582631 Introduction to Machine Learning
Separate examination, June 7th 2017
Examiner: Teemu Roos

Answer all the four (4) problems. The maximum score for the exam is 60 points. **Please note** that prior to attending the exam, you should have completed project work as instructed on the course web page. The deadline for the project work was one week *before* the exam.

You are allowed to have a calculator and a "cheat sheet" with you at the exam. The cheat sheet is a two-sided, handwritten, A4 where you can write any information whatsoever.

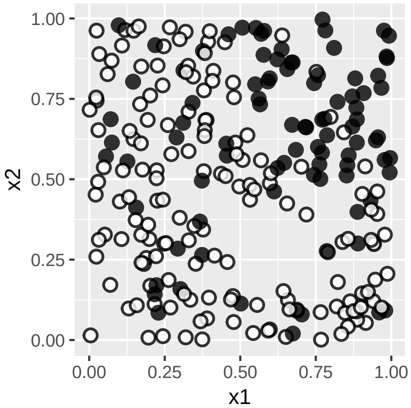
Please write in clear handwriting. You may answer in English, Finnish or Swedish. If you use Finnish or Swedish, it will be helpful to include the English translations to any technical terms that may be ambiguous.

Problem 1. [*15 points*] Explain briefly the following terms and concepts. Your explanation should include, when appropriate, both a precise definition and a brief description of how the concept is useful in machine learning. Your answer to each subproblem should fit to roughly one third of a page of normal handwriting or less.

1. *clustering*
2. *dimension reduction*
3. *test set*
4. *generative classifier*
5. *principal component analysis (PCA)*
6. *decision tree*
7. *logarithmic loss (log-loss)*

Problem 2. [*15 points*] Consider a simple two-dimensional classification problem where the input features x_1 and x_2 are both uniformly distributed between 0.0 and 1.0. The class label $Y \in \{0, 1\}$ takes value 1 (black) with probability $\Pr[Y = 1 \mid x_1, x_2] = 0.8$ if $x_1 > 0.5$ and $x_2 > 0.5$, and with probability 0.2 otherwise.

On the right you can see a sample of $n = 300$ points from the above generating distribution.



1. Give an example of a classification method that would be well-suited for this kind of a task, and explain your choice.

2. Give an example of a classification method that would be less suitable, and again explain why.

3. What is the optimal (Bayes) classification boundary which minimizes the expected zero–one-loss for new test data sampled from the same distribution? Calculate the minimum error probability (i.e., the Bayes error).

Problem 3. [*15 points*] Use the discrete naive Bayes classifier in the following problem. Assume that the class variable Y can take three values $\{0, 1, 2\}$, and that there are two binary features X_1 and X_2.

The training data is as follows:

Y	X_1	X_2
2	0	0
2	0	1
1	0	1
2	0	1
0	1	1
1	0	0
2	0	0

1. Provide estimates of the class-conditional distributions of the form $P(X_j = x \mid Y = c)$ for all j, x, and c. Provide also estimates of the class distribution $P(Y = c)$ for all c. Use no smoothing in the estimators. *Hint:* Use empirical frequencies.

2. Do the same with Laplace smoothing. *Hint:* Add pseudocounts $m_{c,j,x} = 1$ and $m_c = 1$ to the empirical frequencies for all c, j, x.

3. How is the class value predicted for a new test instance (x_1, x_2)? Give a formula.

4. Suppose the test instance is $(1, 0)$. What happens if no smoothing is used?

5. Show that the class that maximizes the posterior probability of the test instance $(1, 0)$ is $Y = 0$ when Laplace smoothing is applied. Apply the smoothing also to the estimation of the class distribution.

Problem 4. [*15 points*] Consider a data set with $n = 100$ observations. Imagine you learn a classification model and find that it classifies 85 of the training examples correctly.

1. What can you say about the performance of your classifier on new test data? Why does the test error typically differ from the training error?

2. What properties of the classification method are most relevant concerning the difference between training and test accuracy?

3. Explain cross-validation.

4. Now suppose that instead of classification, the task would have been to estimate, for example, the median of an unknown distribution from which we have $n = 100$ data points. How would you apply resampling to measure the accuracy of an estimate computed from the given n points?