

## Homework #2

Question 1

Questions 1 and 2 of this homework assignment require you to simulate a mixture model where the mixture components are Normal distributions. The data set for Question 1 consists of 600 data items. Items 1-300 are samples from a two-dimensional Normal distribution whose mean is  $[3 \ 3]^T$  and whose covariance matrix is the identity matrix  $\mathbf{I}$ ; items 301-600 are samples from a two-dimensional Normal distribution whose mean is  $[7 \ 7]^T$  and whose covariance matrix is the identity matrix  $\mathbf{I}$ . I will e-mail the data set to you. Note that it contains three columns. Columns 1 and 2 give the values of a data item along the two dimensions; column 3 indicates whether the item was sampled from the first distribution (indicated by a zero) or the second distribution (indicated by a one). Note that you should not use column 3 when training the mixture model; it should only be used when you evaluate the model's performance.

The first thing that you should do is normalize the data items. Calculate the mean and standard deviation of columns 1 and 2, and then normalize each value in columns 1 and 2 by subtracting the mean and dividing by the standard deviation.

In your first simulation, use two mixture components. The prior distribution over the components should be fixed at a uniform distribution. The elements of the mean vectors of the mixture components should be initialized to random values between -1 and 1. The covariance matrices of the mixture components should be fixed to  $\sigma^2\mathbf{I}$  where  $\sigma^2 = 0.25$ . That is, you only need to adapt the mean vectors, not the covariance matrices and not the prior distribution. Use on-line learning with a learning rate of 0.0001 for 200 epochs (each data item should appear once in each epoch; the data items should be randomly ordered within each epoch and each epoch should use a different random order). At the end of training, compare the posterior distribution over the mixture components for each data item with column 3. Did the mixture model learn well? Explain. Repeat the simulation 5 times. Do you get the same results each time? (It ought to be the case that the mean vector for one component is near  $[-0.9 \ -0.9]^T$  and the mean vector for the other component is near  $[0.9 \ 0.9]^T$  [roughly, these are the normalized means of the original distributions that were used to generate the data]. In addition, it ought to be the case that one component should have a larger posterior probability for all data items that came from one Normal distribution [as indicated by column 3] and that the other component should have a larger posterior probability for all data items that came from the other Normal distribution.)

Now repeat your simulations but use a mixture model with 3 mixture components. What do the results look like? What happens if you use 4 mixture components?

### Question 2

The data set for Question 2 was originally collected by Peterson and Barney (1952), and is a benchmark database in the speech recognition literature. The data items represent instances of vowels spoken by a large number of speakers. The instances were taken from utterances of ten words, each of which began with an “h”, contained a vowel in the middle, and ended with a “d”. These data are from seventy-five speakers who uttered each word twice, with the words in different random orders for each presentation. Thirty-two of the speakers were male adults, twenty-eight were female adults, and fifteen were children. A spectral analysis was performed on each utterance; the portion of the spectrogram corresponding to the vowel was hand segmented, and the first two formants were extracted from the middle portion of the segmented region. Formants are the vocal tract’s resonant frequencies. Different placements of the speech articulators, corresponding to different vowels, alter the vocal tract’s shape and, thus, its frequency response. The first and second columns of the data set give the first and second formant values for each vowel instance. Four classes of vowels are represented in the data set (as indicated in column 3).

As before, the first thing that you need to do is normalize the data items. In your first simulation, use four mixture components. The prior distribution over the components should be fixed at a uniform distribution. The elements of the mean vectors of the mixture components should be initialized to random values between -1 and 1. The covariance matrices of the mixture components should be fixed to  $\sigma^2\mathbf{I}$  where  $\sigma^2 = 0.1111$ . Again, you only need to adapt the mean vectors, not the covariance matrices and not the prior distribution. Use on-line learning with a learning rate of 0.0001 for 400 epochs. At the end of training, compare the posterior distribution over the mixture components for each data item with column 3. Did the mixture model learn well? Explain. Repeat the simulation 5 times. Do you get the same results each time? It ought to be the case that the model’s performance is imperfect in the sense that it is not the case that one component has the largest posterior probability for every item in a vowel class, and that different components have learned to represent different vowel classes. Why do you think that this model’s performance is imperfect?

Repeat the simulations but use a mixture models with 8 mixture components. What do the results look like? What happens if you use 12 mixture components?

### Question 3

A coin has a probability  $p$  of landing heads-up when tossed in the air. The coin is tossed 10 times. Let  $x_1, \dots, x_{10}$  denote the outcomes, where  $x_i = 1$  if the coin landed heads-up on the  $i^{\text{th}}$  toss and  $x_i = 0$  if the coin landed tails-up on the  $i^{\text{th}}$  toss. Of the 10 tosses, it landed heads-up on seven tosses, and landed tails-up on three tosses. You would like to use maximum likelihood estimation to estimate the value of  $p$ . Write down the likelihood function that you would like to maximize.

### Question 4

A rolled die has a probability  $p_i$  of landing so that  $i$  spots are on the uppermost face. That is, with probability  $p_1$  it lands on 1, with probability  $p_2$  it lands on 2, and so on. The die is rolled 70 times. Let  $x_1, \dots, x_{70}$  denote the outcomes, where  $x_i$  denotes the outcome on the  $i^{\text{th}}$  roll. Of the 70 rolls, it landed on 1 eight times, it landed on 2 twenty times, it landed on 3 eight times, it landed on 4 thirteen times, it landed on 5 ten times, and it landed on 6 eleven times. You would like to use

maximum likelihood estimation to estimate the values of  $p_1, p_2, p_3, p_4, p_5$ , and  $p_6$ . Write down the likelihood function that you would like to maximize.

#### Question 5

Suppose that the nature of the world is such that a supreme being determines the temperature on each day during the first week of June by sampling from a Normal distribution. During the first week of June, the following seven temperatures are recorded: 72.3, 68.5, 76.9, 82.1, 84.3, 77.0, and 79.7. Let  $x_i$  denote the temperature on the  $i^{\text{th}}$  day. You would like to use maximum likelihood estimation to estimate the value of the mean of the Normal distribution (let  $\mu$  and  $\sigma^2$  denote the mean and variance of this distribution). Write down the likelihood function that you would like to maximize.

#### Question 6

On one side of a curtain, a person has two coins. At each time frame, he performs the following two-step process. First, he randomly selects a coin (with probability  $\pi_1$  he selects coin 1, and with probability  $\pi_2$  he selects coin 2). Next, he flips the selected coin (coin 1 has probability  $p_1$  of landing heads-up; coin 2 has probability  $p_2$  of landing heads-up). After performing this process, the person lets you know the outcome. That is, you are told that the selected coin landed heads-up or tails-up (you are NOT told which coin was selected). Suppose that this process is repeated 10 times. Let  $x_i$  denote the outcome of the  $i^{\text{th}}$  repetition. Of the 10 repetitions, a head was produced 2 times and a tail was produced 8 times. You would like to use maximum likelihood estimation in order to estimate the values of  $\pi_1, \pi_2, p_1$ , and  $p_2$ . Write down the likelihood function that you would like to maximize.

#### Question 7

Visual perception is inherently ambiguous. Suppose that you see a ball. Based on the size of the image of the ball cast on your retina, you might conclude that its a small ball relatively close to you or a large ball relatively far from you. Let  $I$  be a binary variable indicating the retinal image size ( $I = 1$  means that the image size is big;  $I = 0$  means that its small), let  $S$  be a binary variable indicating the size of the ball ( $S = 1$  means that the ball is big;  $S = 0$  means that its small), and let  $D$  be a binary variable indicating the distance from the ball to the viewer ( $D = 1$  means that the ball is far from the viewer;  $D = 0$  means that its close). Figure 1 illustrates a Bayesian network showing the relationships among  $I$ ,  $S$ , and  $D$ .

(Part 1) Assume the following probability values:

$$p(S = 1) = 0.8 \tag{1}$$

$$p(D = 1) = 0.9 \tag{2}$$

$$p(I = 1|S = 1, D = 1) = 0.6 \tag{3}$$

$$p(I = 1|S = 1, D = 0) = 0.9 \tag{4}$$

$$p(I = 1|S = 0, D = 1) = 0.1 \tag{5}$$

$$p(I = 1|S = 0, D = 0) = 0.4. \tag{6}$$

Suppose that the image size is small. Calculate  $p(D = 0|I = 0)$ .

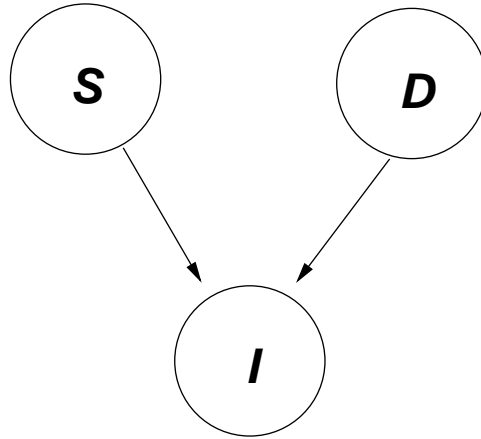


Figure 1: Bayesian network for Question 7.

(Part 2) Suppose that you discover that the ball is small. Calculate  $p(D = 0 | S = 0, I = 0)$ .

(Part 3) Intuitively, it should be that the size of a ball and the distance from a ball to a viewer are unrelated. After all, a big ball can be placed close to a viewer or far from a viewer, and this is also true of a small ball. Yet, in Parts 1 and 2, you calculated the probability that the ball is close to a viewer, and you got different answers. Why is this?