582631 Introduction to Machine Learning
Separate examination, Tuesday 13 September 2016
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 2015. To take this exam, you should either have completed the required minimum amount of homework during the lecture course, or complete a separate programming project. If you were not on the lecture course in Autumn 2015, please send e-mail to `jyrki.kivinen@cs.helsinki.fi` after the exam and explain your situation (if you have not done so already).
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, techniques 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 concept should take *at most* half a page in normal handwriting.

   (a) generative model
   (b) Gini index
   (c) pruning a decision tree
   (d) one-versus-one classification
   (e) precision and recall
   (f) Laplace correction

2. [*16 points*] **Bayes optimal prediction.** We consider predicting a class $Y \in \{1, 2, 3\}$, when the input is just one real value $X \in [0, 9]$. Suppose the distribution of pairs (X, Y) is as follows:

   [TR] This question wouldn't be possible in this form since we didn't practice calculating Bayes classifiers by hand enough.

   (a) First we pick $Y = 1$ with probability $1/2$, $Y = 2$ with probability $1/3$, and $Y = 3$ with probability $1/6$.
   (b) If $Y = 1$ then X is picked from the uniform distribution over $[0, 9]$.
   (c) If $Y = 2$ then X is picked from the uniform distribution over $[5, 8]$.
   (d) If $Y = 3$ then X is picked from the uniform distribution over $[4, 6]$.

   Calculate are the posterior class probabilities $P(Y = i \mid X = x)$ for $i = 1, 2, 3$ as a function of x. What is the Bayes optimal classifier for this distribution? What is the Bayes error?

Continues on the other side!

[TR] Item (b) wouldn't be possible in this form since we didn't discuss the derivation of the univariate least-squares solution. In item (c) I wouldn't require you to remember the formula (X^T X)^{-1} X^T y for the multivariate case.
Otherwise the problem would be fine.

3. [*16 points*] **Linear regression.**

   (a) Explain what is meant by *linear regression*. Give an example of an application in which it might be appropriate, and also a general mathematical formulation.

   (b) Derive the *ordinary least-squares* solution for univariate linear regression (with just one input variable, but including the intercept, or bias, term). Remember to explain what you do and why.

   (c) State also the solution for the general multivariate linear least-squares problem. You do not need to derive it, but explain the meaning of all symbols and any underlying assumptions. What computations one would need to do in practice to apply the formula? Estimate the computational complexity.

   (d) Explain how you would apply the linear regression framework to fit a *polynomial* to a set of data points $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$. Using polynomials as an example, explain the notions of *overfitting* and *underfitting*. What techniques can one use to avoid overfitting and underfitting, both generally and specifically in linear regression?

4. [*16 points*] **K-means algorithm.**

   (a) For what kind of tasks can we use the K-means algorithm? Explain carefully what the inputs and outputs of the algorithm are, and give a very brief intuitive explanation of how the results are to be interpreted.

   (b) Describe the actual K-means algorithm. The description should brief and on a high level.

   (c) Define formally the objective function which the K-means algorithm tries to minimise. Can you give any guarantees about how well the algorithm actually minimises the function?

   (d) *K-medoids* is a variant of K-means. How is the actual algorithm different? In what kind of situation would K-medoids be preferrable over K-means?