T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

3 September 2008 at 9–12.

To pass the course you must also pass the term project. Results of this examination are valid for one year after the examination date.

To get full points you must choose and complete **five of the six problems**. Only the first five answers read by the examiner will be graded.

This examination has six problems (of which you must choose five) and two pages. You can answer in Finnish, Swedish or English. Please write clearly and leave a wide left or right margin. You can have a calculator, with memory erased. No other extra material is allowed.

An important grading criterion is understandability: in addition to being complete and correct, your answer should be understandable to your fellow student who has the necessary prerequisite knowledge but has not yet taken the course.

The results will be announced in Noppa on 3 October 2008, at latest. No other announcements will be sent.

You can keep this paper.

1. *Model selection.* Assume that you have at your disposal a training data set $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t=1}^N$, where $r^t \in \{0, 1\}$ is a binary class and $\mathbf{x}^t \in \mathbb{R}^k$ is a covariate vector of $k$ real variables. Consider the problem of constructing a predictor or classifier $h(\mathbf{x})$ for the class $r$ for data vectors $\mathbf{x}$ that do not appear in the training data.
   (a) Explain concepts "inductive bias", "underfitting", "overfitting", "hypothesis space" and "generalization" and their relation in the framework of this problem.
   (b) Give an example of a realistic hypothesis space for this problem.
   (c) How could you estimate the prediction error for yet unseen data?
2. *Bayesian probability theory.* Consider the problem of finding the probability that a coin flip gives "heads" given a set of observed coin flips (assume that the probability of "heads" or "tails" can also be something else than $\frac{1}{2}$ of a fair coin).
   (a) Demonstrate at least two prior probability densities for this problem, compare them and explain their interpretation.
   (b) Describe (using relevant concepts) how you could find the probability of getting "heads" after observing $N$ coin flips for various choices of prior probability density. Write down the essential formulae.
   (c) Define the maximum likelihood (ML) and maximum a posteriori (MAP) estimates and compare their properties.
3. *Regression.* Consider the problem of linear regression using least squares estimates, given a data set of $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t=1}^N$, where $r^t \in \mathbb{R}$ is the dependent variable and $\mathbf{x}^t \in \mathbb{R}^k$ is the covariate vector of $k$ real variables.



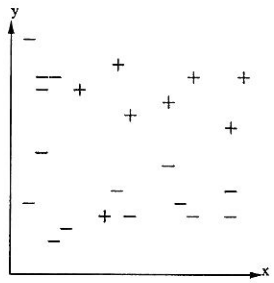


Figure 1: Toy data set for problem 5.

(a) Define a likelihood function and use it to derive the error function to be maximized.

(b) Explain the difference between linear and polynomial regression.

4. *Principal component analysis.* Assume that your data $\mathcal{X}$ is $N$ d-dimensional real vectors, that is, $\mathcal{X} = \{\mathbf{x}^t\}_{t=1}^N$, $\mathbf{x}^t \in \mathbb{R}^d$. Consider the problem of reducing the dimensionality of your data to $k$ dimensions, where $k < d$, using principal component analysis (PCA).

   (a) Write down in pseudocode how you could find the PCA representation of the data in $k$ dimensions. (Hint: it is probably easiest to use matrix representation here. You can assume that you have access to a function that gives eigenvectors and eigenvalues of a matrix.)

   (b) How can you interpret the PCA dimension reduction geometrically?

   (c) How can you choose $k$? List some methods.

5. *Classification trees.*

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

   (b) Describe the ID3 algorithm. What else do you need to take into account when constructing a classification tree using a real world data?

   (c) Sketch the running of the ID3 algorithm with a toy data set of Figure 1 (binary classification task in $\mathbb{R}^2$).

6. *Logistic discrimination.*

   (a) Define logistic discrimination. What can it be used for?

   (b) Derive the error function to be maximized in logistic discrimination.

   (c) Discuss the ways of optimizing this cost function. What do you need to take into account?