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

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 $Pr(A|B)$?
Answer: 0 (Note that: $A \wedge B = \emptyset$)

(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 $Pr(X < 0) =?$
Answer: 0 (Note that: $0 \leq x \leq 1$)

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

Answer: A

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

(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: $P(B_i|A)P(A) = P(B_i \wedge A)$, $\sum_{i=1}^{k} P(B_i) = 1$, $\sum_{i=1}^{k} P(B_i|A) = 1$, $P(B_i \wedge A) + P(B_i \wedge \neg A) = P(B_i)$, $\sum_{i=1}^{k} P(B_i \wedge A) = P(A)$)

Answer: We need to 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 ($i = 2, ..., k$).

Proof: We know that $\sum_{i=1}^{k} Pr(B_i) = 1$ and $\sum_{i=1}^{k} Pr(B_i|A) = 1$,
Suppose that for all i ($i = 2, ..., k$), we have $Pr(B_i|A) \leq Pr(B_i)$, then we can get that
$\sum_{i=1}^{k} Pr(B_i) = Pr(B_1) + \sum_{i=2}^{k} Pr(B_i)$
$> Pr(B_1|A) + \sum_{i=2}^{k} Pr(B_i) > Pr(B_1|A) + \sum_{i=2}^{k} Pr(B_i|A)$
So we get that $1 > 1$. Confliction!.

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

Answer:

The training error tends to increase. As more examples have to be fitted, it becomes harder to 'hit', or even come close, to all of them.

The test error tends to decrease. As we take into account more examples when training, we have more information, and can come up with a model that better resembles the true behavior. More training examples lead to better generalization.

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

Answer: D.

Very similar with the problem 4 in our homework2, we perform Maximum Likelihood estimation.

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

We could write the log likelihood as:

$$LL = log(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_i - log(wx_i))^2}{2\sigma^2})) = \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}} exp(-\frac{(y_i - log(wx_i))^2}{2}))$$

$$\frac{\partial LL}{\partial w} = 0 \Rightarrow \frac{\partial \sum_{i=1}^{n} (y_i - log(wx_i))^2}{\partial w} = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{x_i}{wx_i} * (y_i - log(wx_i)) = 0 \Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} log(wx_i)$$

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*?

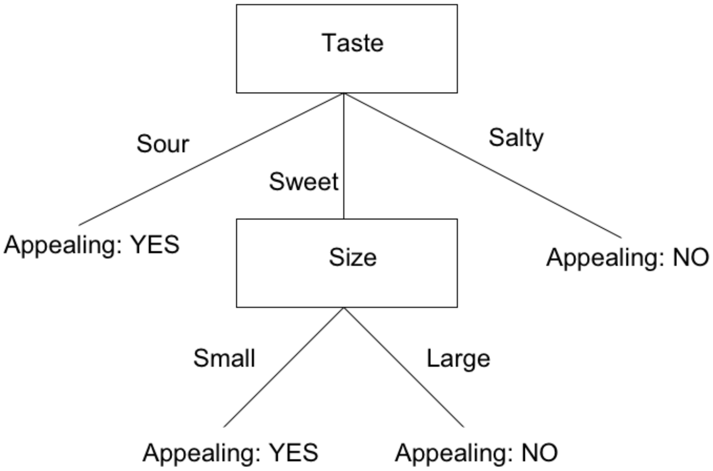
Answer: 1. $(-5/10 \cdot log(5/10) - 5/10 \cdot log(5/10))$**.**

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

Answer: 3/5. $(1 - 4/10 \cdot (2/4 \cdot log(2/4) + 2/4 \cdot log(2/4)) - 6/10 \cdot 0)$**.**

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

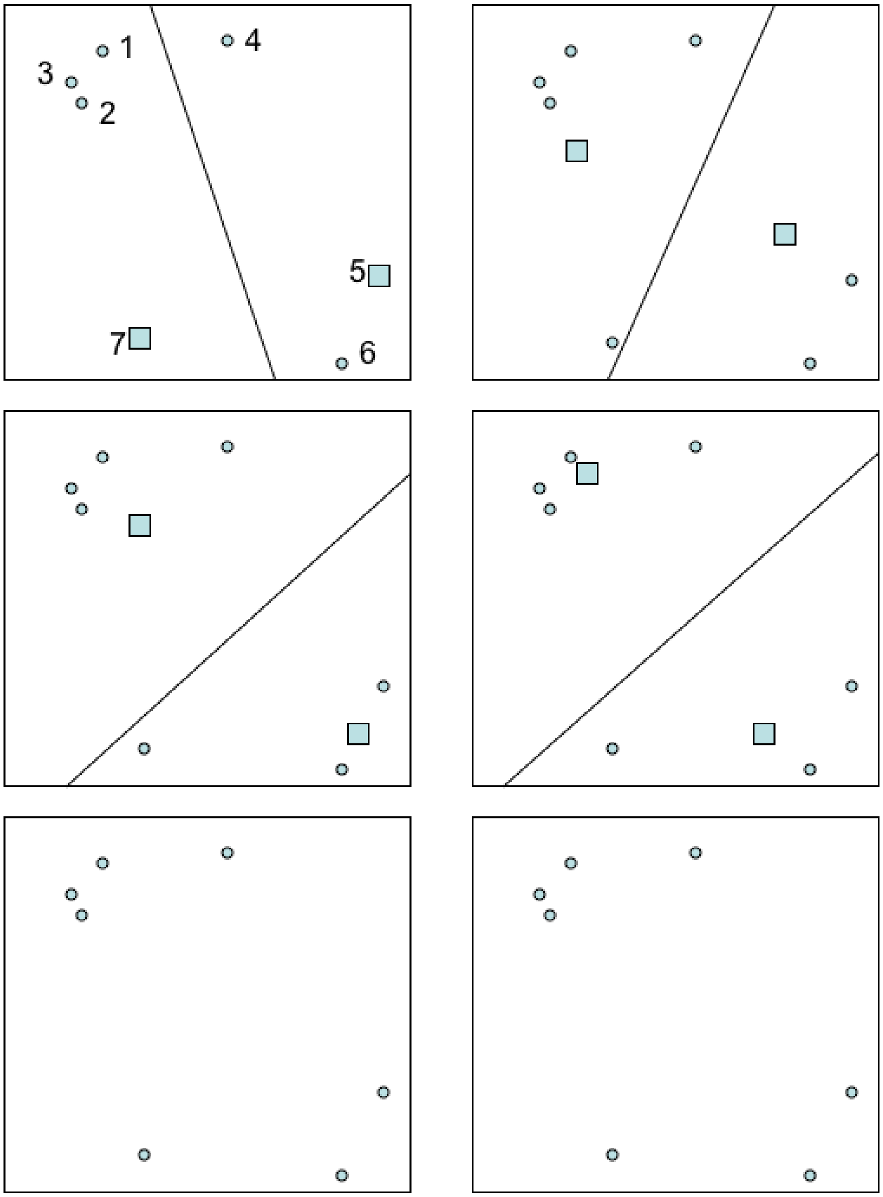
Answer:



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:



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

Answer: Many possibilities.

Some advantages of hierarchical clustering:

1. **Don't need to know how many clusters you're after**
2. **Can cut hierarchy at any level to get any number of clusters**
3. **Easy to interpret hierarchy for particular applications**
4. **Can deal with long stringy data**

Some advantages of K-means clustering:

1. **Can be much faster than hierarchical clustering, depending on data**
2. **Nice theoretical framework**
3. **Can incorporate new data and reform clusters easily**

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.

Answer: $\hat{w} = \max(|X_1|, |X_2|, \ldots, |X_n|)$

Let $\hat{w}$ denote an MLE estimate of w. From MLE principle $\hat{w} = \arg\max_w p(X_1, \ldots, X_n|w)$.
Since $X_1, \ldots, X_n$ are iid: $\hat{w} = \arg\max_w \prod_{i=1}^n p(X_i|w)$.
Let $X_M = \max(|X_1|, |X_2|, \ldots, |X_n|)$.
If $w < X_M$, then $\prod_{i=1}^n p(X_i|w) = 0$ from the equation of $p(x)$.
Thus, $w \geq X_M$.

Given this, we have $p(X_i|w) = \frac{1}{2w}$, and thus
$\hat{w} = \arg\max_{w \geq X_M} \prod_{i=1}^n p(X_i|w)$
$= \arg\max_{w \geq X_M} \frac{1}{(2w)^n} =$
$\arg\max_{w \geq X_M} log(\frac{1}{(2w)^n}) =$
$\arg\max_{w \geq X_M} log1 - nlog(2w) =$
$\arg\max_{w \geq X_M} -nlog(2w) =$
$\arg\min_{w \geq X_M} nlog(2w) =$
$\arg\min_{w \geq X_M} log(w) =$
$\arg\min_{w \geq X_M} w =$
X_M

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.

Answer: The classifier will predict 1

$P(y = 0) = 3/7; P(y = 1) = 4/7$
$P(A = 0|y = 0) = 2/3; P(B = 0|y = 0) = 1/3; P(C = 1|y = 0) = 1/3$
$P(A = 0|y = 1) = 1/4; P(B = 0|y = 1) = 1/2; P(C = 1|y = 1) = 1/2$

Predicted y maximizes $P(A = 0|y)P(B = 0|y)P(C = 1|y)P(y)$
$P(A = 0|y = 0)P(B = 0|y = 0)P(C = 1|y = 0)P(y = 0) = 0.0317$
$P(A = 0|y = 1)P(B = 0|y = 1)P(C = 1|y = 1)P(y = 1) = 0.0357$
Hence, the predicted y is 1.

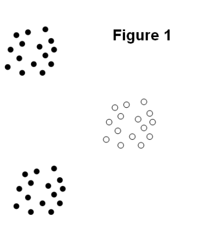
(b) *(5 points)* Suppose you know for fact that A, B, C are independent random variables. In this case is it possible for any other classifier (e.g., a decision tree or a neural net) to do better than a naive Bayes classifier? (The dataset is irrelevant for this question)

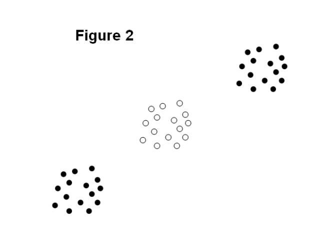
Answer: Yes

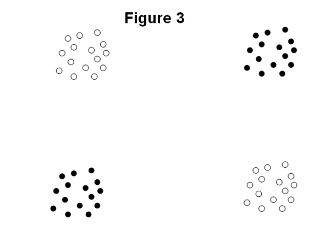
The independency of A, B, C does not imply that they are independent within each class (in other words, they are not necessarily independent when conditioned on y). Therefore, naive Bayes classifier may not be able to model the function well, while a decision tree might.
Thus, for example, $y = A$ XOR B, is an example where A, B might be independent variables, but a naive Bayes classifier will not model the function well since for a particular class (say, $y = 0$), A and B are dependent.

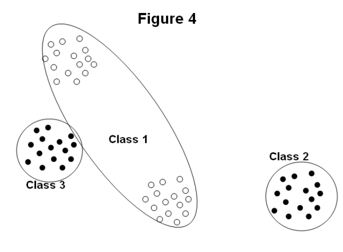
7 Classification (12 points)

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.





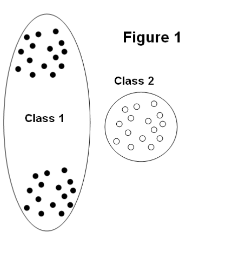


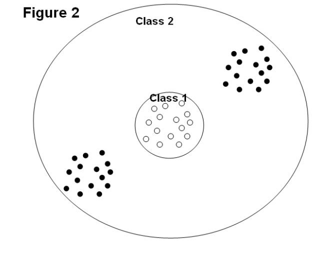


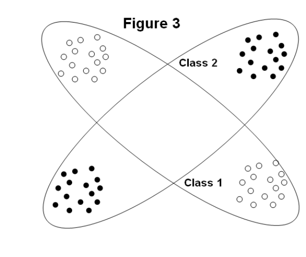
(a) *(6 points)* For a Gaussian Bayes classifier and for each of the three figures state the **minimum** number of classes required to achieve the above goal. For all figures you can assume equal class priors, that is P(A) = P(B).

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

Answer: The number of classes is 2 for all of the cases.







(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.**

Figure 3? **Answer: Full is required. In this case, both the mean and the variance terms are same for both clusters. The only difference is in the covariance terms.**

8 Neural Nets and Regression (12 points)

Suppose we want to learn a quadratic model:

$$\begin{array}{lllllllllllll}
y = & w_0 & + & w_1x_1 & + & w_2x_2 & + & w_3x_3 & + & & \dots & w_kx_k & + \\
 & & & w_{11}x_1^2 & + & w_{12}x_1x_2 & + & w_{13}x_1x_3 & + & & \dots & w_{1k}x_1x_k & + \\
 & & & & & w_{22}x_2^2 & + & w_{23}x_2x_3 & + & & \dots & w_{2k}x_2x_k & + \\
 & & & \vdots & & & & \vdots & & & & \vdots & \\
 & & & & & & & & & w_{k-1,k-1}x_{k-1}^2 & + & w_{k-1,k}x_{k-1}x_k & + \\
 & & & & & & & & & & + & w_{k,k}x_k^2 &
\end{array}$$

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?

Answer: $O(k^6)$

$O(k^6)$ since it is $O([\text{number of basis functions}]^3)$ to solve the normal equations, and the number of basis functions is $\frac{1}{2}(k+1)(k+2)$.

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

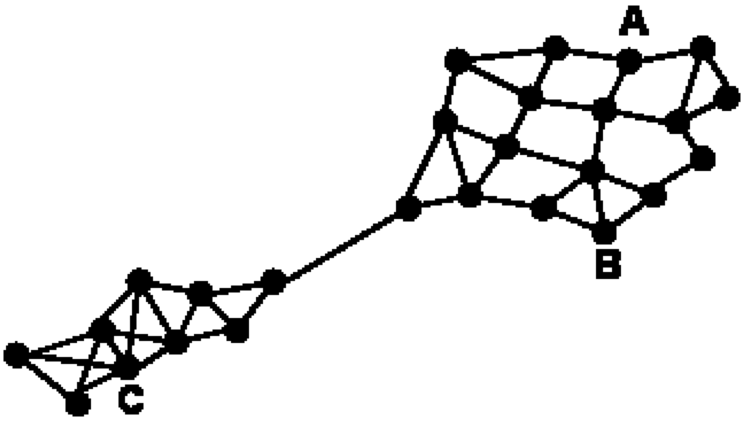
Answer: $O(k^2)$

$O(k^2)$ since work of computing δ_k for each datapoint involves $\frac{1}{2}(k+1)(k+2)$ operations and then there is one weight update for each weight.

Interesting note: If we had also included R as the number of records in the complexity then the answers are:
(a) $O(Rk^4 + k^6)$, where the first term is for building an X^TX matrix, and the second term is for matrix inversion.
(b) $O(Rk^2)$

9 Spectral clustering (12 points)

Consider the graph below. Let W be the distance matrix for this graph where $w_{i,j} = 1$ iff there is an edge between nodes i and j and otherwise $w_{i,j} = 0$. We will define the matrices D and P as we did in class by setting $D_{i,i} = \sum_j w_{i,j}$ and $P = D^{-1}W$. As we mentioned in class, P is the probability transition matrix for this graph. We denote by $P^t_{i,j}$ the i, j entry in the matrix P raised to the power of t.



For each of the expressions below, replace ? with either <, > or = and briefly explain your reasoning.

(a) *(3 points)* $P^{20}_{A,C}$? $P^{100}_{A,C}$

Answer: $P^{20}_{A,C} < P^{100}_{A,C}$
As the power of P increases it is more likely to transition to another cluster. Since A and C are in different clusters, it is more likely to end up in C when we take 100 steps than when we take 20 steps.

(b) *(3 points)* $P^{20}_{A,B}$? $P^{100}_{A,B}$

Answer: $P^{20}_{A,B} > P^{100}_{A,B}$
A and B are in the same cluster. It is more likely to stay in the same cluster when the power of P is low (few steps) than for higher powers of P (many steps).

(c) *(3 points)* $\sum_j P^{20}_{A,j}$? $\sum_j P^{100}_{A,j}$

Answer: $\sum_j P^{20}_{A,j} = \sum_j P^{100}_{A,j}$
P^t for any t is a probability transition matrix and so its rows always sum to 1.

(d) *(3 points)* $P^{\infty}_{B,A}$? $P^{\infty}_{B,C}$

Answer: $P^{\infty}_{B,A} < P^{\infty}_{B,C}$
At the limit, the point we end at is independent of the point we started at. Thus, we need to evaluate the (fixed) probability of ending in A vs. the probability of ending in C. In class, we have shown that this probability is proportional to the components of the first eigenvector of the symmetric matrix we

defined ($D^{-1/2}WD^{-1/2}$). In class (and in the problem set) we have derived the actual values for the entries of this vector. As we showed, these entries are the square root of the sum of the rows of W. Since in our case rows sum up to the out degree (or in degree) of the nodes, the probability that we will end up at a certain point is proportional to the connectivity of that point. Since C is connected to 5 other nodes whereas A is only connected to 3, $P^{\infty}_{B,A} < P^{\infty}_{B,C}$.