Final Exam Machine Learning
Normal Group, Normal Version

December 20, 2007

18.30 – 21.15

Please write down the version of your exam! You are allowed to use a calculator. The exam will be graded as follows: You start with 1 point, and for each of the 12 subquestions you can get 3/4 points. Partial points may be awarded for partially correct answers. Good luck!

1. (a) Figure 1 shows classification data with two classes: Black and White. The two instances with dotted lines, which have been labeled 1 and 2, have not been classified yet. Which class labels would be assigned to them by k-nearest neighbour for $k = 1$, $k = 3$ and $k = 5$?

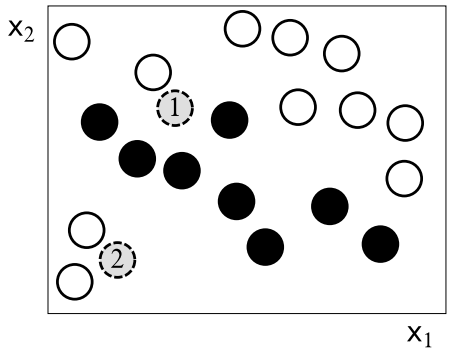

Figure 1: A classification data set

(b) In a different data set, given in Table 1, the feature x_1 has three possible values: Black, White and Brown. The feature x_2 can take on any integer value, and you may assume that it makes sense to look at the difference between two of its values. What would be an appropriate way to represent these features for the k-nearest neighbour algorithm (assuming it uses Euclidean distance between feature vectors)?

Table 1: A data set

x_1 HorseColour	x_2 NumberOfEnemiesDefeated	y GoodOrEvil
Black	1	Good
Black	36	Evil
White	0	Good
Brown	0	Good

(c) Perhaps it would be better to measure NumberOfEnemiesDefeated by the dozen. So suppose we would measure x_2 from Table 1 in units of twelve. For example, $1 \rightarrow 1/12$ and $36 \rightarrow 3$. How would this influence the k-nearest neighbour algorithm?

(d) Consider yet another classification task, in which there are two attributes, x_1 and x_2, that can both take values 1, 2, ..., 100, and the possible classes are again Black and White. Suppose the target function assigns the class label y like on a chess board, where x_1 indexes the rows of the board and x_2 the columns: y is Black if $x_1 + x_2$ is even and y is White if $x_1 + x_2$ is odd. Would it be hard or easy (in terms of the amount of training data required) for 1-nearest neighbour to learn this target function? (Please motivate your answer.)

2. (a) Suppose we want to learn perceptron weights from the following five examples:

$$D = \begin{array}{|c||r|r|r|r|r|} \hline y & -1 & -1 & -1 & 1 & -1 \\ \hline x_1 & -1 & 1 & 0 & 1 & -1 \\ \hline x_2 & 1 & -1 & 0 & 1 & -1 \\ \hline \end{array}$$

Show that there exist weights such that the perceptron classifies all examples correctly.

(b) Give an example of a data set with at least four examples, on which a perceptron would always make at least one classification error.

3. In one extension of the basic gradient descent algorithm, the learning rate is decreased slowly while running the algorithm. Why would this be done?

Please turn page.

4. Given the training data in Table 2, how would naive Bayes classify a new feature vector with both of its components set to True?

Table 2: Some Boolean-valued data

x_1	x_2	y
False	False	False
False	True	True
True	False	True

5. Suppose we want to predict how the following binary sequence continues:

$$D = \begin{array}{|c|c|c|c|c|c|c|c|} \hline y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 \\ \hline 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ \hline \end{array}.$$

We are given a hypothesis space containing two deterministic hypotheses, $\mathcal{H} = \{h_1, h_2\}$, which make the following predictions:

$$h_1 : y_n = 1,$$
$$h_2 : \begin{cases} y_n = 0 & \text{for } n \leq 3, \\ y_n = 1 & \text{for } n > 3. \end{cases}$$

(a) Suppose we think that the probability of observing a measurement error is 1/10 for each outcome y_n in the sequence. Describe the model that incorporates this knowledge.

(b) Which hypothesis would be selected from that model by maximum likelihood parameter estimation based on data D? (Please include sufficient computations to motivate your answer.)

(c) Which hypothesis would be selected from that model by Bayesian MAP estimation if we gave prior probability 1/100 to the hypothesis selected by maximum likelihood and 99/100 to the other hypothesis in the model? (Please include sufficient computations to motivate your answer.)

6. Suppose we have an English text consisting of n words and want to use two-part MDL to choose between the three context-free grammars (CFGs) from class:

- the promiscuous grammar, which accepts any text of any length;
- the ad hoc grammar, which only accepts the training text;
- the 'right' grammar, which provides a good CFG approximation of the real grammar of English.

Why does two-part MDL prefer the 'right' grammar over the other two grammars? In your answer you should mention whether $L(H)$ and $L(D \mid H)$ are small or large for the grammars, relative to the size of the uncompressed text (assuming n is very large).