Neural Networks for Computing
CAP6615, Fall 2004
Midterm II
FEEDS Midterm

1. **[30 points] Relationship between supervised and unsupervised learning:**

Supervised learning: Regardless of whether a radial basis function (RBF) or multi-layer perceptron (MLP) is used for supervised learning, the underlying thread is the same. Given a set of feature vectors $\mathbf{x}_i \in \mathbb{R}^D, i \in \{1, \ldots, N\}$ and a set of outputs $y_i \in \{0, 1\}, i \in \{1, \ldots, N\}$, fit a function $f(\mathbf{x}; \mathbf{w})$ that minimizes the least-squares objective function $\sum_{i=1}^{N} |y_i - f(\mathbf{x}_i; \mathbf{w})|^2$. When a Gaussian radial basis function approximation is used, the function is explicitly written as

$$f(\mathbf{x}; c, \mathbf{z}) = \sum_{a=1}^{K} c_a \exp\{-\frac{\| \mathbf{x} - \mathbf{z}_a \|^2}{2\sigma^2}\} \tag{1}$$

where $f(\mathbf{x}; c, \mathbf{z})$ is a Gaussian radial basis function (GRBF) approximation which depends on the unknown centers $\mathbf{z}_a \in \mathbb{R}^D, a \in \{1, \ldots, K\}$ and the unknown coefficients $c_a \in \mathbb{R}^1, a \in \{1, \ldots, K\}$. The function $f(\mathbf{x}; c, \mathbf{z})$ is just a more elaborate form of $f(\mathbf{x}; \mathbf{w})$. We use the semicolon notation to denote the fact that $\mathbf{w}$ or $(c, \mathbf{z})$ are a set of parameters. The parameter σ is a free parameter but it can also be estimated from the data.

Unsupervised learning: Regardless of whether a mixture model or other density estimation methods are used for unsupervised learning, the underlying thread is the same. Given a set of feature vectors $\mathbf{x}_i \in \mathbb{R}^{D+1}, i \in \{1, \ldots, N\}$, fit a probability density function $p(\mathbf{x}; \mathbf{w})$ that minimizes the negative log-likelihood $-\sum_{i=1}^{N} \log p(\mathbf{x}_i; \mathbf{w})$. When a Gaussian mixture model density approximation is used with isotropic covariances, the density function is explicitly written as

$$p(\mathbf{x}; \pi, \mathbf{z}) = \sum_{a=1}^{K} \pi_a \exp\{-\frac{\| \mathbf{x} - \mathbf{z}_a \|^2}{2\sigma^2}\} \tag{2}$$

where $p(\mathbf{x}; \pi, \mathbf{z})$ is a Gaussian mixture model (GMM) approximation which depends on the unknown centers $\mathbf{z}_a \in \mathbb{R}^{D+1}, a \in \{1, \ldots, K\}$ and the unknown occupancy probabilities $\pi_a \in \mathbb{R}^1, a \in \{1, \ldots, K\}$ and $\sum_{a=1}^{K} \pi_a = 1$, $\pi_a > 0$. The parameter σ is a free parameter but it can also be estimated from the data.

- **[5 points]** Consider the case where σ is very small and where the number of centers K is equal to the number of data points N, i.e. $K = N$. For both the GRBF and the GMM models, what are the optimal choices for $\mathbf{z}$? What are the optimal choices for c and π for the GRBF and GMM models respectively? In both cases, assume a value of σ which can be made as small as possible while remaining positive. Explain your choice for the optimal solution.

- **[5 points]** How do you convert a supervised learning function approximation problem into an unsupervised learning density estimation problem? That is, show how you can incorporate the labels $y_i, i \in \{1, \ldots, N\}$ into the feature vectors $\mathbf{x}_i, i \in \{1, \ldots, N\}$. Qualitatively, explain the difference in perspective as you make this transition from supervised to unsupervised learning using GRBFs for supervised learning and GMMs for unsupervised learning. [In other words, try and explain how and why these two approaches—functional fit and density estimation—are so closely related. Once you understand that supervised learning can be viewed as unsupervised learning but with an extra feature dimension, then you're set on this question.]

- **[5 points]** Once the labels have been incorporated into the feature vectors, what is the expression for $p(y|\mathbf{x})$ in terms of the GMM parameters $(\pi, \mathbf{z}, \sigma)$? [Recall that the unsupervised learning will give you $p(\mathbf{x}, y)$. Can you obtain $p(y|\mathbf{x})$ from this?

- **[5 points]** Having converted a supervised learning problem into an unsupervised learning problem, show how you would classify an incoming pattern. That is, write down a formula that clearly does the job of classifying a new incoming pattern $\mathbf{x}$ whose label y is unknown. Explain your choice.

- **[10 points]** There is no reason why GRBF approximations should be restricted to the case where $y_i \in \{0, 1\}$. Consider the general situation where $y_i \in \mathbb{R}^1$. Using an example where $x_i \in \mathbb{R}^1$ (for the sake of simplicity), show how you can use either the GRBF approximation or the GMM density approximation to fit the set of samples $(x_i, y_i), i \in \{1, \ldots, N\}$. That is, if a GRBF is used, you need to treat it as a supervised learning problem where x is the input and y the "label." If a GMM is used, you are fitting a Gaussian mixture model to the pairs (x_i, y_i). What are the similarities and differences between using these two methods? We want the results from the two methods to be identical. [Obviously the methods themselves are very different.] Write down a formula which guarantees equivalence.

2. **[30 points] Bayesian networks:** Imagine that you are in charge of a political survey where each subject is classified according to four category spaces—a) conventional (C): {liberal (0), conservative (1)}, b) nonlinear (N): {moderate (0), radical (1)}, c) orthogonal (O): {authoritarian (0), libertarian (1)} and d) moral (M): {exclusivist (0), integral (1)}. In each space, please note that we have assigned binary values to the labels. For example, liberal is assigned '0' and conservative '1' etc. We are interested in studying the co-occurrences between these spaces.

- **[5 points]** Using a histogramming scheme, how would you estimate the joint probability $\Pr(C, N, O, M)$ between the four spaces from the data? [Explain qualitatively how you'd build up the four-way probability distribution.]

- **[5 points]** You are given the following: $\Pr(C = 0) = x$, $\Pr(N = 0) = y$, $\Pr(O = 0) = z$, and $\Pr(M = 0) = w$. Also, $\Pr(C = 0, N = 0) = a$, $\Pr(N = 0, O = 0) = b$, $\Pr(O = 0, M = 0) = c$, and $\Pr(M = 0, C = 0) = d$. Evaluate the pairwise joint probabilities $\Pr(C, N), \Pr(N, O), \Pr(O, M)$ and $\Pr(M, C)$ given this information. [You'll need to use basic rules relating two variable probability distributions to single variable probability distributions. Pretend that a, b, c, d and x, y, z, w are numbers. Now, write all the probabilities in terms of these 8 numbers. You'll need to know that $\sum_c p(C = c, N) = p(N)$.]

- **[10 points]** Given the above pairwise probabilities, estimate the full joint probability $\Pr(C, N, O, M)$ two ways. In case 1, remove $\Pr(M, C)$ to get a tree. In case 2, remove $\Pr(C, N)$ to get a tree. List the conditional probability approximations in both cases. Write down all 16 possibilities for both cases. [Warning: This will take some time. However, since the Fall 2002 class botched this question, I'm making sure that if I asked you ten years from now to do this question, you'll do it like a zombie. We're going for permanent memory etching here. If you think this is horse%$#@, please realize that it builds character.]

- **[10 points]** For both cases above, evaluate $\Pr(C, O)$ which was not given to you. [Establish expressions such that both approaches give you the same answer for all four possibilities. You cannot assume that you have the co-occurrences from the data from which to estimate $\Pr(C, O)$.]

3. **[40 points] Topological clustering:** Consider the following objective function for standard point feature clustering.

$$E(M, \mathbf{y}) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{a=1}^{K} M_{ia} \parallel \mathbf{x}_i - \mathbf{y}_a \parallel^2 + \sum_{i=1}^{N} \lambda_i (\sum_{a=1}^{K} M_{ia} - 1) + \sum_{i=1}^{N} \sum_{a=1}^{K} M_{ia} \log M_{ia} \qquad (3)$$

where $M_{ia} \in \mathbb{R}^1$ and $M_{ia} > 0$ is the usual membership matrix (albeit analog). The cluster centers are denoted by $\mathbf{y}_a, a \in \{1, \ldots, K\}$ and the feature vectors are $\mathbf{x}_i, i \in \{1, \ldots, N\}$.

- **[5 points]** Minimize the above objective function w.r.t. $\mathbf{y}$ by differentiating $E(M, \mathbf{y})$ w.r.t. $\mathbf{y}$ and setting the result to zero. What is the closed form solution that you get for $\mathbf{y}$? [You can differentiate separately for each $\mathbf{y}_a$ to solve this problem.]

- **[5 points]** Minimize the above objective function w.r.t. M by differentiating $E(M, \mathbf{y})$ w.r.t. M and setting the result to zero. What is the closed form solution that you get for M? [You can differentiate separately for each M_{ia} to solve this problem.]

- **[5 points]** Eliminate λ from the closed form solution for M by satisfying the constraint $\sum_{a=1}^{K} M_{ia} = 1$. What is the closed form solution that you get for λ? [To solve this, all

you need to do is to satisfy $\sum_{a=1}^{K} M_{ia} = 1$ for the solution you got for M from the previous question.]

- **[5 points]** How do the pair of closed form solutions that you get for $(M, \mathbf{y})$ differ from the solutions that we obtained from the EM algorithm for a Gaussian mixture model.

Let us now extend the above objective function to perform topological clustering. Assume that the clusters $\mathbf{y}_a$ are organized in a graph whose (a, b) adjacency matrix entry is G_{ab} [$G_{ab} \in \{0, 1\}$. If $G_{ab} = 0$ then there is no topological link. If $G_{ab} = 1$ then a and b are topologically connected. For example, if we seek a two dimensional topology, then the graph G would have a mesh structure. The topological clustering objective function is

$$\begin{aligned} E(M, \mathbf{y}) &= \frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{a=1}^{K} \sum_{b=1}^{K} M_{ia} M_{jb} G_{ab} (\| \mathbf{x}_i - \mathbf{y}_a \|^2 + \| \mathbf{x}_j - \mathbf{y}_b \|^2 + \| \mathbf{y}_a - \mathbf{y}_b \|^2) \quad (4) \\ &+ \sum_{i=1}^{N} \lambda_i (\sum_{a=1}^{K} M_{ia} - 1) + \sum_{i=1}^{N} \sum_{a=1}^{K} M_{ia} \log M_{ia}. \end{aligned}$$

- **[5 points]** Explain the first term [the term with the four summations is referred to as the first term] of the objective function using a 1-D topology example. What is G for the 1-D topology? [You may assume an open curve and a suitable number of clusters (such as 10) in order to explain the basic idea. In an open curve, you know the neighborhood structure of the clusters as in who is ahead of me on the curve and who is behind me on the curve, where I am a cluster center on the curve. Using this, qualitatively explain what the objective function is trying to accomplish.]

- **[10 points]** Minimize the above objective function w.r.t. $\mathbf{y}$ by differentiating $E(M, \mathbf{y})$ w.r.t. $\mathbf{y}$ and setting the result to zero. What is the closed form solution that you get for $\mathbf{y}$? [Very tough question and you'll get full points only if you solve for $\mathbf{y}$ in closed form.]

- **[5 points]** Minimize the above objective function w.r.t. M by differentiating $E(M, \mathbf{y})$ w.r.t. M and setting the result to zero. What is the solution that you get for M? You won't be able to do this in closed form but you can express M_{ia} in terms of M_{jb}. [Easier than you think. Just differentiate w.r.t. M_{ia} and set the result to zero. Solve for M_{ia} in terms of M_{jb} and $\mathbf{y}$.]