Solution to 10-701/15-781 Midterm Exam
Fall 2004

– with solutions blacked out so you can practice.

1 Introductory Probability and Statistics (12 points)

(a) *(2 points)* If A and B are disjoint events, and $Pr(B) > 0$, what is the value of $Pr(A|B)$?

(b) *(2 points)* Suppose that the p.d.f of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{4}{3}(1-x^3), & \text{for } 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Then $Pr(X < 0) =?$

(c) *(4 points)* Suppose that X is a random variable for which $E(X) = \mu$ and $Var(X) = \sigma^2$, and let c be an arbitrary constant. Which **one** of these statements is true:

A. $E[(X-c)^2] = (\mu - c)^2 + \sigma^2$

B. $E[(X-c)^2] = (\mu - c)^2$

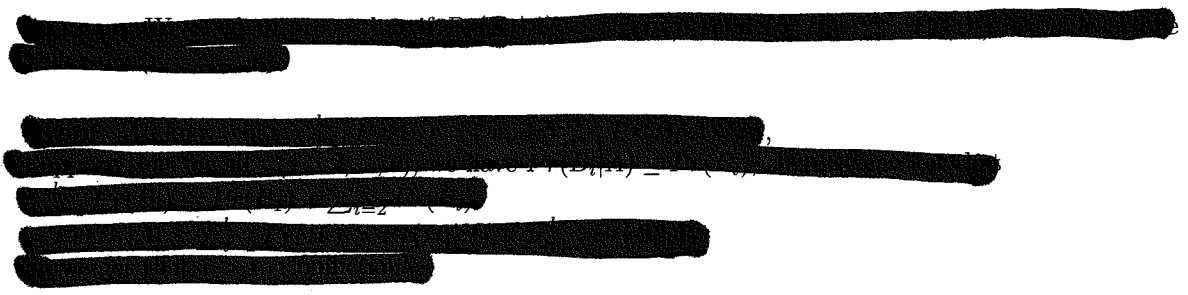
C. $E[(X-c)^2] = (\mu - c)^2 - \sigma^2$

D. $E[(X-c)^2] = (\mu - c)^2 + 2\sigma^2$

E. $E[(X-c)^2] = \mu^2 + c^2 + 2\sigma^2$

F. $E[(X-c)^2] = \mu^2 + c^2 - 2\sigma^2$

(d) *(4 points)* Suppose that k events $B_1, B_2, ..., B_k$ form a partition of the sample space S. For $i = 1, ..., k$, let $Pr(B_i)$ denote the prior probability of B_i. There is another event A that $Pr(A) > 0$. Let $Pr(B_i|A)$ denote the posterior probability of B_i given that the event A has occurred.
Prove that if $Pr(B_1|A) < Pr(B_1)$, then $Pr(B_i|A) > Pr(B_i)$ for at least one value of i $(i = 2, ..., k)$.
(Hint: one or more of these tricks might help: $P(B_i|A)P(A) = P(B_i \wedge A)$, $\sum_{i=1}^{k} P(B_i) = 1$, $\sum_{i=1}^{k} P(B_i|A) = 1$, $P(B_i \wedge A) + P(B_i \wedge \neg A) = P(B_i)$, $\sum_{i=1}^{k} P(B_i \wedge A) = P(A)$)



2 Linear Regression (12 points)

We have a dataset with R records in which the i^{th} record has one real-valued input attribute x_i and one real-valued output attribute y_i.

(a) *(6 points)* First, we use a linear regression method to model this data. To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set.
Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors? (No explanation required)

- Mean Training Error: A. Increase; B. Decrease

- Mean Testing Error: A. Increase; B. Decrease

[redacted]

(b) *(6 points)* Now we change to use the following model to fit the data. The model has one unknown parameter w to be learned from data.

$$y_i \sim N(log(wx_i), 1)$$

Note that the variance is known and equal to one. (no explanation required) Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?

A. $\sum_i x_i \log(wx_i) = \sum_i x_i y_i \log(wx_i)$

B. $\sum_i x_i y_i = \sum_i x_i y_i \log(wx_i)$

C. $\sum_i x_i y_i = \sum_i x_i \log(wx_i)$

D. $\sum_i y_i = \sum_i \log(wx_i)$

[redacted]

3 Decision Trees (11 points)

For this question, you're going to answer a couple questions regarding the dataset shown below. You'll be trying to determine whether Andrew finds a particular type of food appealing based on the food's temperature, taste, and size.

Appealing	Temperature	Taste	Size
No	Hot	Salty	Small
No	Cold	Sweet	Large
No	Cold	Sweet	Large
Yes	Cold	Sour	Small
Yes	H	Sour	Small
No	H	Salty	Large
Yes	H	Sour	Large
Yes	Cold	Sweet	Small
Yes	Cold	Sweet	Small
No	H	Salty	Large

(a) *(3 points)* What is the initial entropy of *Appealing*?

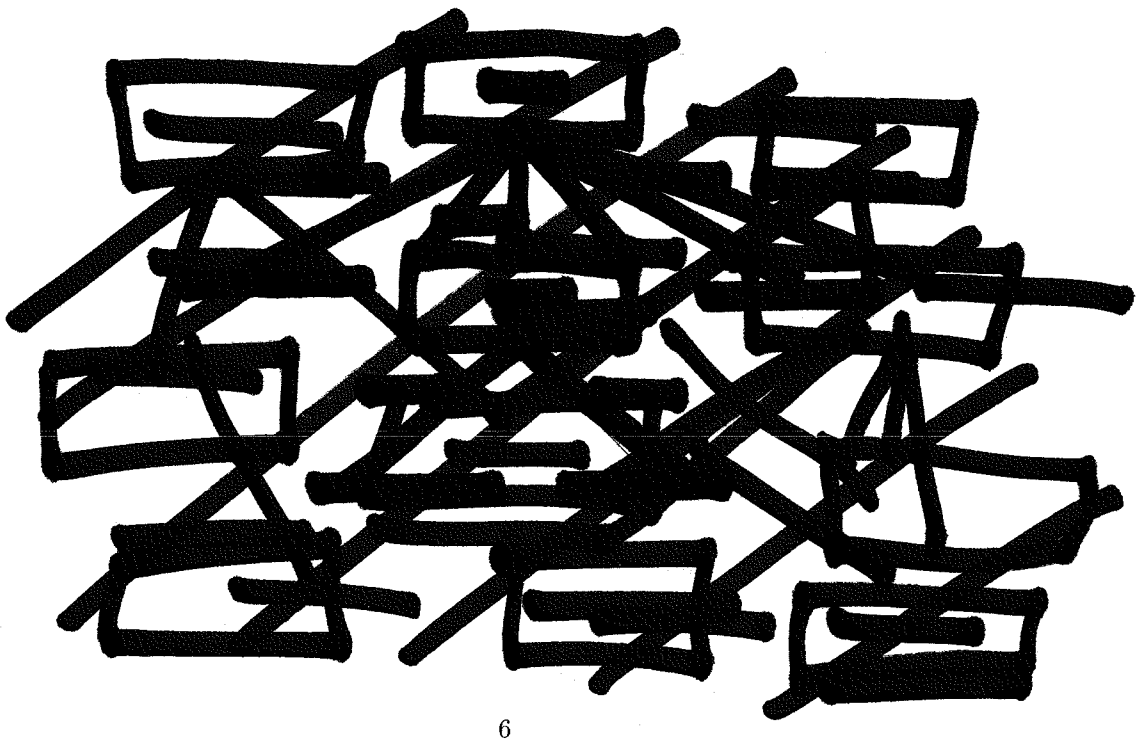
[redacted]

(b) *(3 points)* Assume that *Taste* is chosen for the root of the decision tree. What is the information gain associated with this attribute?

[redacted]

(c) *(5 points)* Draw the full decision tree learned for this data (without any pruning).

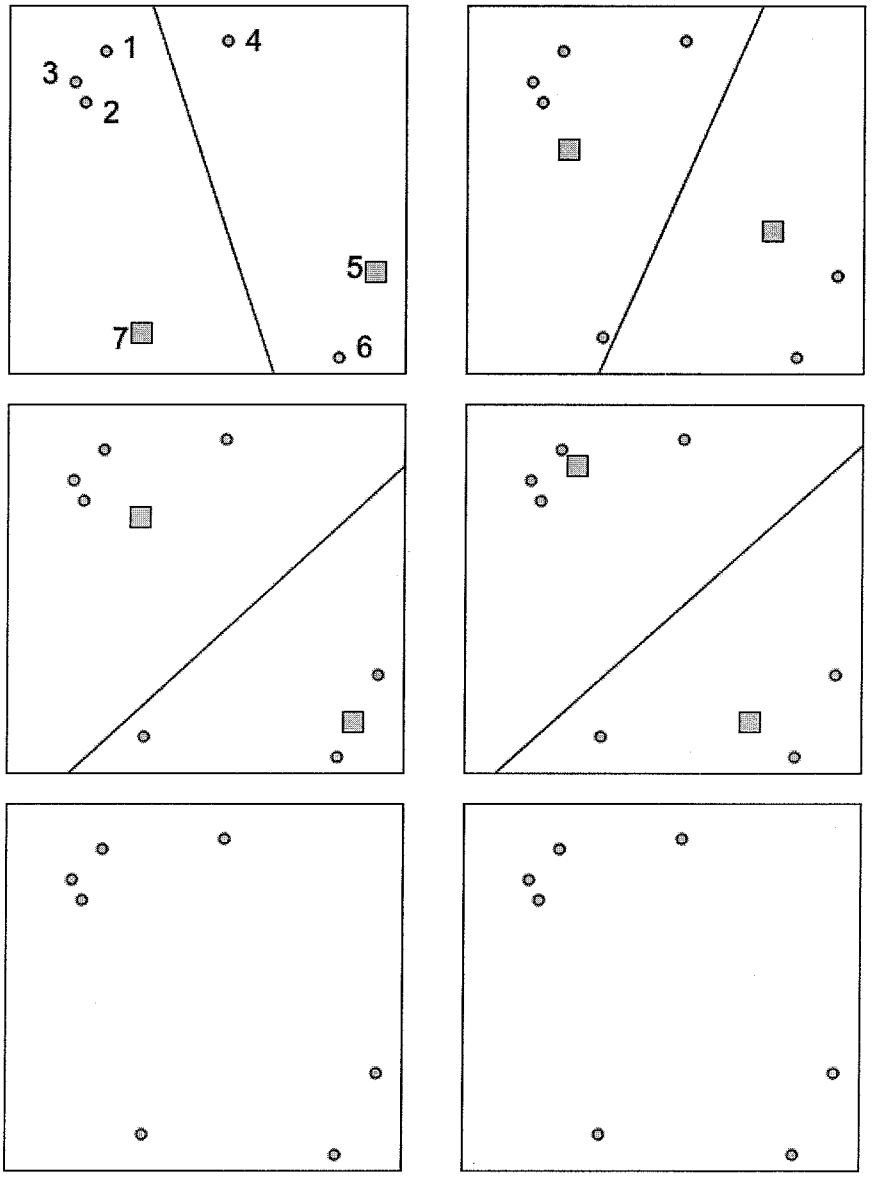
[redacted]



4 K-means and Hierarchical Clustering (10 points)

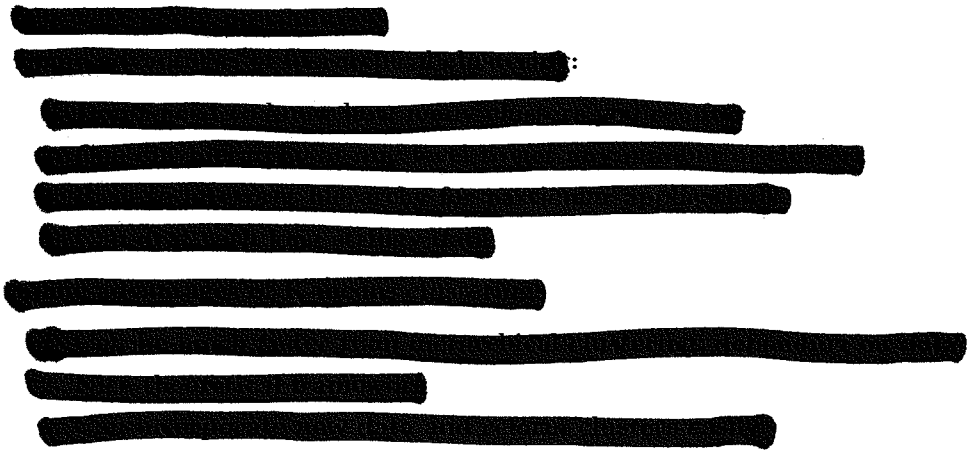
(a) *(6 points)* Perform K-means on the dataset given below. Circles are data points and there are two initial cluster centers, at data points 5 and 7. Draw the cluster centers (as squares) and the decision boundaries that define each cluster. If no points belong to a particular cluster, assume its center does not change. Use as many of the pictures as you need for convergence.

Answer:



Sorry - I can't disguise the answers!

(b) *(4 points)* Give one advantage of hierarchical clustering over K-means clustering, and one advantage of K-means clustering over hierarchical clustering.

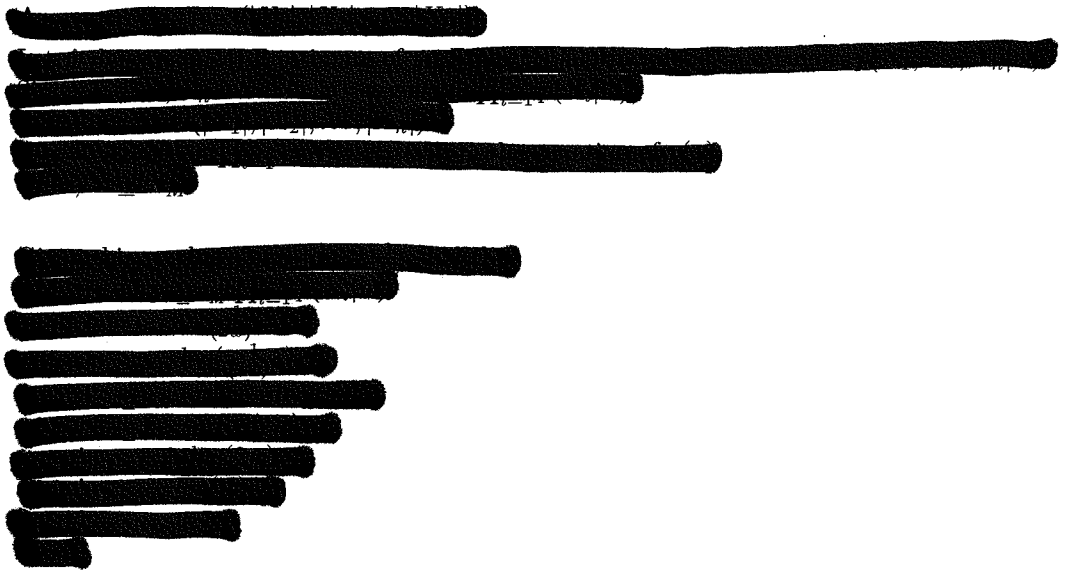


5 Maximum Likelihood Estimates (9 points)

(a) *(9 points)* Suppose $X_1, \ldots, X_n$ are iid samples from $U(-w, w)$ That is,

$$p(x) = \begin{cases} 0, & x < -w \\ \frac{1}{2w}, & -w \leq x \leq w \\ 0, & x > w \end{cases}$$

Write down a formula for an MLE estimate of w.

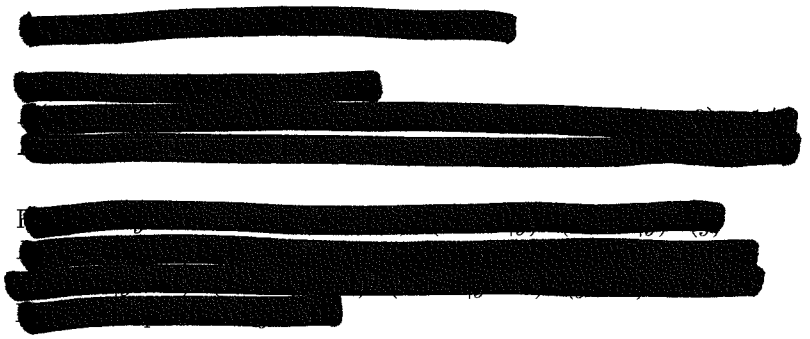


6 Bayes Classifiers (10 points)

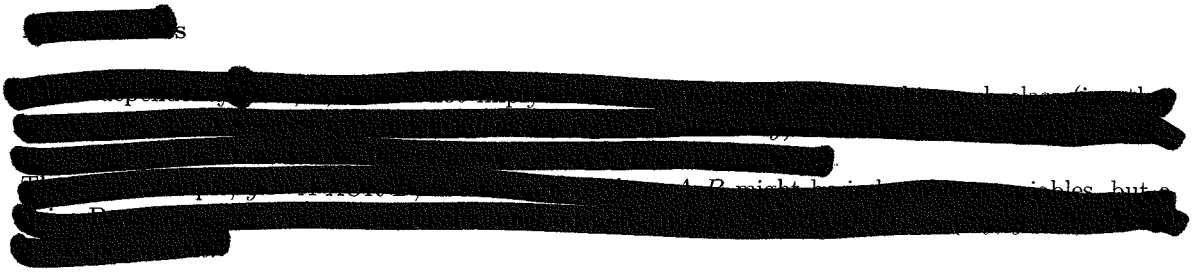
Suppose we are given the following dataset, where A, B, C are input binary random variables, and y is a binary output whose value we want to predict.

A	B	C	y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

(a) *(5 points)* How would a **naive** Bayes classifier predict y given this input: $A = 0, B = 0, C = 1$. Assume that in case of a tie the classifier always prefers to predict 0 for y.

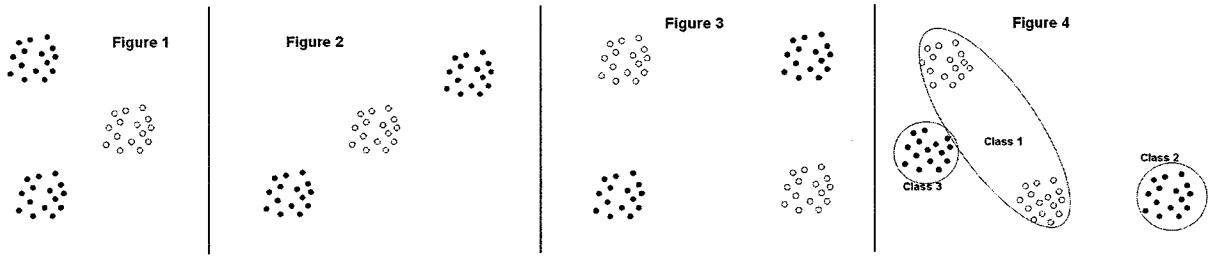


(b) *(5 points)* Suppose you know for fact that A, B, C are independent random variables. In this case is it possible for any other classifier (e.g., a decision tree or a neural net) to do better than a naive Bayes classifier? (The dataset is irrelevant for this question)



7 Classification (12 points)

Figures 1, 2 and 3 present points from two different clusters: A (solid points) and B (hollow points). We would like to learn a classifier that achieves zero training error on this data. To do that we allow each classifier to divide the data into more than two classes, however, for each classifier there must be a subset of the classes that perfectly match class A and the complementary set of classes must match cluster B. For example, in Figure 4 classes 2 and 3 contain all of A's points and class 1 contains all of B's points and so this classification is a legitimate solution to this problem.



(a) *(6 points)* For a Gaussian Bayes classifier and for each of the three figures state the **minimum** number of classes required to achieve the above goal. For all figures you can assume equal class priors, that is P(A) = P(B).

	minimum number of classes
Figure 1	
Figure 2	
Figure 3	
Figure 4	3

(b) *(6 points)* For the following figures, do we need a full covariance matrix for the classification or would a diagonal covariance matrix be enough

Figure 2? **Answer: Diagonal is enough. Note that the variance of the two clusters is different. A has a large variance for both the x and the y axis while B's variance is low in both direction. Thus, even though both have the same mean, the variance terms are enough to separate them.**

8 Neural Nets and Regression (12 points)

Suppose we want to learn a quadratic model:

$$
\begin{array}{lllllllllll}
y = & w_0 & + & w_1 x_1 & + & w_2 x_2 & + & w_3 x_3 & + & \ldots & w_k x_k & + \\
 & & & w_{11} x_1^2 & + & w_{12} x_1 x_2 & + & w_{13} x_1 x_3 & + & \ldots & w_{1k} x_1 x_k & + \\
 & & & & & w_{22} x_2^2 & + & w_{23} x_2 x_3 & + & \ldots & w_{2k} x_2 x_k & + \\
 & & & \vdots & & & & \vdots & & & \vdots & \\
 & & & & & & & & w_{k-1,k-1} x_{k-1}^2 & + & w_{k-1,k} x_{k-1} x_k & + \\
 & & & & & & & & & + & w_{k,k} x_k^2 &
\end{array}
$$

Suppose we have a fixed number of records and k input attributes.

(a) *(6 points)* In big-O notation what would be the computational complexity in terms of k of learning the MLE weights using matrix inversion?

[redacted]

(b) *(6 points)* What would be the computational complexity of one iteration of gradient descent? (The "batch" gradient descent method, NOT the online method).

[redacted]

(Aside note: all this [redacted] censoring is fun! I can see how secretive regimes can go out of control!)