模式识别基础

复习资料

教师：张学工

2009秋季

Xinbenlv@zzxy.org

清华大学 自动化系

2009秋季

Contents

[卷一、模式识别机经 3](#_Toc250310350)

[卷二、2008~2009秋\_模式识别\_张学工\_期末B 4](#_Toc250310351)

[卷三、2007~2008秋\_模式识别\_张学工\_期末B 6](#_Toc250310352)

[卷四、2007~2008秋\_模式识别\_张学工\_期末A 8](#_Toc250310353)

[卷五、2006~2007秋\_模式识别\_张学工\_期末B 9](#_Toc250310354)

[卷六、2005~2006秋\_模式识别\_张学工\_期末 10](#_Toc250310355)

[卷七、2004~2005秋\_模式识别\_张学工\_期末 11](#_Toc250310356)

[卷八、2003~2004秋\_模式识别\_张学工\_期末 12](#_Toc250310357)

[卷九、1999~2000春\_模式识别\_张学工\_期末 13](#_Toc250310358)

[附录、名词以及原理 15](#_Toc250310359)

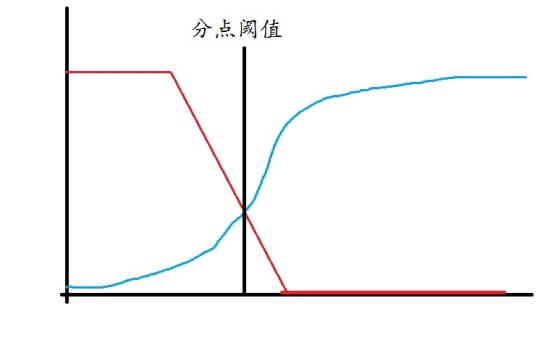
# 卷一、模式识别机经

答：  
１．设计最小错误率分类器：如果则为反之则为（特别的，当，可以接受为任何一类也可以拒绝。在连续情况下这种情况几乎完全不出现。

1.(16分)正态分布N(0,9)与均匀分布[5,10],先验概率1/2  
最小错误率分类器,并画图.最小错误率分类器缺点

，

2．画图如下，红色表示（[5,10]均匀分布的后验概率），蓝色表示（N(0,9)的后验概率）（注意！！！图是错的——作者）

（图是错的）

3．最小错误分类器有许多局限性。其一，其只考虑到错误率的限制，并未考虑到实际情况中不同错误的代价等等，为了克服这类局限性，学者们在最小错误率之外又设计了最小风险分类器、限定一类错误率令另一类错误率最小的分类器等；其二，其只能用于监督学习，要求已知分类情况，并且要满足I.I.D条件即样本与总体分布相符；其三，其要求必须知道先验概率，许多实际问题的先验概率是难以得到的。

2.(22分)8个样品w1 3个...,w2 5个.....  
1)最近邻分类,画图 2)给两个点..,问怎么分  
3)最近邻分类器线性分类器? 4)设计二次曲面分之  
5)压缩近邻法压缩过程及结果 6)压缩近邻法分界面并比较1)

答：（题目不完整）最近邻法（k=1），k近邻（k个最近的点投票），可以考虑加权

3.(5分)证明两个样品构成得协方差矩阵一般是奇异的（不可逆的）

答：设,从而其第i行j列元素为，表示第m个样本第n个维度的值。从而这是一个行列式为零的矩阵（第一行除以(）,第二行除以之后想减得零，故该协方差阵是奇异的。

4.(26分)"距离"不同会有不同结果  
1)给出两种你认为重要的距离,如何定义及物理意义  
2)贝叶斯分类器,近邻法,fisher投影法,c均值法是否受距离度量影响,不受则说明理由,  
受,距离说明

答：距离可以是任意一种范数，设对点

欧几里得距离（2范数）：

每个坐标的绝对值差之和（1范数）：

贝叶斯分类器不受、fisher投影法不受：都属于判别函数与距离度量无关的，故不受影响，近邻法，c均值法与受到距离度量的

5.(15分)AB玩游戏,A问是非问题,B回答是或非.  
问  
1)游戏与模式识别什么内容有关系?用模式识别语言描述  
2)B应如何调整问题  
3)如果做成计算机游戏,写出设计思路,方法,难点,解决方案.游戏受欢迎吗?为什么?

答：???这是一类决策树问题。可以将每一个是非问题看做一个决策树节点，是非答案决定了分支的方向（左或右）。调整问题应该使得在有效分类的情况下树深度最浅，就能在单次游戏最坏情况下最快猜出。（我觉得挺2的，应该不会受欢迎。。不过可以用作公安局犯罪嫌疑人指认的软件）

答：例如k-近邻法的模糊性质。定义隶属函数表示属于类的程度，值域上界为1，表示完全属于，下界为0，表示完全不属于，将对任意一个待判定的样本,选取某个距离度量（如欧几里得距离）最近的k个已知类别的训练样本，从而令得到x对于类的隶属函数。

6.(6分)自己设计模糊数学用于模式识别的其他理论方法(除去聚类).写出思路必要细节  
.分析可能结果.

7.(10分)专家经常根据特殊笔迹或特殊长相分类.问如何在一个人脸自动识别系统或笔迹  
自动识别系统中实现人的这一经验.从数据预处理,特征提取,选择分类器设计等角度描述  
实现这一经验的方法与可能性.优点?缺点?

答：这是一类特殊的图像识别问题，可以采用K-L分类

# 卷二、2008~2009秋\_模式识别\_张学工\_期末B

答：

1.（30'） 简要介绍下列概念  
·ROC曲线、交叉验证、似然函数、特征选择与提取、推广能力、随机森林

1. ROC曲线（Receiver Operational Characteristic Curve）是以False Positive Rate为横坐标，True Positive Rate为纵坐标的绘制的曲线，通过改变分点给出对应的FP和TP绘成。曲线的点表示了在敏感度和特殊性之间的平衡，例如越往左，也就是假阳性越小，则真阳性也越小。曲线下方的面积越大，则表示该方法越有利于区分两种类别。
2. 交叉验证是指交叉验证(Cross-validation)主要用于建模应用中，在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和，这个数越小，表示该方法越好
3. 特征选择与提取：特征是指对事物的观测或者某种运算，用于分类。特征提取是指直接观测到或者经过初步运算的特征——原始特征。特征选择是指从m个特征中选择m1个，m1<m。特征提取（特征压缩、变换）将m个特征变为m2个新特征。
4. 推广能力：在有限的样本上建立的学习机器对未来样本的表现。
5. 随机森林是指通过bootstrapping 样本来生成许多决策树，对于决策树上的每一个节点，选择m个变量来进行决策。计算训练集基于这m个变量的最佳分划。（Wiki:在[机器学习](http://zh.wikipedia.org/zh-cn/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0)中，**随机森林**是一个包含多个[决策树](http://zh.wikipedia.org/zh-cn/%E5%86%B3%E7%AD%96%E6%A0%91)的[分类器](http://zh.wikipedia.org/w/index.php?title=%E5%88%86%E9%A1%9E%E5%99%A8&action=edit&redlink=1)， 并且其输出的类别是由个别树输出的类别的[众数](http://zh.wikipedia.org/zh-cn/%E7%9C%BE%E6%95%B8)而定。 [Leo Breiman](http://zh.wikipedia.org/w/index.php?title=Leo_Breiman&action=edit&redlink=1)和[Adele Cutler](http://zh.wikipedia.org/w/index.php?title=Adele_Cutler&action=edit&redlink=1)发展出推论出随机森林的算法。 而 "Random Forests" 是他们的[商标](http://zh.wikipedia.org/zh-cn/%E5%95%86%E6%A8%99)。 这个术语是1995年由贝尔实验室的[Tin Kam Ho](http://zh.wikipedia.org/w/index.php?title=Tin_Kam_Ho&action=edit&redlink=1)所提出的**随机决策森林**（**random decision forests**）而来的。这个方法则是结合 Breimans 的 "[Bootstrap aggregating](http://zh.wikipedia.org/w/index.php?title=Bootstrap_aggregating&action=edit&redlink=1)" 想法和 Ho 的"random subspace method"" 以建造决策树的集合。）

答：

2.（30’）简述Fisher线性判别方法的基本原理并讨论如何在Fisher线性判别中实现最小风险的分类决策。

1. Fisher线性判别法是指：将多维判别决策问题通过线性组合降低为1维的时候，选择一个最优的一维线性判别函数的方法。该方法考虑了不同类别样本的协方差和均值之差。决策面法向量为。
2. 考虑最小风险分类决策的时候，需要甄选分界阈值点。从而可以根据贝叶斯决策规获得一种一维内的最小风险分类器。例如，经过投影后的分别服从的分布规律…P14页下方

3.（20’）简述C-均值方法的基本原理，并讨论在什么情况下此方法比较适用。

答：

C均值法是一种动态聚类法，通过多次迭代，逐步调整类别划分，最终使某准则达到最优。C均值法的目标是最小化——最小方差划分。用另一个角度来看，就是用C个码本来代表整个样本集，使这种表示带来的总体误差最小——向量量化。

误差平方和聚类准则：见课件9.3(p14)

适用条件：课件9.3 (p22)

4.（20’）SVM与多层感知器都可以实现非线性分类，试讨论它们在原理上的联系与区别。

答：

1. 支持向量机（SVM：Supporting Vector Mechine)原理：课件5.5 P6（特点是：P13），课本P301。

支持向量机的基本思想可以概括为：首先通过非线性变化将输入空间变换到一个高维空间，然后在这个新空间中求取最优先行分类面，而这种非线性变换是通过定义适当的内积函数K(X,Y)实现的。

1. 多层感知器（MLP:Multi-Layer Perception）原理：课件5.4 P10。

是一种前馈人工神经网络。它利用三层或者三层以上多个感知器来讲输入数据映射为复杂输出数据，使得分类器可以产生非线性的或者超平面的决策面，从而成为分线性分类器。(Wiki:A **multilayer perceptron** is a [feedforward](http://en.wikipedia.org/wiki/Feedforward) [artificial neural network](http://en.wikipedia.org/wiki/Artificial_neural_network) model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear [perceptron](http://en.wikipedia.org/wiki/Perceptron) in that it uses three or more layers of neurons (nodes) with nonlinear [activation functions](http://en.wikipedia.org/wiki/Activation_function), and is more powerful than the [perceptron](http://en.wikipedia.org/wiki/Perceptron) in that it can distinguish data that is not linearly separable, or separable by a [hyperplane](http://en.wikipedia.org/wiki/Hyperplane).[[1]](http://en.wikipedia.org/wiki/Multi-layer_perceptron#cite_note-0))

1. 联系与差异：都属于非线性分类器，但是其中一种是通过特征空间的非线性变换来实现最优决策面的，而多层感知器则是通过多层感知器（逻辑单元）的逻辑组合（输入-中间-输出）以及每层之间各数据的线性组合来实现非线性决策面的

5.（没有分）给课程提建设性建议。

1.通过《模式识别基础》课程的学习。。。

# 卷三、2007~2008秋\_模式识别\_张学工\_期末B

 1 .写出英文全称，中文，100字简述 5\*6=30  
a)Hierachical Clustering； b)MLP；c)Nonparametric Estimation； d)SVM  
e)Fisher's Linear Discriminant Analysis；f)PCA；g)Random Forest

1. Hierachical Clustering，分级聚类（课件9.4P45）：从各类只有一个样本点开始，逐级合并，每级只合并两类，直到最后所有样本都归到一类。聚类过程中考察类间相似度，依次决定类别数。
2. MLP，Multi-Layer Perception，多层感知器原理：课件5.4 P10。

是一种前馈人工神经网络。它利用三层或者三层以上多个感知器来讲输入数据映射为复杂输出数据，使得分类器可以产生非线性的或者超平面的决策面，从而成为分线性分类器。(Wiki:A **multilayer perceptron** is a [feedforward](http://en.wikipedia.org/wiki/Feedforward) [artificial neural network](http://en.wikipedia.org/wiki/Artificial_neural_network) model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear [perceptron](http://en.wikipedia.org/wiki/Perceptron) in that it uses three or more layers of neurons (nodes) with nonlinear [activation functions](http://en.wikipedia.org/wiki/Activation_function), and is more powerful than the [perceptron](http://en.wikipedia.org/wiki/Perceptron) in that it can distinguish data that is not linearly separable, or separable by a [hyperplane](http://en.wikipedia.org/wiki/Hyperplane).[[1]](http://en.wikipedia.org/wiki/Multi-layer_perceptron#cite_note-0))

1. Nonparametric Estimation：非参数估计(课件3.5.1，P36)

不假设概率模型，通过样本落在区域的概率来估计出一个概率密度函数。常用的有窗函数法（如近邻估计、Parzen窗估计）。

1. SVM，Supporting Vector Mechine，支持向量机：课件5.5 P6（特点是：P13），课本P301。

支持向量机的基本思想可以概括为：首先通过非线性变化将输入空间变换到一个高维空间，然后在这个新空间中求取最优先行分类面，而这种非线性变换是通过定义适当的内积函数K(X,Y)实现的。

1. Fisher's Linear Discriminant Analysis， Fisher线性判别分析:

Fisher线性判别法是指：将多维判别决策问题通过线性组合降低为1维的时候，选择一个最优的一维线性判别函数的方法。该方法考虑了不同类别样本的协方差和均值之差。决策面法向量为。

1. PCA，Principal Component Analysis，主成分分析：课件8.2P21，目的：在[统计学](http://zh.wikipedia.org/zh-cn/%E7%BB%9F%E8%AE%A1%E5%AD%A6)中，主成分分析（principal components analysis (PCA)）是一种简化数据集的技术。它是一个线性变换。这个变换把数据变换到一个新的坐标系统中，使得任何数据投影的第一大方差在第一个坐标（称为第一主成分）上， 第二大方差在第二个坐标（第二主成分）上，依次类推。主成分分析经常用减少数据集的维数，同时保持数据集的对方差贡献最大的特征。这是通过保留低阶主成 分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。但是，这也不是一定的，要视具体应用而定。
2. Random Forest，随机森林：随机森林是指通过bootstrapping 样本来生成许多决策树，对于决策树上的每一个节点，选择m个变量来进行决策。计算训练集基于这m个变量的最佳分划。（Wiki:在[机器学习](http://zh.wikipedia.org/zh-cn/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0)中，**随机森林**是一个包含多个[决策树](http://zh.wikipedia.org/zh-cn/%E5%86%B3%E7%AD%96%E6%A0%91)的[分类器](http://zh.wikipedia.org/w/index.php?title=%E5%88%86%E9%A1%9E%E5%99%A8&action=edit&redlink=1)， 并且其输出的类别是由个别树输出的类别的[众数](http://zh.wikipedia.org/zh-cn/%E7%9C%BE%E6%95%B8)而定。 [Leo Breiman](http://zh.wikipedia.org/w/index.php?title=Leo_Breiman&action=edit&redlink=1)和[Adele Cutler](http://zh.wikipedia.org/w/index.php?title=Adele_Cutler&action=edit&redlink=1)发展出推论出随机森林的算法。 而 "Random Forests" 是他们的[商标](http://zh.wikipedia.org/zh-cn/%E5%95%86%E6%A8%99)。 这个术语是1995年由贝尔实验室的[Tin Kam Ho](http://zh.wikipedia.org/w/index.php?title=Tin_Kam_Ho&action=edit&redlink=1)所提出的**随机决策森林**（**random decision forests**）而来的。这个方法则是结合 Breimans 的 "[Bootstrap aggregating](http://zh.wikipedia.org/w/index.php?title=Bootstrap_aggregating&action=edit&redlink=1)" 想法和 Ho 的"random subspace method"" 以建造决策树的集合。）

答：

2． 200字简述 4\*5=20  
a)最小距离分类器何时是最优分类器？  
b)写出最小距离分类器，分段线性判别，最近邻法的关系  
c)写出决策树和随机森林的关系，以及为什么要有随机森林  
        记不太清了，貌似有个SVM的原理？  
d)写出K-L变换和PCA的关系  
        写出特征提取和fisher线性判别的关系

1. 最小距离分类器：（课本P120页）最小距离分类器是指在贝叶斯决策规则当类条件概率密度为正态分布且各特征统计独立同方差时，利用贝叶斯决策规则在可以得到线性距离判别函数，特别是当时，决策规则式可以写成



这时的最小距离分类器是满足贝叶斯决策规则的最优分类器。

1. 最小距离分类器是一种分段线性的判别方法。每个样本点的类别由最近的代表点的类别决定。最近邻法是这种分类器的一种应用，特殊之处在于（课本P120页）最近邻法是将每一个样本都当做类别的代表点
2. 随机森林由多个决策树组成。随机森林的优缺点：

Wiki:The advantages of random forest are:

* For many data sets, it produces a highly accurate classifier
* It handles a very large number of input variables
* It estimates the importance of variables in determining classification
* It generates an internal unbiased estimate of the generalization error as the forest building progresses
* It includes a good method for estimating missing data and maintains accuracy when a large proportion of the data are missing
* It provides an experimental way to detect variable interactions
* It can balance error in class population unbalanced data sets
* It computes proximities between cases, useful for [clustering](http://en.wikipedia.org/wiki/Clustering), detecting [outliers](http://en.wikipedia.org/wiki/Outlier), and (by scaling) visualizing the data
* Using the above, it can be extended to unlabeled data, leading to unsupervised clustering, outlier detection and data views
* Learning is fast

The disadvantages of random forest are:

* Random forest are prone to overfitting for some datasets. This is even more pronounced in noisy classification/regression tasks.[[6]](http://en.wikipedia.org/wiki/Random_forest#cite_note-5)
* Random Forest does not handle large numbers of irrelevant features as well as ensembles of entropy-reducing decision trees.[[7]](http://en.wikipedia.org/wiki/Random_forest#cite_note-6)
* It is more efficient to select a random decision boundary than an entropy-reducing decision boundary, thus making larger ensembles more feasible. Although this may seem to be an advantage at first, it has the effect of shifting the computation from training time to evaluation time, which is actually a disadvantage for most applications.

3. 写出K-L变换和PCA的关系:PCA和K-L变换目的都是从高维的原始数据中提取低维度相关性低的特征数据。（白云飞：PCA针对样本协方差矩阵，找出特征值较大的，K-L变换针对产生矩阵，可以根据不同准则挑选特征值和特征向量，二者本质相同）

答：

3．构造两类分类器，100个训练数据，使用K近邻法，如何选择K，给出一个合适的K，说明理由（1\*5=5）

K值不宜太大，太大则太多干扰样本点被纳入，太小则无法适应小规模错误情况，对于100个训练数据，一般来说选择5、7、9比较合适。去奇数是为了避免出现两种类别数相同无法判决的情况。 答：

4．SOM若没有邻域作用，则变成什么算法？写出该算法的目标函数，算法基本步骤（1\*15=15）

SOM思路类似先作向量量化（聚类），用较少的点（子类）代表原样本空间，把这些代表点进行分类。(Wiki: A **self-organizing map (SOM)** or **self-organizing feature map (SOFM)** is a type of [artificial neural network](http://en.wikipedia.org/wiki/Artificial_neural_network) that is trained using [unsupervised learning](http://en.wikipedia.org/wiki/Unsupervised_learning) to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a **map**. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the [topological](http://en.wikipedia.org/wiki/Topology) properties of the input space.)

如果没有邻域作用，就退化一个典型的C均值法。

该算法计算步骤（课件9.3，P17）

5．综合题。文本自动分类的原理性设计，按内容分类为军事，体育，娱乐等。例举本科所学内容和概念能在这个设计上怎么体现。1\*30=30

答：因为是按已知内容分类，是一个监督学习的模式识别问题。

考虑如下监督学习模式

1.随机获取已经定义好类别的样本

1.特征提取：每个样本进行词频统计，词汇来自《辞海》。假设辞海有80000个词，则对每一个样本，得到一个80000维的特征数据。显然，这是一个高维数据。

2.特征处理：（1）首先去所有样本出现频度都小于一定阈值的词汇，比方说出现5次。剩下10000维。（2）对这10000维进行主成分分析，得到一个低维度的特征数据，比如4维。（3）对于这三维，可以进行一下分类：k近邻法、支撑向量机法（广义支撑向量机）、Fisher判别法等。其中Fisher判别再次降维后还可以运用基于最小错误率的贝叶斯决策

3．对于以上监督学习模式的样本可以进行同一个方法的交叉验证，去除一定的干扰。

附加题．说明两类错误率。在SVM线性不可分的情况下，如何把最小风险的概念引入SVM，对目标函数应该怎样修正。（最多10分，全卷不会超过100分）

答：

（1）两类错误率为“拒真纳伪”。

（2）（课本P298）如果线性不可分，就是某些样本不能满足条件时，引入松弛因子，使得条件变成，对于足够小的，只要使最小就能使错分样本数最小。对应线性可分情况下的使分类间隔最大，在线性不可分的时候引入约束，最终得到目标函数，C为某个指定常数，起着控制对错分样本惩罚的程度（风险）的作用。关于风险的讨论在课本P299页

# 卷四、2007~2008秋\_模式识别\_张学工\_期末A（略）

# 卷五、2006~2007秋\_模式识别\_张学工\_期末B

1.  写出英文全称并解释相应的概念或者方法的原理30%    
a)非参数估计    b)非监督模式识别  c)最大似然估计

d)MDS e)推广能力      f)kNN

答：其他见前文卷子。

* 1. MDS：[Multi-Dimensional Scaling](http://en.wikipedia.org/w/index.php?title=Multi-Dimensional_Scaling&action=edit&redlink=1),

Multidimensional scaling (MDS) is a set of related [statistical](http://en.wikipedia.org/wiki/Statistical) techniques often used in [information visualization](http://en.wikipedia.org/wiki/Information_visualization) for exploring similarities or dissimilarities in data. MDS is a special case of [ordination](http://en.wikipedia.org/wiki/Ordination_(statistics)). An MDS algorithm starts with a [matrix](http://en.wikipedia.org/wiki/Matrix_(mathematics)) of item–item [similarities](http://en.wikipedia.org/wiki/Similarity_(mathematics)), then assigns a location to each item in N-dimensional space, where N is specified a priori. For sufficiently small N, the resulting locations may be displayed in a graph or [3D visualisation](http://en.wikipedia.org/wiki/3D_visualisation).

* 1. 这里附上红宝书预测词汇：

### LDF:Linear Discriminant Functions线性判别函数。

### ISOMAP

[Isomap](http://en.wikipedia.org/wiki/Isomap)[[6]](http://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction#cite_note-5) is a combination of Curvilinear Distance Analysis (CDA) and [Multidimensional Scaling](http://en.wikipedia.org/wiki/Multidimensional_Scaling) (MDS). Both CDA and Isomap begin by finding a set of the nearest neighbors of each point. They seek to preserve the geodesic distances between points while projecting the data into fewer dimensions. Geodesic distance is defined as the length of the shortest path between two points that stays on the manifold surface. These algorithms estimate geodesic distances by finding the shortest neighbor-to-neighbor path between every pair of points, which can be computed efficiently using [Dijkstras algorithm](http://en.wikipedia.org/wiki/Dijkstras_algorithm).

**LLE**

Locally-Linear Embedding (LLE)[[8]](http://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction#cite_note-7) was presented at approximately the same time as Isomap. It has several advantages over Isomap, including faster optimization when implemented to take advantage of sparse matrix algorithms, and better results with many problems. LLE also begins by finding a set of the nearest neighbors of each point. It then computes a set of weights for each point that best describe the point as a linear combination of its neighbors. Finally, it uses an eigenvector-based optimization technique to find the low-dimensional embedding of points, such that each point is still described with the same linear combination of its neighbors. LLE tends to handle non-uniform sample densities poorly because there is no fixed unit to prevent the weights from drifting as various regions differ in sample densities. LLE has no internal model.

2.    如果用kNN来解决作业中的男女生训练样本问题，那么你如何选择k？简述理由

答：K=5,7,9,理由见卷子

3. 12%  无邻域相互作用的SOM的相当于是优化什么目标函数，写出目标函数。8%

答：C means，见前文

4. 1)简述两类错误率的概念，并由此谈一谈最小风险错误率Bayes分类器的原理？    8%    
   2)怎么把最小风险的概念引入线性SVM中（考虑线性不可分的情况），尝试写出公式6%  
   3)怎么把最小风险的概念引入Fisher线性判别中，简述一下。6%

答：见前文

5. 现在先要建立一种针对不同用户的文本邮件分类系统，比如分成私人邮件，公务邮件等。描述怎么建立，课上讲的哪些方法和概念可以用到？30%  

答：见前文

# 卷六、2005~2006秋\_模式识别\_张学工\_期末

张老师说：言多必失

1. 10%  简述最小错误率Bayes分类器的原理和适用范围

答：（NBA@zzxy说）：原理：通过观察 x 把状态的先验概率转化为状态的后验概率，比较不同的后验概率并作出决策。 适用范围：各类别总体的概率分布是已知的，要决策分类的类别数是一定的，离散数据，相较其它算法可以解决高维问题。

2. 10%  什么是分类器的错误率？如何估计？  
        简述两类错误率的概念，以及如何控制两类错误率的影响？

答：两类错误率是：拒真纳伪。控制两类错误率（1）选择更好的特征。（2）对特征进行提取。（3）变更分类器。（4）同一种分类器引入风险函数进行权衡（Tradeoff）

3. 10%  写出英文全称  
        模式识别        似然函数        非监督学习      非参数估计  
        SVM     MLP     PCA     MDS(必杀!)      k-NN    SOM

答：Pattern Recognition; Likelihood Function; Unsupervised Learning; Non-parametrical Estimation

4. 10%  无邻域相互作用的SOM的准则函数？(记不清怎么说的了..)

C-means，见前文

5. 15%  学号末两位和为奇数用身高数据，为偶数用体重数据  
        体重：  55      57      67      68      50  
        进行平均距离分级聚类，给出分类树，给出每级的距离矩阵。  
        聚为几类合适？为什么？

(((67,68)),50),(55,57)),分为两类合适。每级距离矩阵：(((67[1]68))[方差\*]50)[方差\*\*](55[1.414]57)),

6. 15%  参数估计和非参数估计的概念？最大似然估计和Parzen窗法的原理？  
        如何估计错误率？

答：参数估计：已知概率模型，根据样本对参数进行估计。非参数估计：未知概率模型，通过样本对概率模型进行直接估计。最大似然估计：在参数估计中选取使样本出现概率最大（Most Likelihood）的参数进行的估计。Parzen窗法是一种非参数估计方法（课件3.5.2P39），每个样本点对分布函数产生贡献用窗函数描述为k(x,xi)

7. 20%  以下线性分类器的设计原理？各自特性，联系？  
        A) 最小距离分类器   B) Fisher线性分类器  
        C) 感知准则函数分类器   D) 线性SVM

除感知准则函数分类器外原理见前文。

感知准则函数分类器：（课本P93）

8. 10%  给出一个模式识别的实例。

答：套用前文的例子

# 卷七、2004~2005秋\_模式识别\_张学工\_期末

见前文

一。给出了一组男生女生的身高体重数据（各十组），作为样本集  
然后给出了6组数据的测试集。给出了决策风险矩阵  
1。用最小风险bayes决策做分类器,并分类（学号单号用身高，双号  
用体重）  18分  
2。对测试集数据用分级聚类的方法聚类，要求写出距离矩阵和聚类  
树（同样单号用身高...）   12分

二。简答  
1。概率密度函数的参数估计和非参数估计的原理和特点。分别说明  
最大似然估计和Panzen窗法的条件和影响因素。  20分

2。C均值聚类和模糊C均值聚类的原理       20分

见前文，模糊C用

见前文

三。综合  
1。描述怎么建立一种方法对文本内容进行分类，比如分成军事，科  
技，教育等。描述从数据获取到特征选择到建立分类器等过程的方法  
等。    30分

# 卷八、2003~2004秋\_模式识别\_张学工\_期末（略，同卷七）

# 卷九、1999~2000春\_模式识别\_张学工\_期末（略，同卷七）

**附某次研究生课模式识别考题，张长水的。**1.一个两类分类问题如下：第一类的先验概率是2/3，第二类的先验概率是1/3，两类的  
概率密度函数为：p1是一个[1，10]的均匀分布，p2是一个[8，12]的均匀分布。请用贝  
叶斯决策方法按照最小错误率准则设计分类器。  
2已知一批样本为：0，2，3，3，3，4，6。  
 a.如果这些样本来自正态分布，请用最大似然方法估计该分布的均直和方差。并请分析  
样本数的多少对所估计的结果的影响。  
 b.如果不知道样本的分布，给出一种估计这些样本的概率密度函数的方法。详细说明估  
计方法、估计步骤，并给出估计结果。  
3已知2类样本如下，请用Fisher准则设计分类器，画出样本分布和投影直线，给出分界  
面方程和图形。简单说明类内协方差矩阵在Fisher判别中的作用。  
   w1:(1,0)T,(0,1)T  
   w2:(3,3)T,(3,5)T,(5,5)T,(5,3)T  
4有下列两类样本，请用1近邻法设计分类器，给出设计好的分类器，并画出样本分布和  
界面。  
   w1:(0,0)T,(1,1)T,(1,-1)T  
   w2:(1,0)T,(0,2)T,(0,-2)T  
5有下列样本。请用多级聚类方法对其聚类。类间距离采用最近距离。给出计算过程，画  
出多级聚类树，说明应该分成几类，并说明理由。用最近距离准则有什么缺点，请举例  
说明。  
  (1,0)T,(0,2)T,(0,4)T,(2,0)T  
6问答题  
  a.利用贝叶斯理论设计分类器对于学习模式识别有什么意义？其缺点是什么？  
  b.分析压缩近邻法的有点和缺点。  
  c.在特征提取和特征选择时，应找什么特征？  
  d.特征提取和特征选择的困难是什么？为什么说特征提取和特征选择是设计分类器的  
关键？  
  e.人工神经网络理论可以用于模式识别吗？就你所了解给出说明。  
7由N个样本构成的集合S被划分成c个互不相交的子集S1,S2,....Sc，如果Si是空集，则  
该集合样本的均值mi不定义。这种情况下，误差平方和仅仅和非空子集有关；  
J=对i求和对属于Si的x求和||x-mi||2  
其中，i是不包含空子集的子集符号。假定N>=c，请证明使得J最小的划分中没有空子集