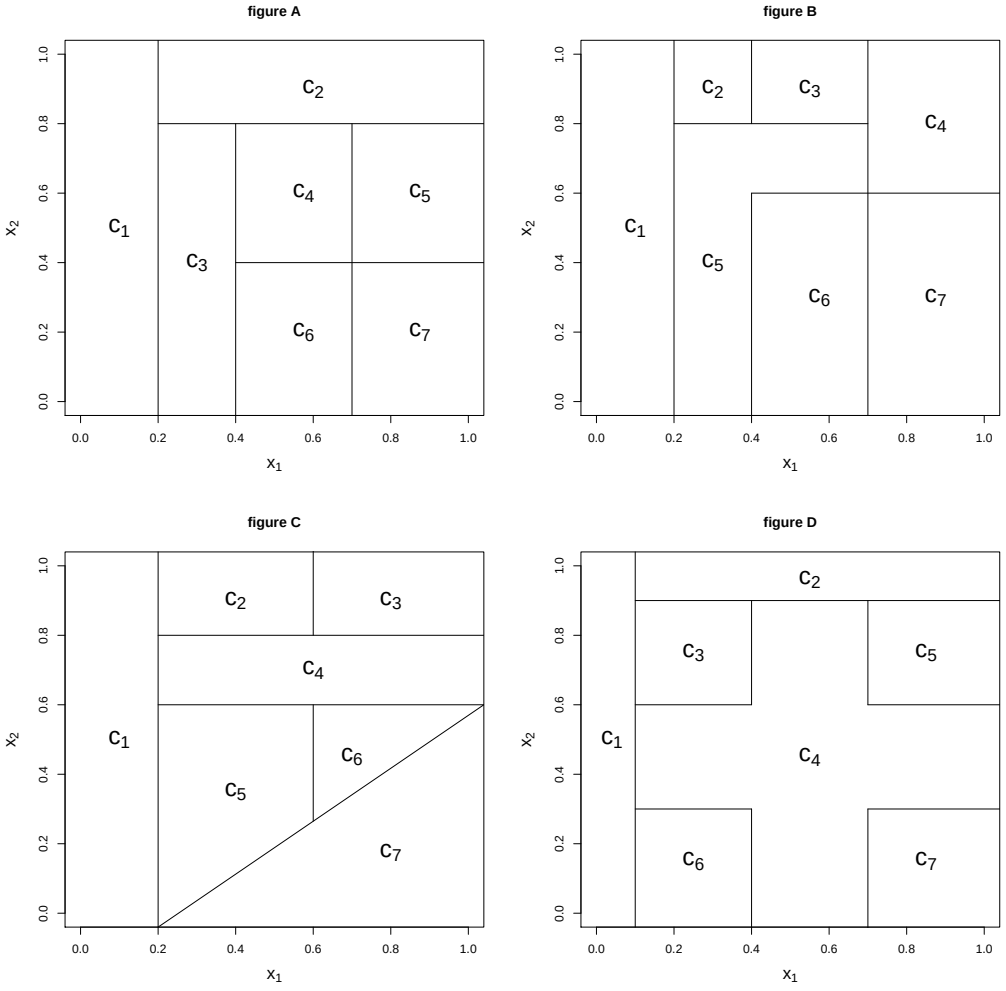
All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments:

uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html

GOOD LUCK!

Problem 1. Consider the following figures:



(a) Which of them can have been produced by a regression tree? Explain why that/those you did not select cannot have been.

(b) For that/those which can have been, write down the corresponding tree, including the formulas in the nodes and the values in the leaves.

(c) Using the tree(s) selected at point (b), provide an estimate for the following values:

(i) $f(x_1 = 0.5,\ x_2 = 0.5)$;

(ii) $f(x_1 = 0.3,\ x_2 = 0.5)$;

(iii) $f(x_1 = 0.5,\ x_2 = 0.3)$;

(iv) $f(x_1 = 0.95,\ x_2 = 0.05)$;

(v) $f(x_1 = 0.05,\ x_2 = 0.95)$;

(vi) $f(x_1 = 0.45,\ x_2 = 0.75)$.

Problem 2. Consider the `Vertebral Column` dataset (Dua & Graff, 2019). It contains information about 310 orthopaedic patients, which belong to three 3 classes (normal, disk hernia or spondilolysthesis). Each patient is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (in this order): pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis. The following convention is used for the class labels: DH (Disk Hernia), Spondylolisthesis (SL), Normal (NO).

Consider initially a two-class problem (healthy patients, NO, versus unhealthy patients, DH and SL). Split the data into a training (2/3 of the observations) and a test set (1/3 of the observations), taking care that the proportion between healthy and unhealthy patients is similar in the two sets.

(a) Fit on the training set a large (more than 20 partitions) classification tree, and report its graphical representation (do not worry if the graphic is not readable).

(b) Use a cross-validation procedure to find the best number of partitions and prune the tree. Comment on the (possible) differences between the two trees.

(c) Compute the misclassification error on the test set for both trees and comment the results. Did you expect this result? Why?

(d) Apply LDA and QDA to the same problem, and contrast their prediction (misclassification) error with those obtained by using the trees. Which method does perform better in this example?

(e) Repeat points (a) to (d) for the three-class problem (DH, SL and NO).

(f) Using the tree obtained at point (b), write a small report for a medical doctor in which you explain him/her how to separate between healthy (NO) and unhealthy (SL and AB) patients based on their biomechanical attributes.

Problem 3. Consider the data in `data.dat`. They contain some values `x` and the corresponding response `y`. Randomly split them into a training and a test set of equal size.

(a) Estimate $f(x)$, the function which transform `x` into `y`, by using the k-nearest-neighbours algorithm, making sure of selecting the right k by 5-fold cross-validation. Plot the training points and add the estimated curve in red. Comment on the choice of k.

(b) Repeat the procedure using cubic splines, with 4 knots, and cubic smoothing splines, the latter after having suitably chosen the penalty term λ by 5-folds cross validation. Add the two estimated curves to the previous graphic in green and blue, respectively. Comment on the differences among the two curves (if any) and the effect of λ on the smoothness of the latter.

(c) Compute the prediction error for the three methods on the test set and comment the results.

Bibliography

Dua, D. & Graff, C. (2019). UCI machine learning repository.