EEL6825 – Fall 00

Exam 2 November 13, 2000

NAME: ________________________________

This exam is open-book and calculator. You may use any books or papers that you like. There are four problems and one extra credit in this exam, you have three full class periods. State your assumptions and reasoning for each problem. Justify your steps and clearly indicate your final answers.

1	**/25**
2	**/25**
3	**/25**
4	**/25**
*****	**/3**
TOTAL	

1. (25 points)

   Three 1D probability distributions are given as follows:

$$p(x|\omega_1) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

$$p(x|\omega_2) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

$$p(x|\omega_3) = \begin{cases} 2 - 2x & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

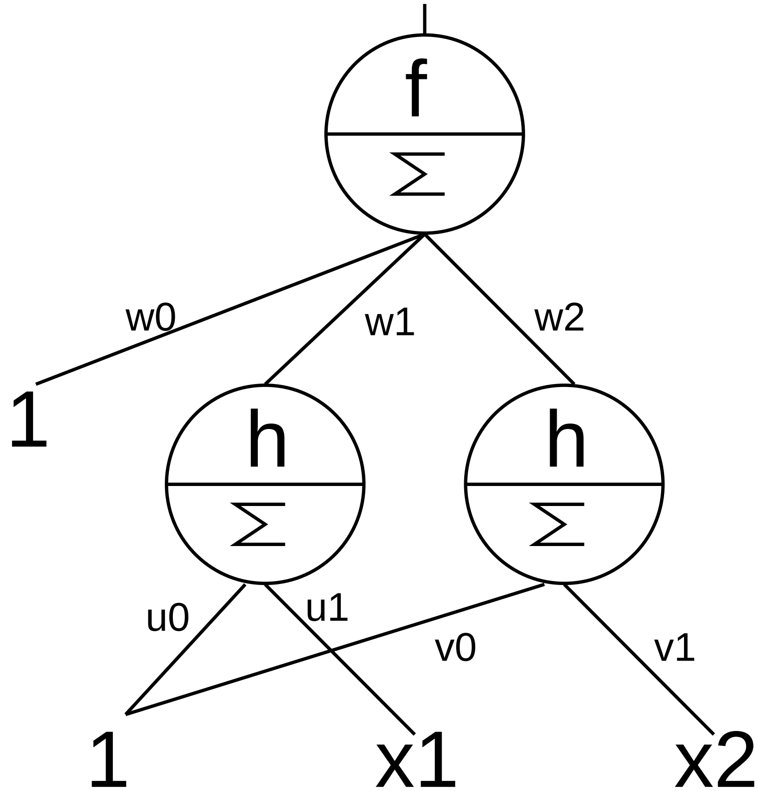
   Answer the following questions about a **one nearest neighbor classifier**. Assume that $P(\omega_1) = P(\omega_2) = P(\omega_3) = \frac{1}{3}$.

   (a) (5 points) What is p(error|x) at $x = 0.5$?

(b) (10 points) What is p(error|x) ? Express your answer as a function of x and make sure that you are consistent with what you wrote in part (a).

(c) (10 points) What is p(error) for this problem? (Of course, your answer should be a numerical value).

2. (25 points)

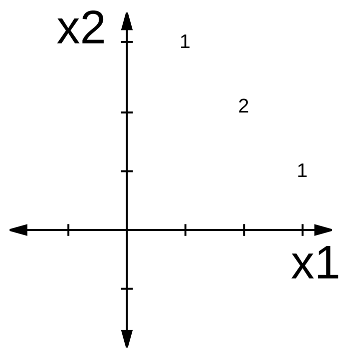


This problem deals with a slightly modified MLP architecture. The first difference to notice is that the input layer is not fully connected to the hidden layer. The second difference is that the hidden layer has a nonlinear function $h(a)$ which may be different from the output nonlinear function $f(a)$. Assume for *now* that the activation functions are the same and are given by the following hard-limiting function:

$$f(a) = h(a) = \begin{cases} 1 & \text{if } a > 0 \\ -1 & a < 0 \\ \text{undefined} & a = 0 \end{cases}$$

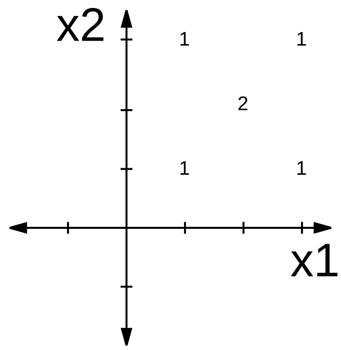
There are only two hidden units and the weights are labeled as shown in the figure ($u_0, u_1, v_0, v_1, w_0, w_1$ and w_2). Answer the following questions regarding the use of this neural network to solve a simple two-class classification problems.

(a) (10 points)



Assuming that class 1 and class 2 data points are given as shown above, can this problem be solved using this architecture with two hidden units? Justify why or why not. If possible, provide all of the necessary weight values for the architecture. As usual, the final output of your neural network should be +1 for class 1 points and -1 for class 2 points.

(b) (15 points)



You are given the above class 1 and class 2 points. Suppose that you can redefine the nonlinear function $h(a)$ to be any arbitrary continuous function you want. Can you come up with an $h(a)$ function and associated weight values that correctly classify the above points? If it is not possible, explain why not. If it is possible, provide all of the necessary weight values for the architecture as well as your definition of $h(a)$ Remember that the final output of your neural network should be +1 for class 1 and -1 for class 2.

3. (25 points)

Two normal distributions are characterized by:

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$\mu_1 = \mu_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 5 & 0 & 4 \\ 0 & 5 & 4 \\ 4 & 4 & 4 \end{bmatrix}$$

Answer the following questions regarding the KL dimensionality reduction of this problem.

(a) (5 points) Compute the combined covariance matrix (Σ) of the data.

(b) (5 points) If you correctly computed part (a) then the eigenvectors of $\hat{\Sigma}$ are

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$$

Note: These eigenvectors are listed in no particular order and are not normalized. Compute the normalized eigenvectors and their respective eigenvalues. If you realize that you did part (a) wrong, clearly state this or, better yet, find your mistake.

(c) (5 points) If you had to drop a single feature, which feature would it be? What is the error that you would then pay in terms of representation and in terms of classifier performance?

(d) (10 points) Map the two distributions to the new two-dimensional space spanned by y_1 and y_2. What are the new values of the 2-dimensional μ_1, μ_2, Σ_1 and Σ_2?

4. (25 points) Short Answer.

   (a) (5 points) Suppose we estimate a Gaussian pdf using a Parzen window estimator with uniform rectangular kernels. Sketch a Gaussian pdf and sketch possible Parzen window estimators for (1) a high bias estimate and (2) a high variance estimate.

   (b) (5 points) It is well-known that it is difficult to choose a proper window size for Parzen windows estimation. Many times the window is too big where the data is dense and too small where the data is sparse. Briefly describe a technique that dynamically adjusts the size of the window function based on the density of the sample points.

(c) (5 points) A student randomly split her data, with half going into a training set and half going into a test set. She then implemented numerous k-nearest neighbor algorithms with k varying from 1 to 100, each one training on the same training set and running on the same test set. She found that the k=13 algorithm achieved the best result on the test set with 11.2% error. Is 11.2% likely to be an optimistic or pessimistic estimate of the true expected error of the 13-NN algorithm on a new data set? Explain.

(d) (5 points) Two students collaborated on a homework assignment that required them to implement an MLP with backpropagation in Matlab for solving a two-class classification problem. They used exactly the same training and test sets, and used *exactly* the same Matlab code which they each ran on their own home computers. However, when they turned in their homeworks they were surprised to find that their programs got different error rates on the test set. What is the likely reason for the difference?

(e) (5 points) Two normal distributions are characterized by:

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0.1 \\ 1 \\ 2 \\ 3 \end{bmatrix}, \Sigma_1 = \Sigma_2 = \begin{bmatrix} .01 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

If you could only keep one of the four original features (x_1, x_2, x_3 or x_4), which one would you keep for lowest classification error? Explain.

5. (3 points, no partial credit) For a two-class classification problem in d-dimensional space with $\mu_1 = \mu_2 = \underline{0}$, $\Sigma_1 = I$, and $\Sigma_2 = 2 * I$. What is the value of the Bayes error as d approaches infinity? Provide a rigorous mathematical proof of your answer.