582631 Introduction to Machine Learning (Autumn 2015)
Course examination, Wednesday 16 December
Examiner: Jyrki Kivinen

Answer all the problems. The maximum score for the exam is 60 points.

You may answer in English, Finnish or Swedish. If you use Finnish or Swedish, it may be helpful to include the English translations to any technical terms you introduce.

To make grading easier, please write your answer to each problem on its own sheet.

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 half a page of normal handwriting.

   (a) *Gini index* and *entropy*

   (b) *one-versus-one* and *one-versus-rest*

   (c) *single linkage* and *complete linkage*

2. [*12 points*] We have a binary classification task over an arbitrary instance space $\mathcal{X}$. Suppose we are given a scoring classifier $\hat{s}\colon \mathcal{X} \to \mathbb{R}$. Hence, large $\hat{s}(x)$ means that the true label of x is considered more likely to be $+1$ than -1. We have a test set of 10 examples. (For simplicity, we consider an unrealistically small set.) The table below gives for each of the 10 data points x_i the score $\hat{s}(x_i)$ and the true label y_i.

data point	1	2	3	4	5	6	7	8	9	10
score	78	23	11	92	65	44	52	48	36	42
label	-1	$+1$	-1	$+1$	$+1$	-1	$+1$	-1	-1	$+1$

   (a) Draw the ROC curve that visualises the performance of the scoring classifier $\hat{s}$ on our test set.

   (b) Define the *ranking accuracy* of a scoring classifier. How can you determine it using the ROC curve? Use the ROC curve from part (a) as an example.

Continues on the reverse side!

3. [*12 points*] We consider predicting a binary class $Y \in \{-1, +1\}$ with two binary input features $X_1, X_2 \in \{-1, +1\}$. Suppose we know that $(X, Y) = ((X_1, X_2), Y)$ is generated as follows.

   (a) First we pick $Y = +1$ with probability $2/3$ and $Y = -1$ with probability $1/3$.

   (b) We then pick $X_1 = +1$ with probability $1/2$ and $X_1 = -1$ also with probability $1/2$.

   (c) Finally, if $Y = +1$ we pick $X_2 = X_1$ with probability $3/4$ and $X_2 = -X_1$ with probability $1/4$; if $Y = -1$ we pick $X_2 = X_1$ with probability $1/4$ and $X_2 = -X_1$ with probability $3/4$

   Calculate the posterior probabilities $P(Y = +1 \mid X = (x_1, x_2))$ for all $(x_1, x_2) \in \{-1, +1\}^2$. What is the Bayes-optimal classifier for this distribution? What is the Bayes error?

4. [*12 points*] What is meant by *generalisation* and *overfitting* in the context of supervised learning? In general, how can we evaluate generalisation, and try to avoid overfitting? You should both explain general methods, and give specific examples of techniques for avoiding overfitting with at least two different types of models (for example, linear models and decision trees).

5. [*12 points*]

   (a) For what kind of tasks can we use the *K-means algorithm*? Explain carefully what the inputs and outputs of the algorithm are, and give a very brief intuitive explanation of how the results are to be interpreted.

   (b) Describe the actual K-means algorithm. The description should brief and on a high level.

   (c) Define formally the objective function which the K-means algorithm tries to minimise. Can you give any guarantees about how well the algorithm actually minimises the function?