Final Exam Machine Learning
2008

December 17, 2008

15.15 – 18.00

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.

Good luck!

Questions

1. **Decision Trees**

   The following dataset will be used to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, colour and odour.

Table 1: Mushroom data

Shape	Colour	Odour	Edible
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

   (a) What is entropy $H(Edible|Odour = 1 \text{ or } Odour = 3)$? 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) Draw the full decision tree that would be learned for this data (no pruning).

(c) Suppose we have a validation set as follows. What will be the training set error and validation set error of the tree? Express your answer as the number of examples that would be misclassified.

Table 2: Mushroom validation set

Shape	Colour	Odour	Edible
C	B	2	No
D	B	2	No
C	W	2	Yes

(d) The ID3 algorithm as mentioned in the book does not specify what to do if all attributes have zero (0.0) information gain. Suggest (and motivate) how to proceed in this case. Use no more than three sentences.

2. Bayesian Classifiers

(a) Consider a **naive** Bayes classifier trained on the same dataset from Table 1. How would that predict `Edible` given the input `Colour = 2, Odour = B`? Show your calculations.

(b) Consider two very simple functions that estimate, given a value x, the likelihood of a positive and a negative value for y:

$$P(y = negative|x) = 0.3 \text{ if } 0.5 \leq x \leq 1,\ 0 \text{ otherwise.}$$
$$P(y = positive|x) = 0.4 \text{ if } 0.8 \leq x \leq 2,\ 0 \text{ otherwise.}$$

We are also told that $P(y = negative) = 0.6$ (and therefore, $P(y = positive) = 0.4$).

What would be the Bayes (or MAP) classification for $x = 0.9$? Show your calculations.

(c) In the same setting, what would be the Maximum Likelihood classification given $x = 0.9$? Show your calculations.

(d) For any given set of likelihood estimators, under which circumstances will the MAP and ML predictions be the same?

3. Neural Nets and Regression

(a) When training perceptrons with gradient descent, one can add momentum: why and how does momentum help?

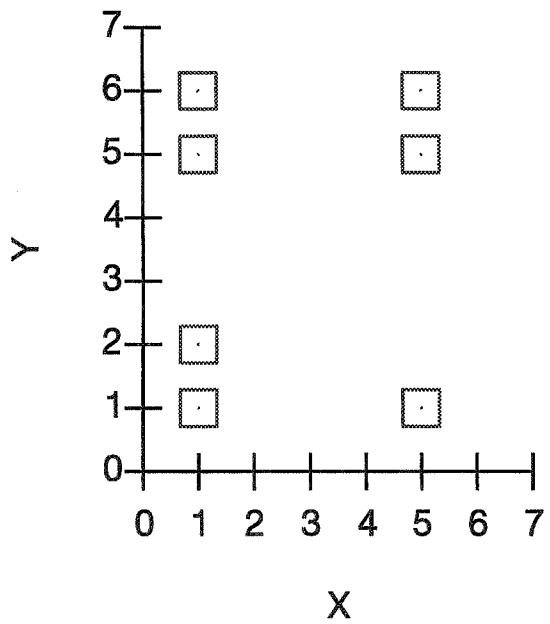
(b) What would be the consequence of setting momentum too high? Too low?

(c) Suppose we have a neural net with a single hidden unit and training and test data-sets of 16 examples each. What would be the effect of adding more hidden units on the training error and on the test set error?

(d) Describe –in no more than three sentences– a procedure to select the optimal number of hidden units.

4. **Instance-Based Learning**

The following picture shows a dataset with one real-valued input x and one real-valued output y. There are seven training points.



Suppose you are training using distance weighted nearest neighbour ("kernel regression" in the lecture) with some unspecified distance weighting (kernel) function. The only thing you know about the kernel function is that it is a monotonically decreasing function of distance that decays to zero at a distance of 3 units (and is strictly greater than zero at a distance of less than 3 units).

(a) What is the predicted value of y when x = 1?

(b) What is the predicted value of y when x = 3?

(c) What is the predicted value of y when x = 4?

(d) What is the role of a kernel function and why does it allow instance-based learners to take all examples into account instead of only the k nearest neighbours?

5. **Hypothesis Comparison**

(a) You have trained a classifier to predict credit card fraud. Over the 120 training examples, it correctly classifies 108. Over a separate set of 50 test examples, it correctly classifies 42. Give an *unbiased* 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). [2]

6. **Genetic Algorithms**

Suppose you decide to use a genetic algorithm to determine the connection weights for some neural net for a regression problem. There are 6 weights to optimise and you have a training set T and a validation set V to measure performance.

(a) Describe a suitable encoding scheme for the input weights.

(b) Describe the fitness function.

(c) What are the benefits (if any) of using a genetic algorithm in this manner compared to gradient descent?

(d) What are the drawbacks (if any) of using a genetic algorithm in this manner compared to gradient descent?

[2]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