In the Name of God, the Compassionate, the Merciful

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

## Final Exam – Spring 2011
## (120 minutes – 100+20 points)

1) **[20 Pts.]** Let f(x) ~ U(0, a) be the probability density function of the data points, and $\{X_1, \ldots, X_n\}$ be an iid sample set from f(x). Further assume that we are going to use the Parzen window method to estimate PDF of the data (with $K(x) = e^{-x}$ for $x \geq 0$ and $K(x) = 0$ for $x < 0$, as the window function).
   a. Show that the mean of such a Parzen-window estimation of the density function will be
   $$E[\hat{p}(x)] = \begin{cases} 0 & x < 0 \\ \frac{1}{a}\left(1 - e^{-\frac{x}{h}}\right) & 0 \leq x \leq a \\ \frac{1}{a}\left(e^{\frac{a}{h}} - 1\right)e^{-\frac{x}{h}} & x > a \end{cases}.$$
   b. Plot $E[\hat{p}(x)]$ versus x for a = 1 and h = 1, $\frac{1}{4}$, $\frac{1}{16}$.

2) **[15 Pts.]** Prove that the following functions are valid kernels, if we know that $k_1$ is a valid kernel and g(x) is an *arbitrary* function :
   a. $k(x, y) = k_1(g(x), g(y))$
   b. $k(x, y) = g(x)\, k_1(x, y)\, g(y)$

3) **[20 Pts.]** Figure 1 shows a two-state HMM. The transition probabilities of the Markov chain are given in the transition diagram. The output distribution corresponding to each state is defined over {1, 2, 3, 4} and is given in the table next to the diagram. The HMM is equally likely to start from either of the two states.
   a. Give an example of an output sequence of length 2 which cannot be generated by the HMM in Figure 1.
   b. We generated a sequence of $6{,}867^{2011}$ observations from the HMM, and found that the last observation in the sequence was 3. What is the most likely hidden state corresponding to that last observation?
   c. Consider an output sequence 3 3. What is the most likely sequence of hidden states corresponding to these observations?
   d. Now, consider an output sequence 3 3 4. What are the *first two states* of the most likely hidden state sequence?
   e. We can try to increase the modeling capacity of the HMM a bit by breaking each state into two states. Following this idea, we created the diagram in Figure 2. Can we set the initial state distribution and the output distributions so that this 4-state model, with the transition probabilities indicated in the diagram, would be equivalent to the original 2-state model? If yes, how? If no, why not?

| s | 1 | 2 |
|---|---|---|
| P(x=1) | 0 | 0.1 |
| P(x=2) | 0.199 | 0 |
| P(x=3) | 0.8 | 0.7 |
| P(x=4) | 0.001 | 0.2 |

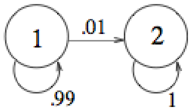


Figure 1. HMM Characteristics

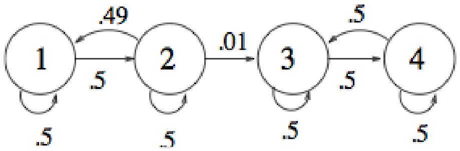


Figure 2. Extended HMM

4) **[20 Pts.]** Figure 3 illustrates a binary classification problem along with our solution using support vector machines (SVMs). We have used a radial basis kernel function given by

$$k(x, x') = \exp\left\{-\frac{\|x - x'\|^2}{2}\right\}$$

where $\|.\|$ denotes the Euclidean distance and x = ($x_1$, $x_2$). The classification decision for any x is made on the basis of the sign of

$$f(x; w, b) = w^T \Phi(x) + b = \sum_{i=1}^{n} \lambda_i y_i k(x, x_i) + b$$

where w and b and $\lambda_i$ are all coefficients estimated from the available data displayed in the figure. The support vectors we obtain for this classification problem (indicated with larger circles in the figure) seem a bit curious. Some of the support vectors appear to be far away from the decision boundary and yet be support vectors. Some of our questions below try to resolve this issue.

a. What happens to our SVM predictions f(x; w, b) with the radial basis kernel if we choose a test point $x_{far}$ far away from any of the training points $x_j$ (distances here measured in the space of the original points)?

b. Let's assume for simplicity that b = 0. What equations do all the training points $x_j$ have to satisfy? Would $x_{far}$ satisfy the same equation?

c. Using the previous part, if we included $x_{far}$ in the training set, would it become a support vector? Briefly justify your answer.

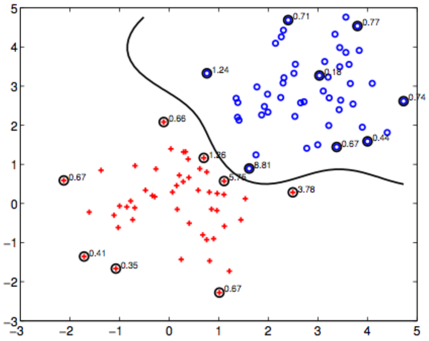
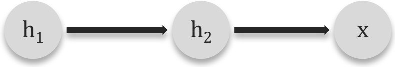


Figure 3

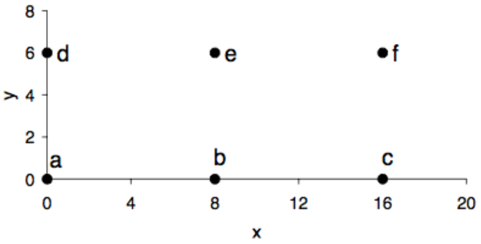
5) **[30 Pts.]** Suppose that we have *n* iid samples $\{X_1, \ldots, X_n\}$ which are generated according to the following Bayesian network model :



where $h_1$ and $h_2$ take their values in $\{1, 2, \ldots, k\}$ and $\{1, 2, \ldots, k+1\}$, respectively. We know the values of $P(h_1=1) = 1/k, \ldots, P(h_1=k) = 1/k$. Furthermore, we know that $P(h_2|h_1 = c) = \begin{cases} \alpha & h_2 = c+1 \\ 1-\alpha & h_2 = c \\ 0 & otherwise \end{cases}$ , where $\alpha$ is unknown. In addition we know that $p(x|h_2 = c) \sim N(\mu_c, 1)$, where $\mu_c$s are uknown. Find the expectation and maximization steps in the EM algorithm to estimate the unknowns based on the observed data.

6) **[15 Pts.]** There is a set S consisting of 6 points in the plane shown as below, a = (0, 0), b = (8, 0), c = (16,0), d = (0,6), e = (8,6), f = (16,6). Now we run the k-means algorithm on those points with k = 3. The algorithm uses the Euclidean distance metric (i.e. the straight line distance between two points) to assign each point to its nearest centroid. Ties are broken in favor of the centroid to the left/down. Two definitions:

- A k-starting configuration is a subset of k starting points from S that form the initial centroids, e.g. {a, b, c}.
- A k-partition is a partition of S into k non-empty subsets, e.g. {a, b, e}, {c, d}, {f } is a 3-partition.



A k-partition is called stable if a repetition of the k-means iteration with the induced centroids leaves it unchanged.

a. How many 3-starting configurations are there? (Remember, a 3-starting configuration is just a subset, of size 3, of the six data points).

b. Fill in the following table:

| 3-partition | Stable? | An example of 3-starting that leads to this 3-partition (none, if there is no one) |
|---|---|---|
| {a,b,e}, {c,d}, {f} | | |
| {a,b}, {d,e}, {c,f} | | |
| {a,d}, {b,e}, {c,f} | | |
| {a}, {d}, {b,c,e,f} | | |
| {a,b}, {d}, {c,e,f} | | |
| {a,b,d}, {c}, {e,f} | | |