T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAM

13 December 2011.

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 in two pages. Each problem is worth 6 points. 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.

You can keep this paper.

1. Write about the terms below in the context of the course, e.g. what is in common and what are the differences. Use full sentences in your answer.
   (a) supervised learning–unsupervised learning (2 points)
   (b) validation data–testing data (2 points)
   (c) estimation bias–estimation variance (2 points)
2. Consider a probabilistic model for parametric regression, a prior distribution over the parameters, and a training set.
   (a) Show that the optimal predictions for new data are given by an integral over the posterior distribution of parameters (Give the derivation). (2 points)
   (b) How can one approximate the integral in the case that we cannot solve the integral analytically? Show the connection of the approximate method to the optimal predictions. (1 point)
   (c) What are the advantage(s) and disadvantage(s) of the above approximate method? (1 point)
   (d) Give definitions of two estimation methods for parametric models. (2 points)
3. Principal Component Analysis (PCA)
   (a) Do the PCA learning using the 2-dimensional data set in the table below. Describe the steps of your solution. (4 points)
   (b) Compute the proportion of variance (PoV) explained by the first principal component. (1 point)
   (c) Find the reconstruction $\hat{\mathbf{x}}$ of point $\mathbf{x} = [3.0 \;\; 2.0]^T$ with the first principal component. (1 point)

t	x_1^t	x_2^t
1	2.0	1.0
2	1.0	7.0
3	4.0	4.0

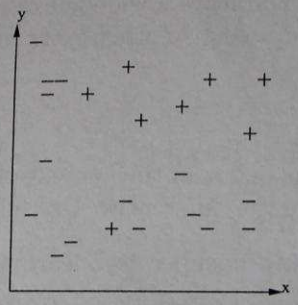
4. Classification tree

(a) What is classification tree? Define it. (1 point)

(b) Sketch the running of the vanilla ID3 algorithm with a toy data set in the figure below (binary classification task in $\mathbb{R}^2$). (4 points)

(c) How to avoid overfitting in the vanilla ID3 algorithm? (1 point)

6



5. Combining classifiers

(a) Describe the rationale of combining multiple base classifiers. (1 point)

(b) What is an important requirement of the base classifiers? (1 point)

(c) Give at least four ways to generate the base classifiers that satisfy the above requirement. (2 points)

(d) Describe the voting scheme for combining classifiers. (1 point)

(e) How is the voting scheme connected to the Bayesian framework? (1 point)

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

28 October 2011.

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 in two pages. Each problem is worth 6 points. 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.

You can keep this paper.

1. Write about the terms below in the context of the course, e.g. what is in common and what are the differences. Use full sentences and give examples.
   (a) generative learning–discriminative learning (2 points)
   (b) parametric methods–nonparametric methods (2 points)
   (c) classification–clustering (2 points)
2. Consider a Bayesian network that has three binary variables M (trip to Mexico), S (swine flu), and F (fever). The joint distribution is $P(M, S, F) = P(M)P(S \mid M)P(F \mid S)$ 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 $P(F = 1 \mid S = 1) = 0.9$.
   (a) Draw the graphical representation of the Bayesian network. (3 points)
   (b) Compute $P(M = 1 \mid F = 1)$, that is, the probability that one has been to Mexico if we know that she have fever. (3 points)
3. Consider a parametric regression problem
   (a) Write a pseudocode function to choose a regression model among $M_1, M_2, \ldots, M_8$ and its parameters θ given a data set of 1000 samples $\{\mathbf{x}^t, \mathbf{r}^t\}_{t=1}^{1000}$. You should implement 10-fold cross validation for model selection in your function. You can use abstract auxiliary functions such as one for estimating parameters, but you should describe each with one sentence and carefully list each function's inputs and outputs. (5 points)
   (b) Mention one advantage and one disadvantage of 10-fold cross validation when compared to basic validation. (1 point)
4. 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.) (4 points)

(b) What is the objective of PCA? (1 point)

(c) What is the relationship between the objective and the eigenvalues? (1 point)

5. Do three iterations of the Lloyd's algorithm for K-means clustering on the 2-dimensional data below. Use $K = 2$ clusters and the initial prototype vectors (=mean vectors) $\mathbf{m}_1 = (0.0, 2.0)$ and $\mathbf{m}_2 = (2.0, 0.0)$. Write down calculation procedure and the cluster memberships as well as mean vectors after each iteration. Draw the data points, cluster means and cluster boundary after each iteration. (6 points)

t	$\mathbf{x}^t$
1	(0.0,1.0)
2	(1.0,2.0)
3	(4.0,5.0)
4	(5.0,3.0)
5	(5.0,4.0)

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

14 December 2010.

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

This examination has five problems each worth 6 points, 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.

The results will be announced in Noppa on 14 January, at latest.

You can keep this paper.

1. Write about the terms below in the context of the course, e.g. what is in common and what are the differences. Use full sentences and give examples.
   (a) hypothesis space–version space (2 points)
   (b) generative learning–discriminative learning (2 points)
   (c) parametric methods–nonparametric methods (2 points)
2. Consider a classification of real-valued numbers $x \in \mathbb{R}$ into two classes $C \in \{1, 2\}$ using a model with $p(x \mid C) = N(C, 1)$ and $P(C = 1) = 2/3$. The utility of a correct classification is zero, classifying a sample with class 1 into class 2 has utility -2, and class 2 into class 1 has utility -4. There is also a "don't know" option whose utility is -1 regardless of the true class. What is the optimal decision for each x? (Hint: If you are not sure about your answer, it is a good idea to draw a figure.)
3. Consider the feature selection in a nonparametric classification problem.
   (a) What is feature selection and why is it useful (at least two reasons)? (1.5 points)
   (b) Explain, also using pseudocode, how you would implement forward and backward selection of features in a real world application. (3 points)
   (c) What can you say about time complexity and the optimality of the solutions produced by the forward and backward selection methods? (1.5 points)

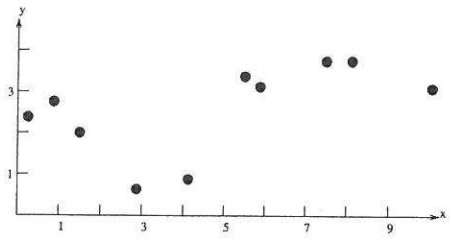


Figure 1: Toy data set for problem 5.

4. Consider the problem of clustering N real valued data vectors into k clusters using the Lloyd's algorithm, also known as the k-means algorithm.

   (a) Write down the Lloyd's algorithm in pseudocode. Pay attention to clearly marking the inputs and outputs of each function. Include an initialization in your algorithm. (4.5 points)

   (b) What can you say about the convergence and solutions found by the Lloyd's algorithm? How could you take this into account in practical data analysis? (1.5 point)

5. *Regression trees.*

   (a) Describe the ID3 algorithm for regression trees by using pseudocode. What is the cost function that the algorithm is optimizing? (3 points)

   (b) Explain pruning in this context. Why and when is the pruning necessary? (1.5 point)

   (c) Sketch the running of the ID3 algorithm with a toy data set of Figure 1 (regression task of predicting y given x). (1.5 points)

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

15 December 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 15 January 2009, at latest.

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.
   (a) supervised learning–unsupervised learning
   (b) feature extraction–feature selection
   (c) generative learning–discriminative learning
   (d) Akaike Information Criterion (AIC)–Bayesian Information Criterion (BIC)
   (e) classification–clustering
   (f) validation data–testing data
2. (a) Consider a regression problem, where you are trying to predict r based on x using some regressor $g(x)$. The expected generalization error at x is $E[(r - g(x))^2 \mid x]$ over the joint distribution of unseen r and x. The error can be divided into different parts. Name and give an example of three conceptually different sources of error. What can you do to minimize each type of error?
   (b) Consider the Bayesian network and the data set in Figure 1. Write the joint distribution $P(A, B)$ and compute the maximum likelihood estimates of the model parameters.

t	A^t	B^t
1	0	1
2	0	0
3	1	0
4	1	0
5	0	1

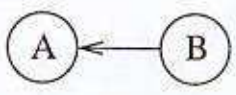


Figure 1: The data set and the Bayesian network structure for Problem 2 (b).

3. *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 two different 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) estimate, the maximum a posteriori (MAP) estimate and the Bayes estimate and compare their properties.

4. *Bayesian multivariate classification.* Consider the problem of classifying real vectors into two classes using a Bayesian classifier 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 reconstruct the data vectors from the principal components? Give an equation.

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

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

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

2 September 2009 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 2 October 2008, at latest.

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 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. *Bias and variance of an estimator.*

(a) Define bias and variance of an estimator.

(b) What is unbiased estimator?

(c) Compute the bias of an estimator of variance, given by $s^2 = \sum_{t=1}^{N} (x^t - m)^2/N$, where $m = \sum_{t=1}^{N} x^t/N$, where the data is given by N real numbers x^t, $t \in \{1, \ldots, N\}$.

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. *Feature selection.* Consider the feature selection in regression problems.

   (a) What is feature selection and why it is needed?

   (b) Assume that you have a regression problem (for example, such as in Problem 1). Explain, also using pseudocode, how you would implement forward and backward selection of features in a real world application.

   (c) What can you say about time complexity and the optimality of the solutions produced by the forward and backward selection methods?

6. *Classification trees.*

   (a) What is 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 $\mathbb{R}^2$).

Figure 1: Toy data set for problem 6.

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.

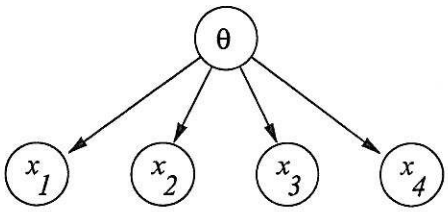


Figure 1: Bayesian network for problem 2.

(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.

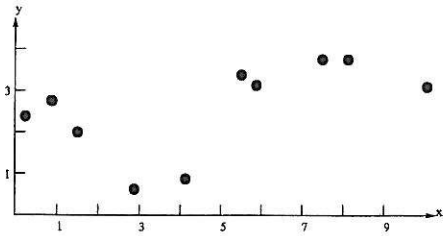


Figure 2: Toy data set for problem 6.

(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?

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.

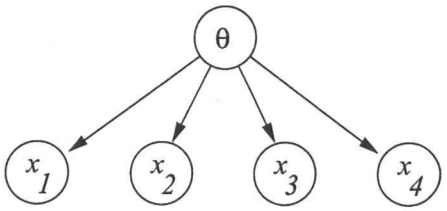


Figure 1: Bayesian network for problem 2.

(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.

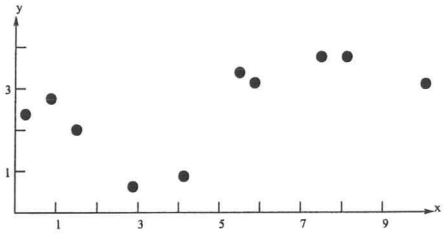


Figure 2: Toy data set for problem 6.

(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?

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES

EXAMINATION 24 October 2014

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 each worth 10 points, and three pages. You must answer in 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.

The results will be announced in Noppa. You can keep this paper.

Problem 1: Multiple-choice questions (10 points total).

The following five questions each have different proposed answers, in each question only one answer is correct. For each question, give both your answer (one of A,B,C,...) and your confidence ("High" or "Low"). For example, "A, Low" is a proper way to answer a question. Each question is graded as follows:

- +2 if the answer is correct and confidence is High
- +1 if the answer is correct and confidence is Low
- 0 if the answer is missing
- −1 if the answer is wrong and confidence is Low
- −2 if the answer is wrong and confidence is High.

The total score for the multiple-choice questions is between 0 and 10 (you cannot get a negative total score).

1) Assume the covariance matrix of your dataset $\mathcal{X}$ is Σ and while doing Principal Component Analysis, you found $\Sigma = \mathbf{CDC}^T$. Here $\mathbf{D}$ is the following diagonal matrix:

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2.1 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & .9 \end{pmatrix}$$

Then the proportion of variance (POV) explained by the first 2 principal components would be (accurate until 2 decimal places):

A) .85

B) .68

C) .31

D) 1

E) None of the above answers is correct.

2) For a dataset distributed as i.i.d $\mathcal{N}(\mu, \sigma)$, after computing the posterior distribution using a normal prior for μ you found the posterior distribution of μ to be $\mathcal{N}(5.24, 4.62)$. Then the MAP-estimate and Bayes-estimate of μ would be:

A) MAP-estimate = 5, Bayes-estimate = 4;

B) MAP-estimate = 2.62, Bayes-estimate = 2.62;

C) MAP-estimate = 5.24, Bayes-estimate = 5.24;

D) MAP-estimate = 5.24, Bayes-estimate = 2.62;

E) None of the above answers is correct.

3) EM-algorithm is a type of fuzzy clustering as:

A) The clusters we get are fuzzy-sets;

B) Each observations are assigned to only one of the K-clusters;

C) For each observation there are different probabilities of it being assigned to K different clusters;

D) The algorithm uses fuzzy-methods;

E) None of the above answers is correct.

4) Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be written in a general form as $\mathcal{L}_{ML}(\theta \mid \mathcal{X})$-regularizer term. The regularizer term of BIC can be expressed as:

A) $\frac{1}{2}N \times$(regularizer term of AIC);

B) $\frac{1}{2}logN \times$(regularizer term of AIC);

C) $\frac{1}{N} \times$(regularizer term of AIC);

D) $\frac{1}{N}N \times$(regularizer term of AIC).

5) In a classification problem, you have two classes 0 and 1. You observed a new data point x_{new}, whose likelihood of belonging to Class 0 and Class 1 are correspondingly .32 and .41. Your prior beliefs (probabilities) for the Class 0 and Class 1 are correspondingly 0.6 and 0.4. Then, following Bayesian decision theory, x_{new} belongs to:

A) Not enough information given to use Bayesian Decision Theory;

B) Class 0;

C) Class 1;

D) It can belong to both Class 0 and 1 with equal probability.

Problem 2: Explanations of concepts (10 points total).

Explain the terms below in the context of the course. If two terms are given, explain them so that it becomes clear what they have in common and what are the differences. Use full sentences.

1. Forward Search, in context of feature selection (2 points)

2. naive estimator—kernel estimator, in context of density estimation (2 points)

3. regression—classification (2 points)

4. Bias/Variance Dilemma (2 points)

5. parametric methods—nonparametric methods (2 points).

Problem 3: Bayesian Decision Theory and Parametric Methods (10 points total).

a) A company has to decide whether to accept or reject a lot of incoming parts. (Label these actions a_1 and a_2 respectively.) The lots are of three types: θ_1 (very good), θ_2 (acceptable), and θ_3 (bad). The loss $L(\theta_i, a_j)$ incurred in making the decision is given in the following table. The prior belief is that $\pi(\theta_1) = \pi(\theta_2) = \pi(\theta_3) = \frac{1}{3}$.

	a_1	a_2
θ_1	0	3
θ_2	1	2
θ_3	3	0

- What is the definition of expected utility - explain with equations? (1 points)
- What is the expected utility of action a_1 for the company? (2 points)
- Find the optimal decision for the company. (2 points)

b) A smart-phone producing company periodically samples smart-phones coming off a production line, in order to make sure the production process is running smoothly. They choose a sample of size 5 and observe the number of defectives (X). They assume number of defectives (X) is distributed as $Binomial(5, \theta)$. Past records show that the proportion of defectives θ varies according to a $Beta(1, 9)$ distribution.

[Hints: (a) $Binomial(x|n, \theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{(n-x)}$; (b) $Beta(\theta|\alpha, \beta) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha,\beta)}$; (c) $B(\alpha, \beta) = \int_0^1 x^{(\alpha-1)}(1-x)^{(\beta-1)}$]

- If they have observed $X = 2$ defective smart-phones in one inspection run, explain how we derive the posterior distribution of θ? (3 points) [Hint: The posterior distribution would be $Beta(\alpha + 2, \beta + (5 - 2))$]
- What would be the Bayes estimate of θ using the posterior distribution dervied in last step? [Hint: You know the mean of $Beta(\alpha, \beta)$ distribution is $\frac{\alpha}{\alpha+\beta}$] (2 points).

Problem 4: Clustering (10 points total).

- What can you say about the convergence and solutions found by the Lloyd's algorithm? How could you take this into account in practical data analysis? (3 points)
- In EM-algorithm if the probability of assignment of the i-th observation to one of the K clusters ($\{h_i^k\}_{k=1,\ldots,K}$) becomes 1 for all observations (e.g. for $i = 8, h_8^3 = 1$, for $i = 5, h_5^2 = 1$ and so on), then show that the EM-algorithm becomes Lloyd's algorithm for K-means clustering. (3 points)
- Do three iterations of Lloyd's algorithm for K-means clustering on the 2-dimensional data below. Use $K = 2$ clusters and the initial prototype vectors (=mean vectors) $m_1 = (0.0, 2.0)$ and $m_2 = (2.0, 0.0)$. Write down calculation procedure and the cluster memberships as well as mean vectors after each iteration. Draw the data points, cluster means and cluster boundary after each iteration.

t	x^t
1	(0.0, 1.0)
2	(1.0, 2.0)
3	(4.0. 5.0)
4	(5.0, 3.0)
5	(5.0, 4.0)

(4 points).

Problem 5: Principal Component Analysis (10 points total).

You have a dataset of N two-dimensional points $\mathbf{y}^t$. You want to perform Principal Component Analysis (PCA) on the dataset. You have already estimated that the data is zero-mean and has the covariance matrix

$$\mathbf{S} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

and you know the covariance matrix can be diagonalized as $\mathbf{C}^\top\mathbf{S}\mathbf{C} = \mathbf{D}$ where

$$\mathbf{C} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix}.$$

Perform the following tasks:

a) Explain how the matrices $\mathbf{C}$ and $\mathbf{D}$ are related to Principal Component Analysis. (2 points)
b) In 2-dimensional space, plot the PCA coordinates (directions of the largest and second-largest variance of the data) (3 points)
c) Compute the proportion of variance explained by the first principal component. (2 points)
d) Define the principal components of the data by $\mathbf{z}^t = \mathbf{C}^T\mathbf{y}^t$. What is the covariance matrix of $\mathbf{z}^t$? (3 points).

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.

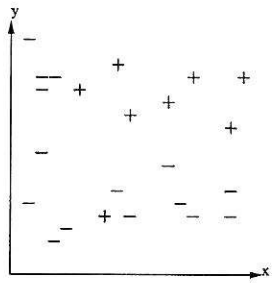


Figure 1: Toy data set for problem 5.

(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?

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES

EXAMINATION

B, 19 December 2007 at 16–19.

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 posted to the (blue binder at the) notice board on 19 January 2008, at latest, and also emailed to an address of form `12345X@students.hut.fi`, where `12345X` is your student number.

Please fill the course feedback form (open until 7 January 2008) at `http://www.cs.hut.fi/Opinnot/Palaute/kurssipalaute.html`

You can keep this paper.

1. *Model selection.* Assume that you have at your disposal a data set $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t=1}^N$, where r^t is a class and $\mathbf{x}^t$ is a covariate; and a set of k black box classification algorithms A_i, $i \in \{1, \ldots, k\}$, which try to predict the class r, given the covariate $\mathbf{x}$ and the training data. More formally, you can think A_i as a known arbitrary function $r_{PREDICTED} = A_i(\mathbf{x}, \mathcal{X}_{TRAIN})$, where $r_{PREDICTED}$ is the predicted class, given $\mathbf{x}$, and $\mathcal{X}_{TRAIN}$ is the data used to train the classifier. Your task is to choose and train the classification algorithm that would work best for yet unseen data. Describe, in detail, different ways how you could accomplish this (and why). How do you expect the various classification errors to behave?
2. *Bayesian networks.*

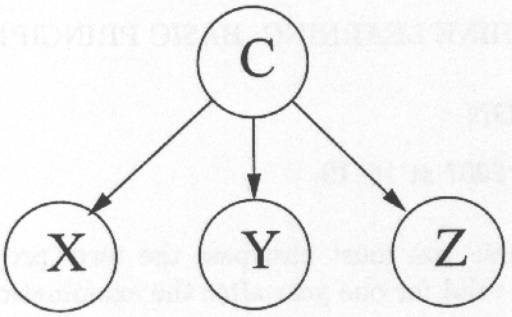


Figure 1: Bayesian network for problem 2.

(a) Define the concept of Bayesian network.

(b) Find an expression for probability $P(C \mid X, Y, Z)$, given the network in Figure 1, where C, X, Y and Z are binary random variables.

(c) What is the type of a classifier defined by the item (2b) above commonly called?

3. *Bayesian probability theory.* Consider the problem of finding the probability that a coin flip gives "heads" (assume that the probability of "heads" can also be something else than $\frac{1}{2}$ of a fair coin).

   (a) Using the concepts of prior and posterior probability density, describe (using formulae and figures) how you could find this probability after observing N coin flips for various choices of prior probability density.

   (b) Define the maximum likelihood (ML) and maximum a posteriori (MAP) estimates. What would ML and MAP estimates be in your coin flipping example?

4. *Bias and variance dilemma.* Explain the bias and variance dilemma, with the relevant formulae, in the context of linear regression.

5. *Linear discriminant analysis.*

   (a) Define the concept of linear discriminant analysis (LDA), and derive the formulae for the case of two classes.

   (b) What is the main difference between principal component analysis (PCA) and LDA? Demonstrate this difference by sketching

out how PCA and LDA would work with a toy data set of your choosing.

6. *Clustering.* Consider the problem of clustering N real valued data vectors into k clusters using the Lloyd's algorithm, also known as the k-means algorithm.

   (a) Write the Lloyd's algorithm in pseudocode.

   (b) What can you say about the convergence and solutions found by the Lloyd's algorithm? How could you take this into account in practical data analysis?

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

26 May 2014.

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 each worth 10 points, and three pages. You must answer in 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.

The results will be announced in Noppa. You can keep this paper.

Problem 1: Explanations of concepts (10 points total)

Explain the terms below in the context of the course. If two terms are given, explain them so that is becomes clear what they have in common and what are the differences. Use full sentences.

1. consistent hypothesis—version space (2 points)
2. receiver operating characteristics (ROC) curve (2 points)
3. Bayesian model averaging (2 points)
4. decision tree (2 points)
5. generalization error of a classifier (2 points)

Problem 2: Model Selection (10 points total)

- Explain what is the K-fold Cross Validation Scheme and write down one advantage and disadvantage of it. (2 points)
- Consider a parametric regression (Bayesian regression) scenario where we try to regress target values r from one-dimensional inputs x using the regressor $r \approx g(x|\theta) = a_0 + a_1 x + a_2 x^2$, where the parameters are $\theta = (a_0, a_1, a_2)$. We assume Gaussian noise $p(r|x, \theta) \sim N(g(x|\theta), \sigma^2)$ where σ^2 is the noise variance which is a known constant.

  Assume we have observed N input points x_i and their corresponding targets r_i. Write down the likelihood function of the model. (3 points)
- Now suppose that we will also assume a Gaussian prior for the parameters, $a_0 \sim N(0, 1/\lambda)$, $a_1 \sim N(0, 1/\lambda)$, $a_2 \sim N(0, 1/\lambda)$, where $1/\lambda$ is the prior variance of the parameters. Write down the equation you need to maximize in order to find the maximum a posteriori (MAP) estimate of the parameters. (3 points)
- Explain the role of the constant λ in the prior distribution, and how it causes regularization of the learned parameters θ. (2 points)

Problem 5: Nonparametric Density Estimation (10 points total)

a) In the context of probability density estimation, explain the naive estimator and give its mathematical definition. (2 points)

b) In the context of density estimation, explain the kernel estimator and give its mathematical definition. (2 points)

c) You have observed seven one-dimensional data points whose coordinates are shown in the figure below. Use the naive estimator with bin width $h = 4$ to estimate the probability density over the interval $[-6, 6]$. Draw the density function. (3 points)

d) Using the same naive estimator as in part c), give the numerical value of the probability density at the locations -4.25, -0.75, and 1.5. (3 points)

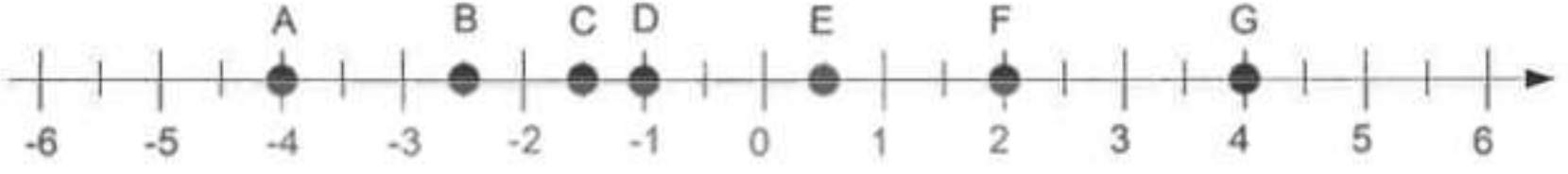


Figure 1: Observed data points for problem 5 d). For convenience the data points are labeled A-G.

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

25 October 2013.

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 each worth 10 points, and two pages. You must answer in 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.

The results will be announced in Noppa. You can keep this paper.

Problem 1: Explanations of concepts (10 points total)

Explain the terms below in the context of the course. If two terms are given, explain them so that is becomes clear what they have in common and what are the differences. Use full sentences.

1. consistent hypothesis—version space (2 points)
2. overfitting—underfitting (2 points)
3. histogram estimator—naive estimator, in context of density estimation (2 points)
4. expected utility in classification (2 points)
5. k-fold cross-validation (2 points)

Problem 2: Bayesian Decision Theory and Parametric Methods (10 points).

a) Consider $X_1, X_2, \ldots, X_n$ are i.i.d. observations from a model $P(X|\theta)$ with unknown parameter θ. If you want to estimate θ following Bayes Theorem, answer the following first:

- What do the concepts prior, likelihood and posterior mean in the above problem? Write with mathematical notation. You do not have to define specific functions for the prior, likelihood, and posterior, just explain what the concepts are. (2 points)
- How can you compute the posterior if you know the prior and likelihood? (2 points)
- If you know the posterior density of θ, how can you compute the Bayes estimate of θ? You do not have to perform the computation, just explain how it would be done. (2 points)

b) Suppose in a), the observations $X_1, X_2, \ldots, X_n$ are light bulbs where each bulb X_i is either working ($X_i = 0$) or broken ($X_i = 1$), and we have observed $n = 10000$ bulbs. We assume the model $P(X|\theta)$ is a Bernoulli process, where the parameter θ, $0 < \theta < 1$, is the probability for a bulb being broken. Then answer the following:

- We want to use a flat prior (also known as a uniform prior) for θ. Write the equation of the prior. (1 point)
- Write the expression of the posterior density of θ. (3 points)

Problem 3: Clustering (10 points).

- Give one example of Hard and Fuzzy Clustering (also called Soft Clustering). Explain the differences between these two types of clustering. (2 points)
- Suppose you are performing an iteration of K-means clustering, and you know the set of K cluster means m_i. What would be the error function that you need to minimize to assign observations to the clusters? (2 points)
- Write the Lloyd's algorithm in pseudocode. (4 points)
- Does the solution of K-means depend on the initial location of the cluster means? If yes, how can you try to get better solutions? If not, why not? (2 points)

Problem 4: Principal Component Analysis (10 points total)

You have a data set of the following five two-dimensional points:

$$\mathcal{X} = \left\{ \begin{bmatrix} -5 \\ -3 \end{bmatrix}, \begin{bmatrix} 0 \\ -4 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 5 \\ 3 \end{bmatrix} \right\}$$

You want to reduce the dimensionality of the data points to one, using Principal Component Analysis. You have already estimated that the data is zero-mean and has a covariance matrix of

$$\mathbf{S} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

and you know the covariance matrix can be diagonalized as $\mathbf{C}^\top \mathbf{S}\mathbf{C} = \mathbf{D}$ where

$$\mathbf{C} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \text{ and } \mathbf{D} = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix}.$$

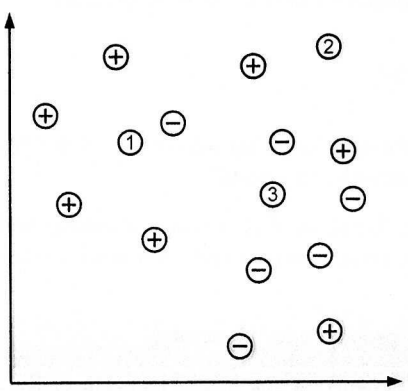
Perform the following tasks.

a) Explain how the matrices C and D are related to Principal Component Analysis. (1 points)
b) Reduce the dimensionality of the data to one, by computing the projections of the five data points onto the first principal component. It is enough to do the computation for the first two data points. (3 points)
c) Compute the proportion of variance explained by the first principal component. (2 points)
d) Reconstruct the original data points approximately, by projecting the coordinates computed in step a) back into the original space. It is enough to do the computation for the first two data points. (2 points)
e) Compute the reconstruction error. If you reconstructed just the first two data points in step c), it is acceptable to use only those two points in this step. (2 points)

Problem 5: Nonparametric Classification (10 points).

You have acquired the training data shown in the scatter plot below, where circles are locations of data points, '+' signs are data from the positive class and '-' signs are data from the negative class. You also have three validation points marked as 1, 2, and 3 in the scatter plot. You know that validation point 1 comes from the positive class and validation points 2 and 3 come from the negative class.

a) Explain the principle of k-nearest neighbor classification. Write the necessary equations for the case $k = 1$. (3 points)
b) Classify the three validation points based on the training set, using k-nearest neighbor classification with $k = 1$. (2 points)
c) Classify the three validation points based on the training set, using k-nearest neighbor classification with $k = 5$. (3 points)
d) Compute the classification errors on the validation set, and choose the best complexity for the classifier (choose $k = 1$ or $k = 5$). (2 points)



T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

25 October 2013.

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 each worth 10 points, and two pages. You must answer in 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.

The results will be announced in Noppa. You can keep this paper.

Problem 1: Explanations of concepts (10 points total)

Explain the terms below in the context of the course. If two terms are given, explain them so that is becomes clear what they have in common and what are the differences. Use full sentences.

1. consistent hypothesis—version space (2 points)
2. overfitting—underfitting (2 points)
3. histogram estimator—naive estimator, in context of density estimation (2 points)
4. expected utility in classification (2 points)
5. k-fold cross-validation (2 points)

Problem 2: Bayesian Decision Theory and Parametric Methods (10 points).

a) Consider $X_1, X_2, \ldots, X_n$ are i.i.d. observations from a model $P(X|\theta)$ with unknown parameter θ. If you want to estimate θ following Bayes Theorem, answer the following first:

- What do the concepts prior, likelihood and posterior mean in the above problem? Write with mathematical notation. You do not have to define specific functions for the prior, likelihood, and posterior, just explain what the concepts are. (2 points)
- How can you compute the posterior if you know the prior and likelihood? (2 points)
- If you know the posterior density of θ, how can you compute the Bayes estimate of θ? You do not have to perform the computation, just explain how it would be done. (2 points)

b) Suppose in a), the observations $X_1, X_2, \ldots, X_n$ are light bulbs where each bulb X_i is either working ($X_i = 0$) or broken ($X_i = 1$), and we have observed $n = 10000$ bulbs. We assume the model $P(X|\theta)$ is a Bernoulli process, where the parameter θ, $0 < \theta < 1$, is the probability for a bulb being broken. Then answer the following:

- We want to use a flat prior (also known as a uniform prior) for θ. Write the equation of the prior. (1 point)
- Write the expression of the posterior density of θ. (3 points)

Problem 3: Clustering (10 points).

- Give one example of Hard and Fuzzy Clustering (also called Soft Clustering). Explain the differences between these two types of clustering. (2 points)
- Suppose you are performing an iteration of K-means clustering, and you know the set of K cluster means m_i. What would be the error function that you need to minimize to assign observations to the clusters? (2 points)
- Write the Lloyd's algorithm in pseudocode. (4 points)
- Does the solution of K-means depend on the initial location of the cluster means? If yes, how can you try to get better solutions? If not, why not? (2 points)

Problem 4: Principal Component Analysis (10 points total)

You have a data set of the following five two-dimensional points:

$$\mathcal{X} = \left\{ \begin{bmatrix} -5 \\ -3 \end{bmatrix}, \begin{bmatrix} 0 \\ -4 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 5 \\ 3 \end{bmatrix} \right\}$$

You want to reduce the dimensionality of the data points to one, using Principal Component Analysis. You have already estimated that the data is zero-mean and has a covariance matrix of

$$\mathbf{S} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

and you know the covariance matrix can be diagonalized as $\mathbf{C}^\top \mathbf{S}\mathbf{C} = \mathbf{D}$ where

$$\mathbf{C} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \text{ and } \mathbf{D} = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix}.$$

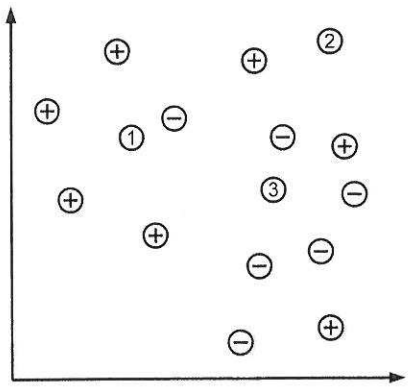
Perform the following tasks.

a) Explain how the matrices C and D are related to Principal Component Analysis. (1 points)
b) Reduce the dimensionality of the data to one, by computing the projections of the five data points onto the first principal component. It is enough to do the computation for the first two data points. (3 points)
c) Compute the proportion of variance explained by the first principal component. (2 points)
d) Reconstruct the original data points approximately, by projecting the coordinates computed in step a) back into the original space. It is enough to do the computation for the first two data points. (2 points)
e) Compute the reconstruction error. If you reconstructed just the first two data points in step c), it is acceptable to use only those two points in this step. (2 points)

Problem 5: Nonparametric Classification (10 points).

You have acquired the training data shown in the scatter plot below, where circles are locations of data points, '+' signs are data from the positive class and '-' signs are data from the negative class. You also have three validation points marked as 1, 2, and 3 in the scatter plot. You know that validation point 1 comes from the positive class and validation points 2 and 3 come from the negative class.

a) Explain the principle of k-nearest neighbor classification. Write the necessary equations for the case $k = 1$. (3 points)
b) Classify the three validation points based on the training set, using k-nearest neighbor classification with $k = 1$. (2 points)
c) Classify the three validation points based on the training set, using k-nearest neighbor classification with $k = 5$. (3 points)
d) Compute the classification errors on the validation set, and choose the best complexity for the classifier (choose $k = 1$ or $k = 5$). (2 points)



T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

27 April 2013.

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

This examination has five problems each worth 10 points, and 4 pages. You have to answer in 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.

The results will be announced in Noppa on May 26, at latest.

You can keep this paper.

1. Multiple choices questions (10 points). The following five questions have different proposed answers. Only one of them is correct. You have to give your answer along with your confidence ("High" or "Low") for each answer. Grading for each of these questions is then:

   - +2 if the answer is correct and confidence High
   - +1 if the answer is correct and confidence Low
   - 0 if the answer is missing
   - −1 if the answer is wrong and confidence Low
   - −2 if the answer is wrong and confidence High

   Write on your answer sheet the correct answer A, B, C, D,...) along with the confidence you have (High or Low) for that question. For example, "A, Low" is a proper way of answering a question. No need to justify your answers. Total score for this question is between 0 and 10 (you cannot get a negative score for the whole question).

   1) For a binary classification problem, each class is modeled using a Multivariate Normal (Gaussian) Distribution. A Bayes classifier is calculated.
      A) The boundary is always linear.
      B) The boundary is nonlinear (not purely linear).
      C) The boundary is independent from the priors of the classes.
      D) The boundary can always separate the classes perfectly (for the training set).
      E) None of the previous answers is correct
   2) For a multidimensional dataset, a Principal Component Analysis (PCA) is performed.
      A) The average reconstruction error is always increasing with the dimension of projection.
      B) The projection is independent from the variances of the input variables.

C) The average reconstruction error is never increasing with the dimension of projection.

D) The projection dimension has to be larger than the number of points and the number of variables (samples).

E) None of the previous answers is correct

3) The Lloyd's algorithm is used to perform clustering.

A) This algorithm will never converge and has to be stopped after an arbitrary number of iterations.

B) The error function which is minimized can increase for some iterations but is globally decreasing.

C) The Lloyd's algorithm will always converge to the best clustering solution.

D) The Lloyd's algorithm is dependent from the initialization.

E) None of the previous answers is correct

4) For a binary classification problem, a K-Nearest-Neighbor (KNN) Classifier is built.

A) The classification error is always decreasing with respect to the parameter K.

B) The best value for K is always 3.

C) The parameter K cannot be optimized using validation.

D) The performances of the KNN classifier are independent from the distance metric which is used.

E) None of the previous answers is correct

5) A k-fold cross-validation is used to determine the optimal complexity of a regression model.

A) The cross-validation error is a perfect estimate of the generalization performances of the regression model.

B) The best value for k is always 2.

C) The best value for k is always 10.

D) The complexity selected by the k-fold cross-validation is always larger than the complexity selected using a Bayesian Information Criterion (BIC) regularization.

E) None of the previous answers is correct

2. *Model selection.* Assume that you have at your disposal a data set $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t=1}^N$, where r^t is a class and $\mathbf{x}^t$ is a covariate; and a set of k black box classification algorithms A_i, $i \in \{1, \ldots, k\}$, which try to predict the class r, given the covariate $\mathbf{x}$ and the training data. More formally, you can think A_i as a known arbitrary function $r_{PREDICTED} = A_i(\mathbf{x}, \mathcal{X}_{TRAIN})$, where $r_{PREDICTED}$ is the predicted class, given $\mathbf{x}$, and $\mathcal{X}_{TRAIN}$ is the data used to train the classifier. Your task is to choose and train the classification algorithm that would work best for yet unseen data. Describe, in detail, different ways how you could accomplish this (and why). How do you expect the various classification errors to behave? (10 points)

3. (a) *Maximum Likelihood* (4 points). Consider a univariate data set $\mathcal{X} = (x^1, x^2, \ldots, x^N)$ that has a *log-normal* distribution. Find the maximum likelihood estimates of the mean μ and variance σ^2. The probability density function is given by

$$p(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

(b) *Naïve Bayes* (6 points). Consider binary classification for multivariate data $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t\in\{1,\ldots,N\}}$, where $r^t \in \{0,1\}$ and $\mathbf{x}^t \in \mathbb{R}^d$. Assume that

- r is Bernoulli distributed with $P(r = 1) = \pi$.
- Variable x_i, $i = 1, \ldots, d$ is continuous and normally distributed with $P(x_i|r = k) = \mathcal{N}(\mu_{ik}, \sigma_i^2)$. The variance σ_i^2 is class independent!
- All variables are independent of each other given the class label r (Naïve Bayes assumption).

Show that the posterior distribution $P(r = 1|\mathbf{x})$ can be written in logistic form, i.e.

$$P(r = 1|\mathbf{x}) = \frac{1}{1 + \exp(w_0 + \sum_{j=1}^{d} w_j x_j)}.$$

and write down the expressions for w_0 and w_j, $j = 1, \ldots, d$.

4. *Feature selection.* Consider the feature selection in classification problems.

(a) What is feature selection and why is it needed? (4 points)

(b) Assume that you have a binary classification algorithm. Explain, also using pseudocode, how you would implement forward and backward selection of features (in a real world application). (4 points)

(c) What can you say about time complexity and the optimality of the solutions produced by the forward and backward selection methods? (2 points)

5. Consider the problem of clustering N real valued data vectors into k clusters using the Lloyd's algorithm, also known as the k-means algorithm.

   (a) Write down the Lloyd's algorithm in pseudocode. Pay attention to clearly marking the inputs and outputs of each function. Include an initialization in your algorithm. (6 points)

   (b) What can you say about the convergence and solutions found by the Lloyd's algorithm? How could you take this into account in practical data analysis? (4 points)

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

26 October 2012.

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

This examination has five problems each worth 10 points, and 4 pages. You have to answer in 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.

The results will be announced in Noppa on November 25, at latest.

You can keep this paper.

1. Multiple choices questions (10 points). The following five questions have different proposed answers. Only one of them is correct. You have to give your answer along with your confidence ("High" or "Low") for each answer. Grading for each of these questions is then:

   - +2 if the answer is correct and confidence High
   - +1 if the answer is correct and confidence Low
   - 0 if the answer is missing
   - −1 if the answer is wrong and confidence Low
   - −2 if the answer is wrong and confidence High

   Write on your answer sheet the correct answer A, B, C, D,...) along with the confidence you have (High or Low) for that question. For example, "A, Low" is a proper way of answering a question. No need to justify your answers. Total score for this question is between 0 and 10 (you cannot get a negative score for the whole question).

   1) For a binary classification problem, each class is modeled using a Multivariate Normal (Gaussian) Distribution. A Bayes classifier is calculated.
      A) The boundary is always linear.
      B) The boundary is always nonlinear.
      C) The boundary is independent from the priors of the classes.
      D) The boundary can never separate the classes perfectly (for the training set).
      E) None of the previous answers is correct
   2) For a multidimensional dataset, a Principal Component Analysis (PCA) is performed.
      A) The average reconstruction error is never increasing with the dimension of projection.
      B) The projection is independent from the variances of the input variables.

C) The average reconstruction error is always increasing with the dimension of projection.

D) The projection dimension has to be larger than the number of points and the number of variables (samples).

E) None of the previous answers is correct

3) The Lloyd's algorithm is used to perform clustering.

A) This algorithm will never converge and has to be stopped after an arbitrary number of iterations.

B) The error function which is minimized can increase for some iterations but is globally decreasing.

C) The Lloyd's algorithm will always converge to the best clustering solution.

D) The Lloyd's algorithm is independent from the initialization.

E) None of the previous answers is correct

4) For a binary classification problem, a K-Nearest-Neighbor (KNN) Classifier is built.

A) The classification error is always decreasing with respect to the parameter K.

B) The best value for K is always 1.

C) The parameter K can be optimized using validation.

D) The performances of the KNN classifier are independent from the distance metric which is used.

E) None of the previous answers is correct

5) A k-fold cross-validation is used to determine the optimal complexity of a regression model.

A) The cross-validation error is a perfect estimate of the generalization performances of the regression model.

B) The best value for k is always 2.

C) The best value for k is always 10.

D) The complexity selected by the k-fold cross-validation is always larger than the complexity selected using a Bayesian Information Criterion (BIC) regularization.

E) None of the previous answers is correct

2. *Model selection.* Assume that you have at your disposal a data set $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t=1}^N$, where r^t is a class and $\mathbf{x}^t$ is a covariate; and a set of k black box classification algorithms A_i, $i \in \{1, \ldots, k\}$, which try to predict the class r, given the covariate $\mathbf{x}$ and the training data. More formally, you can think A_i as a known arbitrary function $r_{PREDICTED} = A_i(\mathbf{x}, \mathcal{X}_{TRAIN})$, where $r_{PREDICTED}$ is the predicted class, given $\mathbf{x}$, and $\mathcal{X}_{TRAIN}$ is the data used to train the classifier. Your task is to choose and train the classification algorithm that would work best for yet unseen data. Describe, in detail, different ways how you could accomplish this (and why). How do you expect the various classification errors to behave? (10 points)

3. (a) *Maximum Likelihood* (4 points). Consider a univariate data set $\mathcal{X} = (x^1, x^2, \ldots, x^N)$ that has a *log-normal* distribution. Find the maximum likelihood estimates of the mean μ and variance σ^2. The probability density function is given by

$$p(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

(b) *Naïve Bayes* (6 points). Consider binary classification for multivariate data $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t\in\{1,\ldots,N\}}$, where $r^t \in \{0, 1\}$ and $\mathbf{x}^t \in \mathbb{R}^d$. Assume that

- r is Bernoulli distributed with $P(r = 1) = \pi$.
- Variable x_i, $i = 1, \ldots, d$ is continuous and normally distributed with $P(x_i|r = k) = \mathcal{N}(\mu_{ik}, \sigma_i^2)$. The variance σ_i^2 is class independent!
- All variables are independent of each other given the class label r (Naïve Bayes assumption).

Show that the posterior distribution $P(r = 1|\mathbf{x})$ can be written in logistic form, i.e.

$$P(r = 1|\mathbf{x}) = \frac{1}{1 + \exp(w_0 + \sum_{j=1}^{d} w_j x_j)}.$$

and write down the expressions for w_0 and w_j, $j = 1, \ldots, d$.

4. *Feature selection.* Consider the feature selection in classification problems.

(a) What is feature selection and why is it needed? (4 points)

(b) Assume that you have a binary classification algorithm. Explain, also using pseudocode, how you would implement forward and backward selection of features (in a real world application). (4 points)

(c) What can you say about time complexity and the optimality of the solutions produced by the forward and backward selection methods? (2 points)

5. *Combining classifiers* (a) Explain why is it a good idea to teach several different classifiers and use majority voting as the final classification. (2 points) (b) Why does this approach work better if the individual base-learners are as different as possible? (2 points) (c) Give at least four ways to make them different. (4 points) (d) Assuming each base learner gives a correct classification with probability p and the classification errors are independent of each other, what is the probability that a majority vote over L classifiers gives the correct answer? (2 points)

T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

6 August 2012.

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 in two pages. Each problem is worth 6 points. 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.

You can keep this paper.

1. Write about the terms below in the context of the course, e.g. what is in common and what are the differences. Use full sentences in your answer.
   (a) supervised learning–unsupervised learning (2 points)
   (b) classification–clustering (2 points)
   (c) empirical error–generalization error (2 points)
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).
   (a) 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. (2 points)
   (b) 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.) (2 points)
   (c) How could you estimate the prediction error for yet unseen data? (2 points)
3. Principal Component Analysis (PCA)
   (a) Do the PCA learning using the 2-dimensional data set in the table below. Describe the steps of your solution. (4 points)
   (b) Compute the proportion of variance (PoV) explained by the first principal component. (1 point)
   (c) Find the reconstruction $\hat{\mathbf{x}}$ of point $\mathbf{x} = [4.0 \;\; 7.0]^T$ with the first principal component. (1 point)

t	x_1^t	x_2^t
1	2.0	2.0
2	3.0	4.0
3	5.0	6.0

4. Consider the problem of clustering N real valued data vectors into k clusters using the Lloyd's algorithm, also known as the k-means algorithm.

   (a) Write down the Lloyd's algorithm in pseudocode. Pay attention to clearly marking the inputs and outputs of each function. Include an initialization in your algorithm. (4.5 points)

   (b) What can you say about the convergence and solutions found by the Lloyd's algorithm? How could you take this into account in practical data analysis? (1.5 point)

5. Classification tree

   (a) What is classification tree? Define it. (1 point)

   (b) Sketch the running of the vanilla ID3 algorithm with a toy data set in the figure below (binary classification task in $\mathbb{R}^2$). (4 points)

   (c) How to avoid overfitting in the vanilla ID3 algorithm? (1 point)