**In The Name of God, The Compassionate, The Merciful**

Name: ..............................................

Student ID#: ..............................................

# Statistical Pattern Recognition (CE-725)
# Department of Computer Engineering
# Sharif University of Technology

## Final Exam - Spring 2012
## (150 minutes – 100+5 points)

### 1) Basic Concepts (15 points)

a) True or false questions: For each of the following parts, specify that the given statement is true or false. In the case of true, provide a brief explanation, otherwise, propose a counter example.

i- ............. The kernel $K(x_i, x_j)$ is symmetric, where $x_i$ and $x_j$ are the feature vectors for *i*-th and *j*-th examples.

ii- ............. Any decision boundary that we get from a generative model with class-conditional Gaussian distributions could in principle be reproduced with an SVM and a polynomial kernel.

iii- ............. After training an SVM, we can discard all examples which are not support vectors and can still classify new examples.

b) What would happen if the activation function at hidden and output layer in MLP be linear? Explain why this simpler activation function is not normally used in MLPs although it would simplify and accelerate the calculations for the back-propagation algorithm?

### 2) Support Vector Machines (20 points)

Consider the given data points and labels. Suppose that we use the following embedding function to separate two classes by a large margin classifier:

$$\phi(x) = (x_1^2 + x_2^2, x_1 - x_2)$$

| Data point | Label |
|---|---|
| $x_1 = (1,1)$ | 1 |
| $x_2 = (2,1)$ | 1 |
| $x_3 = (2,0)$ | 1 |
| $x_4 = (1,2)$ | -1 |
| $x_5 = (2,2)$ | -1 |
| $x_6 = (1,-3)$ | -1 |

a) Find the support vectors, visually.

b) Find the parameters of the SVM classifier ($w$, $w_0$, $\lambda_i$).

c) Introduce an embedding function from 2-D to 1-D that separates original data points linearly.

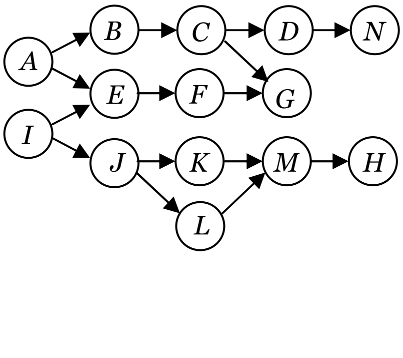
### 3) Graphical Methods (25 points)

a) Consider an HMM with three states $\{S_1,S_2,S_3\}$, two outputs $\{A,B\}$, initial state probabilities {1,0,0}, state transition probability matrix *A*, and output probability matrix *B*.

$$A = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 1 & 0 \end{pmatrix}$$

Compute $P(O_1=B, O_2=B, \ldots, O_{200}=B)$ in the given HMM.

b) Given the following Bayesian network, which of the following statements are true, regardless of the conditional probability distributions?



b1) $P(D,H) = P(D)P(H)$

b2) $P(A, I) = P(A)P(I)$

b3) $P(A, I|G) = P(A|G)P(I|G)$

b4) $P(J,G|F) = P(J|F)P(G|F)$

b5) $P(J,M|K,L) = P(J|K,L)P(M|K,L)$

b6) $P(E,C|A,G) = P(E|A,G)P(C|A,G)$

## 4) Expectation and Maximization (20 points)

Consider a random variable *x* that is categorical with *M* possible values {1,2,...,M}. Suppose that *x* is represented as a vector such that *x(i)=1* if *x* takes the *i*-th value, and $\sum_i x(i) = 1$. The distribution of *x* is represented by a mixture of *K* discrete multinomial distributions such that:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x \mid \mu_k) \text{ and } p(x \mid \mu_k) = \prod_{j=1}^{M} \mu_k(j)^{x(j)}$$

$\pi_K$ denotes the mixing coefficient for the *k*-th component (or the prior probability that the hidden variable *z* = *k*), and $\mu_k(j)$ represents the probability *P(x(j)=1|z=k)*. Observed data points $\{x_i\}$, *i*=1,...,*n* derive the E and M steps to estimate $\pi_K$ and $\mu_k(j)$ for all values of *k* and *j*.

## 5) Clustering (15 points)

a) Assume we are trying to cluster the points $2^0, 2^1, 2^2, \ldots, 2^n$ (a total of *n*+1 points where $n+1=2^N$) using hierarchical clustering. We break ties by combining the two clusters in which the lowest number resides. For example, if the distance between clusters *A* and *B* is the same as the distance between clusters *C* and *D* we would choose *A* and *B* as the next two clusters to combine if *min{A,B} < min{C,D}* where *{A,B}* are the set of numbers assigned to *A* and *B*.

a1) If we are using Euclidian distance, draw a sketch of the hierarchical clustering tree we would obtain for each of single/complete linkage methods.

a2) Now assume that we are using the distance function *d(p,q) = max(p,q)/min(q,q)*. Which of the single/complete linkage methods will result in a different tree from the one obtained in (a) when using this distance function? If you think that one or more of these methods will result in a different tree, sketch the new tree as well.

b) Consider the following algorithm to partition the data points to *K* clusters:

1. Calculate the pairwise distance $d(P_i, P_j)$ between every two data points $P_i$ and $P_j$ in the set of data points to be clustered, and build a complete graph on the set of data points with edge weights = corresponding distances.
2. Generate the Minimum Spanning Tree of the graph i.e. Choose the subset of edges *E'* with minimum sum of weights such that *G' = (P,E')* is a single connected tree.
3. Throw out the *K-1* edges with the heaviest weights to generate *K* disconnected trees corresponding to the *K* clusters.

Identify which of the clustering algorithms you saw in the class corresponds to the mentioned algorithm.

## 6) Semi-Supervised Learning (10 points)

Consider the figure (a) which contains labeled (class 1 filled circles, class 2 hollow circles) and unlabeled (squares) data. We would like to use two methods (co-training and re-weighting) in order to utilize the unlabeled data when training a Gaussian classifier.

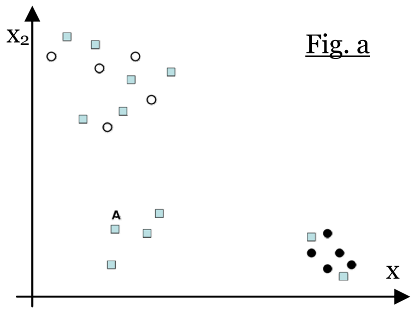


Fig. a

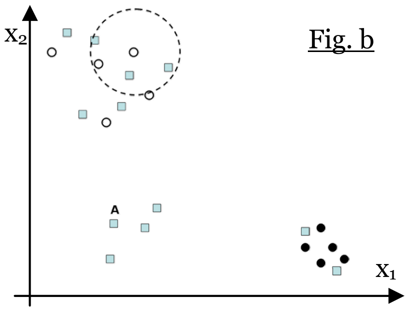


Fig. b

a) How can we use co-training in this case (what are the two classifiers)?

b) We would like to use re-weighting of unlabeled data to improve the classification performance. Re-weighting method is done by placing the dashed circle (shown in figure b) on each of the labeled data points and counting the number of unlabeled data points in that circle. Next, a Gaussian classifier is run with the weights computed.

b1) To what class (hollow circles or full circles) would we assign the unlabeled point *A* if we were training a Gaussian classifier using only the labeled data points (with no re-weighting)?

b2) To what class (hollow circles or full circles) would we assign the unlabeled point A is we were training a classifier using the re-weighting procedure described above?

**Good Luck!**