<<模式分类>>整理与心得

第一章

模式识别:输入原始数据并根据其类别采取相应行为的能力

一个例子:鲑鱼和鲈鱼的区分

特征:双方在物理特性上的一些差异，比如长度、光泽、宽度等要素

模型:可以用数学模型表达的不同特征的描述

模型分类-感知数据并处理以滤除干扰-选择与数据最接近的类别

根据最小平均误差率和总体代价来确定分类所遵照的阈值

特征可以精简为多维特征向量

根据分类性能和分界面复杂度选择最优折中来确定分类器

统计模式识别:对模式的统计分类方法，把模式类看成是用某个随机向量实现的集合

句法模式识别:用模式的基本组成元素及其相互间的结构关系对模式进行描述和识别的方法

以上两个识别光看定义并没有分得很清楚，希望学长能解释区分一下

嵌入特定领域的背景知识，比如发音过程和椅子功能

这两个例子我觉得举得很好，当设计模式分类系统不知道从哪方面入手的时候，思考其本质和潜在共同点

回归分析:对输入数据找到合适的函数表示，来预测新数据的值

函数内插:从已知的一定范围内输入数据对应的函数值，来设法求出这些输入点之间的数据点的函数值

把已知的各点用曲线连起来得到连续函数不就可以了吗

密度函数估计:求解具有某种特定特征的类别成员出现的概率密度问题

模式识别系统:预处理-特征提取-分类

传感器-分割和组织-特征提取(结合具体领域知识)-分类器(其最简单的性能度量是分类误差率)-后处理(结合上下文)

设计一个模式识别系统:数据采集-特征选择-模型选择-训练和评估

可能出现过拟合的现象，考虑计算复杂度

新样本不令人满意也不一定由于过拟合，还有可能训练样本集不好

有监督学习:存在一个教师信号，对训练样本集中的每个输入样本能提供类别标记和分类代价，并寻找降低总体代价的方向

无监督学习:自动形成聚类或自然的组织，根据聚类系统采用的准则

何为聚类系统的准则，是科学家设计的很多数学公式构成的准则吗

强化学习:给定一个输入样本，计算它的输出类别，把它和已知的类别做比较，根据差异来改善分类器的性能

第二章

贝叶斯决策论是解决模式分类问题的一种基本统计途径,其出发点是利用概率的不同分类决策与相应的决策代价之间的定量折中

类条件概率密度函数即某类别状态下某特定特征值的概率密度

后验概率主要是由先验概率和似然函数的乘积决定的

贝叶斯决策规则将先验概率和似然概率两个因子结合起来以获得最小的误差概率

贝叶斯风险是最小化后的总风险值

对称损失或 0-1 损失函数默认所有误判都是等代价的

以极小化极大准则设计分类器，使先验概率取任何值时所引起的总风险最坏的情况尽可能小

Neyman-Pearson 准则规定对某个特定类别状态中做出误判的次数不允许超过某个限定值

针对多类情况的分类器可以视为一个计算 c 个判别函数并选取与最大判别值对应的类别的网络或机器，二分分类器则可视为计算一个简单判别函数并根据结果符号对特征向量分类的机器

正态分布和熵有密切关系，后者是用来描述从一种分布中随机选取的样本点值的不确定性

正态分布的判别函数:情况 1 发生在各特征统计独立且具有相同方差；情况 2 是所有类的协方差矩阵都相等但各自均值向量是任意的；情况 3 是每类协方差矩阵都不相同

使用线性判别函数的分类器称为线性机器，判定面是一些超平面

正态密度的误差上界有 Chernoff 界和 Bhattacharyya 界

检测器利用一个阈值来判定是否存在外部脉冲

ROC 曲线以虚警率为横坐标，击中率为纵坐标，可由这些概率决定判决能力

丢失特征:对于某个待识别的模式而言，其中的一种特征丢失了的情况

噪声特征:某个特定的特征量受到统计独立的噪声干扰

贝叶斯置信网采用有向无环图的拓扑形式

朴素贝叶斯规则或傻瓜贝叶斯规则可由简单的置信网来表示

上下文信息可用来帮助找到连续出现的类别之间的统计相关性

第三章

在得不到有关问题概率结构的全部知识时，通常利用训练样本来估计问题中所涉及的先验概率和条件密度函数，并把估计的结果当作实际的，从而设计分类器，后者的估计存在很大困难，但是如果已知参数的个数并能将条件概率密度进行参数化则可以降低难度。

参数估计问题最常用的两种方法是最大似然估计和贝叶斯估计。前者是把待估计的参数看作是确定性的量，取值未知，最佳估计是使得产生训练样本的概率为最大的值；后者是把带估计参数看成是符合某种先验概率分布的随机变量，观测样本的过程就是先把先验概率密度转化为后验概率密度从而修正对参数的初始估计值，后验概率密度函数会随着新观测样本的得到而变得更加尖锐，即为“贝叶斯学习”过程。

无论用何种参数估计方法，完成后都使用后验概率作为分类准则。

有监督学习对每一个样本所属的自然状态都是已知的；无监督学习中每个样本的自然状态都是未知的，所以更为困难。

存在一些非参数化的方法，先变换特征空间再采用参数化的方法，在判别分析法中，最重要的是 Fisher 线性判别函数。

最大似然估计分高斯情况(u 未知)，高斯情况(u 和 Σ 均未知)，对方差的最大似然估计是有偏的估计。

贝叶斯分类方法的核心是后验概率的计算。

贝叶斯参数估计：高斯情况（单变量情况，多变量情况）。

对于一般理论，即参数估计的贝叶斯方法，也是增量学习或在线学习算法的例子，其特点是学习过程随着观察数据的不断获得而不断进行最大似然方法和贝叶斯方法何时有哪个，若所使用的标准是计算复杂度或可理解性，则最大似然方法较好；若 p 的波形较宽或在参数向量附近是不对称的，明显的不对称性显然表示了分布本身的某种特点，最大似然方法忽略了这些特点，但贝叶斯方法能利用这些特点，它对偏差和方差之间的折中研究得更加透彻。

分类误差有几种：贝叶斯误差是由于不同的类条件概率密度函数之间的互相重叠引起的，永远无法消除；模型误差是由于选择了不正确的模型导致的分类误差；估计误差是由于采用了有限样本进行估计所带来的误差，其影响可用增加训练样本个数的方法来减小。

理论上，贝叶斯估计方法有很强的理论和算法基础，实际应用中最大似然估计更加简便。

无信息先验：假设每一类概率相同，能够达到客观性，样本本身能发挥出最大的作用，贝叶斯方法中不变形的考虑是非常有用的。

吉布斯算法：依据 p 仅选取一个参数向量，并且就把它当作真实值充分统计量（一个基本理论是因式分解定理）和核密度对于参数估计很重要，参数估计总涉及概率密度函数且这些概率密度常拥有简单形式的充分统计量和简单形式的核密度，对于任何分类规则，总能找到对应的基于充分统计量的分类方法，这些分类器常有相等或更好的分类效果。

维数问题研究的是特征维数和训练样本集大小对于分类精度有何影响，其次特征维数对于设计分类器时的计算复杂度有何影响。