Final Exam Machine Learning
Normal Group,
Resit-of-intermediate-exam 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. Consider the hypothesis space described in Chapter 2 of Mitchell for the EnjoySport concept learning example:

$$\begin{aligned}\mathcal{H} = \{&\langle ?, ?, ?, ?, ?, ? \rangle, \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle,\\ &\langle \text{Cloudy}, ?, ?, ?, ?, ? \rangle, \ldots, \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle\},\end{aligned}$$

To remind you: Each hypothesis consists of a list of six constraints, one for each attribute. If a feature vector satisfies all constraints, then it is classified as Yes (part of the concept), otherwise as No (not part of the concept). The first attribute has three possible values, the other attributes two. Each constraint can take the following values:

Constraint Value	Description
?	Any value is acceptable for the corresponding attribute.
$\emptyset$	No value is acceptable.
Warm	Single required value for attribute (e.g. Warm)

(a) For each of the following hypotheses determine whether they are a member of this hypothesis space $\mathcal{H}$. In case a hypothesis is a member of $\mathcal{H}$, express it in Mitchell's notation as a list of six constraints. Otherwise motivate why it cannot be expressed in this form.

$$h_1(\mathbf{x}) = \begin{cases} \text{Yes} & \text{if } x_2 = \text{`Warm' and } x_3 = \text{`Normal'},\\ \text{Yes} & \text{if } x_2 = \text{`Cold' and } x_3 = \text{`Normal'},\\ \text{No} & \text{otherwise.} \end{cases}$$

$$h_2(\mathbf{x}) = \begin{cases} \text{Yes} & \text{if } x_1 = \text{`Sunny' and } x_2 = \text{`Cold'},\\ \text{Yes} & \text{if } x_1 = \text{`Sunny' and } x_2 = \text{`Warm' and } x_3 = \text{`High'},\\ \text{No} & \text{otherwise.} \end{cases}$$

(b) Give an example of a hypothesis not contained in $\mathcal{H}$. (No explanation is required.)

2. How does ID3 decide which attributes to put near the root of the tree?

3. Fitting a 14th degree polynomial to 15 data points using least squares regression will result in overfitting. For each of the following two cases, argue whether they would help against overfitting:

   (a) Suppose that instead of minimizing the sum of the squares of the errors, we would minimize the sum of the absolute values of the errors.

   (b) Suppose that instead of 15 data points we had 100 000.

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 $k = 5$?

   (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)?

5. Give an example of a data set with at least four examples, on which a perceptron would always make at least one classification error. (N.B. Thus your answer should be of the form:

$$\begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \begin{pmatrix} y_2 \\ \mathbf{x}_2 \end{pmatrix}, \begin{pmatrix} y_3 \\ \mathbf{x}_3 \end{pmatrix}, \begin{pmatrix} y_4 \\ \mathbf{x}_4 \end{pmatrix}, \ldots,$$

Please turn page.

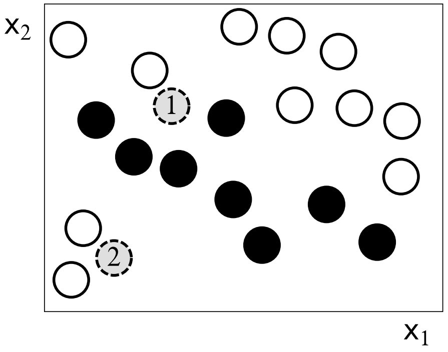

Figure 1: A classification data set

Table 1: A data set

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

with specific numbers filled in for each y and the components of each $\mathbf{x}$.)

6. 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

7. 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 model containing two probabilistic hypotheses, $\mathcal{M} = \{P_1, P_2\}$, which make the following predictions:

$$P_1(y_n = 1) = 0.9$$
$$P_2(y_n = 1) = 0.1$$

(a) 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.)

(b) 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.)

8. 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).