Resit Exam Machine Learning

February 14, 2008

18.30 – 21.15

Please include sufficient motivation for your answers. 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!

Table 1: Classification data

x_1	x_2	y
0	2	+1
2	2	−1
1	2	−1
0	0	+1

1. Consider a classification problem where the class label y can take the values -1 and $+1$ and there are two features, x_1 and x_2, which both have possible values 0, 1 and 2. Let $\mathcal{H} = \{h_1, h_2, h_3\}$ be a hypothesis space for this problem that contains the following three hypotheses[1]:

$$h_1(\mathbf{x}) = \begin{cases} +1 & \text{if } x_1 \cdot x_2 = 0, \\ -1 & \text{otherwise.} \end{cases}$$

$$h_2(\mathbf{x}) = \begin{cases} +1 & \text{if } x_1 \neq x_2, \\ -1 & \text{otherwise.} \end{cases}$$

$$h_3(\mathbf{x}) = \begin{cases} +1 & \text{if } x_1 = 0, \\ -1 & \text{otherwise.} \end{cases}$$

   (a) Give a decision tree that makes the same classifications as h_1.

[1] $x_1 \cdot x_2$ denotes the product of x_1 and x_2.

   (b) Using hypothesis space $\mathcal{H}$, how would the List-Then-Eliminate algorithm classify the new instance $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ based on the data in Table 1? (Remember to motivate your answer.)

   (c) Give an example of a hypothesis for this classification problem that is consistent with the data in Table 1, but is not a member of $\mathcal{H}$.

   (d) Which of the hypotheses in $\mathcal{H}$ can be implemented by a perceptron by choosing suitable weights?

   (e) Give an example of a prefix code, as used in minimum description length learning, to encode the elements of $\mathcal{H}$.

2. Given the training data in Table 1, how would naive Bayes classify the new instance $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$? (Please include sufficient computations to motivate your answer.)

   A mistake was found in this question during the exam: Naive Bayes cannot classify the given instance x = $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$, because in Table 1 the feature x_2 does not take the value 1 for any of the classes. Therefore the question was changed: In the corrected version you were asked how naive Bayes would classify x = $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ instead.

3. Suppose we get a new dataset D and run the ID3 algorithm on it.

   (a) If we use the tree selected by ID3 to classify new data, can we expect it to achieve approximately the same accuracy as on data D? (Motivate your answer.)

   (b) Suppose that after running ID3, we perform reduced-error pruning, but we use the same data D to decide which nodes of the tree to prune. What would be the effect of pruning in this case?

4. (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 $k = 5$?

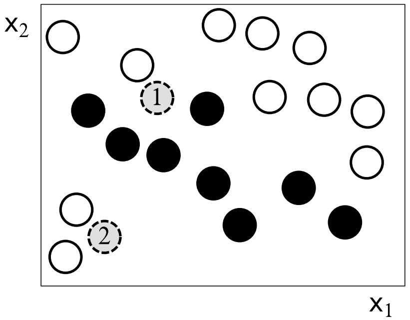

Figure 1: A classification data set

(b) Suppose we multiply all feature values by the same number, what will be the effect on the k-nearest neighbour algorithm (assuming it uses Euclidean distance between feature vectors)? (Motivate your answer.)

(c) Consider 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 by the following rule: y is Black if $x_1 + x_2 > 100$ and y is White otherwise. Would it be hard or easy (in terms of the amount of training data required) for 5-nearest neighbour to learn a close approximation of this target function? In your answer please discuss how good the learned approximation would be: Are there some instances on which it would be more likely to make mistakes? (Please motivate your answers.)

5. Suppose we have the following data consisting of a's and b's:

$D =$

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
a	b	a	b	a	b	a	b

We are given a model containing two probabilistic hypotheses, $\mathcal{M} = \{P_\theta \mid \theta \in \{1,2\}\}$, which make the following predictions:

$$P_1(y_n = a) = 0.3 \qquad P_2(y_n = a) = 0.8$$
$$P_1(y_n = b) = 0.7 \qquad P_2(y_n = b) = 0.2$$

Come up with a prior distribution on θ such that for data D Bayesian MAP would select a different hypothesis from maximum likelihood parameter estimation. (Please include computations showing which hypotheses are selected by maximum likelihood and MAP.)