AI534 — Written Homework Assignment 3 —

1. (Naive Bayes Classifier) (7 pts) Consider the following training set:

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

   (a) (3 pts) Learn a Naive Bayes classifier by estimating all necessary probabilities (there should be 7 independent probabilities to be estimated in total).
   Your answer goes here.

   (b) (3 pts) Compute the probability $P(y = 1|A = 1, B = 0, C = 0)$.
   Your answer goes here.

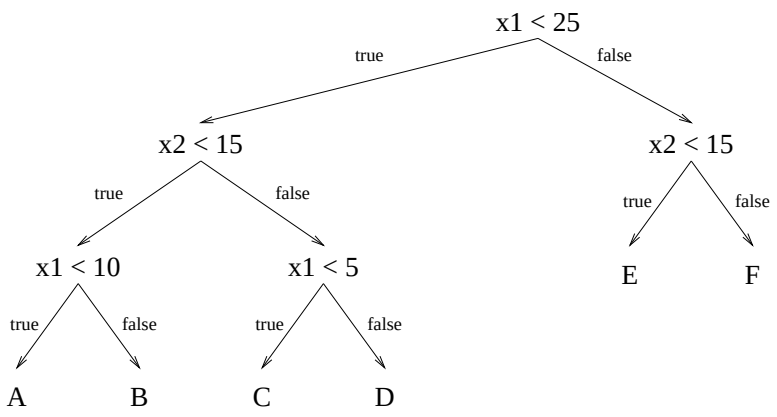
   (c) (1 pts) Suppose we know that the three features A, B and C are independent from one another, can we say that the Naive Bayes assumption is valid? (Note that the particular data set is irrelevant for this question). If your answer is yes, please explain why; if you answer is no please give an counter example.
   Your answer goes here.

2. (Naive Bayes learns linear decision boundary.) (10 pts) Show that the following naive Bayes classifiers learn linear decision boundary $w_0 + w_1x_1 + w_2x_2 + ... + w_dx_d = 0$. Express the weights using the corresponding Naive Bayes parameters. Hint: start with the decision boundary defined by $\log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = 0$.

   (a) Bernoulli Naive Bayes model, where features $x_1, x_2, ..., x_d$ are binary indicating the presence/absence of words in the vocabulary.
   Your answer goes here.

   (b) Multinomial Naive Bayes Model, where $x_1, ..., x_d$ representing counts of words $w_1, ..., w_d$ in the vocabulary. Express the weights using the Naive Bayes parameters: the class priors $P(y = 1), P(y = 0)$ and the class conditionals: $p(w_i|y = 1)andp(w_i|y = 0)$
   Your answer goes here.

3. (6 pts) Consider the following decision tree:



(a) (2 pts) Draw the decision boundaries defined by this tree. Each leaf of the tree is labeled with a letter. Write this letter in the corresponding region of input space.

Your answer goes here.

(b) (2 pts) Give another decision tree that is syntactically different but defines the same decision boundaries. This demonstrates that the space of decision trees is syntactically redundant.

Your answer goes here.

(c) (2pts) How does this redundancy influence learning (does it make it easier or harder to find an accurate tree)?

your answer goes here.

4. (6 pts) In the basic decision tree algorithm (assuming we always create binary splits), we choose the feature/value pair with the maximum information gain as the test to use at each internal node of the decision tree. Suppose we modified the algorithm to choose at random from among those feature/value combinations that had non-zero mutual information, and we kept all other parts of the algorithm unchanged.

(a) (2 pts) What is the maximum number of leaf nodes that such a decision tree could contain if it were trained on m training examples?

Your answer goes here.

(b) (2 pts) What is the maximum number of leaf nodes that a decision tree could contain if it were trained on m training examples using the original maximum mutual information version of the algorithm? Is it bigger, smaller, or the same as your answer to (b)?

Your answer goes here.

(c) (2 pts)How do you think this change (using random splits vs. maximum information mutual information splits) would affect the testing accuracy of the decision trees produced on average? Why?

Your answe goes here.

5. (8 pts) Consider the following training set:

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

Learn a decision tree from the training set shown above using the information gain criterion. Show your steps, including the calculation of information gain (you can skip $H(y)$ and just compute $H(y|\mathbf{x})$) of different candidate tests. You can randomly break ties (or better, choose the one that give you smaller tree if you do a bit look ahead for this problem).

Your answer goes here.

6. (8pts) Prove that

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

.

Hint: you should start with the definition $H(X,Y) = -\sum_{x,y} P(x,y) \log P(x,y)$. Here we use X, Y to denote the random variables and x, y denote the values X and Y take, $P(x,y)$ is a short hand notation denoting $P(X = x, Y = y)$.

Your answer goes here.