# Machine Learning Exam

21 December 2006

*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** **2**

**2 Bayes Rule** **4**

**3 Gradient Search** **5**

**4 Locally Weighted Regression** **6**

**5 Misc** **7**

**6 Naive Bayes** **8**

**7 Short Questions** **9**

# 1 Decision Trees

Master Yoda is concerned about the number of Jedi apprentices that have turned to the Dark Side, so he's decided to train a decision tree on some historical data to help identify problem cases in the future. The following table summarizes whether or not each of 12 initiates turned to the Dark Side based on their age when their training began, whether or not they completed their training, their general disposition and their species.

| Dark Side | Age Started Training | Completed Training | Disposition | Species |
|---|---|---|---|---|
| 0 | 5 | 1 | Happy | Human |
| 0 | 9 | 1 | Happy | Gungan |
| 0 | 6 | 0 | Happy | Wookie |
| 0 | 6 | 1 | Sad | Mon Calamari |
| 0 | 7 | 0 | Sad | Human |
| 0 | 8 | 1 | Angry | Human |
| 0 | 5 | 1 | Angry | Ewok |
| 1 | 9 | 0 | Happy | Ewok |
| 1 | 8 | 0 | Sad | Human |
| 1 | 8 | 0 | Sad | Human |
| 1 | 6 | 0 | Angry | Wookie |
| 1 | 7 | 0 | Angry | Mon Calamari |

1. What is the initial entropy of *Dark Side*?
2. Which attribute would the decision-tree building algorithm choose to use for the root of the tree?
3. What is the information gain of the attribute you chose to split on in the previous question?
4. Draw the full decision tree that would be learned for this data (with no pruning). *(Hint: The tree will have no more than three splits. The correct split at each point should be clear from just the groups it splits the data into, without having to actually compute the information gain for each possible split)*
5. Consider the possibility that the input data above is noisy and not completely accurate, so that the decision tree you learned may not accurately reflect the function you want to learn. If you were to evaluate the three initiates represented by the data points below, on which one would you be most confident of your prediction, and why?

| Dark Side | Age Started Training | Completed Training | Disposition | Species |
|---|---|---|---|---|
| Ardath | 5 | 0 | Angry | Human |
| Barbar | 8 | 0 | Angry | Gungan |
| Caldar | 8 | 0 | Happy | Mon Calamari |

6. Assume we train a decision tree to predict Z from A,B, and C using the following data (with no pruning).

| Z | A | B | C |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 |
| 0 | 1 | 0 | 0 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 |

What would be the training set error for this dataset? Express your answer as the number of records out of 12 that would be misclassified.

7. Consider a decision tree built from an arbitrary set of data. If the output is discrete-valued and can take on $k$ different possible values, what is the maximum training set error (expressed as a fraction) that any data set could possibly have?

## 2 Bayes Rule

1. I give you the following fact

$$P(A|B) = 2/3$$

Do you have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

2. Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|\neg B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

3. Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|\neg B) &= 1/3 \\ P(B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

4. Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|\neg B) &= 1/3 \\ P(B) &= 1/3 \\ P(A) &= 4/9 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

# 3 Gradient Search

1. Derive a gradient descent training algorithm that minimizes the sum of squared errors for a variant of a perceptron where the output $o$ of the unit depends on its inputs $x_i$ as follows:

$$o = w_0 + w_1 x_1 + w_1 x_1^3 + w_2 x_2 + w_2 x_2^3 + \ldots + w_n x_n + w_n x_n^3$$

Give your answer in the form $w_i \leftarrow w_i + \ldots$ for $1 \leq i \leq n$. You do not need to give the update rule for $w_0$

# 4 Locally Weighted Regression

Here's an argument made by a misguided practitioner of Locally Weighted Regression

Suppose you have a dataset with $R_1$ training points and another dataset with $R_2$ test points. You must predict the output for each of the test points. If you use a kernel function that decays to zero beyond a certain Kernel width then Locally Weighted Regression is computationally cheaper than regular linear regression. This is because with locally weighted regression you must do the following for each query point in the test set,

- Find all the points that have non-zero weight for this particular query.
- Do a linear regression with them (after having weighted their contribution to the regression appropriately).
- Predict the value of the query

whereas with regular linear regression you must do the following for each query point:

- take all the training set datapoints.
- Do an unweighted linear regression with them.
- Predict the value of the query

The locally weighted regression frequently finds itself doing regression on only a tiny fraction of the datapoints because most have zero weight. So most of the local method's queries are cheap to answer. In contrast, regular regression must use every single training points in every single prediction and so does at least as much work, and usually more

1. This argument has a serious error. Even if it is true that the kernel function causes almost all points to have zero weight for each LWR query, the argument is wrong. What is the error?

# 5 Misc

Amazing Web-2.0 Technologies Ltd, has hired you as a consultant for their latest genetic algorithm project: learning to distinguish which web home pages belong to Republican versus Democrats. Due to the proprietary nature of their software product, they are unable to reveal to you the details of the hypothesis encoding used by the GA. However, they are willing to tell you that the GA encodes its hypothesis using a bitstring containing exactly 20 bits. They also reveal that this algorithm is allowed to run indefinitely until it outputs a hypothesis that classifies every training example correctly.

1. Their question to you is this: how many training examples of the boolean target function "Republican web pages" must they provide in order to assure that with 85% probability their GA will find a hypothesis whose true error is less than 15%, assuming that the genetic algorithm will find an answer.

2. They now run their GA and produce a hypothesis. When they test it on a set of 130 new instances they find it commits 20 errors. What is the 90% confidence interval (two-sided) for the true error rate of this hypothesis. Give a one-sentence justification for your answer.

3. What is the 95% one-sided interval (i.e., what is the upper bound $U$ such that $error_D(h) \leq U$ with 95% confidence)? Give a one-sentence justification.

# 6 Naive Bayes

Given this dataset of 16 records:

| A | B | C |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |

1. Write down the probabilities needed to make a naive Bayes classifier
2. Write the classification that the naive Bayes classifier would make for C given A=0, B=1
3. How many parameters would have to be estimated if the naive Bayes assumption is not made, and we wish to learn the full joint distribution of $A$, $B$ and $C$.

# 7 Short Questions

1. Describe the difference between a *maximum likelihood* hypothesis and a *maximum a posteriori* hypothesis.

2. Consider a learning problem defined over a set of instances $X$. Assume the space of possible hypothesis $H$, consists of all possible disjunctions over instances in $X$. E.g.., the hypothesis $x_1 \vee x_6$ labels these two instances positive, and no others, and the hypothesis $False$ labels no instances positive. What is the VC dimension of $H$?

3. True or False? If $P(A|B) = P(A)$ then $P(A \wedge B) = P(A)P(B)$

4. True or False? Because decision trees learn to classify discrete-valued outputs instead of real-valued functions, it is impossible for them to overfit.

5. True or False? The error of a hypothesis measured over the training set provides a pessimistically biased estimate of the true error of the hypothesis.