582631 Introduction to Machine Learning
Separate examination, Friday 17 April 2015
Examiner: Jyrki Kivinen

Answer all the problems. The maximum score for the exam is 60 points.

This exam is based on the lecture course of Autumn 2014. To take this exam, you should either have completed the required minimum amount of homework during the lecture course, or have completed a separate programming project. If you have not done either, please send e-mail to `jyrki.kivinen@cs.helsinki.fi` after the exam and explain your situation.

You may answer in English, Finnish or Swedish. If you use Finnish or Swedish, it may be helpful to include the English translations for any technical terms you introduce.

1. [*12 points*] Explain briefly the following terms and concepts. Your explanation should include, when appropriate, both a precise definition and a brief description of how the concept is useful in machine learning. The explanation of a single item should take *at most* half a page.

   (a) *generative* and *discriminative* classification

   (b) *Gini index*

   (c) *pruning* a decision tree

   (d) *ridge regression*

[TR] We didn't discuss ridge regression so skip that.

2. [*12 points*] **Bayes and Naïve Bayes.** We consider probabilistic binary classification of data represented by two binary attributes. In other words, we are assigning labels + and − to data points of the form (x_1, x_2) where $x_1, x_2 \in \{0, 1\}$.

   Assume that the data consists of following numbers of instances with various attribute and class values:

$y = +$	$x_1 = 0$	$x_1 = 1$
$x_2 = 0$	3	5
$x_2 = 1$	4	8

$y = -$	$x_1 = 0$	$x_1 = 1$
$x_2 = 0$	12	9
$x_2 = 1$	6	3

   (For example, there are 4 positive and 6 negative examples with the attribute values $(x_1, x_2) = (0, 1)$.)

   (a) What is the probabilistic Bayes classifier for this data set?

   (b) How would you convert the probabilistic classifier into a forced-choice ("hard") classifier? We assume that the basic 0/1 cost is used. What is then the Bayes error?

   (c) Using construct the Naïve Bayes classifier for this data. Explain the mathematical model on which it is based.

   In all subproblems, show your intermediate calculations.

[TR] Problem 2 will not be used as such since we didn't practice calculating Bayes classifiers and Bayes errors by hand except for one exercise. Something a bit like this could still be asked. We did study the naive Bayes classifier, so you should definitely know how it works.

Continues on the other side!

3. [*12 points*] **Overfitting and underfitting.** Explain what is meant by overfitting and underfitting. How can one detect that they have happened, or try to avoid them? Explain both generally, and using the k-nearest neighbor classifier as an example case.

4. [*12 points*] **Classification algorithms.** The *perceptron algorithm* and *Hunt's algorithm* (also known as TDIDT) are two very commonly used learning algorithms for classification. Describe the main properties of these two algorithms, and compare them, considering for example

   [TR] Just replace Hunt's algorithm by decision tree learning as discussed this year and something similar to this could be used. However, we didn't discuss variations or generalizations or computational complexity.

   - what kind of classifiers the algorithms produce
   - can they handle different kinds of input variables (categorical vs. numerical etc.)
   - what are the most important variations and possible generalizations of the algorithms
   - how efficient they are computationally.

   Based on this, can you give some ideas on which kind of problems one or the other algorithm would be more appropriate?

   (You do *not* need to give pseudocode or other detailed explanation of *how* the algorithms work.)

5. [*12 points*] **Gaussian mixtures.**

   (a) What are Gaussian mixtures? In what kind of machine learning problems are they used: what are the inputs and outputs?

   [TR] Skip this one: we didn't discuss mixture models or the EM algorithm at all.

   (b) How is the *Expectation Maximisation* (EM) algorithm used in the context of Gaussian mixtures? Explain the mathematical problem that the EM algorithm is, in this context, intended to solve. What can you say about how efficiently EM does solve the problem?

   (c) Give a detailed (pseudocode) description of the EM algorithm when applied to Gaussian mixtures.