Problem Set 3

MAS 622J/1.126J: Pattern Recognition and Analysis

Due Monday, 16 October 2006

[Note: All instructions to plot data or write a program should be carried out using either Python accompanied by the `matplotlib` package or Matlab. Feel free to use either or both, but in order to maintain a reasonable level of consistency and simplicity we ask that you do not use other software tools.]

Problem 1: (DHS 2.6) Optimal Decision Boundaries

Your friend has built a system to recognize into which of two categories, ω_1 or ω_2, her advisor's email can be classified. She has brilliantly identified two features such that her training data is well approximated by two Gaussians:

$$p(\mathbf{x}|\omega_1) \sim \mathcal{N}(\mu_1, \Sigma_1)$$
$$p(\mathbf{x}|\omega_2) \sim \mathcal{N}(\mu_2, \Sigma_2)$$

where $\mu_1 = [8\ 9]^T$, $\mu_2 = [0\ 9]^T$, $\Sigma_1 = \mathrm{I}$, and $\Sigma_2 = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix}$, where I is the identity matrix.

a. Plot the one-sigma ellipses for these two classes in the place $\mathbf{x} = [x_1\ x_2]^T$.

b. Your friend finds that choosing a threshold at $x_1 = 4$ perfectly separates the training examples she has; thus, she proposes that this should be the best classifier. Show her an expression, in terms of $\mathbf{x}$, which can improve her classifier with respect to minimizing the Bayes probability of error. Assume that email from class ω_2 is twice as likely as email from class ω_1.

c. The shape of this optimal decision boundary is:

- a line
- a parabola
- a hyperbola
- a circle
- an ellipse
- none of the above (explain)

Be sure to justify your answer.

d. She tells you new information: if she accidently treats email from class ω_1 as if it came from ω_2, then this will be terrible for her career. She estimates this error will cost her ten times as much as the cost of misclassifying email from ω_2 as ω_1. She adds that there's no cost to choosing correctly in either case. Describe qualitatively how this changes your result above. Make a sketch of the change (it doesn't have to be precisely plotted). Justify your answer.

Problem 2: Reinforcement Learning

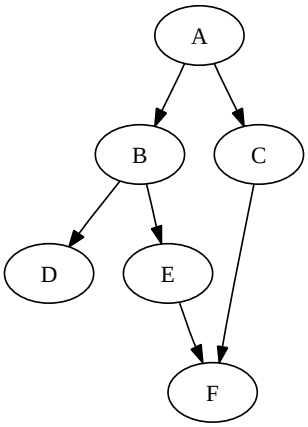
This problem is an online activity. Go to the following web page and play the reinforcement learning game:

`http://robotpatrol.media.mit.edu/mas622j/`

So that we can keep track of who completed the activity, you will receive credit for this problem if you complete the survey at the end of the activity. This game should probably take you about 45 minutes.

Problem 3: (DHS 2.11, 3.7) Bayes Nets and Computational Complexity

Use the following Bayes belief net over categorical variables when answering the parts of this problem:



a. Take advantage of the independence assumptions in the Bayes belief net in order to factor and simplify the following expression for $P(B = b)$:

$$P(B = b) = \sum_{a \in A} \sum_{c \in C} \sum_{d \in D} \sum_{e \in E} \sum_{f \in F} P(a, b, c, d, e, f)$$

b. For this part assume that you have a computer that can perform multiplication and addition operations in one time step with memory operations taking no time. Assume also that the number of categories for each variable (A, B, C, D, E, and F) is k. give the computational complexity "big oh" expression for the order of the original *unsimplified* expression for $P(B = b)$ in part (a) in terms of the number of addition and multiplication operations required to compute the answer. Also, give the numbers of additions and multiplications that are necessary.

c. Using the same computer and category number assumptions in part (b), give the "big oh" expression for the order of your *simplified* expression for $P(B = b)$, again, in terms of the number of addition and multiplication operations required to compute the answer. Also, give the numbers of additions and multiplications that are necessary.

d. Explain the complexity order relationship between part (b) and part (c). Why are they the same or different?

e. Now, in addition to the above Bayes belief net, you are told that there are only two categories for each variable, *true* and *false*. You are also given these specific values for the conditional relationships:

$$P(A) = 1/2$$

$$P(B|A) = \begin{cases} 2/3 & \text{if } A \\ 1/2 & \text{otherwise} \end{cases}$$

$$P(C|A) = \begin{cases} 1/4 & \text{if } A \\ 2/3 & \text{otherwise} \end{cases}$$

$$P(D|B) = \begin{cases} 3/5 & \text{if } B \\ 3/4 & \text{otherwise} \end{cases}$$

$$P(E|B) = \begin{cases} 1/3 & \text{if } B \\ 3/7 & \text{otherwise} \end{cases}$$

$$P(F|C,E) = \begin{cases} 2/3 & \text{if } C \text{ and } E \\ 5/6 & \text{if } C \text{ and not } E \\ 3/4 & \text{if not } C \text{ and } E \\ 1/2 & \text{otherwise} \end{cases}$$

Using these specific values for the conditional distributions, calculate a numerical value for $P(B)$.

f. Using the same specific values for the conditional distributions in part (e), you are told that $F =$ true. What is the new value of $P(B)$ given this evidence?

Problem 4: (DHS 3.8) PCA Dimensionality Reduction

Standard Principle Component Analysis (PCA) transforms the data using only the eigenvectors with the largest associated eigenvalues, thus reducing dimensionality.

a. Standard PCA is not generally considered optimal for pattern classification. Why not? What is the typical optimality criterion that is violated? Draw a picture or give an example to argue where PCA fails.

b. Let's explore a variation on standard PCA. Let R_1 and R_2 be correlation matrices for two classes of observation vectors. Thus $R_i = E[\mathbf{x}\mathbf{x}^T]$ for $\mathbf{x} \in \omega_i$. Now define $R = R_1 + R_2$, and let S be the linear transformation that whitens R. Define $R_1' = S^T R_1 S$ and $R_2' = S^T R_2 S$. Let $\mathbf{e}$ be an eigenvector of R_1' and let λ be the associated eigenvalue.

   i Write an expression for $S^T R S$ in terms of R_1', R_2', and the identity matrix I.

   ii Show that if $\mathbf{e}$ is an eigenvector of R_1' then it is also an eigenvector of R_2'. Write an expression for $R_2'\mathbf{e}$ in terms of λ, I, and $\mathbf{e}$. If λ is the associated eigenvalue for R_1' then what is the associated eigenvalue for R_2'? As the eigenvalues are ordered from largest to smallest for one class, what happens to the eigenvalues for the other class?

   iii Suppose that data $\mathbf{x}$ is linearly transformed so that $\mathbf{y} = Q^T S^T \mathbf{x}$, where each column of Q can be chosen to be an eigenvector of R_1 (and thus, via what you have shown above, also an eigenvector of R_2). Which eigenvectors are likely to be best to use for representing data from both ω_1 and ω_2? (Hint: Consider the relationship between the eigenvalues of R_1' and those of R_2'.)

c. On the course website there is a dataset that contains 1000 five-dimensional points that are samples from a five-dimensional probability distribution.

   i Use a program to calculate the first two principle components (PCs) of this dataset.

   ii Project the dataset onto these normalized vectors and graph the result.

   iii Whiten this two dimensional distribution and graph the result.

   As usual, label your axes for both graphs. Please include a printed copy of your program in your solution.