10-601 Machine Learning, Fall 2011: Homework 3

Machine Learning Department
Carnegie Mellon University

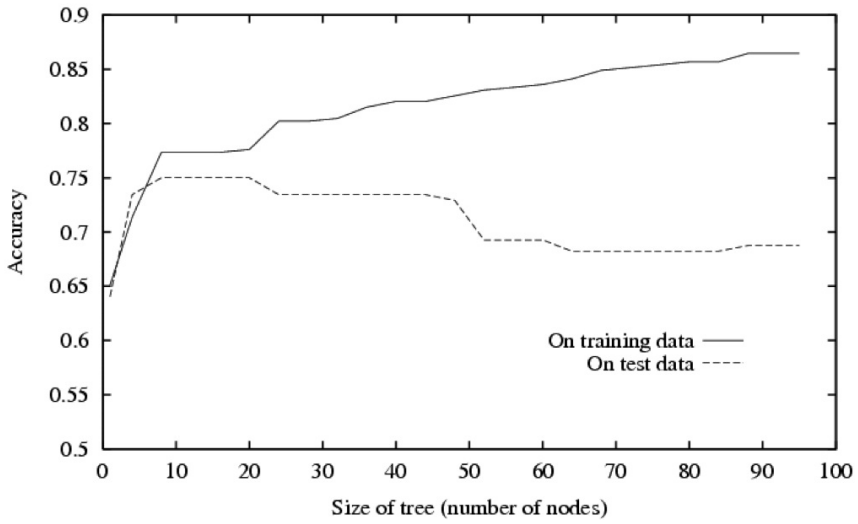
Due: October 17, 5 PM

Instructions There are 3 questions on this assignment. Please submit your completed homework to Sharon Cavlovich (GHC 8215) by 5pm, Monday, October 17. Submit your homework as 3 **separate** sets of pages, one for each question (so the TA's can easily split it up for grading). Include your name and email address on each set.

1 Short Questions [Shing-hon Lau, 10 points]

Here are some short questions to check your basic understanding of course material.

1. [2 pts] True or False? If we train a Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions, then it will achieve zero *training error* over these training examples. Please justify your answer in one sentence.
2. [2 pts] Prove that $P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$. (*Hint:* This is a two-line proof.)
3. [2 pts] True or False? After we train a logistic regression classifier, we can translate its learned weights W into the parameters of an equivalent GNB classifier for which we assume $\sigma_{ik} = \sigma_i$. Give a precise *one sentence* justification for your answer.
4. [4 pts] Consider the plot below showing training and test set accuracy for decision trees of different sizes, using the same set of training data to train each tree. Describe in one sentence how the training data curve (solid line) will change if the *number of training examples* approaches infinity. In a second sentence, describe what will happen to the test data curve under the same condition.



2 Sources of Error [Mladen Kolar, 30 points]

1. Suppose that we are given an independent and identically distributed sample of n points $\{y_i\}$ where each point $y_i \sim \mathcal{N}(\mu, 1)$ is distributed according to a normal distribution with mean μ and variance 1. You are going to analyze different estimators of the mean μ.

   (a) [5 points] Suppose that we use the estimator $\hat{\mu} = 1$ for the mean of the sample, ignoring the observed data when making our estimate. Give the bias and variance of this estimator $\hat{\mu}$. Explain in a sentence whether this is a good estimator in general, and give an example of when this is a good estimator.

   (b) [4 points] Now suppose that we use $\hat{\mu} = y_1$ as an estimator of the mean. That is, we use the first data point in our sample to estimate the mean of the sample. Give the bias and variance of this estimator $\hat{\mu}$. Explain in a sentence or two whether this is a good estimator or not.

   (c) [4 points] In the class you have seen the relationship between the MLE estimator and the least squares problem. Sometimes it is useful to use the following estimate

   $$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2 + \lambda \mu^2$$

   for the mean, where the parameter $\lambda > 0$ is a known number. The estimator $\hat{\mu}$ is biased, but has lower variance than the sample mean $\bar{\mu} = n^{-1} \sum_i y_i$ which is an unbiased estimator for μ. Give the bias and variance of the estimator $\hat{\mu}$.

2. In class we discussed the fact that machine learning algorithms for function approximation are also a kind of estimator (of the unknown target function), and that errors in function approximation arise from three sources: bias, variance, and unavoidable error. In this part of the question you are going to analyze error when training Bayesian classifiers.

   Suppose that Y is boolean, X is real valued, $P(Y = 1) = 1/2$ and that the class conditional distributions $P(X|Y)$ are uniform distributions with $p(X|Y = 1) = \text{uniform}[1, 4]$ and $p(X|Y = 0) = \text{uniform}[-4, -1]$. (we use uniform[a,b] to denote a uniform probability distribution between a and b, with zero probability outside the interval [a,b]).

   (a) [1 point]. Plot the two class conditional probability distributions $p(X|Y = 0)$ and $p(X|Y = 1)$.

   (b) [4 points]. What is the error of the optimal classifier? Note that the optimal classifier knows $P(Y = 1)$, $p(X|Y = 0)$ and $p(X|Y = 1)$ perfectly, and applies Bayes rule to classify new examples. Recall that the error of a classifier is the probability that it will misclassify a new x drawn at random from $p(X)$. The error of this optimal Bayes classifier is the unavoidable error for this learning task.

   (c) [5 points] Suppose instead that $P(Y = 1) = 1/2$ and that the class conditional distributions are uniform distribution with $p(X|Y = 1) = \text{uniform}[0, 4]$ and $p(X|Y = 0) = \text{uniform}[-3, 1]$. What is the unavoidable error in this case? Justify your answer.

   (d) [5 points] Consider again the learning task from part (a) above. Suppose we train a Gaussian Naive Bayes (GNB) classifier using n training examples for this task, where $n \to \infty$. Of course our classifier will now (incorrectly) model $p(X|Y)$ as a Gaussian distribution, so it will be biased: it cannot even represent the correct form of $p(X|Y)$ or $P(Y|X)$.

   Draw again the plot you created in part (a), and add to it a sketch of the learned/estimated class conditional probability distributions the classifier will derive from the infinite training data. Write down an expression for the error of the GNB. (hint: your expression will involve integrals - please don't bother solving them).

   (e) [2 points]. So far we have assumed infinite training data, so the only two sources of error are bias and unavoidable error. Explain in one sentences how your answer to part (d) above would change if the number of training examples was finite. Will the error increase or decrease? Which of the three possible sources of error would be present in this situation?

3 Bayes Nets [William Bishop, 30 points]

1. (a) [6 points]. Please draw the directed graph corresponding to the following distribution:

$$P(A, B, C, D, E, F) = P(A)P(B)P(C)P(D|A)P(E|A)P(F|B, D)P(G|D, E)$$

(b) [6 points]. Please write down the factored joint distribution represented by the graph below.

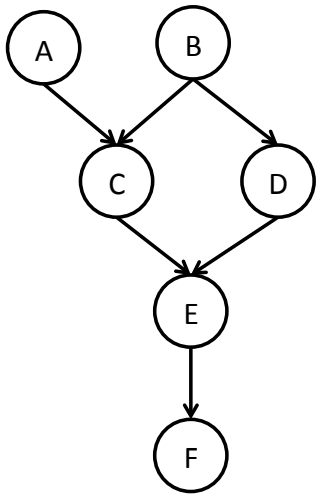


Figure 1: Bayes net for question parts (b) and (c).

(c) [6 points]. Assume the random variables in the graph shown above are Boolean. How many parameters are needed in total to fully specify this Bayesian network? Justify your answer.

(d) [12 points]. Based on the graph shown in part (b), state wether the following are true or false:

i. $A \perp\!\!\!\perp B$
ii. $A \perp\!\!\!\perp B|C$
iii. $C \perp\!\!\!\perp D$
iv. $C \perp\!\!\!\perp D|E$
v. $C \perp\!\!\!\perp D|B, F$
vi. $F \perp\!\!\!\perp B$
vii. $F \perp\!\!\!\perp B|C$
viii. $F \perp\!\!\!\perp B|C, D$
ix. $F \perp\!\!\!\perp B|E$
x. $A \perp\!\!\!\perp F$
xi. $A \perp\!\!\!\perp F|C$
xii. $A \perp\!\!\!\perp F|D$