CSE555: Introduction to Pattern Recognition

Midterm Exam Solution (100 points, Closed book/notes)

There are 5 questions in this exam.

The last page is the Appendix that contains some useful formulas.

- 1. (15pts) Bayes Decision Theory.
 - (a) (5pts) Assume there are c classes w_1, \dots, w_c , and one feature vector \mathbf{x} , give the Bayes rule for classification in terms of a priori probabilities of the classes and class-conditional probability densities of \mathbf{x} .

Answer:

Bayes rule for classification is

Decide
$$\omega_i$$
 if $p(x|\omega_i)P(\omega_i) > p(x|\omega_j)P(\omega_j)$ for all $j \neq i$ and $i, j = 1, \dots, c$.

(b) (10pts) Suppose we have a two-classes problem $(A, \sim A)$, with a single binary-valued feature $(\mathbf{x}, \sim \mathbf{x})$. Assume the prior probability P(A) = 0.33. Given the distribution of the samples as shown in the following table, use Bayes Rule to compute the values of posterior probabilities of classes.

	A	$\sim A$
x	248	167
$\sim {f x}$	82	503

Answer:

By Bayes formula, we have

$$P(A|x) = \frac{p(x|A)P(A)}{p(x)}$$

we also know that

$$p(x) = p(x|A)P(A) + p(x| \sim A)P(\sim A) \text{ and}$$

$$p(x|A) = \frac{248}{248 + 82} \approx 0.7515$$

$$p(x| \sim A) = \frac{167}{167 + 503} \approx 0.2493$$

$$P(A) = 0.33$$

$$P(\sim A) = 1 - P(A) = 0.67$$

thus

$$P(A|x) = \frac{0.7515 \times 0.33}{0.7515 \times 0.33 + 0.2493 \times 0.67} \approx 0.5976$$

Similarly, we have

$$P(\sim A|x) \approx 0.4024$$

 $P(A|\sim x) \approx 0.1402$
 $P(\sim A|\sim x) \approx 0.8598$

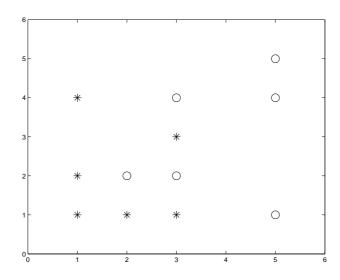
- 2. (25pts) Fisher Linear Discriminant.
 - (a) (5pts) What is the Fisher linear discriminant method?

Answer:

The Fisher linear discriminant finds a good subspace in which categories are best separated in a least-squares sense; other, general classification techniques can then be applied in the subspace.

(b) Given the 2-d data for two classes:

$$\omega_1=[(1,1),(1,2),(1,4),(2,1),(3,1),(3,3)]$$
 and $\omega_2=[(2,2),(3,2),(3,4),(5,1),(5,4),(5,5)]$ as shown in the figure:



i. (10pts) Determine the optimal projection line in a single dimension.

Answer:

Let w be the direction of the projection line, then the Fisher linear discriminant method finds that the best w is the one for which the criterion function $J(w) = \frac{w^t S_B w}{w^t S_w w}$ is maximum, as follows

$$\mathbf{w} = \mathbf{S}_{\mathbf{W}}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

where

$$\mathbf{S}_{\mathbf{W}} = \mathbf{S}_1 + \mathbf{S}_2$$

and

$$\mathbf{S}_i = \sum_{\mathbf{X} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \quad i = 1, 2$$

Thus, we first compute the sample means for each class and get

$$\mathbf{m}_1 = \left[\begin{array}{c} \frac{11}{6} \\ 2 \end{array} \right] \quad \mathbf{m}_2 = \left[\begin{array}{c} \frac{23}{6} \\ 3 \end{array} \right]$$

Then we subtract the sample mean from each sample and get

$$\mathbf{x} - \mathbf{m}_1 = \begin{bmatrix} -\frac{5}{6} & -\frac{5}{6} & -\frac{5}{6} & \frac{1}{6} & \frac{7}{6} & \frac{7}{6} \\ -1 & 0 & 2 & -1 & -1 & 1 \end{bmatrix}$$
$$\mathbf{x} - \mathbf{m}_2 = \begin{bmatrix} -\frac{11}{6} & -\frac{5}{6} & -\frac{5}{6} & \frac{7}{6} & \frac{7}{6} & \frac{7}{6} \\ -1 & -1 & 1 & -2 & 1 & 2 \end{bmatrix}$$

therefore

$$\mathbf{S}_{1} = \begin{bmatrix} \frac{25+25+25+1+49+49}{36} & \frac{5+0-10-1-7+7}{6} \\ \frac{5+0-10-1-7+7}{6} & 1+0+4+1+1+1 \end{bmatrix} = \begin{bmatrix} \frac{29}{36} & -1 \\ -1 & 8 \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} \frac{121 + 25 + 25 + 49 + 49 + 49}{36} & \frac{11 + 5 - 5 - 14 + 7 + 14}{6} \\ \frac{11 + 5 - 5 - 14 + 7 + 14}{6} & 1 + 1 + 1 + 4 + 1 + 4 \end{bmatrix} = \begin{bmatrix} \frac{53}{6} & 3 \\ 3 & 12 \end{bmatrix}$$

and then

$$\mathbf{S}_{\mathbf{W}} = \mathbf{S}_1 + \mathbf{S}_2 = \begin{bmatrix} \frac{41}{3} & 2\\ 2 & 20 \end{bmatrix}$$

$$\mathbf{S}_{\mathbf{W}}^{-1} = \frac{1}{|\mathbf{S}_{\mathbf{W}}|} \begin{bmatrix} 20 & -2\\ -2 & \frac{41}{3} \end{bmatrix} = \frac{1}{\frac{808}{3}} \begin{bmatrix} 20 & -2\\ -2 & \frac{41}{3} \end{bmatrix} = \begin{bmatrix} \frac{15}{202} & -\frac{3}{404}\\ -\frac{3}{404} & \frac{41}{808} \end{bmatrix}$$

Finally we have

$$\mathbf{w} = \mathbf{S}_w^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$$

$$\mathbf{w} = \begin{bmatrix} \frac{15}{202} & -\frac{3}{404} \\ -\frac{3}{404} & \frac{41}{808} \end{bmatrix} \begin{bmatrix} -2 \\ -1 \end{bmatrix} = \begin{bmatrix} -\frac{57}{404} \\ -\frac{29}{808} \end{bmatrix} \approx \begin{bmatrix} -0.1411 \\ -0.0359 \end{bmatrix}$$

ii. (10pts) Show the mapping of the points to the line as well as the Bayes discriminant assuming a suitable distribution.

Answer:

The samples are mapped by $\mathbf{x}' = \mathbf{w}^t \mathbf{x}$ and we get

$$w'_1 = [-0.1770, -0.2129, -0.2847, -0.3181, -0.4592, -0.5309]$$

 $w'_2 = [-0.3540, -0.4950, -0.5668, -0.7413, -0.8490, -0.8849]$

and we compute the mean and the standard deviation as

$$\mu_1 = 0.3304 \quad \sigma_1 = 0.1388$$

$$\mu_2 = 0.6485 \quad \sigma_2 = 0.2106$$

If we assume both $p(x|\omega_1)$ and $p(x|\omega_2)$ have a Gaussian distribution, then the Bayes decision rule will be

Decide
$$\omega_1$$
 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2

where

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} exp \left[-\frac{1}{2} \left(\frac{x - \mu_i}{\sigma_i} \right)^2 \right]$$

If we assume the prior probabilities are equal, i.e. $P(\omega_1') = P(\omega_2') = 0.5$, then the threshold will be about -0.4933. That is, we decide ω_1 if $\mathbf{w}^t \mathbf{x} > -0.4933$, otherwise decide ω_2 .

3. (20pts) Suppose $p(x|w_1)$ and $p(x|w_2)$ are defined as follows:

$$p(x|w_1) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$
, $\forall x$
 $p(x|w_2) = \frac{1}{4}$, $-2 < x < 2$

(a) (7pts) Find the minimum error classification rule g(x) for this two-class problem, assuming $P(w_1) = P(w_2) = 0.5$.

Answer:

(i) In case of -2 < x < 2, because $P(\omega_1) = P(\omega_2) = 0.5$, we have the discriminant function g(x) as

$$g(x) = \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} = \ln \frac{4}{\sqrt{2\pi}} - \frac{x^2}{2}$$

The Bayes rule for classification will be

Decide
$$\omega_1$$
 if $g(x) > 0$; otherwise decide ω_2

or

Decide
$$\omega_1$$
 if $-0.9668 < x < 0.9668$; otherwise decide ω_2

- (ii) In case of $x \geq 2$ or $x \leq -2$, we always decide ω_1 .
- (b) (10pts) There is a prior probability of class 1, designated as π_1^* , so that if $P(w_1) > \pi_1^*$, the minimum error classification rule is to always decide w_1 regardless of x. Find π_1^* .

Answer:

According to the question, π_1^* will satisfy the following equation

$$p(x|\omega_1)\pi_1^* = p(x|\omega_2)(1-\pi_1^*)$$
 when $x = 2$ or $x = -2$

Therefore, we have

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{4}{2}}\pi_1^* = \frac{1}{4}(1 - \pi_1^*)$$
$$\pi_1^* \approx 0.8224$$

(c) (3pts) There is no π_2^* so that if $P(w_2) > \pi_2^*$, we would always decide w_2 . Why not?

Answer:

Because $p(x|\omega_2)$ is only defined for -2 < x < 2, therefore we would always decide w_1 for $x \ge 2$ or $x \le -2$, no matter what is the prior probability $p(w_2)$.

- 4. (20pts) Let samples be drawn by successive, independent selections of a state of nature w_i with unknown probability $P(w_i)$. Let $z_{ik} = 1$ if the state of nature for the kth sample is w_i and $z_{ik} = 0$ otherwise.
 - (a) (7pts) Show that

$$P(z_{i1}, \dots, z_{in}|P(w_i)) = \prod_{k=1}^{n} P(w_i)^{z_{ik}} (1 - P(w_i))^{1 - z_{ik}}$$

Answer:

We are given that

$$z_{ik} = \begin{cases} 1 & if the state of nature for the k^{th} sample is \omega_i \\ 0 & otherwise \end{cases}$$

The samples are drawn by successive independent selection of a state of nature w_i with probability $P(w_i)$. We have then

$$Pr[z_{ik} = 1|P(w_i)] = P(w_i)$$

and

$$Pr[z_{ik} = 0|P(w_i)] = 1 - P(w_i)$$

These two equations can be unified as

$$P(z_{ik}|P(w_i)) = [P(w_i)]^{z_{ik}} [1 - P(w_i)]^{1 - z_{ik}}$$

By the independence of the successive selection, we have

$$P(z_{i1}, \dots, z_{in}|P(w_i)) = \prod_{k=1}^{n} P(z_{ik}|P(w_i))$$
$$= \prod_{k=1}^{n} [P(w_i)]^{z_{ik}} [1 - P(w_i)]^{1 - z_{ik}}$$

(b) (10pts) Given the equation above, show that the maximum likelihood estimate for $P(w_i)$ is

$$\hat{P}(w_i) = \frac{1}{n} \sum_{k=1}^{n} z_{ik}$$

Answer:

The log-likelihood as a function of $P(w_i)$ is

$$l(P(w_i)) = \ln P(z_{i1}, \dots, z_{in} | P(w_i))$$

$$= \ln \left[\prod_{k=1}^{n} [P(w_i)]^{z_{ik}} [1 - P(w_i)]^{1 - z_{ik}} \right]$$

$$= \sum_{k=1}^{n} [z_{ik} \ln P(w_i) + (1 - z_{ik}) \ln(1 - P(w_i))]$$

Therefore, the maximum-likelihood values for the $P(w_i)$ must satisfy

$$\nabla_{P(w_i)} l(P(w_i)) = \frac{1}{P(w_i)} \sum_{k=1}^n z_{ik} - \frac{1}{1 - P(w_i)} \sum_{k=1}^n (1 - z_{ik}) = 0$$

We solve this equation and find

$$(1 - \hat{P}(w_i)) \sum_{k=1}^{n} z_{ik} = \hat{P}(w_i) \sum_{k=1}^{n} (1 - z_{ik})$$

which can be rewritten as

$$\sum_{k=1}^{n} z_{ik} = \hat{P}(w_i) \sum_{k=1}^{n} z_{ik} + n\hat{P}(w_i) - \hat{P}(w_i) \sum_{k=1}^{n} z_{ik}$$

The final solution is then

$$\hat{P}(w_i) = \frac{1}{n} \sum_{k=1}^{n} z_{ik}$$

(c) (3pts) Interpret the meaning of your result in words.

Answer:

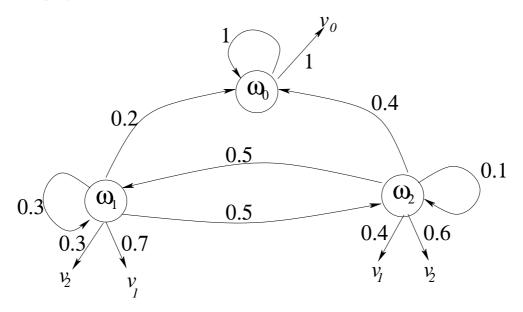
In this question, we apply the maximum-likelihood method to estimate the prior probability. From the result in part (b), it can be observed that the estimate of the probability of category w_i is merely the probability of obtaining its indicatory value in the training data, just as we would expect.

5. (20pts) Consider an HMM with an explicit absorber state w_0 and unique null visible symbol v_0 with the following transition probabilities a_{ij} and symbol probabilities b_{jk} (where the matrix indexes begin at 0):

$$a_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0.2 & 0.3 & 0.5 \\ 0.4 & 0.5 & 0.1 \end{pmatrix} \qquad b_{jk} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.7 & 0.3 \\ 0 & 0.4 & 0.6 \end{pmatrix}$$

(a) (7pts) Give a graph representation of this Hidden Markov Model.

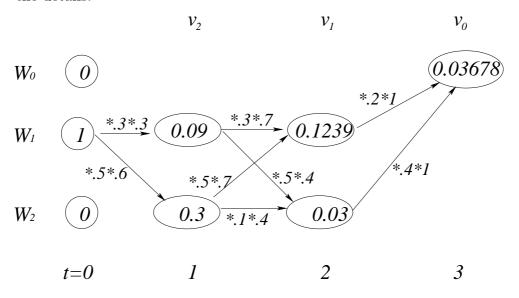
Answer:



(b) (10pts) Suppose the initial hidden state at t = 0 is w_1 . Starting from t = 1, what is the probability it generates the particular sequence $\mathbf{V}^3 = \{v_2, v_1, v_0\}$?

Answer:

The probability of observing the sequence \mathbf{V}^3 is 0.03678. See the figure below for the details.



(c) (3pts) Given the above sequence \mathbf{V}^3 , what is the most probable sequence of hidden states?

Answer:

From the figure above and by using the decoding algorithm, one can observe that the most probable sequence of hidden states is $\{w_1, w_2, w_1, w_0\}$.

Appendix: Useful formulas.

• For a 2×2 matrix,

$$A = \left[\begin{array}{cc} a & b \\ c & d \end{array} \right]$$

the matrix inverse is

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

 \bullet The scatter matrices \mathbf{S}_i are defined as

$$\mathbf{S}_i = \sum_{\mathbf{X} \in D_i} (\mathbf{X} - \mathbf{m}_i) (\mathbf{X} - \mathbf{m}_i)^t$$

where \mathbf{m}_i is the d-dimensional sample mean.

The within-class scatter matrix is defined as

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

The between-class scatter matrix is defined as

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$$

The solution for the W that optimizes $J(\mathbf{W}) = \frac{\mathbf{W}^t \mathbf{S}_B \mathbf{W}}{\mathbf{W}^t \mathbf{S}_W \mathbf{W}}$ is

$$W = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$