模式识别的理论与方法

Pattern Recognition

裴继红

Chapter 3 (Part 3):
Maximum Likelihood and Bayesian Parameter Estimation

(Sections 3.6-3.8)

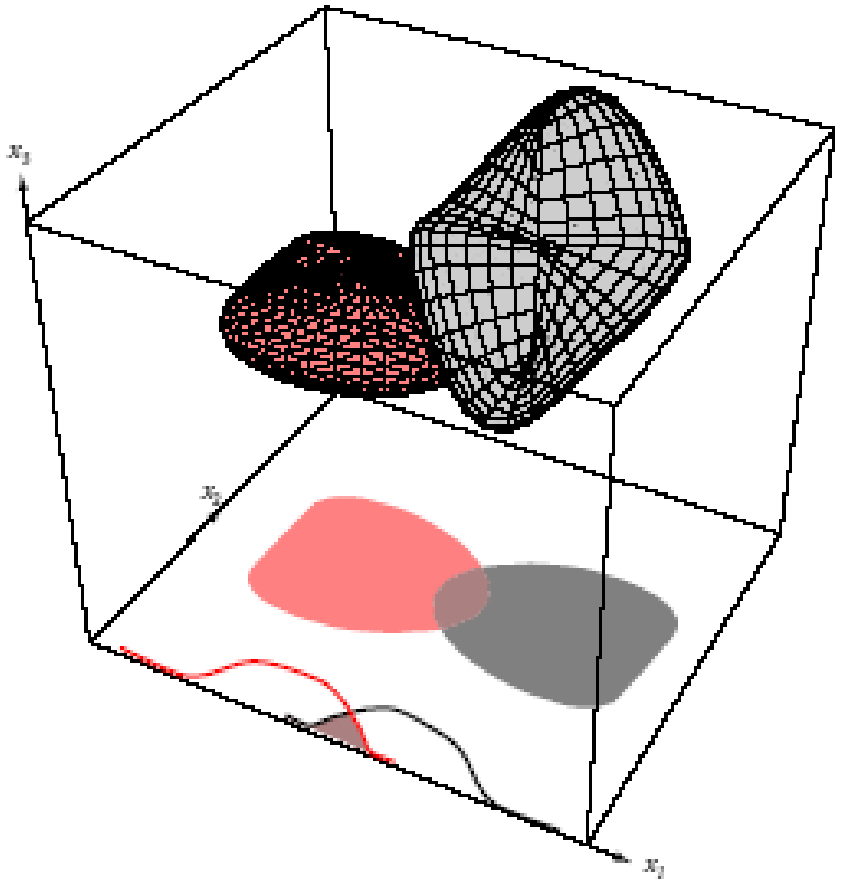
本讲内容：PCA

- 维数问题
- 计算复杂度
- 主成分分析 PCA

Figure3.3

- 两个在三维空间中无相互交叠分布密度，在三维空间中贝叶斯误差概率为0。

- 当将其投影到子空间：二维x1 – x2子空间，或一维x1子空间，分布密度的投影出现了较大的交叠，因此贝叶斯误差增大。



维数问题

- 在实际问题中，特征空间的维数一般很高

1. 包含50 到100维的特征空间并不难见到，尤其是在二进制特征下。

2. 分类精度依赖于维数和训练样本的数量

- 维数对分类精度有什么影响？维数对分类器设计的复杂度有何影响？

维数问题：计算复杂度

在两类、且具有相同协方差矩阵的多元正态分布情况下

$$P(error) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-u^2/2} du$$

$$where: \quad r^2 = (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)$$

$$\lim_{r \to \infty} P(error) = 0$$

维数对分类器设计的影响

- 两类、正态分布的情况，若特征之间是独立的，则:

$$\Sigma = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$$

$$r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$

- 最有用的特征是那些两类之间的均值距离大于类标准差的特征

- 在实际问题中经常可以看到，在特征空间的维数超过某个临界数量后，进一步加入更多的特征后，分类器性能反而会变坏。

计算复杂性

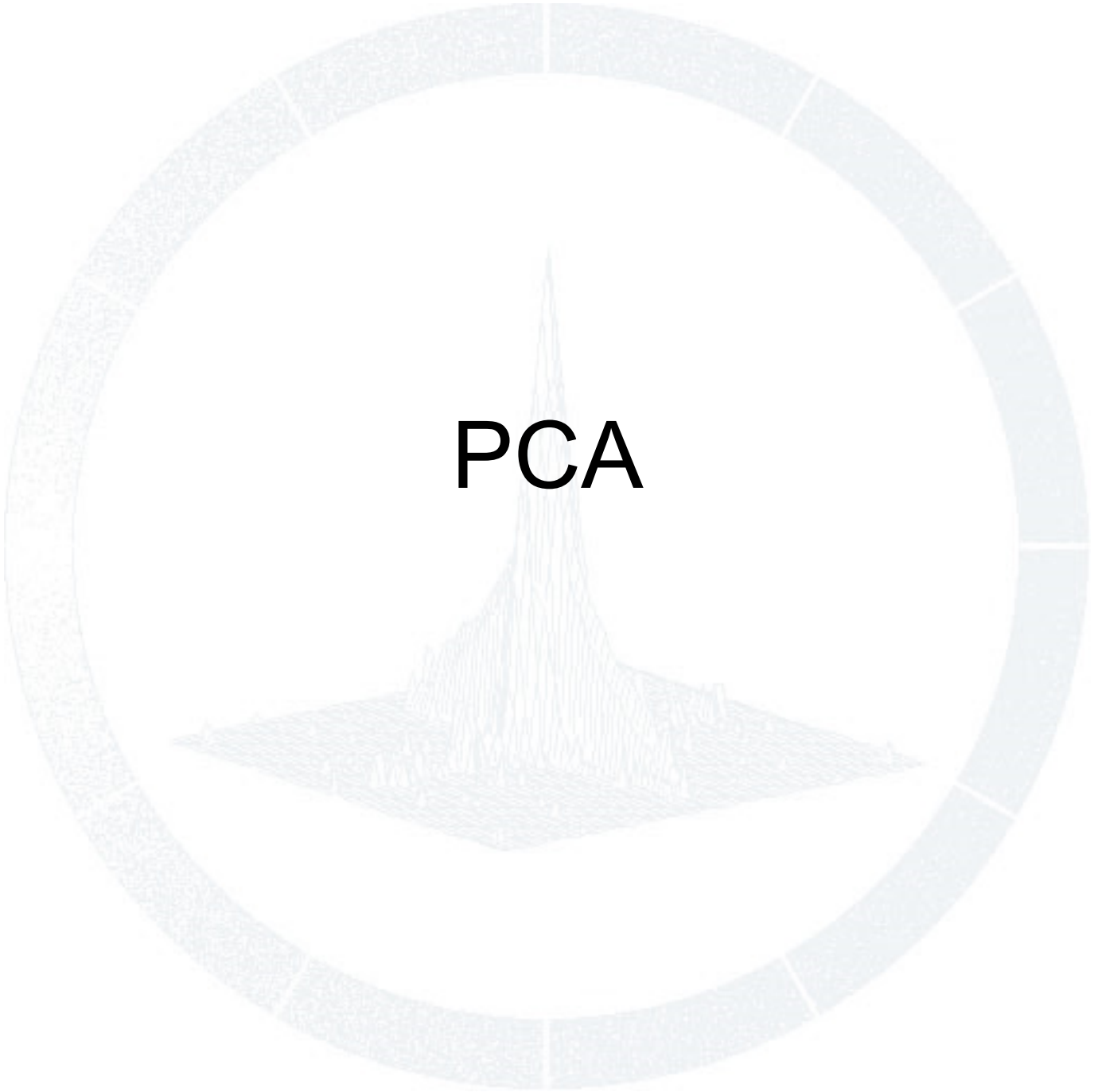
- 分类器设计，以及参数估计的方法的可计算性常常会受到计算复杂度的影响。

- 降低计算复杂度的方法之一：特征空间降维

特征空间降维的方法

- 可以采用特征线性组合的方法减少特征空间中特征的维数。将高维数据投影到较低维空间上。主成分分析和判别分析是两类有效的线性组合变换方法

1. 主成分分析（ Principal Component Analysis，PCA ）
   - 基本思路：寻找在最小均方误差意义下最能代表数据特性的投影方向，用这些方向矢量表示数据。
2. Fisher判别分析（Fisher Discriminant Analysis ，FDA）
   - 基本思路：在最小均方误差意义一下，寻找最能够分开各个类别数据的最佳方向。



PCA

主成分分析（也称主分量分析）
（PCA，Principal Component Analysis）

- 主分量分析的目的：通过特征的投影来降低线性特征空间的维数

  例如：假设一个由 n 个样本组成的数据集合：$\{ x_1, x_2, \ldots, x_n \}$，每个样本是一个 d 维特征矢量；

  - 希望使用模式空间的一个单一的点 $\mathbf{x}_0$来表示整个样本集合；
  - 并希望所选取的模式空间的点 $\mathbf{x}_0$ 与数据集合中各个样本点之间的距离都尽可能地小。

即最小化：
$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{x}_k\|^2$$

令：
$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$

为样本集合的均值。

样本集合的单点表达： 0 维表示

$$J_0(\mathbf{x}_0)=\sum_{k=1}^{n}\|\mathbf{x}_0-\mathbf{x}_k\|^2$$

$$=\sum_{k=1}^{n}\|(\mathbf{x}_0-\mathbf{m})-(\mathbf{x}_k-\mathbf{m})\|^2$$

$$=\sum_{k=1}^{n}\|\mathbf{x}_0-\mathbf{m}\|^2+\underbrace{\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2}_{\text{不依赖于}\mathbf{x}_0}$$

使均方误差最小的矢量 $\mathbf{x}_0$ 是均值矢量 $\mathbf{m}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$

- 样本均值的优点是表达简单，
- 缺点是其不能表达集合中的样本之间的差异。
- 样本集合的 0 维表示对了解数据集合可以提供的信息很少。

样本集合的 1 维表达：过均值矢量的直线

- 将样本集合的全部样本向通过均值矢量的一条直线作投影，可以得到样本空间中表示样本集合的一条直线，称之为样本集合的 1 维表达。

$$\mathbf{x} = \mathbf{m} + a\,\mathrm{e}$$

其中，e 表示通过均值的直线方向上的单位矢量；

a 是一个实数，表示直线上的点 $\mathbf{x}$ 距离均值点 $\mathbf{m}$ 的距离。

- 假设使用该直线上的一个点 $\mathbf{m} + a_k\,\mathbf{e}$ 来表示样本点 $\mathbf{x}_k$，则可以构造样本集合 的平方误差准则函数：

$$J_1\left(a_1, a_2, \cdots, a_n, \mathrm{e}\right) = \sum_{k=1}^{n} \left\| \left(\mathrm{m} + a_k \mathrm{e}\right) - \mathrm{x}_k \right\|^2$$

样本集合的 1 维表达：求最优 a_k 集合

- 求解准则函数的最优解

$$J_1(a_1,a_2,\cdots,a_n,\mathrm{e})=\sum_{k=1}^{n}\left\|(\mathrm{m}+a_k\mathrm{e})-\mathrm{x}_k\right\|^2$$

$$=\sum_{k=1}^{n}\left\|a_k\mathrm{e}+(\mathrm{m}-\mathrm{x}_k)\right\|^2$$

$$=\sum_{k=1}^{n}a_k^2\left\|\mathrm{e}\right\|^2-2\sum_{k=1}^{n}a_k\mathrm{e}^{\mathrm{t}}(\mathrm{x}_k-\mathrm{m})+\sum_{k=1}^{n}\left\|\mathrm{x}_k-\mathrm{m}\right\|^2$$

由于 ||e||=1，通过对 a_k 求偏导，并令结果为 0，可以得到

$$a_{\mathrm{k}}=\mathrm{e}^{\mathrm{t}}(\mathrm{x}_{\mathrm{k}}-\mathrm{m})$$

为向量 $(\mathrm{x_k}-\mathrm{m})$ 在直线 $\mathbf{x}=\mathbf{m}+a\,\mathbf{e}$ 方向上的投影长度。

样本集合的散布矩阵

- 设样本数据集合{ $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ }的均值矢量为$\mathbf{m}$，则该样本集合的散布矩阵（Scatter Matrix）定义为

$$S = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t$$

由上述定义可以看出，散布矩阵是样本协方差矩阵的 n-1倍

样本集合的1维表达：求直线的最优方向e

- 寻找直线的最优方向**e**，使得平方误差准则函数J_1值最小

将 $a_k = \mathrm{e}^{\mathrm{t}}(\mathrm{x}_k - \mathrm{m})$代入的 $J_1(a_1,a_2,\ldots,a_n, \mathrm{e})$ 公式，得

$$J_1(\mathbf{e}) = \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2$$

$$= -\sum_{k=1}^{n} \left[\mathrm{e}^{\mathrm{t}}(\mathrm{x}_k - \mathrm{m})\right]^2 + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2$$

$$= -\sum_{k=1}^{n} \mathbf{e}^{\mathrm{t}}(\mathrm{x}_k - \mathrm{m})(\mathrm{x}_k - \mathrm{m})^t \mathbf{e} + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2$$

$$= -\mathbf{e}^{\mathrm{t}}\mathbf{S}\mathbf{e} + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2$$

这里S是散布矩阵

样本集合的1维表达：求直线的最优方向e

由于$\|x_k-m\|^2$与e无关，所以使

$$J_1(e) = -e^tSe + \sum_{k=1}^{n}\|x_k - m\|^2$$

最小化的向量 **e** 应该使得 e^tSe 最大化，考虑到 $\|e\|=1$，

采用拉格朗日乘数优化方法

$$U = e^tSe - \lambda(e^te - 1)$$

令 $$\frac{\partial U}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} = 0$$

得 $\mathbf{S}\,\mathbf{e} = \lambda\,\mathbf{e}$ 即 **e** 是散布矩阵 **S** 的对应于特征值 λ 的特征矢量

$$e^tSe = \lambda\, e^te = \lambda$$

直线的最优方向 **e** 是散布矩阵S对应于最大特征值 λ 的特征矢量方向

样本集合的最优 d′ 维表达： d′ 维超平面

- 将上述结论从1 维推广到高维，使用d′<d维的特征矢量来近似原空间中的 d 维矢量

$$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$$

其中， $e_1, e_2, \ldots e_{d'}$是散布矩阵 S 的对应于d′个最大特征值的特征矢量;

这些矢量之间相互正交，构成了d′维空间的基矢量。

而系数 a_i 是矢量 x 对应于基 e_i 的系数，称为

主分量——Principal Component

主分量分析

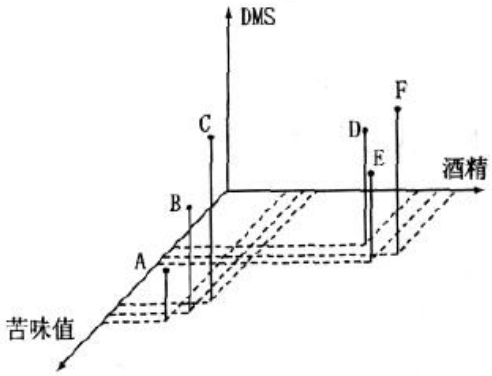
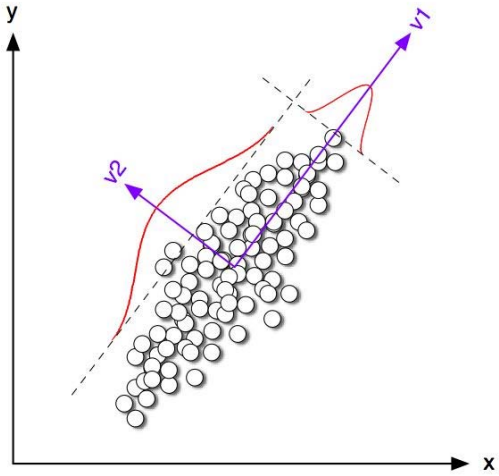
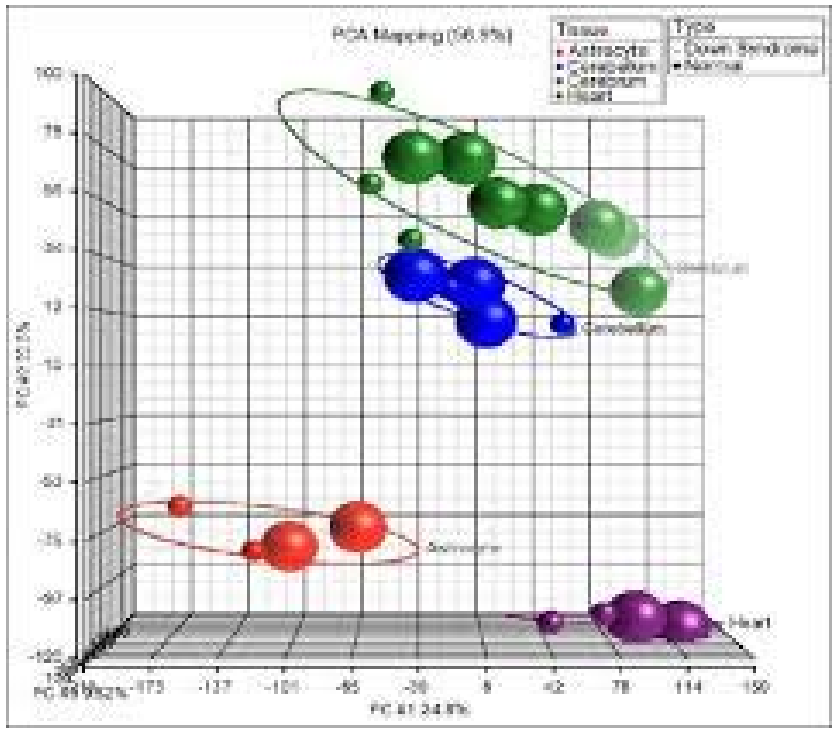


图 1　啤酒的三维坐标图





通过在较高维上的投影得到　$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$