T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

C, 31 October 2008 at 13–16.

You must have passed the term project 2007 or part 1 of the term project 2008 to participate to this examination.

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 three 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 1 December 2008, at latest. No other announcements will be sent.

Please fill the course feedback form (open until 9 November 2008) at `http://tieto.tkk.fi/Opinnot/kurssipalaute.html` (in Finnish) or at `http://www.tkk.fi/Units/CSE/Studies/feedback.html` (in English).

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 $r^t \in \mathbb{R}$ is a real number and $\mathbf{x}^t \in \mathbb{R}^d$ is a covariate vector of d real variables. Consider the problem of constructing a regressor $g(\mathbf{x})$ to approximate 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 examples of realistic hypothesis spaces for this problem.

   (c) How could you estimate the prediction error for yet unseen data?

   (d) Generally in supervised learning: explain how the prediction error on training data and yet unseen data is related?

2. *Bayesian networks.*

   (a) Define the concept of Bayesian network.

   (b) Find an expression for probability $P(x_4 \mid x_1, x_2, x_3)$, given the network in Figure 1. You can assume that θ, x_1, x_2, x_3 and x_4 are discrete random variables.

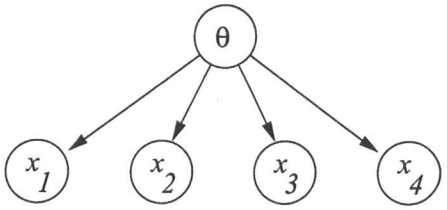

Figure 1: Bayesian network for problem 2.

(c) If x_i, $i \in \{1, \ldots, 4\}$, are observations and θ are parameters of a probabilistic model that has been assumed to have generated the observations then what is $P(\theta \mid x_1, x_2, x_3, x_4)$ commonly called?

3. *Bayesian probability theory.* Consider the problem of finding mean of N real numbers, $\mathcal{X} = \{x^t\}_{t=1}^N$ where $x^t \in \mathbb{R}$.

   (a) Define a feasible probabilistic model for this problem that has a mean as a sole parameter.

   (b) Define a feasible prior probability density for your problem and use it to derive an expression for posterior probability density.

   (c) Use your results to derive maximum likelihood (ML) and maximum a posteriori (MAP) estimates for the mean.

4. *Bayesian multivariate classification.* Consider the problem of classifying real vectors into two classes using Bayesian classifiers with class densities taken to be multivariate normal distributions, given the training data $\mathcal{X} = \{(r^t, x^t)\}_{t=1}^N$, where $r^t \in \{0, 1\}$ and $x^t \in \mathbb{R}^d$.

   (a) Write down the likelihood function.

   (b) How can you tune the complexity of your model?

   (c) What is Naive Bayes assumption? Derive the discriminant function for Naive Bayes classifier.

5. *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.

6. *Decision trees.*

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

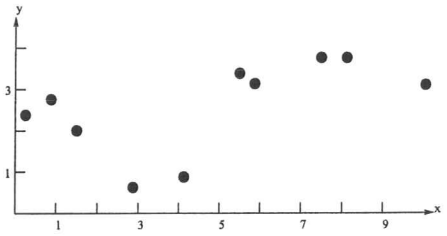

Figure 2: Toy data set for problem 6.

(b) Describe the ID3 algorithm by using pseudocode. Explain pruning in this context. Why and when is the pruning necessary?

(c) Sketch the running of the ID3 algorithm with a toy data set of Figure 2 (regression task of predicting y given x). What is the cost function that the algorithm is optimizing?