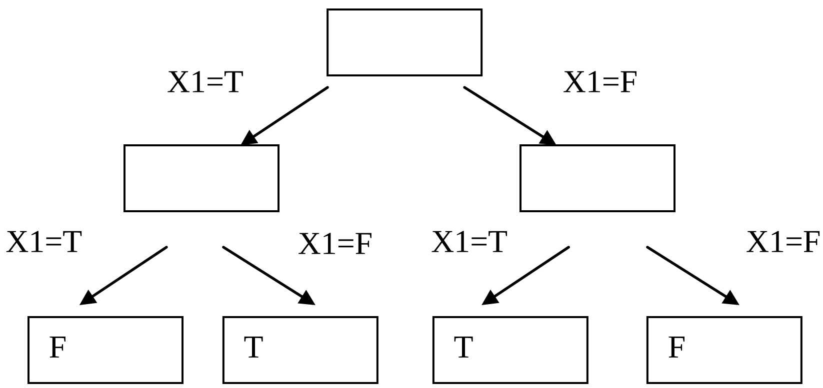
NAME:____ CSE 446 midterm 2/15/12

Please answer clearly and succinctly. Show your work clearly for full credit. If an explanation is requested, think carefully before writing. Points will be removed for rambling answers with irrelevant information (and may be removed in cases of messy and hard to read answers). If a question is unclear or ambiguous, feel free to make the additional assumptions necessary to produce the answer. State these assumptions clearly; you will be graded on the basis of the assumption as well as subsequent reasoning. There are 9 problems on 4 pages worth 30 points.

Problem 1 (1 point) Write your name on the top of each page.

Problem 2 (2 points) Draw the decision tree for X_1 XOR X_2



Problem 3 Consider the following set of training examples:

Instance	**Classification**	**X_1**	**X_2**
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

A. (3 points) What is the entropy of this collection of training examples with respect to the target function classification?

   Denote the target classification as Y.
   $H(Y) = - 0.5 * \log_b 0.5 – 0.5 * \log_b 0.5 = \log_b 2$ ($b > 0$). When $b=2$, $H(Y) = 1$.

B. (3 points) What is the information gain of X_2 relative to these training examples?

   $IG(X_2) = H(Y) – H(Y|X_2) = 0$
   where $H(Y|X_2) = 4/6 * (- 2/4 * \log_b 2/4 – 2/4 * \log_b 2/4)$
   $+ 2/6 * (- 1/2 * \log_b 1/2 – 1/2 * \log_b 1/2)$
   $= \log_b 2$

Problem 4:

A. (2 points) Let *p* be the probability of landing head of a coin. You flip the coin 3 times and note that it landed 2 times on tails and 1 time on heads. Suppose *p* can only take two values: 0.3 or 0.6. Find the Maximum likelihood estimate of *p* over the set of possible values {0.3,0.6}

$Lp = p*(1-p)^2$
$L_{0.3} = 0.3*(0.7)^2 = 0.147$
$L_{0.6} = 0.6*(0.4)^2 = 0.096$
Therefore MLE estimate of $p = 0.3$

B. (2 points) Suppose that you have the following prior on the parameter *p*: P(*p*=0.3)=0.3 and P(*p*=0.6)=0.7; Given that you flipped the coin 3 times with the observations described above, find the MAP estimate of *p* over the set {0.3, 0.6}, using the prior.

$Lp' = p * (1-p)^2 * P(p)$
$L_{0.3}' = 0.3*(0.7)^2*0.3 = 0.0441$
$L_{0.6}' = 0.6*(0.4)^2*0.7 = \cancel{0.00672}$ correction:0.0672
Therefore MAP estimate of $p = 0.6$

Problem 5: Consider learning a function X →Y where Y is Boolean, where X = <X_1, X_2> such that X_1 is a Boolean variable and X_2 is a Real number.

A. (2 points) State the parameters that must be estimated to define a (Gaussian) Naive Bayes classifier in this case.

θ_y =P(Y=0) , we can derive P(Y=1) = 1-P(Y=0)
$\theta_{x1}^{y=0}$ =P(X_1=0|Y=0) , we can derive P(X_1=1|Y=0) = 1-P(X_1=0|Y=0)
$\theta_{x1}^{y=1}$ =P(X_1=0|Y=1) , we can derive P(X_1=1|Y=1) = 1-P(X_1=0|Y=1)
u_0 and v_0 as the mean and variance of the Gaussian for P(X_2|Y=0)
u_1 and v_1 as the mean and variance of the Gaussian for P(X_2|Y=1)

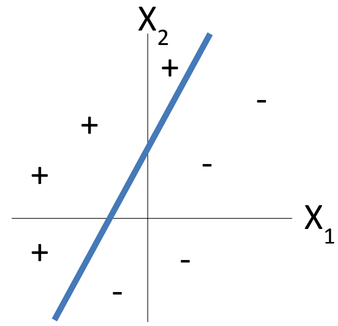
B. (3 points) Give the formula for computing P(Y | X), in terms of these parameters and the feature values X_1 and X_2.

P(Y|X) = P(X|Y) * P(Y) / P(X)
Where P(X)= P(X|Y=0) * P(Y=0) + P(X|Y=1) * P(Y=1)
and P(X|Y=0) * P(Y=0) = P(X_1|Y=0) * P(X_2|Y=0)* P(Y = 0)

$$=(\theta_{x1}^{y=0})^{X1}(1-\theta_{x1}^{y=0})^{1-X1}[\frac{1}{\sqrt{2\pi v0}}e^{-\frac{(x2-u0)^2}{2v_0^2}}]\,\theta_y$$

and similarly for P(X_1|Y=1) * P(X_2|Y=1)* P(Y = 1)

Problem 6: (3 points) What are the weights w_0, w_1, and w_2 for the perceptron whose decision surface is illustrated below? You should assume that the decision surface crosses the X_1 axis at -1 and crosses the X_2 axis at 2.



W_0= -2

W_1= -2

W_2= 1

(or proportional to the above parameters)

Problem 7: (3 points) Suppose you are running a learning experiment on a new algorithm for Boolean classification. You have a data set consisting of 100 positive and 100 negative examples, each with k discrete features. You plan to use leave-one-out cross validation (ie cross validation based on a partition into 200 singleton sets). As a baseline, you decide to compare your algorithm to a simple majority classifier (ie one which predicts whichever class was found to be most common in the training data, choosing randomly in the case of a tie, regardless of the input features). You expect the majority classifier to do about 50% on leave-one-out cross validation, but instead it performs very differently. What does it do and why? (Be concise)

For each run, the majority label of the training data will be different from the label of the validation data instance. For example, if the validation data instance is negative, the majority label of the training data (100 positives and 99 negatives) will be positive. Therefore, the accuracy will always be 0% for all runs.

Problem 8: (3 points) Your friend, Joe, pulled you aside after class. "In order to reduce overfitting, I added a depth bound to my decision-tree code from PS1. It seems to work much better when I test it on the dataset from PS2" he said. "But I'm not sure why it's helping; what do you think?" Pick one of the following:

_______Having a depth bound reduces bias but increases variance.

_______Having a depth bound reduces bias and also variance.

___x___Having a depth bound increases bias but reduces variance.

_______Having a depth bound increases both bias and variance.

_______Having a depth bound changes neither bias nor variance. Something else is going on.

Problem 9: (3 points) Which of the following statements about regularization are true? Check all that apply.

___x__ Using too large a value of λ can cause your hypothesis to underfit the data.

______Because regularization causes $J(\theta)$ to no longer be convex, gradient descent may not always converge to the global minimum (when $\lambda>0$, and when using an appropriate learning rate α).

______Using too large a value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ.

______Using a very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems (such as instability) during gradient descent.