Resit Final Exam Machine Learning 2008

February 19, 2009

18.30 – 21.15

This exam is **open book**: you can use Tom Mitchell's "Machine Learning" as well as prints of the lecture slides and any notes you've taken. You can use a calculator.
Answers are allowed in Dutch and English.
There are 20 sub-questions, all equally weighted.

Good luck!

Questions

1. **Decision Trees**

   Imagine you do a newspaper round to help you get through these lean times. On your round, you encounter a number of dogs that either bark or (try to) bite. The dogs are described by the following attributes: *Heavy*, *Smelly*, *Big* and *Growling*. Consider the following set of examples:

Table 1: Biting dog data

Heavy	Smelly	Big	Growling	Bites
No	No	No	No	No
No	No	Yes	No	No
Yes	Yes	No	Yes	No
Yes	No	No	Yes	Yes
No	Yes	Yes	No	Yes
No	No	Yes	Yes	Yes
No	No	No	Yes	Yes
Yes	Yes	No	No	Yes

   (a) What is the entropy of *Bites*? Show your calculations[1].

[1]Note: if your calculator can't do log_2, use one of the following equations: $log_2(x) = 1.44 \cdot ln(x)$ or $log_2(x) = 3.32 \cdot log(x)$. If you didn't bring a calculator, you may give the answer as an expression.

(b) Which attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)? Hint: you can figure this out by looking at the data without explicitly calculating the information gain for each of the attributes.

(c) What is the information gain of the attribute you chose in the previous question?

(d) Draw the full decision tree that would be learned for this data using ID3 (no pruning).

(e) Suppose three new dogs appear in your round as listed in the table below. Classify them using the decision tree from the previous question.

Table 2: New dogs set

Dog	Heavy	Smelly	Big	Growling
Buster	Yes	Yes	Yes	Yes
Pluto	No	Yes	No	Yes
Zeus	Yes	Yes	No	No

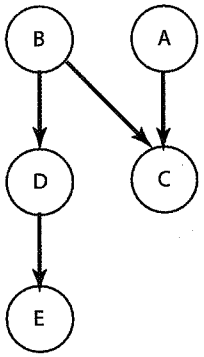
2. Bayesian Classifiers

(a) Consider a **naive** Bayes classifier trained on the same dataset from Table 1. How would that predict `Bites` given the input `Smelly = Yes, Big = Yes`? Show your calculations.

(b) Someone has, using an unknown technique, developed a program that estimates the likelihood of a dog biting on the basis of data like that listed in Table 2. For a certain dog, let's call her Herta, this program outputs a 58% likelihood of her biting based on her data. If you decide from only this number, that Herta bites, is that a Maximum Likelihood (ML) or a Maximum A Posteriori (MAP) estimate?

(c) If you answered ML in the previous question, what is the MAP esitimate? If you answered MAP, what is the ML estimate? (Note: from Table 1, we know that the a priori likelihood that a dog bites equals $\frac{5}{8}$).

(d) Write all the conditional independencies you can read from the graph below the form ($X \perp Y|Z$) ("*X is conditionally independent of Y given Z*"),where X and Y are single variables and Z is any set of variables, including the empty set:

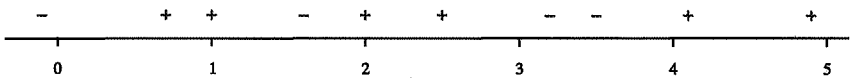


3. Instance-Based Learning

Consider the following dataset with one real-valued input x and one binary output y. We are going to use k-nearest neighbour with unweighted Euclidean distance to predict y for a given x.

Table 3: Instance-based learning data set

X	Y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+



(a) What is the predicted class of 1-NN for a new data-point $x = 1.5$?

(b) What is the predicted class of 3-NN for a new data-point $x = 1.5$?

(c) What is the leave-one-out cross-validation [2] error of 1-NN on this dataset? Give your answer as the number of misclassifications.

(d) What is the leave-one-out cross-validation error of 3-NN on this dataset? Give your answer as the number of misclassifications.

[2]Remeber: Leave-one-out cross-validation = cross-validation using each separate data-point as the hold-out sample once, i.e., first use $\langle -0.1, + \rangle$ as the test set, then $\langle 0.7, + \rangle$, and so on.

4. **Hypothesis Comparison and Cross Validation**

   (a) Explain in your own words why you cannot use training set error as a reliable estimate of hypothesis performance.

   (b) In a binary classification setting, some classifier correctly classifies 1750 out of 2000 test examples. Give an estimate of the true error of this classifier, and a 90% confidence interval around that estimate (you may give your confidence interval in the form of an expression). [3]

   (c) Describe a 10-fold cross-validation procedure in pseudo-code.

   (d) Does it make sense to calculate a confidence interval over the average error over n-fold cross-validation samples? Why (not)?

5. **Understanding a novel learning method: Naive Bayes/Decision tree hybrids**

   Suppose you were to augment a decision tree by adding a Naive Bayes classifier at each leaf. That is, consider learning a decision tree of depth k, (where k is smaller than the number of variables n), where each leaf contains not a class, but a Naive Bayes classifier, which is based on the $n - k$ variables not appearing on the path to that leaf.

   (a) Briefly outline a learning algorithm for this representation, assuming that k is a fixed parameter.

   (b) Will this result in having the same Naive Bayes classifier at each leaf? Why or why not? If not, how will they be different?

   (c) Briefly describe a plausible means of selecting a value for k.

[3] The table of z-values:

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58