ECE 8110 Machine Learning Midterm 2

April 5, 2014

Salvatore Giorgi
salvatore.giorgi@temple.edu

Exercise 1. *Consider two probability distributions defined by*

$$p(x|\omega_1) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \textit{elsewhere} \end{cases}$$

$$p(x|\omega_2) = \begin{cases} 1, & \frac{1}{2} \leq x \leq \frac{3}{2} \\ 0, & \textit{elsewhere} \end{cases}$$

and assume equal priors.

(a) Draw two points at random from each class. Design a nearest-neighbor classifier based on these points. Compute the probability of error.

(b) Exmplain what happens as you allow the number of points drawn to increase. Show that your result in (a) converges to the correct result.

Solution for (a):

We arbitrarily choose the following points

$$p(x|\omega_1) \ni \begin{cases} x_{11} = 0.5 \\ x_{12} = 0.4 \end{cases}$$

$$p(x|\omega_2) \ni \begin{cases} x_{21} = 0.6 \\ x_{22} = 1.4 \end{cases} .$$

Let x be an unclassified data point. A nearest neighbor classifier centers a cell at x and looks for the nearest neighbor. Based on the above 4 data points, any point $x < 0.55$ will be classified in ω_1 and any $x > 0.55$ will be classified in ω_2. This is equivalent to having a decision boundary at 0.55 and gives us the following:

$$P(\text{error}) = \int_{0.55}^{1} p(x|\omega_1)P(\omega_1)dx + \int_{0.5}^{0.55} p(x|\omega_2)P(\omega_2)dx = P(\omega_1)0.5 = 0.25.$$

We simulate the above, by generating N data points, half of which are from $p(x|\omega_2)$ while the other half are from $p(x|\omega_2)$ (since we have equal priors). We classify each point as follows: any point $x < 0.55$ will be classified in ω_1 and any $x > 0.55$ will be classified in ω_2. We let $N \in \{10, 100, 1000, 10000, 100000\}$, and calculate the probability of error as

$$P(\text{error}) = \frac{\text{number of incorrect classifications}}{N}.$$

Table 1 shows the results of our simulation. From this, we see that as N grows, $P(\text{error})$ approaches 0.25.

Table 1: Probability of Error for Nearest Neighbor Classifier with 4 Original Data Points

N	P(error)
10	0.1000
100	0.2400
1000	0.2270
10000	0.2538
100000	0.2510

Solution for (b):

Note that from the problem statement, we are assuming equal priors and also, we are dealing with uniform distributions. Thus, we are equally likely to encounter a data point at any point in the intervals defined by the distributions.

As the number of data points $n \to \infty$ you will be equally likely to be closed to a point from either $p(x|\omega_1)$ and $p(x|\omega_2)$ on the interval $[\frac{1}{2}, 1]$. This is equivalent to having a decision surface which is equi-distant from the means of the two distributions, which in this case is $\frac{3}{4}$. From this, we have the probability of error is

$$P(\text{error}) = \int_{0.75}^{1} p(x|\omega_1)P(\omega_1)dx + \int_{0.5}^{0.75} p(x|\omega_2)P(\omega_2)dx = P(\omega_1)0.5 = 0.25.$$

We simulate the above situation. To do this, we generate N_D data points to use as our classifier, half from each distribution above. For each choice of N_D we choose N_C points to classify (again, half drawn from each distribution above). We find the minimum distance between each of the N_C points and the N_D points and classify the N_C point as coming from the same class as the N_D point corresponding to this minimum distance. We consider $N_D, N_C \in \{10, 100, 1000, 10000\}$ and for simplicity constrain $N_D = N_C$. The results of our simulation are shown in Table 2.

Table 2: Probability of Error for Nearest Neighbor Classifier with N_D Original Data Points

N_C	P(error)
10	0.3000
100	0.2600
1000	0.2600
10000	0.2483

Exercise 2. *Consider the following models for a system that outputs a sequence of characters "$" and "%".*

(a) Compute the probability that model A produced the sequence "%$%".
(b) Which model most likely produced the sequence "%$%". Explain.
(c) Which state sequence most likely produced the sequence "%$%". hat was the probability of that state sequence.
(d) Give at least two reasons why the probabilities in (a) and (c) differ.

Solution for (a):

Figs. 1 and 2 show the results of our calculation. We get that the probability that model A produced "%$%" is equal to

$$P_A = \alpha_3(4) = 0.75\big(0.25a_{11}\big)\big(0.75a_{12}\big) + 0.75\big(0.25a_{12}\big)\big(0.75a_{22}\big) = 0.140625\big(a_{11}a_{12} + a_{12}a_{22}\big), \tag{1}$$

while the probability that model B produced "%$%" is

$$P_B = \beta_3(4) = 0.25\big(0.75b_{11}\big)\big(0.25b_{11}\big) + 0.25\big(0.75b_{12}\big)\big(0.25b_{11}\big) = 0.046875\big(b_{11}b_{12} + b_{12}b_{22}\big). \tag{2}$$

Solution for (b):

From the above, and assuming both models have the same transition probabilities, we see that model A will most likely produce the sequence "%$%" since

$$b_{11}b_{12} + b_{12}b_{22} = a_{11}a_{12} + a_{12}a_{22}$$

and therefore $P_A > P_B$. This follows intuitively, since we have more "%" than "$" and the probability that we see a "%" is higher for model A than model B.

If we do not assume that both models have equal state transition probabilities then model A will most likely produce the sequence if

$$a_{11}a_{12} + a_{12}a_{22} > \frac{0.046875}{0.140625}b_{11}b_{12} + b_{12}b_{22}$$

and model B will most likely produce the sequence otherwise.

Solution for (c):

We note that we have two possible state sequences through either model:

$$\omega_0\omega_1\omega_1\omega_2\omega_3$$

$$\omega_0\omega_1\omega_2\omega_2\omega_3.$$

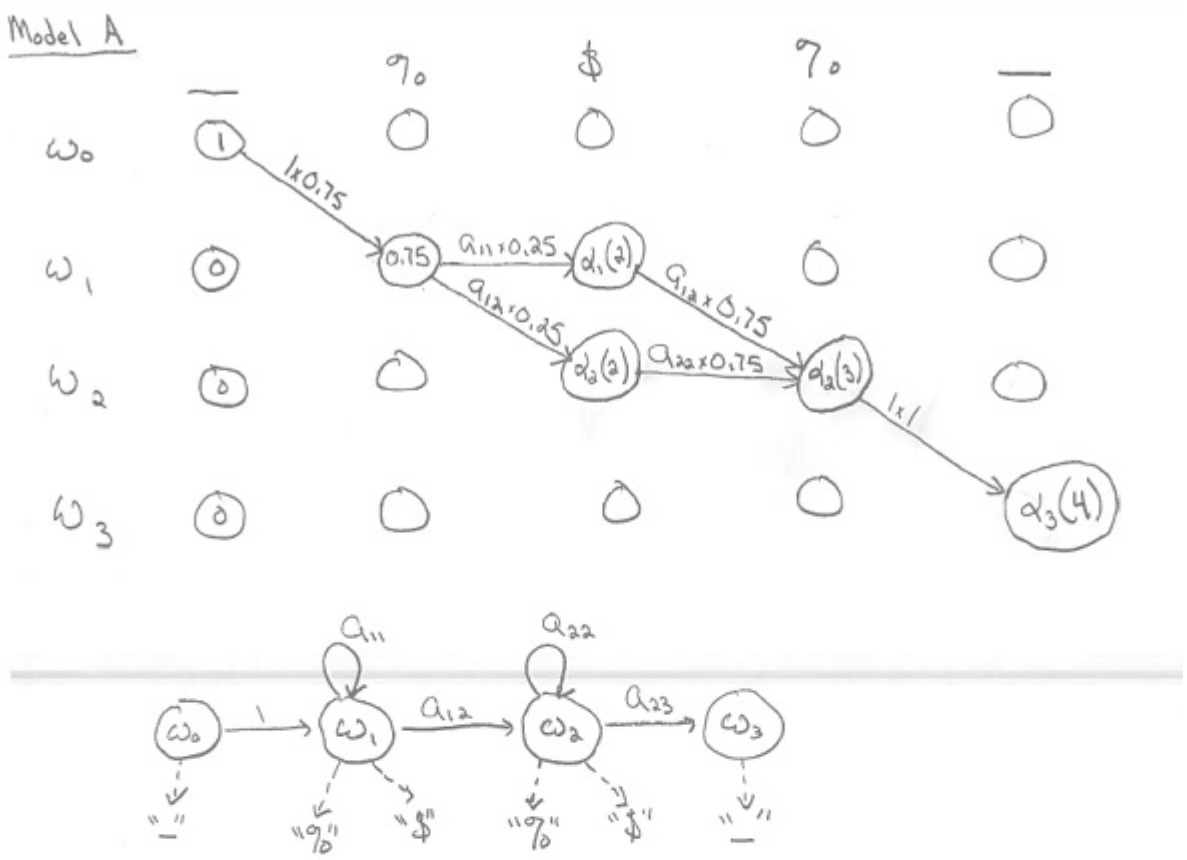
From this, we see that our path is uniquely determined from our second state transition. We also note that the optimal path through will also be the locally optimal path. Thus, we only need consider $\alpha_1(2), \alpha_2(2)$ and $\beta_1(2), \beta_2(2)$. Since $a_{12} = 1 - a_{11}$ and $b_{12} = 1 - b_{11}$, we have that

$$\alpha_1(2) \geq \alpha_2(2) \quad \text{if} \quad a_{11} \geq 0.5$$

$$\beta_1(2) \geq \beta_2(2) \quad \text{if} \quad b_{11} \geq 0.5.$$

It follows that the optimal path through model A and model B, respectively, is

$$\omega_0\omega_1\omega_j\omega_2\omega_3, \;\; j = \arg\max_k\{a_{1k}\}$$



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & a_{11} & a_{12} & 0 \\ 0 & 0 & a_{22} & a_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{matrix}$$

$$B = \begin{bmatrix} \text{"\_"} & \text{"\%"} & \text{"\$"} \\ 1 & 0 & 0 \\ 0 & 0.75 & 0.25 \\ 0 & 0.75 & 0.25 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\alpha_1(2) = 0.75\left[a_{11} \times 0.25\right]$$

$$\alpha_2(2) = 0.75\left[a_{12} \times 0.25\right]$$

$$\alpha_2(3) = \alpha_1(2)\left[a_{12} \times 0.75\right] + \alpha_2(2)\left[a_{22} \times 0.75\right]$$

$$\alpha_3(4) = \alpha_2(3)$$

Figure 1: Model A

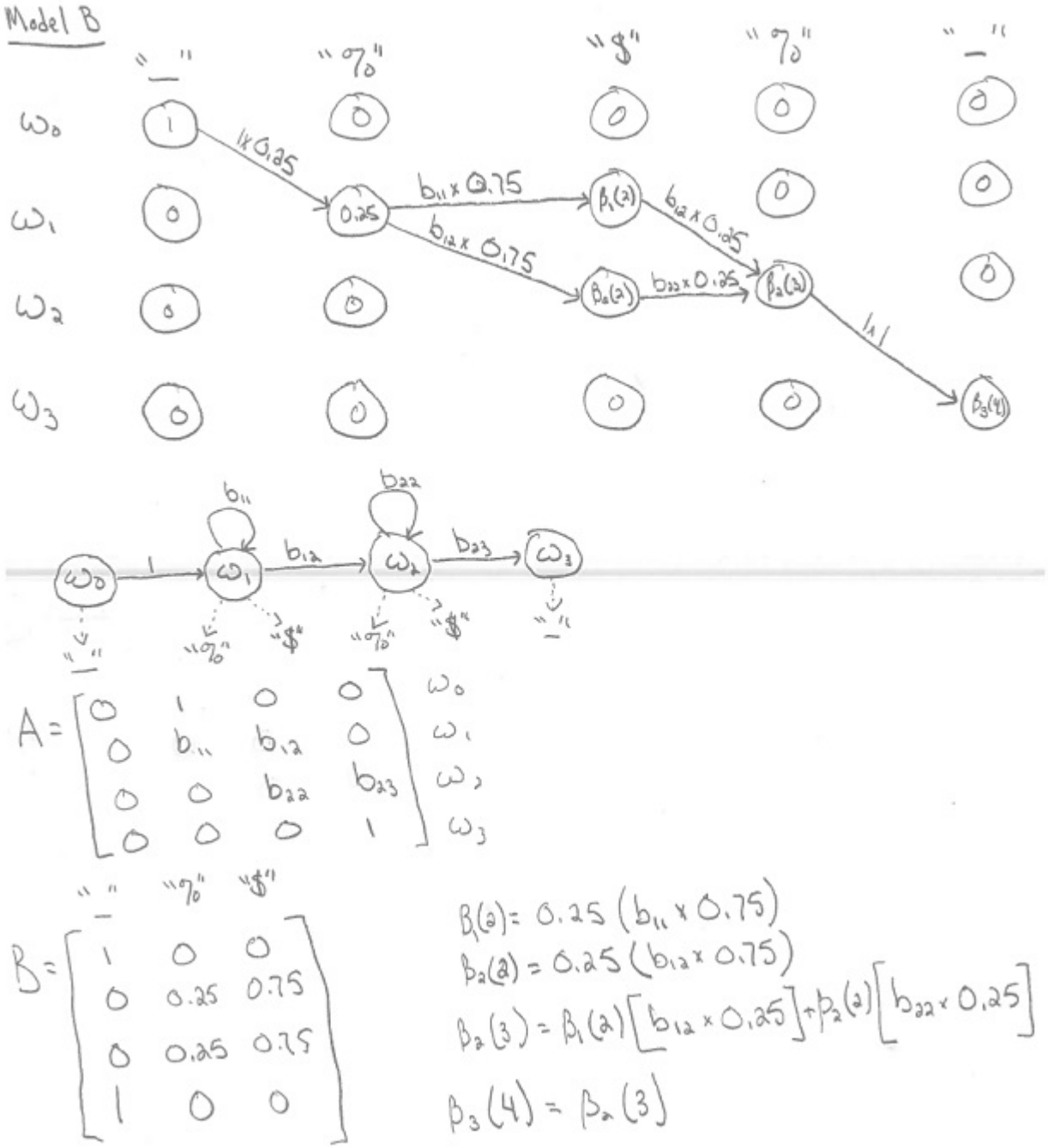


Figure 2: Model B

$$\omega_0\omega_1\omega_i\omega_2\omega_3, \;\; i = \arg\max_k\{b_{1k}\}.$$

The probability of each of these state sequences, for model A and B respectively, is

$$0.75\big(0.25a_{1j}\big)\big(0.75a_{j2}\big), \;\; j = \arg\max_k\{a_{1k}\} \tag{3}$$

$$0.25\big(0.75b_{1i}\big)\big(0.25b_{i2}\big), \;\; i = \arg\max_k\{b_{1k}\}. \tag{4}$$

If we consider both models then the most likely state sequence is the state sequence corresponding to the maximum of (3) and (4).

Solution for (d):

The probability in (a) considers all possible state sequences while the probability in (c) considers each state sequence from each model separately. Thus, one could have a model with very high probability of a single state sequence producing the output, while the other model could have a higher probability that the sum of ALL state sequences produce the given output. These probabilities are determined by both the output probabilities and the state transition probabilities.