Resit Exam Machine Learning
2009

8 February 2010

This exam is **open book**: you can use Tom Mitchell's "Machine Learning" as well as the lecture slides and any notes you've taken. You can use a calculator.
Answers are allowed in Dutch and English.

Good luck!

Questions

1. **Short answers** *(18 Points)*

   (a) (**True** or **False**): Neural nets cannot model XOR.

   (b) (**True** or **False**): Cross-validation can help determine the appropriate value for k in k-nearest neighbour.

   (c) (**True** or **False**): Join-the-dots has a higher representational bias than linear regression.

   (d) (**True** or **False**): Gradient descent is guaranteed to find a local minimum.

   (e) (**True** or **False**): Any type of probability density estimator can serve as a basis for a MAP classification.

   (f) (**True** or **False**): The k means algorithm can only differentiate between two clusters.

2. **Decision Trees** *(25 Points)*

   Next Sunday is Valentine's day, so the dataset in Table 1 will be used to learn a decision tree for predicting whether a bouquet of flowers is suitable to give to the object of your secret affection based on its flowers, colour and odour.

   (a) What is entropy $H(Suitable|Odour = musky)$? 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.

Table 1: Bouquet data

Flower	Colour	Odour	Suitable
Roses	Red	sweet	Yes
Tulips	Red	sweet	Yes
Tulips	White	sweet	Yes
Tulips	White	musky	Yes
Roses	Red	musky	Yes
Tulips	Red	musky	No
Tulips	Black	musky	No
Roses	Yellow	musky	No
Roses	Red	spoilt	No
Roses	White	spoilt	No
Tulips	White	spoilt	No

(b) Which attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)?

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

(d) 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: Bouquet validation set

Flower	Colour	Odour	Suitable
Roses	Red	musky	No
Tulips	Red	musky	No
Tulips	White	musky	Yes

(e) Suggest how ID3 could be extended to take the noisiness of particular input attributes into account when considering them for a split. *Hint*: think of noise as a kind of cost.

3. **Bayesian Classifiers** *(20 Points)*

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

(b) Consider a naive Bayes classifier with 3 boolean input variables, $X_1; X_2$ and X_3, and one boolean output, Y. How many parameters must be estimated to train such a naive Bayes classifier? (you need not list them unless you wish to, just give the total)

(c) Briefly describe the difference between a *maximum likelihood* (ML) hypothesis and a *maximum a posteriori* (MAP) hypothesis.

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

4. **Neural Nets and Regression** *(10 Points)*

(a) For a neural network, which one of these structural assumptions is the one that most affects the trade-off between underfitting (i.e. a high bias model) and overfitting (i.e. a high variance model):

i. The number of hidden nodes

ii. The learning rate

iii. The initial choice of weights

iv. The use of a constant-term unit input

(b) When training perceptrons with gradient descent, one can add momentum. As the name implies, this adds momentum to the descent: the search doesn't stop at a minimum, but 'overshoots' it. This overshooting is hoped to get the descent out of local minima. Suggest a procedure to find a good setting for the momentum.

5. **Instance-Based Learning** *(20 Points)*

(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 $k = 5$? (Assume Euclidean distance)

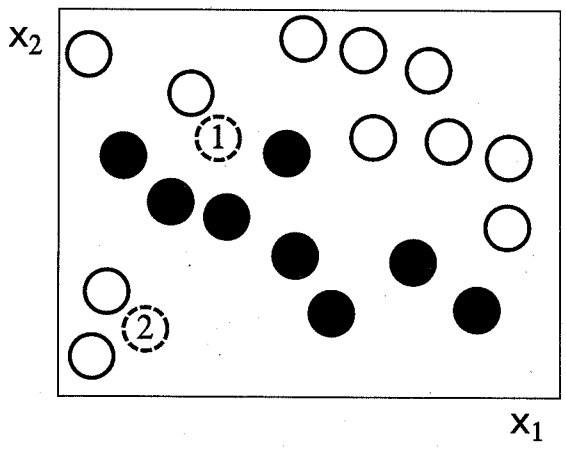

Figure 1: A classification data set

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

(c) Techniques such as decision trees are known as batch learners that require the availability of all training data to build their hypothesis. Thus the arrival of additional training data needs to be handled carefully. Does KNN suffer from this problem and why (not)?

(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?