DATA11002 Introduction to Machine Learning
Course examination, 17 December 2019
Examiner: Kai Puolamäki

Answer all of the 4 problems to obtain the maximum score of 60 points. This examination has 2 pages.

You must have passed the Exercise Set 0 to participate to this examination. You must also have *either* obtained 50% of the exercise points from Exercise Sets 1–6 *or* passed the project work to participate to this examination.

You are allowed to have a "cheat sheet" with you at the exam. The cheat sheet is one two-sided handwritten A4-sized paper where you can write any information whatsoever. No other extra material is allowed.

An important grading criterion is understandability: in addition to giving factually correct and complete answer to the question asked, with all claims made substantiated, your answer should additionally be clear, accurate, and understandable to your fellow student who has the necessary prerequisite knowledge but has not yet taken the course.

Please write in clear handwriting and leave a wide left or right margin. You may answer in English, Finnish, or Swedish. If you use Finnish or Swedish, it will be helpful to include the English translations of any technical terms that may be ambiguous.

1. [*12 points*] Explain briefly the following terms and concepts. Your explanation should include, when applicable, both a *precise definition* and a *brief description* of how the concept is useful in machine learning. Your answer to each subproblem should fit to roughly one third of a page of normal handwriting or less.

   (a) *logistic regression*

   (b) *discriminative model*

   (c) *naive Bayes classifier*

   (d) *support vector machine*

   (e) *linear discriminant analysis*

   (f) *principal component analysis*

2. [*16 points*] Consider a training data set with $n = 10$ observations. Imagine you learn a regression model and find that it estimates the values of the dependent variable in the training examples exactly, with no error whatsoever.

   (a) What can you say with certainty about the performance of your regression model on new test data? Explain what makes generalisation hard. What properties of a regression model are relevant when trying to analyse its generalisation error?

   (b) Explain how k-fold cross-validation is done. Give an example of how it can be used to find a good regression model.

   (c) Now suppose that instead of regression, the task would have been to estimate, for example, the median of an unknown distribution from which we have $n = 10$ data points. How would you apply resampling to measure the accuracy of an estimate computed from the given n points?

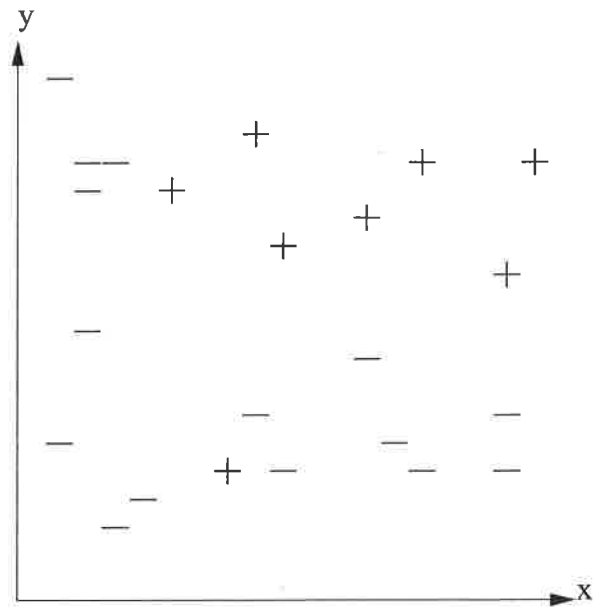

Figure 1: Toy data set for Problem 3. The covariates $(x, y) \in \mathbb{R}^2$ are given by the x and y axis, respectively, and the class $c \in \{-1, +1\}$ to be predicted by plus ("+") and minus ("-") signs.

3. [*16 points*]

   (a) What is a classification tree? Define it.

   (b) Describe the algorithm used to build classification trees (that was covered in the lectures and in the course text book).

   (c) How can this algorithm be modified to address overfitting to the training data?

   (d) Sketch a run of the algorithm with the toy data set in Figure 1 (binary classification task in $\mathbb{R}^2$) and draw the resulting classification tree. (You do not need to worry about overfitting here: the resulting classification tree can fit the training data with no error.)

4. [*16 points*]

   (a) What kind of tasks can we use the Lloyd's (k-means) algorithm for? Explain what the *inputs* and *outputs* of the algorithm are. How to interpret the results?

   (b) Define the objective (or cost) function that the Lloyd's algorithm tries to minimize. What can be said about the value of the objective function during the two stages of each iteration of Lloyd's algorithm?

   (c) Consider the following set of data points in $\mathbb{R}^2$: $x_1 = (0, 1)$, $x_2 = (1, 2)$, $x_3 = (4, 5)$, $x_4 = (5, 3)$, $x_5 = (5, 4)$. Run the Lloyd's algorithm using $K = 2$ and initial prototype (mean) vectors $\mu_1 = (0, 2)$ and $\mu_2 = (2, 0)$. Draw the data points, cluster prototype vectors, and cluster boundary after each iteration until convergence. Also, write down calculation procedure and the cluster memberships as well as prototype vectors after each iteration.