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.

   ★ **SOLUTION:** This statement is false since there will still be unavoidable error. If the true probability of $P(X_1 = 1, X_2 = 1|Y = 0) = 0.1$ and $P(X_1 = 1, X_2 = 1|Y = 1) = 0.2$, then we will predict $Y = 1$ if we see $X_1 = 1, X_2 = 1$. However, we will misclassify points that have $X_1 = 1, X_2 = 1, Y = 0$.

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.)

   ★ **SOLUTION:** $P(X_1|X_2)P(X_2) = P(X_1 \wedge X_2) = P(X_2|X_1)P(X_1)$

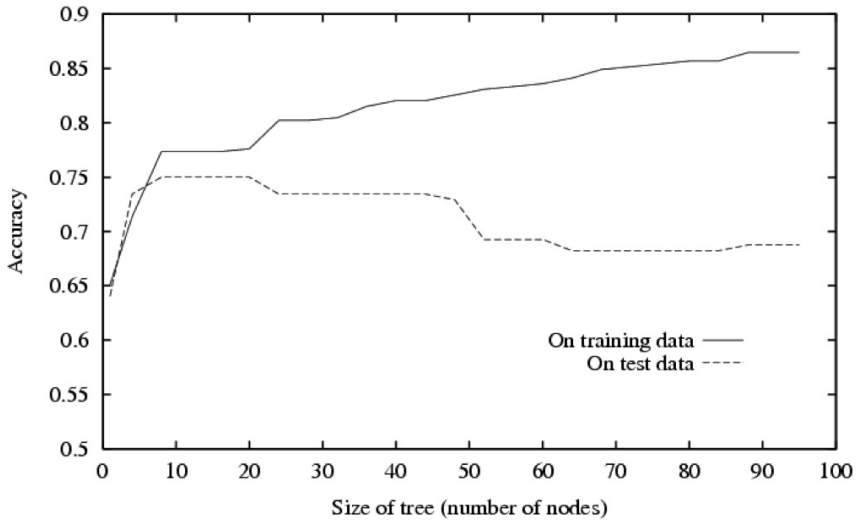
3. [2 pts] True or False? After we train a logistic regression classifier, we can translate its learned weights W into the parameters of an equivalent GNB classifier for which we assume $\sigma_{ik} = \sigma_i$. Give a precise *one sentence* justification for your answer.

   ★ **SOLUTION:** This is true. Logistic regression produces a linear classification boundary and Gaussian Naive Bayes (with the equivalent variance assumption) is capable of producing any linear classification boundary. From a more mathematical perspective, we noted in the class slides (9-29-2011 lecture, page 10) that we can take a GNB classifier (with the variance assumption) and translate its parameters into the parameters of a logistic regression classifier. We saw here that setting

   $$w_i = \sum_i (\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2})$$

   produces a logistic regression classifier that is identical to the original GNB classifier. To go in the opposite direction, choose $\sigma_i = 1$ for convenience. Then, we need only choose μ_{i0} and μ_{i1} such that the expression on the righthand-side is w_i. This can always be done since we have two parameters we can choose.

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.



★ **SOLUTION:** The training data curve will go down. The test data curve will go up. With infinite data, we would actually expect these two curves to be on top of each other. The training data curve will go down since it becomes harder and harder to overfit to statistical concidences in the data as the amount of training data increases. Thus, our training accuracy will decrease. Once we have infinite data, all of the statistical concidences no longer exist. The test data curve will increase since we learn a better classifier (i.e., one that isn't overfit to the training data).

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.

   ★ **SOLUTION:** The bias of an estimator is defined as $E[\hat{\mu}] - \mu$. Since we have that $E[\hat{\mu}] = 1$, the bias is $1 - \mu$.

   The variance of an estimator is defined as $\mathrm{Var}(\hat{\mu}) = E[(\hat{\mu} - E[\hat{\mu}])^2]$. Therefore, plugging in $\hat{\mu} = 1$, we have that $\mathrm{Var}(\hat{\mu}) = 0$.

   This is not a good estimator, since the bias is large when the true value of μ is not 1. Usually we don't have any information about the true value of μ, so it is unreasonable to assume it is equal to 1.

   (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.

★ **SOLUTION:** We have $E[\hat{\mu}] = \mu$. Therefore, the bias is 0. This is an unbiased estimator.
The variance of this estimator is $\text{Var}(\hat{\mu}) = \text{Var}(y_1) = 1$.
Although this estimator is unbiased, this is not a good estimator since its variability does not decrease with the sample size. For example, variance of the sample mean decreases as $1/n$.

(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 $\hat{\mu}$.

★ **SOLUTION:** First, we need to find a closed form for $\hat{\mu}$. Taking the derivative of the objective with respect to μ and setting it equal to zero, we obtain that

$$-2\sum_{i=1}^{n}(y_i - \mu) + 2\lambda\mu = 0.$$

Solving for μ we have that

$$\hat{\mu} = \frac{1}{n+\lambda}\sum_i y_i = \frac{n}{n+\lambda}\bar{\mu}.$$

For this estimator

$$E[\hat{\mu}] = \frac{1}{n+\lambda}E[\sum_i y_i] = \frac{n}{n+\lambda}\mu.$$

Therefore

$$\text{bias} = \frac{-\lambda\mu}{n+\lambda}.$$

The bias decreases to 0 with the sample size.
For variance, we have that

$$\text{Var}(\hat{\mu}) = \text{Var}(\frac{1}{n+\lambda}\sum_i y_i) = \frac{1}{(n+\lambda)^2}\sum_i \text{Var}(y_i) = \frac{n}{(n+\lambda)^2}.$$

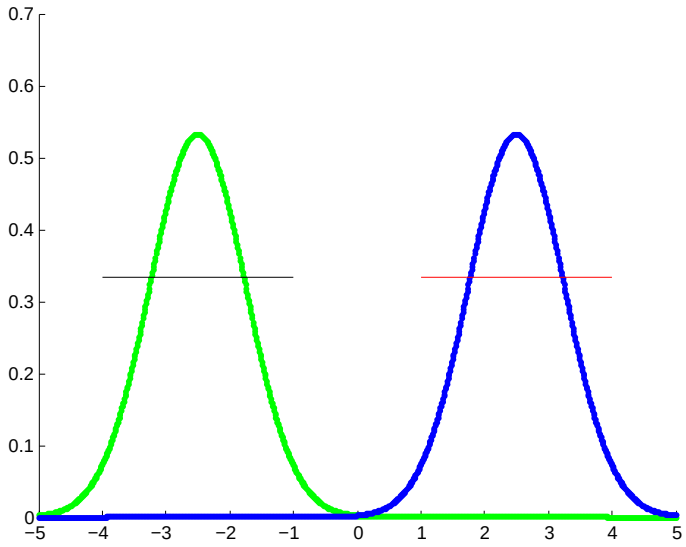
Comparing to the variance of the sample mean, which is $1/n$, we observe that the variance of $\hat{\mu}$ is smaller.

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 $p(X|Y = 1)$.

★ **SOLUTION:**



The black line denotes $P[X|Y=0]$ and the red line denotes $P[X|Y=1]$. Green and blue lines denote $\hat{P}[X|Y=0]$ and $\hat{P}[X|Y=1]$, approximations using the Gaussian distribution in part d.

(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.

★ **SOLUTION:** The error of this classifier is equal to 0. We observe that supports of $p(X|Y=0)$ and $p(X|Y=1)$ do not overlap. Therefore, we can perfectly classify a new example, just by looking whether it is in the interval $[-4,-1]$ or in the interval $[1,4]$.

(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.

★ **SOLUTION:** In this case we will make an error if $x \in [0,1]$. An error of a perfect classifier in when $x \in [0,1]$ is equal to $1/2$. Therefore,

$$\begin{aligned} P[\text{error}] &= P[x \in [0,1]] * P[\text{error}|x \in [0,1]] \\ &= (P[x \in [0,1]|y=0]P[y=0] + P[x \in [0,1]|y=1]P[y=1]) * P[\text{error}|x \in [0,1]] \\ &= (\frac{1}{4}\frac{1}{2} + \frac{1}{4}\frac{1}{2}) * \frac{1}{2} \\ &= \frac{1}{8} \end{aligned}$$

(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).

★ **SOLUTION:** Given that we have infinite amount of data, we can compute $E[X|Y=0]=-2.5$ and $\text{Var}[X|Y=0]=3/4$ (using formula for variance of the uniform distribution) and $E[X|Y=1]=2.5$ and $\text{Var}[X|Y=1]=3/4$. Since we are approximating $p(X|Y)$ with the Normal distribution, we have that $\hat{p}(X|Y=0)=N(-2.5,0.75)$ and $\hat{p}(X|Y=1)=N(2.5,0.75)$.

Using this, we have that for $x<0$, $\hat{p}(X|Y=0) > \hat{p}(X|Y=1)$ and for $x>0$, we have that $\hat{p}(X|Y=0) < \hat{p}(X|Y=1)$. Therefore, the classifier will make no error when classifying new points. This example illustrates that even with incorrect model assumption, we can perform well when classifying examples.

(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?

★ **SOLUTION:** Given finite amount of data we will not perfectly learn mean and variance of $p(X|Y)$. Therefore, the error of the classifier will increace given finite amount of data. We would have bias and error in this situation.

3 Bayes Nets [William Bishop, 30 points]

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

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

Answer:

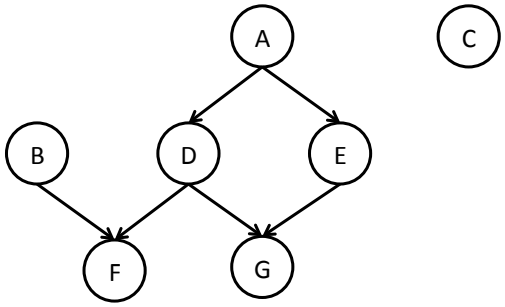


Figure 1: Answer to part (a).

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

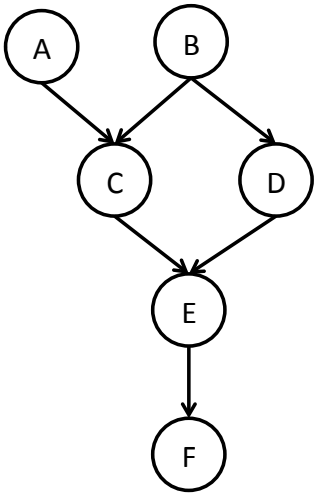


Figure 2: Bayes net for question parts (b) and (c).

Answer: $P(A)P(B)P(C|A, B)P(D|B)P(E|C, D)P(F|E)$

(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.

Answer: In this solution we make use of the fact that probabilities must sum to 1. For example, if we have a parameter for $P(A = \text{true})$, we don't need another parameter for $P(A = \text{false})$ since $P(A = \text{false}) = 1 - P(A = \text{true})$. This same principle holds for conditional distributions. Using this principle, see below for a detailed breakdown, but in total, we nee 14 parameters.

- 1 for $P(A)$
- 1 for $P(B)$
- 4 for $P(C|A,B)$ - this is 1 parameter for each combination of values A and B can take on.
- 2 for $P(D|B)$ - this is 1 parameter for each value B can take on
- 4 for $P(E|C,D)$ - this is 1 parameter for each combination of values C and D can take on.
- 2 for $P(F|E)$ - this is 1 parameter for each combination of values E can take on.

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

Answers indicated in bold font.

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