Exercises

1. How to use the prior and likehood to calculate the posterior ? What is the formula ?
怎么用先验概率和似然函数计算后验概率？公式是什么？

$P(\omega_j \mid x) = p(x \mid \omega_j) \cdot P(\omega_j) / p(x)$

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

$$\sum P(\omega_j) = 1, \ \sum P(\omega_j \mid x) = 1$$

2. 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?
最小误差贝叶斯决策和最小风险贝叶斯决策的概念的差别是什么？什么情况下最小误差贝叶斯决策和最小风险贝叶斯决策是一致的（相同的）？

答：在两类问题中，若有 $\lambda_{12} - \lambda_{22} = \lambda_{21} - \lambda_{11}$，即所谓对称损失函数的情况，则这时最小风险的贝叶斯决策和最小误差的贝叶斯决策方法显然是一致的。

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.

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$ action α_1: "decide ω_1" is taken

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

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

$$\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

3. A person takes a lab test of nuclear radiation and the result is positive. The test returns a correct positive result in 99% of the cases in which the nuclear radiation is actually present, and a correct negative result in 95% of the cases in which the nuclear radiation is not present. Furthermore, 3% of the entire population are radioaetively eontaminated. Is this person

eontaminated?

一人在某实验室做了一次核辐射检测，结果是阳性的。当核辐射真正存在时，检测结果返回正确的阳性概率是 99%；当核辐射不存在时，结果返回正确的阴性的概率是 95%。而且，所有被测人群中有 3%的人确实被辐射污染了。那么这个人被辐射污染了吗？

答： 被辐射污染概率 $P(\omega_1)=0.03$

未被辐射污染概率 $P(\omega_2)=0.97$

X 表示阳性， $\bar{X}$ 表示阴性，则有如下结论：

$P(X\mid\omega_1)=0.99$ ，

$P(\bar{X}\mid\omega_2)=0.95$ 。

则 $P(\omega_1\mid X)=\dfrac{P(X\mid\omega_1)P(\omega_1)}{\sum_{i=1}^{2}P(X\mid\omega_i)P(\omega_i)}=\dfrac{0.99\times0.03}{0.99\times0.03+(1-0.95)\times0.97}\approx0.38$

$P(\omega_2\mid X)=1-P(\omega_1\mid X)=0.62$

根据贝叶斯决策规则有：

$P(\omega_2\mid X)>P(\omega_1\mid X)$

所以这个人未被辐射污染。

4. Please present the basic ideas of the maximum likehood 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？

答：I. 设有一个样本集 χ ，要求我们找出估计量 $\hat{\theta}$ ，用来估计 χ 所属总体分布的某个真实参数 θ 使得带来的贝叶斯风险最小，这就是贝叶斯估计的概念。

(另一种说法：把待估计的参数看成是符合某种先验概率分布的随机变量；对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，这样就利用样本的信息修正了对参数的初始估计值)

II. 最大似然估计法的思想很简单：在已经得到试验结果的情况下，我们应该寻找使这个结果出现的可能性最大的那个 θ 作为真 θ 的估计。

III.在训练样本数目接近无穷时，使用贝叶斯估计方法获得的平均值估计几乎和使用最大似然估计的方法获得的平均值一样

题外话：

从哲学的角度讲，最大似然估计属于经典学派，是唯物主义的思想，即将参数看作是确定的量，只是其具体值未知。

从哲学的角度讲，贝叶斯本人做过神甫，必然属于唯心主义的思想，他将参数看作是不确定的量，是受主观控制的。

Prior + samples

I. Maximum-likelihood view the parameters as quantities whose ~~vales~~ 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.

$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{P(D)}=\frac{\prod_{i=1}^{n}P(x_i|\theta)P(\theta)}{\int_\theta \prod_{i=1}^{n}P(x_i|\theta)P(\theta)d\theta}$$

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 likehood estimation method.

$$P(D)=\int_\theta P(D|\theta)P(\theta)d\theta=\int_\theta \prod_{i=1}^{n}P(x_i|\theta)P(\theta)d\theta$$

5. Please present the nature of principal component analysis.
请描述主成分分析法的本质
答：主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标。

- Capture the component that varies the most.(变化最大)
- The component that varies the most contains main information of the samples（信息量最大）
- We also say that PCA is the optimal representation method, which allows us to obtain the minimum reconstruction error.（最小重构误差）
- As the transform axes of PCA are orthogonal, it is also referred to as an orthogonal transform method.（正交变换）
- PCA is also a de-correlation method.（不相关法）
- PCA can be also used as a compression method and is able to obtain a high compression ratio.（高压缩比）

6. Describe the basic idea and possible advantage of Fisher discriminant analysis.
描述 Fisher 判别分析的基本概念和可能的优势
答：Fisher 准则是典型的模式识别方法，它强调将线性方法中的法向量与样本的乘积看做样本向量在单位法向量上的投影。

所获得的结果与正态分布协方差矩阵等的贝叶斯决策结果类似，这说明如果两类分布围绕各自均值的确相近，Fisher 准则可使错误率较小。

Supervised

Maximize the between-class distance and minimize the within-class distance

Exploit the training sample to produce transform axes.

……(number of effective Fisher transform axes, c-1; how to avoid singular within-class scatter matrix---PCA+FDA)

7. What is the K nearest neighbor classifier ? Is it reasonable ?

什么是 K 近邻分类器，它合理吗？

答：近邻法的基本思想是在测试样本 x 的 k 个近邻中，按出现最多的样本类别来作为 x 的类别，即先对 x 的 k 个近邻一一找出它们的类别，然后最 x 类进行判别。

在 k 近邻算法中，若样本相对较稀疏，只按照前 k 个近邻样本的顺序而不考虑其距离差别以决策测试样本 x 的类别是不适当的，尤其是当 k 取值较大时。

K nearest neighbor classifier view satisfy the k nearest neighbor rule ,the rule classifies x by assigning it the label most fequently represented among the k nearest samples; in other words, a decision is made b examining the labels on the k nearest neighbors and taking a vote.

8. Is it possible that a classifier can obtain a higher accuracy for any dataset than any other classifier?

一个分类器比其他分类器在任何数据集上都能获得更高的精度，可能吗？

答：显然不可能的。这个理由很多。

NO,

9. Please describe the over-fitting problem.

请描述过度拟合的问题

答：过拟合：为了得到一致假设而使假设变得过度复杂称为过拟合。想像某种学习算法产生了一个过拟合的分类器，这个分类器能够百分之百的正确分类样本数据（即再拿样本中的文档来给它，它绝对不会分错），但也就为了能够对样本完全正确的分类，使得它的构造如此精细复杂，规则如此严格，以至于任何与样本数据稍有不同的文档它全都认为不属于这个类别！

过拟合问题就是分类器分的太细了，太具体，

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.

10. Usually a more complex learning algorithm can obtain a higher accuracy in the training stage. So, should a more complex learning algorithm be favored ?

通常一个更复杂的学习算法在训练阶段能获得更高的精度。那么我就该选择更复杂的学习算法吗？

答：不

No context-independent or usage-independent reasons to favor one learning or classification method over another to obtain good generalization performance.

When confronting a new pattern recognition problem, we need focus on the aspects — prior information, data distribution, amount of training data and cost or reward functions.

Ugly Duckling Theorem: an analogous theorem, addresses features and patterns. shows that

in the absence of assumptions we should not prefer any learning or classification algorithm over another.

11. 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 likehood estimation method. Is this statement correct ?
在训练样本数目接近无穷时，使用贝叶斯估计方法获得的平均值估计几乎和使用最大似然估计的方法获得的平均值一样。这种情况正确吗？

答：理由同第 4 题，没找到。

YES

12. Can the minimum squared error procedure be used for binary classification ?
最小平方误差方法能用于 2 维数据的分类吗

答：略

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

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

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

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 $-\frac{n}{n_2}$.

13. Can you devise a minimum squared error procedure to perform multiclass classification ?
你能设计出一个能多级别识别的最小平方误差方法吗？

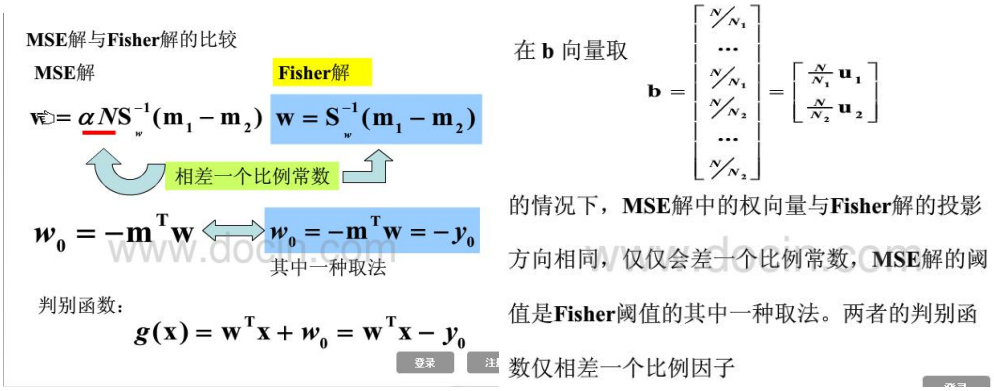
14. Which kind of applications is the Markov model suitable for ?

Markov 模型适合哪类应用？

答：Markov model has found greatest use in such problems， for instance speech recognition or gesture recognition.（语音、手势识别）

- The evaluation problem
- The decoding problem
- The learning problem

15. For minimum squared error procedure based on Ya=b (Y is the matrix consisting of all the training samples), if we have proper b and criterion function, then this minimum squared error procedure might be equivalent to Fisher discriminant analysis. Is this presentation correct ?
对于基于 Ya=b 的最小平方误差方法，如果我们有合适的 b 和判别函数，那么最小平方误差方法就会和 Fisher 判别方法等价。这么说对吗？
答：中文书 198 页，英文书 pdf 的 289 页，章节 5.8.2。
豆丁上的课件



16. Suppose that the number of the training samples approaches to the infinity, then the minimum error Bayesian decision will perform better than any other classifier achieving a lower classification error rate. Do you agree on this ?
假设训练样本的数目接近无穷，那么最小误差贝叶斯决策会比其他分类器的分类误差率更小。你同意这种观点吗？

答：待定

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

18. Can you demonstrate that a statistics-based classifier usually cannot lead to a classification accuracy of 100% ?

你能演示下基于统计的分类器不能导致 100%的准确度吗？

19. What is representation-based classification? Please present the characteristics of representation-based classification.
基于表征的分类是什么？请给出基于表征分类的特点？

20. A simple representation-based classification method is presented as follows:
一个简单的基于表征的分类方法如下

This method seeks to represent the test sample as a linear combination of all training samples and uses the representation result to classify the test sample:

这个方法寻求使用训练样本线性组合方法来表达测试样本，而且使用表征结果来分类测试样本：

$$y = b_1\tilde{x}_1 + ... + b_M\tilde{x}_M , \quad (1)$$

where $\tilde{x}_i$ ($i = 1,2,...,M$) denote all the training samples and b_i ($i = 1,2,...,M$) are the coefficients. We rewrite Eq.(1) into

$$y = \tilde{X}B , \quad (2)$$

where $B = [b_1...b_M]^T$, $\tilde{X} = [\tilde{x}_1...\tilde{x}_M]$. If $\tilde{X}$ is not singular, we can solve B using $B = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y$; otherwise, we can solve it using

$$B = (\tilde{X}^T\tilde{X} + \gamma I)^{-1}\tilde{X}^T y , \quad (3)$$

where γ is a positive constant and I is the identity matrix. After we obtain B , we refer to $\tilde{X}B$ as the representation result of our method. We can convert the representation result into a two-dimensional image having the same size of the original sample image.

We exploit the sum of the contribution, to representing the test sample, of the training samples from a class, to classify the test sample. For example, if all the training samples from the r th ($r \in C$) class are $\tilde{x}_s$ ... $\tilde{x}_t$, then the sum of the contribution, to representing the test sample, of the r th class will be

$$g_r = a_s\tilde{x}_s + ... + a_t\tilde{x}_t . \quad (4)$$

We calculate the deviation of g_r from y using

$$D_r = \| y - g_r \|^2 , r \in C . \quad (5)$$

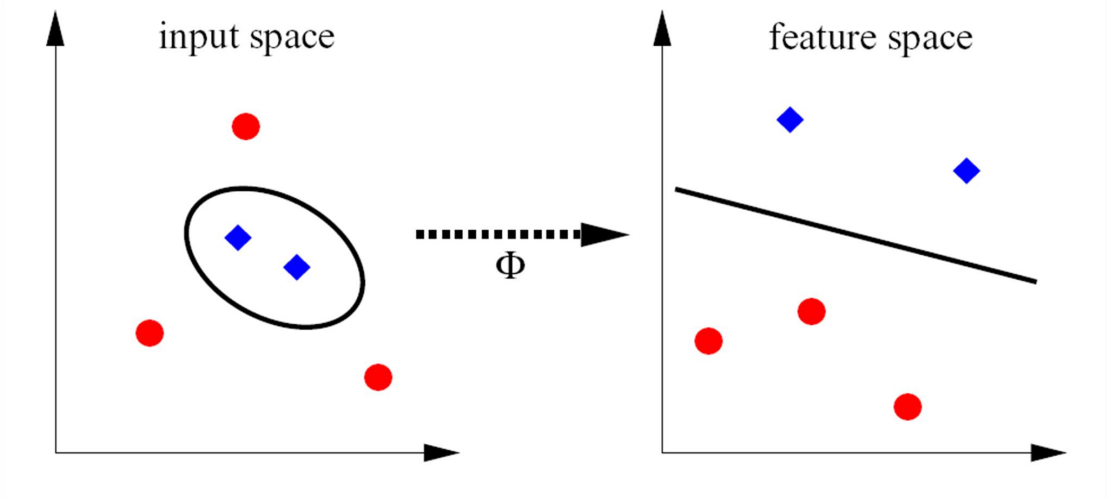
We can also convert g_r into a two-dimensional matrix having the same size of the original sample

image. If we do so, we refer to the matrix as the two-dimensional image corresponding to the contribution of the r th class. The smaller the deviation D_r , the greater the contribution to representing the test sample of the r th class. In other words, if $D_q = \min D_r$ ($q, r \in C$), the test sample will be classified into the q th class.

From the above presentation, we know that representation-based classification method is a novel method and totally different from previous classifiers ! It performs very well in image-based classification, such as face recognition and palmprint recognition.

We should understand its nature and advantages.

21. Please describe the difference between linear and nonlinear discriminant functions? What potential advantage does nonlinear discriminant function have in comparison with linear discriminant function?
请描述线性非线性判别函数的差别？非线性判别函数和线性判别函数比较有什么潜在的优势？
答：I. 简单的说线性判别函数就是其函数图形是直线、平面，非线性判别函数则相反，函数图形是曲线、曲面，不是直线、平面。
II. 在实际中有许多模式识别问题并不是线性可分的，应采用非线性分类器进行设计。例如当两类样本分布具有多峰性质并互相交错时，简单的线性判别函数往往会带来较大的分类错误。



The above figure is just auxiliary for the question !

22. What is the naïve Bayes rule ?

什么是朴素贝叶斯准则

答：朴素贝叶斯分类是一种十分简单的分类算法，叫它朴素贝叶斯分类是因为这种方法的思想真的很朴素，朴素贝叶斯的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。通俗来说，就好比这么个道理，你在街上看到一个黑人，我问你你猜这哥们哪里来的，你十有八九猜非洲。为什么呢？因为黑人中非洲人的比率最高，当然人家也可能是美洲人或亚洲人，但在没有其它可用信息下，我们会选择条件概率最大的类别，这就是朴素贝叶斯的思想基础。

~~23. What is the difference between supervised and unsupervised learning methods? Please show two examples of supervised and unsupervised learning methods.~~

~~监督学习方法和非监督学习方法的差别是什么？请分别给出监督学习方法和非监督学习方法的例子？~~

~~24. In some special real-world classification applications, the Bayesian decision theory might perform badly. What are possible reasons ?~~

~~在一些特殊的真实世界分类的应用中，贝叶斯决策理论可能表现很糟糕，可能的原因是什么？~~

25. Suppose that we are applying a linear discriminant function to a nonlinear separable problem, what means can we adopt to obtain an optimal solution?

    假如我们将一个线性判别函数应用到了一个非线性分割问题，为了获得一个最优解我们可以采取什么方法？

26. Please present possible generalization capability in the sample space of a method.

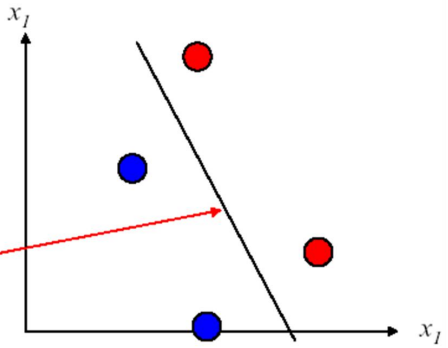
    请表达出在一个方法的样本空间里的可能的泛化能力？

27. Apply model Ya=b to perform classification.

    应用 Ya=b 模型来实施分类。

Example of Linear Classifier by Pseudoinverse

- ω_1: $(1,2)^t$ and $(2,0)^t$
- ω_2: $(3,1)^t$ and $(2,3)^t$



Sample Matrix (d = 1+2, n = 4)

$$Y = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \\ -1 & -3 & -1 \\ -1 & -2 & -3 \end{bmatrix}$$

$$a^t \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = 0$$

Pseudo-inverse

$$Y^* = (Y^t Y)^{-1} Y^t = \begin{bmatrix} 5/4 & 13/12 & 3/4 & 7/12 \\ -1/2 & -1/6 & -1/2 & -1/6 \\ 0 & -1/3 & 0 & -1/3 \end{bmatrix}$$

Assuming $b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

our solution is $a = Y^t b = \begin{bmatrix} 11/3 \\ -4/3 \\ -2/3 \end{bmatrix}$

Spring 2005

28. How to extend the binary minimum squared error procedure to the multiclass minimum squared error procedure?
怎么将 2 维最小平方误差方法扩展到多维最小误差平方方法？