T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

30 October 2009.

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

This examination has five problems 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 30 November 2009, at latest.

Please fill the course feedback form (open until 6 November 2009) at `http://www.cs.hut.fi/Opinnot/Palaute/kurssipalaute-en.html`.

You can keep this paper.

1. Write a couple of sentences about the terms below in the context of the course, e.g. what is in common and what are the differences.
   - hypothesis space–version space
   - overfitting–underfitting
   - probability–probability density
   - feature selection–feature extraction
   - generative learning–discriminative learning
   - parametric methods–nonparametric methods

2. Consider the problem of linear regression using least squares estimates, given a data set of $\mathcal{X} = \{(r^t, x^t)\}_{t=1}^N$, where $r^t \in \mathbb{R}$ is the output (variate) to be predicted and $x^t \in \mathbb{R}$ is the input (covariate).
   - Write the model equation $r^t \approx g(x^t \mid \boldsymbol{\theta}) = \ldots$ and the error function $E(\boldsymbol{\theta} \mid \mathcal{X})$ to be minimized.
   - Give the solution of the parameters $\boldsymbol{\theta}$ either as mathematical equations or as pseudocode. (If you have memorized the solution, explain with a few words how you could have derived it.)
   - Is it possible solve polynomial regression with linear algebra? Why?

3. Consider a Bayesian network that has three binary variables M (trip to Mexico), S (swine flu), and F (fever). The joint distribution is $P(M,S,F) = P(M)P(S \mid M)P(F \mid M)$ and the parameters are: $P(M=1) = 0.05$, $P(S=1 \mid M=0) = 0.01$, $P(S=1 \mid M=1) = 0.05$, $P(F=1 \mid S=0) = 0.01$, and $P(F=1 \mid S=1) = 0.9$.

   - Draw the graphical representation of the Bayesian network.
   - Compute $P(M=1 \mid F=1)$, that is, the probability that one has been to Mexico if we know that she have fever.

4. Consider principal component analysis (PCA) for the 2-dimensional data below.

   - Find the direction of maximal variance (or the first eigenvector). Describe the steps of your solution.
   - Compute the proportion of variance explained by the first principal component.

   Hint: Finding the eigenvectors and eigenvalues of a diagonal matrix is easy, but if you cannot find them, you can solve the rest of the problem in pseudocode style.

t	x_1^t	x_2^t
1	0.0	0.0
2	2.0	0.0
3	1.0	3.0

5. Clustering.

   - Run E, M, and E steps of the Lloyd's algorithm for k-means clustering on the 1-dimensional data below. Use $k = 2$ clusters and the initial prototype vectors (=reference vectors) $m_1 = 0.0$ and $m_2 = 1.0$. Explain the steps.
   - Fit a mixture of Gaussians by taking one M-step, using the cluster assignments from your k-means clustering solution. Remember to estimate the parameters describing both $P(G^t)$ and $p(x^t \mid G^t)$, where G^t are the cluster assignments. Hint: You can think of the cluster assignments G^t as classes, so that the problem becomes equivalent to estimating the parameters of a parametric classifier.

t	x^t
1	0.0
2	1.0
3	3.0
4	4.0
5	5.0