Machine Learning Exam

15 Februari 2007

This exam is open book.
Important: make sure you write down your name and student number on each sheet of paper that you use.

Contents

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, color and odor.

Shape	Color	Odor	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

1. What is entropy $H(\text{Edible}|\text{Odor} = 1 \text{ OR Odor} = 3)$?
2. Which attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)?
3. Draw the full decision tree that would be learned for this data (no pruning)
4. 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.

Shape	Color	Odor	Edible
C	B	2	No
D	B	2	No
C	W	2	Yes

2 Bayes Classifiers

Suppose you have the following training set with three boolean inputs x, y, and z, and a boolean output U.

x	y	z	U
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	1
0	0	1	1
0	1	0	1
1	1	0	1

Suppose you have to predict U using a naive Bayes classifier.

1. After learning is complete what would be the predicted probability $P(U = 0|x = 0, y = 1, z = 0)$?
2. Using the probabilities obtained during the naive Bayes Classifier training, what would be the predicted probability $P(U = 0|x = 0)$?

3 Neural Nets

Consider a single sigmoid threshold unit with three inputs, x_1, x_2, and x_3.

$$y = g(w_0 + w_1x_1 + w_2x_2 + w_3x_3) \text{where} g(z) = \frac{1}{1 + exp(-z)}$$

We input values of either 0 or 1 for each of these inputs.

1. Assign values to weights w_0, w_1, w_2 and w_3 so that the output of the sigmoid unit is greater than 0.5 if and only if (x_1 AND x_2) OR x_3

Answer the following true or false (No explanation required)

2. One can perform linear regression using either matrix algebra or using gradient descent.
3. The error surface followed by the gradient descent Backpropagation algorithm changes if we change the training data.
4. Incremental gradient descent is always a better idea than batch gradient descent.
5. Given a two-input sigmoid unit with weights w_0, w_1, w_2, we can negate the value of the unit output by negating all three wieghts.
6. The gradient descent weight update rule for a unit whose output is $w_0 + w_1(x_1 + 1) + w_2(x_1^2)$ is:

$$\begin{aligned}
\Delta w_0 &= \eta \sum_d (t_d - o_d) \\
\Delta w_1 &= \eta \sum_d [(t_d - o_d)x_{d1} + (t_d - o_d)] \\
\Delta w_2 &= \eta \sum_d [(t_d - o_d)2x_{d2}]
\end{aligned}$$

where

- t_d is the target output for the dth training example
- o_d is the unit output for the dth example
- x_{1d} is the value of x_1 for the dth training example
- x_{2d} is the value of x_2 for the dth training example

4 VC Dimension

Given the *sign* function, defined as:

$$sign(x) = \begin{array}{ll} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{array}$$

1. Suppose we have one input variable x and one output variable y. We are using the machine $f_1(x, \alpha) = \text{sign}(x + \alpha)$. What is the VC dimension of f_1? No proof needed, show your reasoning.

2. Suppose we have one input variable x and one output variable y. We are using the machine $f_2(x, \alpha) = \text{sign}(\alpha x + 1)$. What is the VC dimension of f_2.? No proof needed, show your reasoning.

5 Miscellaneous

1. Suppose H is a set of possible hypotheses and D is a set of training data. We would like our program to output the most probable hypothesis h from H, given the data D. Under what conditions does the following hold?

$$\text{argmax}_{h \in H} P(h|D) = \text{argmax}_{h \in H} P(D|h)$$

2. Explain in your own words why learning from examples is futile without some form of inductive bias.

3. For each of the following algorithms, (a) state the objective that the learning algorithm is trying to optimize, and (b) indicate whether the algorithm is guaranteed to find the global optimum hypothesis.

   - Backpropagation with multi-layer networks
   - The perceptron learning rule, applied to a single perceptron
   - The FindS algorithm from Chapter 2

4. In one sentence each, give

   - one advantage of ID3 over Backpropagation
   - one advantage of Backpropagation over ID3
   - one advantage of FOIL over ID3
   - one advantage of Multi-layer networks over Perceptrons.