Leren: Written Assignment 5

Week 5 - Semester 1 2016/17

To be made individually

1. (1 punt) *Decision Trees*
Do one iteration of the algorithm for decision tree learning by hand, using the data set below. Use information gain for choosing a variable and its threshold value. You do not have to calculate the information gain for all variable and threshold combinations, only for the the 2 or 3 that seem most predictive. (The result of this is called a ”decision stump”.)

ID	**X1**	**X2**	**Y**
1	1	3	p
2	1	4	p
3	3	3	p
4	7	5	q
5	9	2	q
6	12	8	p
7	10	5	?

2. (1 points) *Gaussian Naive Bayes*

   Apply Gaussian Naive Bayes to the data above to classify example 7.

3. (1 points)

   Write a step by step method to compare two learning algorithms for classification (for example Decision Tree Learning and Neural Nets) using a given dataset.

4. (1 punt) *Bayes rule*
Suppose that in the example that Mitchell gives of Bayes rule (ch. 6, sectie 6.2.1) the patient is tested again, The test is again positive. What is the probability that he has cancer after the result of the second test?

5. (2 points)

   We run a learning algorithm on dataset D. It produces hypothesis H which we test on testset T. H makes E errors on the N(T) data in the testset T.

   (a) Calculate the variance of the proportion of errors on T. Use the internet to find the variance of a proportion.

   (b) Suppose that we would collect extra data and double the size of T. Would this affect the variance of the proportion of errors on T? If yes, how? If ”no” or if ”cannot tell”, explain.

   (c) Suppose that we would use the same data instead to double the size of D. Would this affect the variance of the proportion of errors on T? IIf ”no” or if ”cannot tell”, explain.

   (d) Suppose that our learning algorithm has a learning bias that is not correct for the domain. It makes the algorithm find hypotheses that cause on average 10% classification errors. Approximate the distribution of errors with a Gaussian.

   (e) Suppose that we increase the size of D quite substantially: what will be the effect on the expected proportion of errors?