# Problem Set 2

MAS 622J/1.126J: Pattern Recognition and Analysis

Due: 5:00 p.m. on October 12

[Note: All instructions to plot data or write a program should be carried out using Matlab. In order to maintain a reasonable level of consistency and simplicity we ask that you do not use other software tools.]

If you collaborated with other members of the class, please write their names at the end of the assignment. Moreover, you will need to write and sign the following statement: "In preparing my solutions, I did not look at any old homeworks, copy anybody's answers or let them copy mine."

## Problem 1: Discriminability and ROC [20 points]

Please download the datasets (dataset1.txt, dataset2.txt, dataset3.txt, dataset4.txt) from the course website. Each dataset includes 1-D data samples from two classes $w_1$ and $w_2$. The first column and the second column of each dataset specify 1000 samples from class $w_1$ and 1000 samples from class $w_2$, respectively.

a. For each dataset, compute the discriminability $d' = \frac{|\mu_2 - \mu_1|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ where $\mu_1$ and $\sigma_1$ are the mean and standard deviation of the distribution of class $w_1$, and $\mu_2$ and $\sigma_2$ are the mean and standard deviation of the distribution of class $w_2$.

b. Now we compute the ROC curve for each dataset. Please plot four ROC curves in the same figure. To do this, we approximate $\mathbf{P_{TP}} = \mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_2)$ by $\mathbf{N}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_2)/1000$, and $\mathbf{P_{FP}} = \mathbf{P}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_1)$ by $\mathbf{N}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_1)/1000$. Here, for $i$ =1 or 2, $\mathbf{N}(\mathbf{x} > \mathbf{x}^* | \mathbf{x} \in w_i)$ is denotes the number of samples in class $w_i$ whose value is greater than $\mathbf{x}^*$. Note that $\mathbf{N}$ doesn't denote a normal distribution!

c. For each dataset, plot the two *approximated* probability density functions. Note that the probability density function is the derivative of the cumulative density function. (Do NOT just approximate distributions by Gaussians and draw those approximated Gaussians in this problem.) *Hint:* Use the same method we use in (b) to get the cumulative distribution. This is called *Monte Carlo* method. Also, note that the derivative (i.e., probability density function $p(x)$) relates to the increment of the cumulative density function $P_X(x)$. That is, $p(x) = \Delta P_X(x)/\Delta x$.

d. How does the discriminability relate to the ROC curve?

# Problem 2: (DHS 2.6) Optimal Decision Boundaries [20 points]

A Spanish campany called Goorrel has launched an application to recognize important e-mails ($\omega_1$) vs unimportant e-mails ($\omega_2$.) The company is using two secret features such that their training data is well approximated by two Gaussians:

$$p(\mathbf{x}|\omega_1) \sim \mathcal{N}(\mu_1, \Sigma_1)$$
$$p(\mathbf{x}|\omega_2) \sim \mathcal{N}(\mu_2, \Sigma_2)$$

where $\mu_1 = [0\ 0]^T$, $\mu_2 = [5\ 0]^T$, $\Sigma_1 = \mathrm{I}$, and $\Sigma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$, where I is the identity matrix.

a. Plot the one-sigma ellipses for these two classes in the place $\mathbf{x} = [x_1\ x_2]^T$.

b. The company finds that choosing a threshold at $x_1 = 3$ perfectly separates the training examples they have; thus, they propose that this should be the best classifier. Show them an expression, in terms of $\mathbf{x}$, which can improve their classifier with respect to minimizing the Bayes probability of error. Assume that unimportant emails are three times as likely as important emails.

c. The shape of this optimal decision boundary is:

- a line
- a parabola
- a hyperbola
- a circle
- an ellipse
- none of the above (explain)

Be sure to justify your answer.

d. If Gorrel accidently misses relevant e-mails, then this will be terrible for the users. The company estimates this error will cost them twenty times as much as the cost of missclasifying unimportant e-mails (there's no cost to choosing correctly in either case.) Describe qualitatively how this changes your result above. Make a sketch of the change (it doesn't have to be precisely plotted). Justify your answer.

# Problem 3: Principal Component Analysis [10 points]

Consider the covariance matrix for a Gaussian with mean = (0,0) and variance = $\sigma^2 \times I_2$ where $\sigma^2$ is a positive constant, and $I_2$ is a 2 × 2 identity matrix.

a. What are the two principle components for this matrix? What are their eigenvalues?

b. Given a data point (x,y) from this distribution, what is the reconstructed data using the projection onto the first principal component of this matrix?

c. For this reconstructed value, what is the expected value of the reconstruction error (squared error between the true value and reconstructed value).

# Problem 4: EigenFaces [30 points]

In this exercise we provide a dataset of face images (450x400 pixels) to explore the concept of eigenfaces and some of its applications. Please include MATLAB code and images to support your answers.

a. Find the eigenfaces of the dataset. Show the first three components. The images can be loaded by using the helper function "[X] = load_imgs('training');"' in MATLAB. (Hint: svd(X,0))

b. Reduce the dimensionality of the images by projecting them onto a lower dimensional space. How many basis are you using? Justify your answer.

c. We received a new image but we lost some of its information. How do you suggest to automatically fix it? Show your solution as well as the result. ("[X] = load_imgs('corrupted');"')

d. We received another image but this time we do not know the name of the person. Use the eigenfaces to perform face recognition. For the sake of simplicity, you can report the closest image in the training set instead of the name of the person. ("[X] = load_imgs('testing');"')

e. Suggest a different way to use eigenfaces.

# Problem 5: Hidden Markov Models [20 points]

We have two 2-state Hidden Markov Models, where both states have two possible output symbols A and B:
The output probabilities are given by:

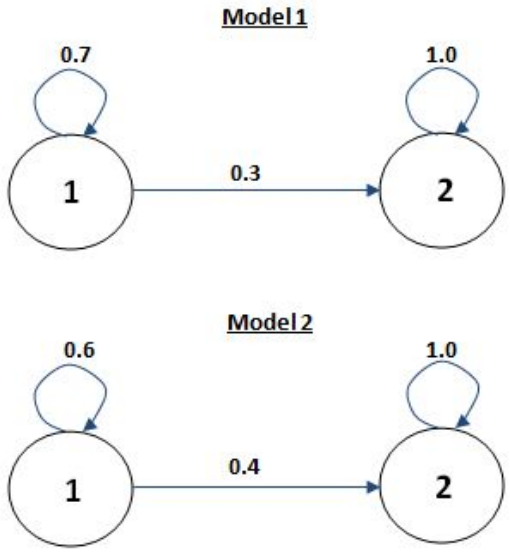


Figure 1: Two HMMs

Model 1: $b_1(A)=0.85$ $b_1(B)=0.15$ $b_2(A)=0.4$ $b_2(B)=0.6$

Model 2: $b_1(A)=0.2$ $b_1(B)=0.8$ $b_2(A)=0.1$ $b_2(B)=0.9$

The initial probabilities are given by:

$$\text{Model 1: } \pi_1 = 0.75 \quad \pi_2 = 0.25$$

$$\text{Model 2: } \pi_1 = 0.5 \quad \pi_2 = 0.5$$

a. For Model 1, what is the probability of an observation sequence { B A B } being generated?

b. For Model 1, given that this HMM produced an observation sequence { B A B }, what is the most likely sequence of hidden states that led to those observations?

c. Which model is more likely to produce the observation sequence { B A B }?