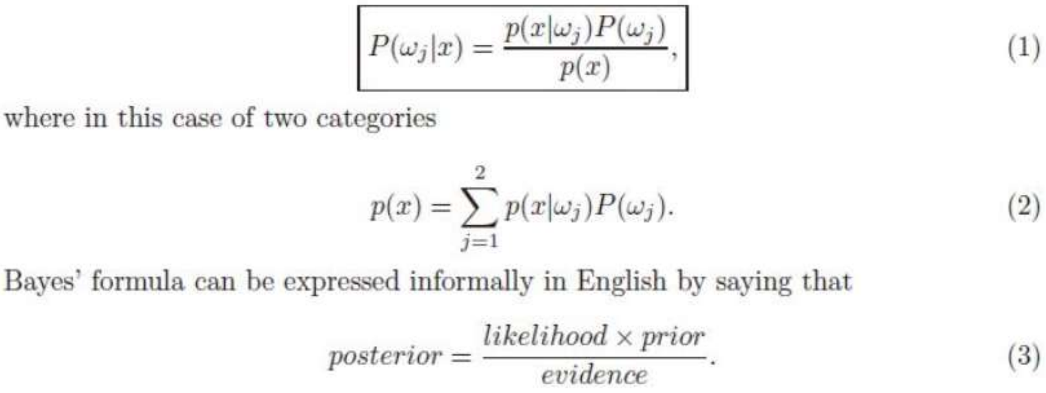
1. How to use the **prior** and **likehood** to calculate the **posterior**? What is the formula?

We can use the Bayes' formula to answer the question:



$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},} \qquad (1)$$

where in this case of two categories

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j). \qquad (2)$$

Bayes' formula can be expressed informally in English by saying that

$$posterior = \frac{likelihood \times prior}{evidence}. \qquad (3)$$

$\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j.

action α_i assign the sample into any class-

Conditional **risk** $R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j)P(\omega_j \mid x)$ for i = 1,…,a

Select the action α_i for which $R(\alpha_i \mid x)$ is minimum

R is minimum and R in this case is called the Bayes risk = best reasonable result that can be achieved!

What's the difference in the ideas of the **minimum error Bayesian decision** and **minimum risk Bayesian decision**? What's the condition that makes the minimum error Bayesian decision identical to the minimum risk Bayesian decision?

the minimum error Bayesian decision: to minimize the classification error of the Bayesian decision.

the minimum risk Bayesian decision: to minimize the risk of the Bayesian decision.

$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$

$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$ "decide ω_1" is taken

$(\lambda_{21}-\lambda_{11})P(\omega_1 \mid x) > (\lambda_{12}-\lambda_{22})P(\omega_2 \mid x)$ "decide ω_1" is taken

Condition: factor $(\lambda_{21}-\lambda_{11})$, $(\lambda_{12}-\lambda_{22})$ both are positive, and $(\lambda_{12}-\lambda_{22}) > (\lambda_{21}-\lambda_{11})$

λ_{ij} :loss incurred for deciding ω_i when the true state of nature is ω_j

$g_i(x) = -R(\alpha_i \mid x)$ 最小化风险

max. discriminant corresponds to min. risk

$g_i(x) = P(\omega_i \mid x)$ 最大化概率

max. discrimination corresponds to max. posterior

$g_i(x) \equiv p(x \mid \omega_i)P(\omega_i)$ $g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$ 最大化后验概率

问题由估计似然概率变为估计正态分布的参数问题

极大似然估计和贝叶斯估计结果接近相同，但方法概念不同

Please present the basic ideas of the maximum likelihood estimation method and Bayesian estimation method. When do these two methods have similar results？
请描述最大似然估计方法和贝叶斯估计方法的基本概念。什么情况下两个方法有类似的结果？

I. Maximum-likelihood view the parameters as quantities whose values are fixed but unknown. The best estimate of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed.

II. Bayesian methods view the parameters as random variables having some known prior distribution. Observation of the samples converts this to a posterior density, thereby revising our opinion about the true values of the parameters.

III. Under the condition that the number of the training samples approaches to the infinity, the estimation of the mean obtained using Bayesian estimation method is almost identical to that obtained using the maximum likelihood estimation method.

Please present the nature of principal component analysis.

1) PCA is an unsupervised method.
2) PCA is a good dimension reduction method. Usually, even if the dimension of the sample is greatly reduced, the main information can be still stored.
3) PCA is a de-correlation method. After the PCA transform, the obtained components are statistically uncorrelated.
4) PCA can be used as a compression method. PCA can achieve a high compression ratio.
5) PCA is the optimal representation method, which allows us to obtain the minimum reconstruction error.
6) As the transform axes of PCA are orthogonal, it is also referred to as an orthogonal transform method.

Disadvantage: need to a training phase.

Describe the basic idea and possible advantage of Fisher discriminant analysis.

Basic idea:

1) Fisher discriminant analysis's task is to find the best ~~such direction~~ w to obtain accurate classification. 不行么
2) Fisher discriminant analysis is to maximize the between-class distance and minimize the within-class distance 高内低
3) Fisher discriminant analysis exploits the training sample to ~~produce~~ get transform axes.

Advantage:

1) The time complex is ~~O(d^2n)~~ $O(d^2n)$. It is linear for n
2) Fisher discriminant analysis is a linear discriminant method, thus it is easy to generalize.
3) Fisher discriminant analysis is supervised; it can exploit the training sample to produce the transform axes.

What is the **K nearest neighbor classifier**? Is it reasonable?

We can center a cell about **x** and let it grow until it captures *kn* samples, where *kn* is some specified function of *n*. These samples are the *kn nearest-neighbors* of **x**. If the density is high near **x**, the cell will be relatively small, which leads to good resolution. If the density is low, it is true that the cell will grow large, but it will stop soon after it enters regions of higher density.

- Classify x by assigning it the label most frequently represented among the k nearest

samples ~~and use~~ a voting scheme

It is reasonable. The nearest-neighbor rule leads to an error rate greater than the minimum possible value of the Bayes error rate; If the number of prototype is large (unlimited), the error rate of the nearest-neighbor classifier is never worse than twice the Bayes rate.

9. Please describe the **over-fitting problem**.

Over-fitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model which has been over-fit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

1) Over-fitting: The classification performs good on training dataset but bad on test datadset.
2) Reason: (i)The training sample is inadequate;(ii)There are difference between training samples and test samples;(iii)The classification focus on detail;(iv)Samples exist noise.
3) How to handle: (i)reduce dimensionality(ii)suppose all classes share same covariance matrix(iii) look for a better estimate for covariance matrix.

$$error_{train}(f_1) < error_{train}(f_2) \quad \overset{test}{>}$$

Can **the minimum squared error procedure** be used for binary classification ?

Yes, the minimum squared error procedure can be used for binary classification.

$$\begin{pmatrix} y_{10} & y_{11} & \cdots & y_{1d} \\ y_{20} & y_{21} & \cdots & y_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n0} & y_{n1} & \cdots & y_{nd} \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \\ \cdots \\ a_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}$$

$$Ya = b,\ Y = \begin{bmatrix} Y_1^T \\ . \\ . \\ . \\ Y_n^T \end{bmatrix}, Y_i^T = \begin{bmatrix} y_{i0} \\ . \\ . \\ . \\ y_{id} \end{bmatrix}.$$

$$\boxed{J_s(a) = ||Ya - b||^2 = \ldots}$$

$$\frac{\partial J_s(a)}{\partial a} = 0 \Rightarrow a = (Y^T Y)^{-1} Y^T b$$

A simple way to set b : if Y_i is from the first class, then b_i is set to 1; if Y_i is from the second class, then b_i is set to -1.

Another simple way to set b : if Y_i is from the first class, then b_i is set to $\frac{n}{n_1}$; if Y_i is from the second class, then b_i is set to $-\frac{n}{n_2}$.

17. What are the upper and lower bound of the classification error rate of the K nearest neighbor classifier ?↵

K 近邻方法的分类误差上界与下界是什么？↵

答：不同 k 值的 k 近邻法错误率不同，k=1 时为最近邻法的情况（上、下界分别为贝叶斯错误率 P^*和 $P^*(2-\frac{c}{c-1}P^*)$）。当 k 增加时，上限逐渐靠近下限---贝叶斯错误率 P^*。当 k 趋于无穷时，上下限重合，P= P^*，此时 k 近邻法已趋于贝叶斯决策方法达到最优。↵

The Bayes rate is p* , the lower bound on p is p* itself.↵

The upper bound is about twice the Bayes rate.s↵

(1)The nearest-neighbor rule leads to an error rate greater than the minimum possible value of the Bayes rate

(2)If the number of prototype is large (unlimited), the error rate of the nearest-neighbor classifier is never worse than twice the Baves rate (it can be demonstrated!)

MSE

27. Apply model **Ya=b** to perform classification.

$$\begin{pmatrix} y_{10} & y_{11} & \cdots & y_{1d} \\ y_{20} & y_{21} & \cdots & y_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n0} & y_{n1} & \cdots & y_{nd} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \cdots \\ a_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix} \Leftrightarrow Ya = b$$

If we define the error vector e by

$$\mathbf{e} = \mathbf{Ya} - \mathbf{b}$$

then one approach is to try to minimize the squared length of the error vector. This is equivalent to minimizing the sum-of-squared-error criterion function

$$J_s(\mathbf{a}) = \|\mathbf{Ya} - \mathbf{b}\|^2 = \sum_{i=1}^{n}(\mathbf{a}^t\mathbf{y}_i - b_i)^2.$$

A simple closed-form solution can also be found by forming the gradient

$$\nabla J_s = \sum_{i=1}^{n} 2(\mathbf{a}^t\mathbf{y}_i - b_i)\mathbf{y}_i = 2\mathbf{Y}^t(\mathbf{Ya} - \mathbf{b})$$

and setting it equal to zero. This yields the necessary condition

$$\mathbf{Y}^t\mathbf{Ya} = \mathbf{Y}^t\mathbf{b}$$

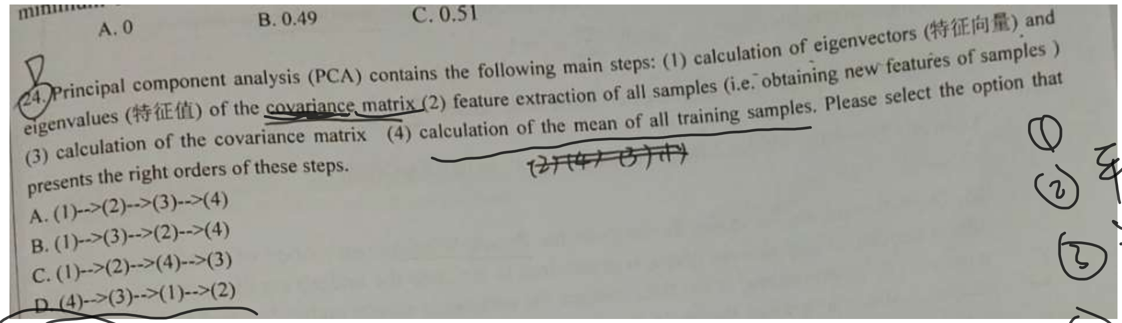
and in this way we have converted the problem of solving Ya = b to that of solving $\mathbf{Y}^t\mathbf{Ya} = \mathbf{Y}^t\mathbf{b}$.

$$\mathbf{a} = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{b}$$

$\mathbf{a} = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{b}$ is an MSE solution to Ya = b.

最小风险决策通常有一个更低的分类准确度相比于最小错误率贝叶斯决策。然而，最小风险决策能够避免可能的高风险和损失。

贝叶斯参数估计方法。



minimum...

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.

~~(2)(4)(3)(1)~~

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

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

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

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

① ② 平 ③ 差 ④ 协, co ⑤ 特征

Vectorize the samples.

Calculation of the mean of all training samples.

Calculation of the covariance matrix

Calculation of eigenvectors and eigenvalue of the covariance matrix. Build the feature space. eigen

Feature extraction of all samples. Calculation the feature value of every sample. ⑥

Calculation of the test sample feature value.

Calculation of the samples of training samples like the above step.

Find the nearest training sample as the result.

写在最后（彩蛋）： 2017 年考试题最后两个大题，一个是关于一阶马尔科夫模型，另一个是 LDA 目标及步骤。