CSE555: Introduction to Pattern Recognition Spring, 2007

Mid-Term Exam (100 points, Closed book/notes)

The last page contains some formulas that might be useful.

1. Part(i) (10 pts)

Suppose a bank classifies customers as either good or bad credit risks. On the basis of extensive historical data, the bank has observed that 1% of good credit risks and 10% of bad credit risks overdraw their account in any given month. A new customer opens a checking account at this bank. On the basis of a check with a credit bureau, the bank believe that there is a 70% chance the customer will turn out to be a good credit risk.

- (a) (5 pts) Suppose that this customer's account is overdrawn in the first month. How does this alter the bank's opinion of this customer's creditworthiness?
- (b) (5 pts) Given (a), what would be the bank's opinion of the customer's creditworthiness at the end of the second month if there was not an overdraft in the second month?

Part(ii) (10 pts)

Consider a two-category classification problem with one-dimensional Gaussian distributions $p(x|w_i) \sim N(\mu_i, \sigma^2)$, i = 1, 2 (i.e. they have same variance but different means).

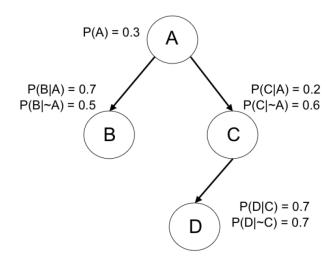
- (a) (3 pts) Sketch the two densities $p(x|w_1)$ and $p(x|w_2)$ in one figure.
- (b) (7 pts) Sketch the two posterior probabilities $P(w_1|x)$ and $P(w_2|x)$ in one figure, assuming same prior probabilities.

Part(iii) (5 pts)

Suppose there is a two-class one-dimensional problem with the Gaussian distributions: $p(x|w_1) \sim N(-1,1)$, and $p(x|w_2) \sim N(4,1)$ and equal prior probabilities. In order to find a good classifier, you are given as much training data as you would like and you are free to pick any method. What is the best error on test data that any learning algorithm can attain and why? (Try to keep your reasoning to two sentences or fewer.)

- (a) 0% error
- (b) more than 0% but less than 10%
- (c) more than 20%
- (d) can't tell from this information

- 2. (15 pts) Compute the following probabilities from the Bayes network below.
 - (a) P(A, B, C, D)
 - (b) P(A|B)
 - (c) P(C|B)
 - (d) P(B|D) (Hint: there is a shortcut for this one.)



3. (20 pts) Assume a two-class problem with equal a priori class probabilities and Gaussian class-conditional densities as follows:

$$p(x|\omega_1) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a & c \\ c & b \end{bmatrix}\right) \qquad p(x|\omega_2) = N\left(\begin{bmatrix} d \\ e \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$
 where $a \times b - c \times c = 1$.

- (a) (7 pts) Find the equation of the decision boundary between these two classes in terms of the given parameters, after choosing a logarithmic discriminant function.
- (b) (5 pts) Determine the constraints on the values of a, b, c, d and e, such that the resulting discriminant function results with a linear decision boundary.
- (c) (8 pts) Let a=2, b=1, c=0, d=4, e=4. Draw typical equal probability contours for both densities and determine the projection line for the Fisher's discriminant, such that these two classes are separated in an optimal way.

2

- 4. (i) (5 pts) When is maximum a posteriori (MAP) estimation equivalent to maximum-likelihood (ML) estimation? (Try to keep your answer to two sentences or fewer.)
 - (ii) (15 pts) Suppose that n samples x_1, x_2, \dots, x_n are drawn independently according to the Erlang distribution with the following probability density function

$$p(x|\theta) = \theta^2 x e^{(-\theta x)} u(x)$$

where u(x) is the unit step function as follows

$$u(x) = \begin{cases} 1 & if & x > 0 \\ 0 & if & x < 0 \end{cases}$$

find the maximum likelihood estimate of the parameter θ .

- 5. (20 pts) Suppose we have a HMM with N hidden states and M visible states, and a particular sequence of visible states V^T is observed, where T is the length of the sequence.
 - (a) (13 pts) Describe a smart algorithm to compute the probability that V^T was generated by this model, i.e. using either *HMM Forward algorithm* or *HMM Backward algorithm*. You need to give the definition of $\alpha_j(t)$ or $\beta_j(t)$ and explain the meaning of it in words.
 - (b) (3 pts) What is the computational complexity of the algorithm above and why?
 - (c) (4 pts) How to compute the probability above without using this smart algorithm? What is the computational complexity and why?

Appendix: Useful formulas.

• For a 2×2 matrix,

$$A = \left[\begin{array}{cc} a & b \\ c & d \end{array} \right]$$

the matrix inverse is

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

ullet The scatter matrices \mathbf{S}_i are defined as

$$\mathbf{S}_i = \sum_{\mathbf{X} \in D_i} (\mathbf{X} - \mathbf{m}_i)(\mathbf{X} - \mathbf{m}_i)^t$$

where \mathbf{m}_i is the d-dimensional sample mean.

The within-class scatter matrix is defined as

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

The between-class scatter matrix is defined as

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$$

The solution for the W that optimizes $J(W) = \frac{W^t \mathbf{S}_B W}{W^t \mathbf{S}_W W}$ is

$$W = S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$