DATA11002 Introduction to Machine Learning
Separate examination, January 31st 2018
Examiner: Teemu Roos

Answer all the four (4) problems. The maximum score for the exam is 60 points.

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 c l e a r 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.

1. [*12 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.

   (a) *underfitting*

   (b) *interaction (in regression)*

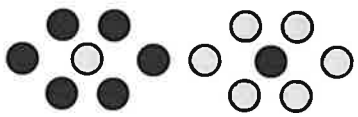
   (c) *model complexity*

   (d) *independent and identically distributed (i.i.d.)*

   (e) *complete linkage (in clustering)*

   (f) *dimension reduction*

2. [*16 points*] Here we will apply the linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) classifiers to a simple two-dimensional binary classification problem. There are two classes (dark gray and light gray) and the training data points are located as shown here:



   Items (a) and (b) are about LDA and QDA generally, and items (c) and (d) are about this particular data set.

   (a) Explain how the two classifiers are learned from data. You can draw diagrams to illustrate your explanation, if you think it is useful.)

   (b) Explain the procedure for classifying new test data points. What is maximized?

   (c) Sketch a diagram showing the classification boundaries of LDA and QDA as well as you can.

   (d) Which one do you think would be more appropriate in this case? Explain.

3. [*16 points*] Use the discrete naive Bayes classifier in the following problem. Assume that the class variable is binary Y, and that there are three input features X_1, X_2, and X_3 which are also binary.

   The training data is as follows:

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

   (a) Provide estimates of the class-conditional distributions of the form $P(X_j = x \mid Y = c)$ for all j, x, and c without smoothing. *Hint:* Use empirical frequencies.

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

   (c) Assume that the required probabilities have been estimated from data. How is the class value predicted for a new test instance (x_1, x_2)? Provide a formula.

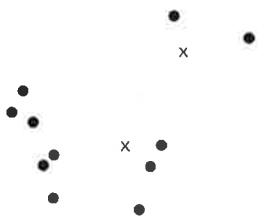
   (d) Suppose the test instance is $(0, 0)$. Calculate the posterior distribution of the class variable Y when Laplace smoothing is applied, to show that the more probable class value is 0. (Remember to apply smoothing to the class probabilities $P(Y = c)$ too.)

4. [*16 points*]

   (a) What kind of tasks can we use the K-means algorithm? Explain what the *inputs* and *outputs* of the algorithm are. Also explain how the results are to be interpreted.

   (b) Define the objective (or cost) function that the K-means algorithm tries to minimize. What can be said about the value of the objective function during the two stages of each iteration of Lloyd's algorithm.

   (c) Consider the following set of data points:



   The black dots are two-dimensional data points, and the crosses indicate the positions of the $K = 2$ centroids in the beginning. (Here the algorithm is initialized by giving the positions of the centroids.) The centroids are *not* data points.

   Simulate the algorithm until convergence. In each iteration, draw the cluster assignments and the positions of the centroids.

   (d) With the above data, give an example of initial centroid positions (with $K = 2$) that would lead to more equal-sized clusters. You don't have to present the iterations. Giving the initial centroid positions is enough.