哈尔滨工业大学深圳研究生院

2016 年 秋 季学期期末考试试卷

HIT Shenzhen Graduate School Examination Paper

Course Name: Pattern Recognition **Lecturer:** Yong Xu (徐勇)

Question	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten	Total
Mark	78	22									100

Part One (Score of each exercise is 3)

C 1. There are c classes, $\omega_1, \omega_2, ..., \omega_c$. $P(\omega_1), ..., P(\omega_c)$ stand for prior probabilities of these classes. $P(\omega_1 \mid x), ..., P(\omega_c \mid x)$ denote probabilities of sample x belonging to these classes. Which proposition is not true?

A. $P(\omega_1 \mid x) + ... + P(\omega_c \mid x) = 1$.

B. $P(\omega_1) + ... + P(\omega_c) = 1$.

C. $P(x \mid \omega_1) + ... + P(x \mid \omega_c) = 1$.

A 2. A pattern classification system has the following main components: ① classification procedure; ②data collection; ③ feature extraction or feature selection ④ data pre-processing. Which term presents the proper chronological sequence (时间顺序) of these components in a pattern classification system ?

A. ② ④ ③ ①

B. ④ ② ③ ①

C. ③ ② ④ ①

D. ③ ② ① ④

(3) For a typical pattern classification system for real-world applications, there are several propositions on the training and testing stages (训练与测试阶段). Which proposition is not true?

A. In the training stages, a model, classifier or other things are obtained by exploiting all training samples. In the testing stage, the obtained things are applied to test samples to predict their class labels.

B. It is supposed that all training samples have known class labels and these labels can be exploited in the training stage.

C. In the testing stage, class labels of test samples cannot be used by the classification procedure.

D. In the testing stage, class labels of test samples can be used by the classification procedure.

A (4) Which judgement on the nearest neighbor classifier (NNC) is wrong?

A. Because its theoretical error rate is greater than that of the Bayesian minimum error classifier, NNC is a poor classifier.

B. NNC first calculates distances between a test sample and all training samples. Then NNC assigns the class label of the training sample with the greatest distance to the test sample to it.

C. NNC can be applied to a multi-class problem.

B (5) Which of the following propositions on principal component analysis (PCA) is right?

A. When PCA transforms samples into a new space, the dimensionality of the samples in the new space must be smaller than or equal to that of the original samples.

B. The naïve PCA method can be directly applied to the samples in the form of images and any additional step is not needed for the images.

C. The eigenvectors obtained using PCA is the eigenvectors of the mean matrix of the samples.

D. The eigenvectors obtained using PCA are not orthogonal (正交) and the components of the transform results may be statistically correlated.

6. Both the maximum likelihood estimation and Bayesian parameter estimation are important. Which option is incorrect ?

B

A. The maximum likelihood estimation does not take the prior probability of a class into account when it evaluates parameters of the normal distribution which is assumed to be the likelihood of this class.
B. Under any condition, the maximum likelihood estimation is impossible to obtain the same result as Bayesian parameter estimation.
C. Based on the maximum likelihood estimation we can conclude that the expectation of the likelihood of a class is the mean of all known samples of this class.

7. We define that action α_i assigns the sample into class ω_j. For the minimum-risk decision rule, which statement is right?

B

A. In all cases the result of minimum-risk decision cannot be the same as the result of minimum-error Bayesian decision.
B. Minimum-risk decision usually has a lower classification accuracy than minimum-error-rate Bayesian decision. However, minimum-risk decision is able to avoid possible high-risk or cost.
C. The following zero-one loss function is the best loss function:

$$\lambda_{ij}=\lambda(\alpha_i|\omega_j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases} \qquad i,j=1,\ldots,c.$$

D

8. For evaluating the likelihood of a class, the Bayesian parameter estimation method can be employed in the case where mean μ is the only unknown parameter and $P(x|\mu)\sim N(\mu,\sigma^2)$, $P(\mu)\sim N(\mu_0,\sigma_0^2)$. Moreover, the estimation result of μ is the following formulae.

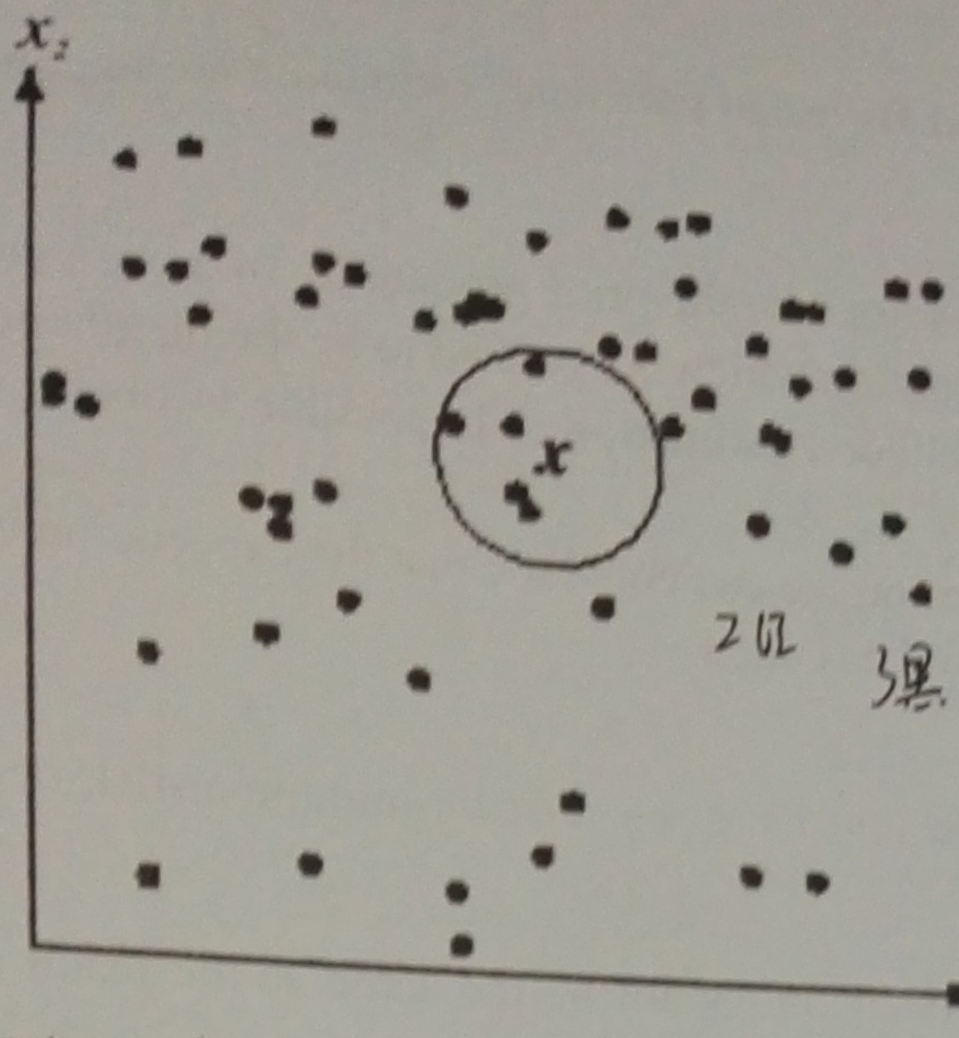
$$P(\mu|D)\sim N(\mu_n,\sigma_n^2) \quad , \quad \mu_n=\left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n+\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\cdot\mu_0, \quad \sigma_n^2=\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}, \hat{\mu}_n=\frac{1}{n}(x_1+\ldots+x_n) \quad , \quad D$$

means the sample set. Please select the correct sentence that provides right analysis.

Q

A. Uncertainty of μ increases with sample number n. In other words, if sample number n increases, then σ_n^2 increases.
B. It is impossible that μ_n equals to $\hat{\mu}_n$, the average of all samples $x_1,\ldots,x_n$.
C. The Bayesian parameter estimation method is absolutely better than the maximum likelihood estimation method.
D. When sample number n approaches the intensity (样本数n趋近无穷), μ_n equals to $\hat{\mu}_n$, the average of all samples $x_1,\ldots,x_n$.

9. In the following figure, denotation "x" stands for a test sample and other points denote training samples. The red training samples belong to a class and the black training samples belong to another class. The center of the circle is the location of "x", i.e. the test sample. Two red training samples and three black training samples are located within the circle. If the 5 nearest neighbor classifier (i.e. the k nearest neighbor classifier with k=5), then which conclusion is right?



A. The test sample should be classified into the class of the red training samples.
B. The test sample should be classified into the class of the black training samples.
C. The test sample should be rejected because no confident decision can be made.

10. For minimum-error-rate Bayesian decision, a discriminant function can be defined for every class and a sample is assigned to the class with the maximum discriminant function value. If $g_i(x)$ stands for the discriminant function of the i-th class and likelihood of each class is normally distributed(每类的似然均服从正态分布), then which definition of the discriminant function is improper.

A. $g_i(x) = -\frac{1}{2}(x-\mu_i)^t(\sum_i)^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$

B. $g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$

C. $g_i(x) = P(x|\omega_i) + P(\omega_i)$

11. The following two figures illustrate that "excessive training" or too complex model or algorithm may cause "overfitting". Please select the sentence that is wrong.

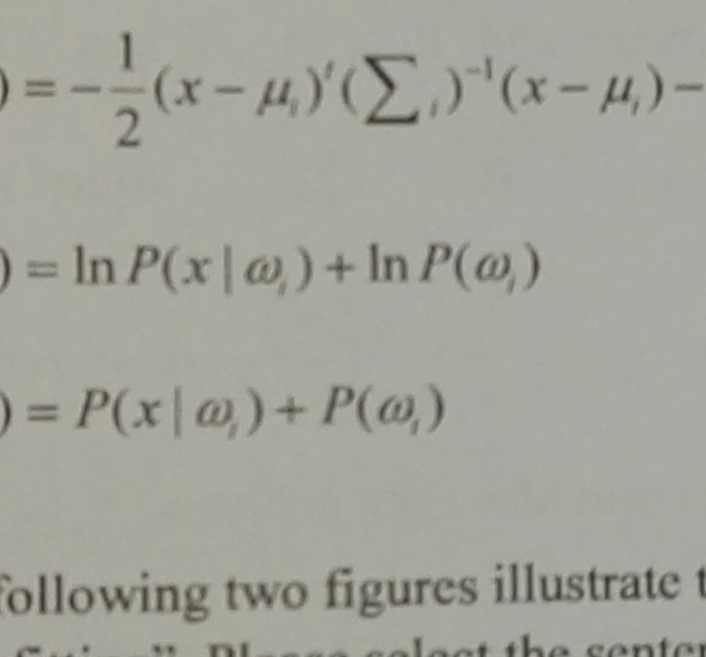
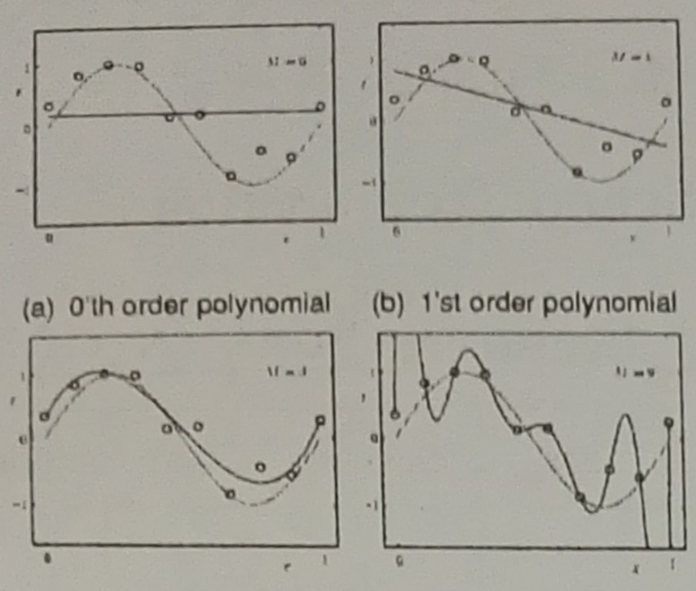


Figure 1: Expected generalization behavior due to overfitting.



(a) 0'th order polynomial (b) 1'st order polynomial

(c) 3'rd order polynomial (d) 9'th order polynomial

Figure 19: Polynomial curve fitting: plots of polynomials having various orders, shown as red curves, fitted to the set of 10 sample points.

A. If there are very many samples or data points, then the extent of "overfitting" may be alleviated.

B. Few samples or data points is beneficial to overcome "overfitting".

C. If samples or data points are sufficient enough, then we can select relatively complex model or algorithm.

12. If there are c classes, $\omega_1, \omega_2, ..., \omega_c$. $P(\omega_1 \mid x), ..., P(\omega_c \mid x)$ denote probabilities of sample x belonging to these classes. Action α_i assigns the sample into class ω_i. $\lambda(\alpha_i|\omega_j)$ denotes the loss (cost) in the case where the class label is j but action is adopted. Which proposition on the minimum-risk decision rule is not true?

A. $R(\alpha_i|x)=\sum_{j=1}^{c}\lambda(\alpha_i|\omega_j)P(\omega_j|x)$

B. $R(\alpha_i|x)=\sum_{j=1}^{c}\lambda(\alpha_i|\omega_j)P(\omega_j)$

C. If $R(\alpha_k|x)$ is the minimum among all $R(\alpha_1|x),...,R(\alpha_c|x)$, then x should be classified into k-th class.

13. Function $f(x)$ is defined as $f(x)=2+3x+4x^2$. Which presentation on its computational complexity is right?

A. "big oh" notation of $f(x)$ is sole and should be written as $f(x)=O(x^2)$.

B. "big θ" notation of $f(x)$ is sole and should be written as $f(x)=\theta(x^2)$.

C. "big θ" notation of $f(x)$ is sole and should be written as $f(x)=\theta(4x^2)$.

14. For the definition of the transition probabilities in the first-order hidden Markov model : $a_{ij}=P(\omega_j(t+1)|\omega_i(t))$, please choose the sentence that give right explanation.

A. The transition probability may be greater than 1.

B. $P(\omega_j(t+1)\mid\omega_i(t))$ means that the probability of the visible state changing from ω_i at time t to ω_j

time $t+1$.

C. $P(\omega_j(t+1)|\omega_i(t))$ means that the probability of the hidden state changing from ω_i at time t to ω_j at time $t+1$.

15. For the first-order hidden Markov models, please select the incorrect proposition.

A. Transition probabilities a_{ij} satisfy $\sum_i a_{ij}=1$.

B. The probability of the emission of a visible state satisfy $\sum_j b_{jk}=1$.

C. In any first-order hidden Markov model, $a_{ii}=0$ must be satisfied.

16. The following figure shows the result of the Parzen Windows estimation of distributions. In other words, after a number of samples are given, it is required to estimate the underlying distribution. n is the number of samples. The graphs shown in the third row are the estimation results under the condition that there are infinite samples. Also, the graphs shown in the third row can be viewed as true distributions. According to your experience, which column has the smallest "window"?

A. The first column.
B. The second column.
C. The third column.

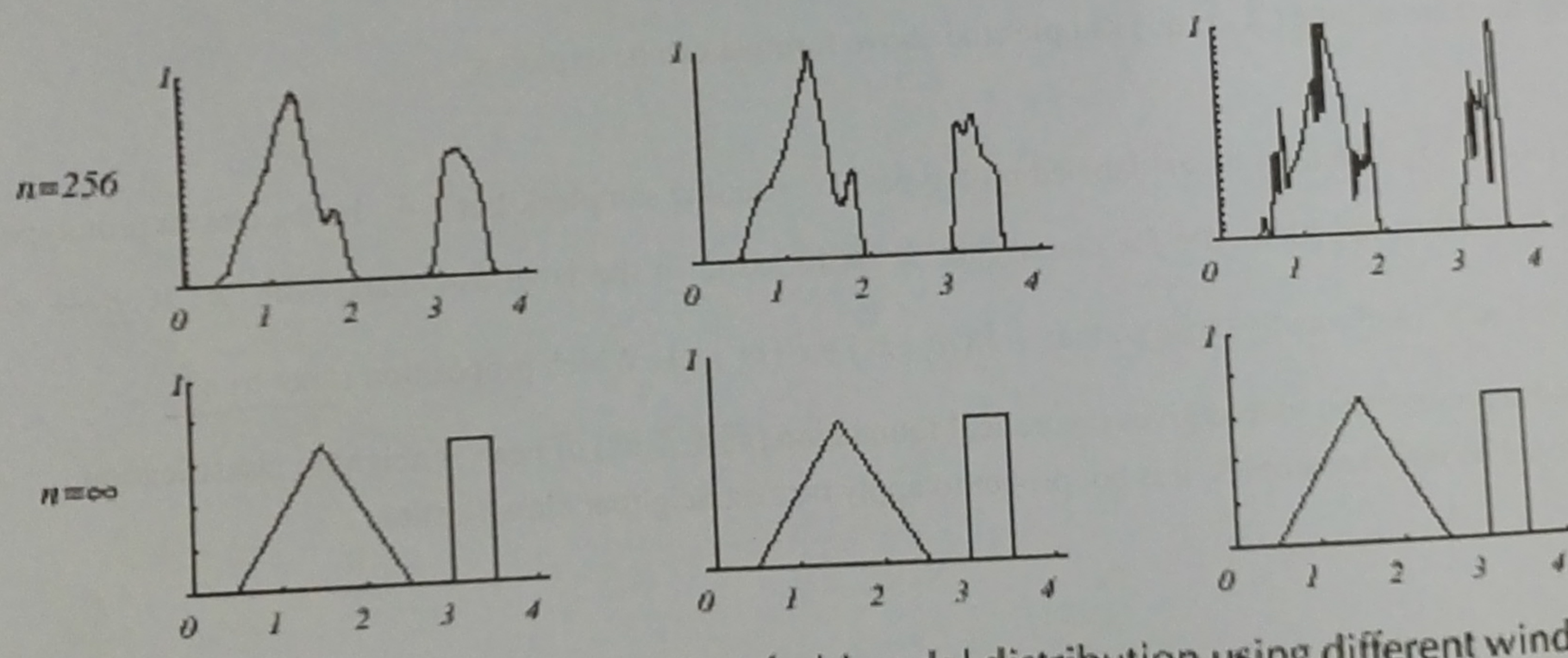
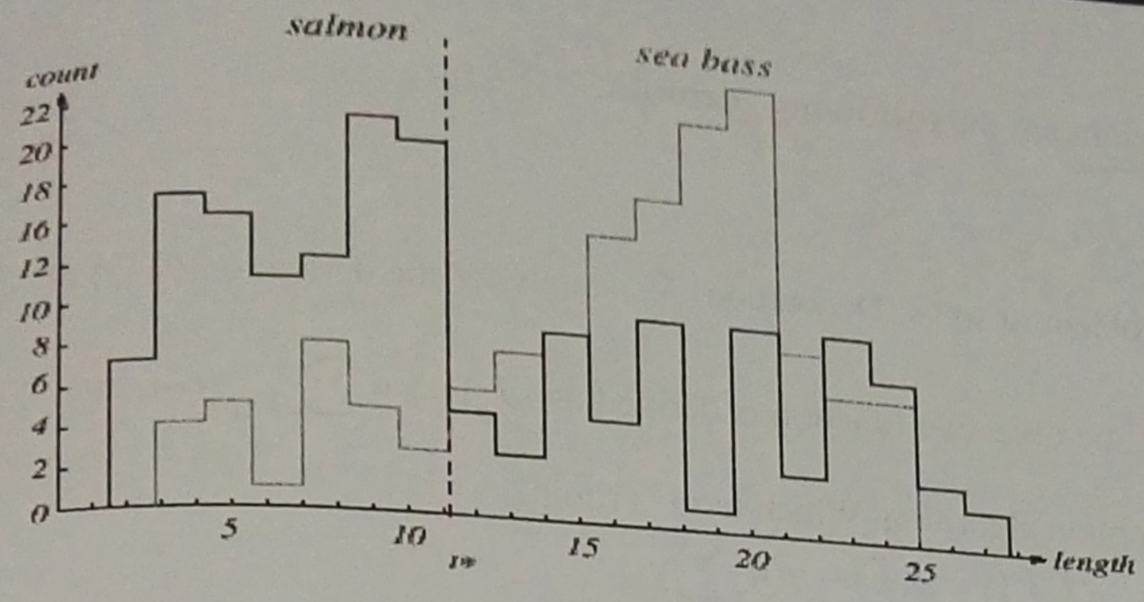


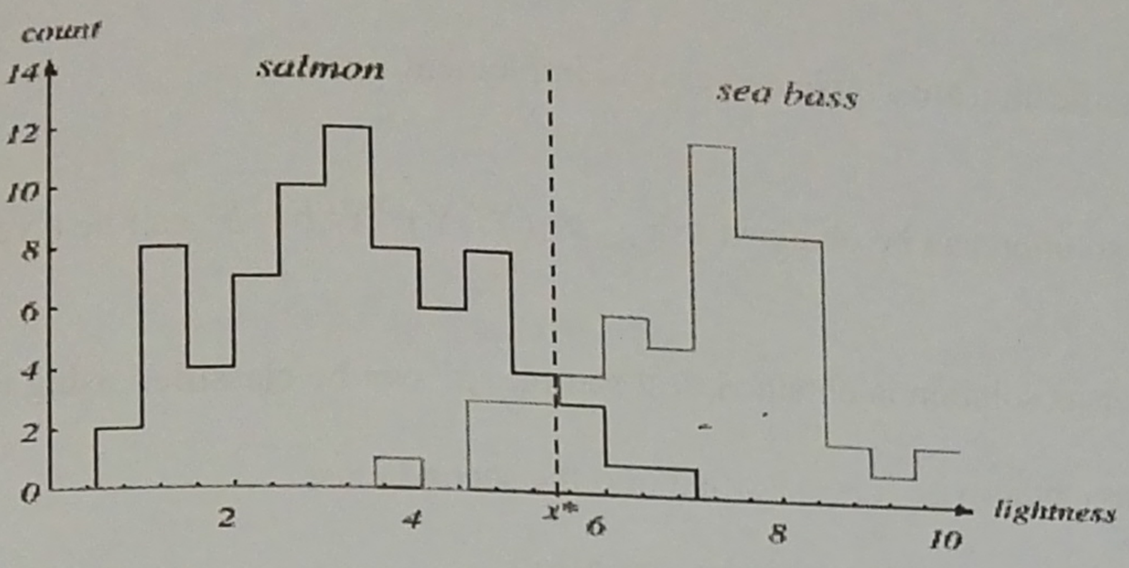
FIGURE 4.7. Parzen-window estimates of a bimodal distribution using different window widths and numbers of samples. Note particularly that the $n=\infty$ estimates are the same (and match the true distribution), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

17. The following two figures show two features (i.e. length and lightness) of "salmon" and "sea bass". Do you think which feature (i.e. length or lightness) is better for classification of "salmon" and "sea bass"?

A. length

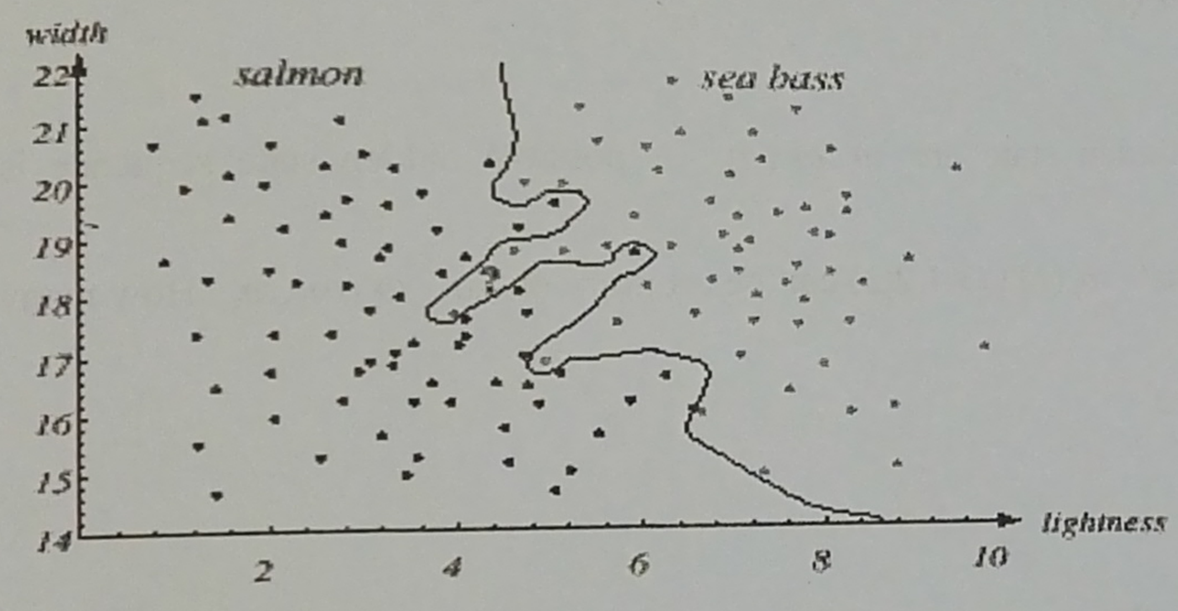
B. lightness



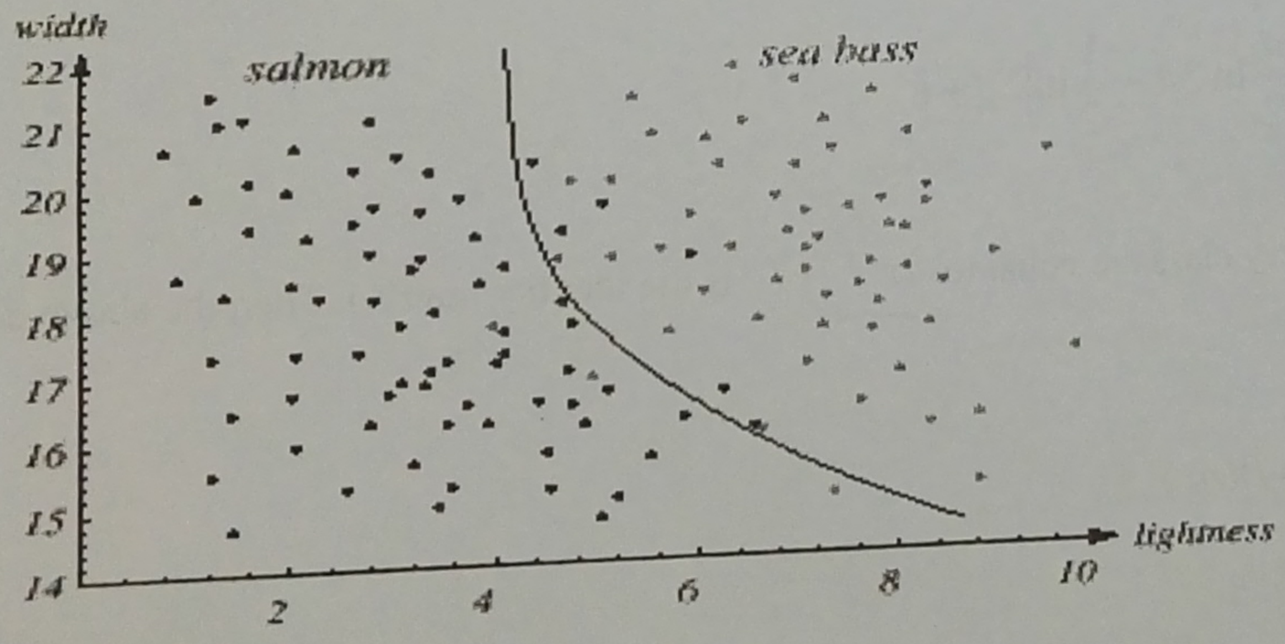


B

18. The following two figures show class boundaries obtained using two classification algorithms on a two-class problem. Points in two colors respectively stand for samples of two classes. We see that there are a limited number of samples. Which algorithm is better for classification ?



(1)



(2)

A. The algorithm to determine (1) is better, because the class boundary can lead to 100% classification accuracy.

B. The algorithm to determine (2) is better. Because the class boundary in (1) is too complex, the corresponding algorithm of (1) may generalize badly and its classification accuracy on test samples may be low.

A

19. As we know, the gradient descent algorithm has the following formula.

$$a(k+1) = a(k) - \eta(k)\nabla J(a(k))$$

$J(a(k))$ is the objective function on the problem to solve. Denotation ∇ stands for the derivation(偏导) of $J(a(k))$ with respect to a. For the Perception algorithm, the objective function is defined as $J_p(a) = \sum_{y \in Y(a)}(-a^T y)$. $Y(a)$ is the set of the samples mis-classified by the Perception algorithm. Which term is right ?

A. $a(k+1) = a(k) + \eta(k)\sum_{y \in Y} y$

B. $a(k+1) = a(k) - \eta(k)\sum_{y \in Y} y$

C

20. Minimum squared error procedures for classification are simple and easy to implement. Which presentation is not right ?

A. Suppose that there are only two classes. Its solution can be obtained using $a=(Y^T Y)^{-1} Y^T b$. b can be a vector with all entries being 1.

B. Suppose that there are only two classes. After its solution is obtained, test sample z can be classified using the following rule: if $z^T a > 0$, it is classified into the first class, otherwise it is classified into the second class.

C. The minimum squared error procedure cannot be extended to multi-class problems.

21. The formula to calculate probability of a visible state sequence $V = \{v(1),...,v(T)\}$ is

D

$$P(V^T) = \sum_{r=1}^{r_{max}} \prod_{t=1}^{t=T} P(v(t) \mid \omega_r(t))\, P(\omega_r(t) \mid \omega_r(t-1))$$

If $T = 3$ and all possible values of the hidden state are $\omega_1, \omega_2, \omega_3$. A possible hidden state sequence is denoted by $\omega_r = \{\omega_r(1), \omega_r(2), \omega_r(3)\}$. It is assumed that $\omega_r(t)$ (t=1,2,3) can be arbitrary one of $\omega_1, \omega_2, \omega_3$. How many hidden state sequences are there in total ?

A. 3 B. 6 C. 9 D. 27

22. When the likelihood of a class is normally distributed (正态分布), then the discriminant function of the Bayesi minimum error classification is

B

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t (\sum_i)^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + 1$$

If covariance matrices (协方差矩阵) of every class are equal to $\sigma^2 I$ (I is the identity matrix). Then the above discrim function can be simplified as

A. $$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i) + \ln P(\omega_i)$$

B. $$g_i(x) = -\frac{\| X-\mu_i \|^2}{2\sigma^2} + \ln P(\omega_i)$$

B

23. If there are two classes, ω_1, ω_2. It is known that $P(\omega_1 \mid x)=0.51$, $P(\omega_2 \mid x)=0.49$. Then the error rate of minimum-error Bayesian decision (classifier) on x is

A. 0　　B. 0.49　　C. 0.51

B

24. Principal component analysis (PCA) contains the following main steps: (1) calculation of eigenvectors (特征向量) and eigenvalues (特征值) of the covariance matrix (2) feature extraction of all samples (i.e. obtaining new features of samples) (3) calculation of the covariance matrix (4) calculation of the mean of all training samples. Please select the option that presents the right orders of these steps.

A. (1)-->(2)-->(3)-->(4)

B. (1)-->(3)-->(2)-->(4)

C. (1)-->(2)-->(4)-->(3)

D. (4)-->(3)-->(1)-->(2)

D

25. Which explanation on the following formula is not true?

$$p_n(\omega_i \mid x)=\frac{p_n(x,\omega_i)}{\sum_{j=1}^{j=c} p_n(x,\omega_j)}=\frac{k_i}{k}$$

A. Posteriori probabilities can be simply evaluated by the following formula.

B. k means that k neighbor of x are select ed.

C. k_i means that among the k neighbors k_i samples belong to the i -th class.

D. Only there are infinite (无穷的) samples,the above formula can be exploited.

B

26. Let $\{x_1, \cdots, x_n\}$ be a set of n labeled prototypes (i.e. training samples). Let x' be the closest prototype to a test point x, then the nearest-neighbor rule for classifying x is to assign it the label associated with x. If $n \to \infty$, it is always possible to find x' sufficiently close so that $P(\omega_i \mid x')=P(\omega_i \mid x)$. Which proposition is not true?

A. The above presentation indeed gives theoretical foundation (理论基础) of nearest neighbor classification.

B. If there are no infinite samples, it is not proper to apply nearest neighbor classification.

Part Two.

27. A patient takes a lab test and the result is positive. The test returns a correct positive result in 98% of the cases in which the cancer disease is actually present, and a correct negative result in 97% of the cases in which the disease is not present. Furthermore, 0.8% of the entire population have this cancer disease. Does the patient suffers from the cancer?