Chapter 3 ML & 贝叶斯参数估计

Author： 中山大学 17数据科学与计算机学院 YSY

https://github.com/ysyisyourbrother

在典型的有监督模式识别问题中($P(w|x) \propto p(x|w)P(w)$)，估计先验概率通常没有太大的困难，但是估计类条件概率密度则存在两个问题：

- 已有的训练样本数不足
- 特征向量x维度较大，产生计算复杂度问题。

ML

把待估计的参数看作是确定性的量，但是取值未知。最佳估计是使得产生已观测到的样本的概率最大的值。

假设

- i.i.d
- **类条件概率密度形式已知(需要估计参数)**
- 不同类别参数相互不提供信息

样本集D中有n个样本，$x_1, x_2, \cdots, x_n$，则有

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

$$\hat{\theta} = \underset{\theta}{argmax}\, lnp(D|\theta)$$

注意：导数为0的点不等同最值点，可能是局部极值，也可能只是拐点

maximum a posteriori (MAP) : $l(\theta)p(\theta)$

先验概率$p(\theta)$为均匀分布时为ML。对参数空间做某些任意的非线性变换，概率密度$p(\theta)$发生变化，ML失效。

ML是渐进无偏的，无法获得最优解。

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t = \frac{n-1}{n}\Sigma$$

如果对于样本分布的数学模型及其参数向量θ的建模都是可靠的，那么ML就能有很好的结果。然而，不正确的样本概率分布模型带来的误差的影响会非常巨大。

最大似然估计估计均匀分布的参数举例：

$[0,\theta]$ 区间上的均匀分布为例，独立同分布地采样样本 $x_1,x_2,\ldots,x_n$，我们知均匀分布的期望为：$\frac{\theta}{2}$。

首先我们来看，如何通过最大似然估计的形式估计均匀分布的期望。均匀分布的概率密度函数为：$f(x|\theta)=\frac{1}{\theta},0\leq x\leq\theta$。不失一般性地，将 $x_1,x_2,\ldots,x_n$ 排序为顺序统计量：$x_{(1)}\leq x_{(2)}\leq\cdots\leq x_{(n)}$。则根据似然函数定义，在此样本集合上的似然函数为：

$$L(\theta|\mathbf{x})=\prod_{i=1}^{n}\frac{1}{\theta}=\theta^{-n}\quad(*)$$

对 $x_{(1)}\geq 0,x_{(n)}\leq\theta$，否则为 0。然后求其对数形式关于 θ 的导数：

$$\frac{\mathrm{d}\ln L(\theta|\mathbf{x})}{\mathrm{d}\theta}=-\frac{n}{\theta}<0.$$

导数小于 0，因此可以说 $L(x|\theta)$ 是单调减函数 $\theta\geq x_{(n)}$，因此当 $\theta=x_{(n)}$（θ 能取到的最小值），也即 $\theta=\max\{x_1,x_2,\ldots,x_n\}$ 时，$L(x|\theta)$ 值最大，则关于 θ 的最大似然估计为：

$$\hat{\theta}=x_{(n)}=\max\{x_1,x_2,\ldots,x_n\}$$

贝叶斯参数估计

假设

- i.i.d
- 我们通常认为先验概率可以事先得到，故简化为$P(w_i|D)=P(w_i)$
- 因为是有监督的学习，每个样本都有自己的类别标签，我们可以将样本分到各自不同的类别下，不同类别参数相互不提供信息 $p(x|w_i,D)=p(x|w_i,D_i)$
- 条件概率密度$p(x|\theta)$是完全已知的，虽然参数向量θ的数值未知
- 参数向量θ的先验概率密度函数$p(\theta)$包含了我们对于θ的全部先验知识
- 其余的关于参数向量θ的信息包含在观察到的独立样本$x_1,x_2,\ldots,x_n$中，他们服从未知的概率密度函数$p(x)$

在贝叶斯估计中，为了突出训练样本在估计过程的重要性，把$P(w_i|x)$写作$P(\omega_i|x,D)$

$$P(w_i|x,D)=\frac{p(x|w_i,D)P(w_i|D)}{\sum_{j=1}^{c}p(x|w_j,D)P(w_j|D)}=\frac{p(x|w_i,D_i)P(w_i)}{\sum_{j=1}^{c}p(x|w_j,D_j)P(w_j)}$$

若D一共有c类，则上式可看成c个形式相同的独立的问题。核心问题是估计$p(x|w_i,D_i)$。基于问题的独立性，我们将$p(x|w_i,D_i)$简写为$p(x|D)$。

于是贝叶斯学习的**核心问题**就被简化为：**已知一组训练样本D，这些样本都是从固定但未知的概率密度函数$p(x)$中抽取的，要求根据这些样本估计$p(x|D)$**

参数的分布

我们假设了概率密度函数$p(x)$是已知的，未知的只有参数θ 的值。

把待估计的参数看作是服从某种先验概率分布的随机变量。**对样本的观测过程中，把先验概率密度$p(\theta)$转化为后验概率密度$p(\theta|D)$，**利用样本信息修正对参数的初始估计值。在典型的情况，每得到新的观测样本，后验概率密度函数就会变得更加尖锐。我们希望后验概率密度$p(\theta|D)$在θ的真实值附近有非常显著的尖峰。

由于测试样本x和训练样本集D的选取是独立进行，得$p(x|\theta,D)=p(x|\theta)$，**由此构建类条件概率密度$p(x|D)$和后验概率密度$p(\theta|D)$的桥梁**：

$$p(x|D)=\int p(x,\theta|D)d\theta=\int p(x|\theta,D)p(\theta|D)d\theta=\int p(x|\theta)p(\theta|D)d\theta$$

可以看出，当后验概率在某个值$\hat{\theta}$上取得了尖峰（接近真实值），即$p(\hat{\theta}|D) \to \infty$，则$p(x|D) \approx p(x|\hat{\theta})$

贝叶斯参数估计：高斯情况（具体例子：p74-78）

单变量情况

假如知道x的分布是一个高斯分布，只知道方差不知道均值，并假设均值是满足一个特定的正态分布，应用贝叶斯公式可以知道：

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\,d\mu} = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

经过化简发现，$p(\mu|D)$是一个指数函数，并且指数部分是一个二次型，因此它仍然是一个正态分布，不管样本数有多少仍然保持：

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

我们把$p(\mu|D)$成为**复制密度函数**，$P(\mu)$称为**共轭先验**。也可以写成：

prior)。如果写成下面的形式：$p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$，也就是

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] \tag{30}$$

那么对公式(29)和公式(30)应用对应项相等的原则，就可以求得 μ_n 和 σ_n^2

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{31}$$

和

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \tag{32}$$

其中，$\hat{\mu}_n$ 是样本均值

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{33}$$

进一步求解 μ_n 和 σ_n^2，我们得到

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \tag{34}$$

和

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} \tag{35}$$

根据公式可以看到，随着观测的样本越来越多，对未知参数的估计也会越来越准。用图表示，当n增加时，$p(\mu|D)$的波形会变得越来越尖，在n趋于无穷时为一个狄拉克函数。**这一现象就是贝叶斯学习的过程。**

基本过程

1. 利用$p(D|\theta)=\prod_{k=1}^{n}p(x_k|\theta)$计算后验概率密度函数$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$。参数的先验概率一般知道，其实就是利用了最大似然估计的方法来估计参数。
2. 估计出了参数的分布后，就可以利用$p(x|\omega_i,D_i)=p(x|D)=\int p(x|\theta)p(\theta|D)d\theta$获得类条件概率密度函数，相当于对不同可能的 θ 加权求和
3. 有了类条件概率，就可以计算模式x属于某个类别的后验概率$p(\omega_i|x)=\alpha p(x|\omega_i)p(\omega_i)$。

上式也阐明了贝叶斯估计和最大似然估计的关系。假设$p(D|\theta)$在$\theta=\hat{\theta}$处有一个非常尖的峰值，如果先验概率$p(\theta)$在$\theta=\hat{\theta}$非零，并且在周围区域变化不大，那么 $p(\theta|D)$也在同一个地方有峰值

递归的贝叶斯方法

因为样本是独立同分布的，根据极大似然估计的公式，总的概率等于各自概率连乘：

$$p(D^n|\theta)=p(x_n|\theta)p(D^{n-1}|\theta)$$

基本过程1可用增量学习法：

$$p(\theta|D^n)=\frac{p(x_n|\theta)p(D^{n-1}|\theta)p(\theta)}{\int p(x_n|\theta)p(D^{n-1|}|\theta)p(\theta)d\theta}=\frac{p(x_n|\theta)p(\theta|D^{n-1})}{\int p(x_n|\theta)p(\theta|D^{n-1})d\theta}$$

本质就是增加样本，不断更新对参数的估计。当尚未有样本的时候$p(\theta|D^0)=p(\theta)$

举例说明：

$$\text{假设：}\ p(x\mid\theta)\sim U(0,\theta)=\begin{cases}1/\theta & 0\le x\le\theta\\ 0 & \text{其他}\end{cases}$$

$$p(\theta)\sim U(0,10),\quad D=\{4,7,2,8\}$$

$$p(D^0\mid\theta)=p(\theta)\sim U(0,10)$$

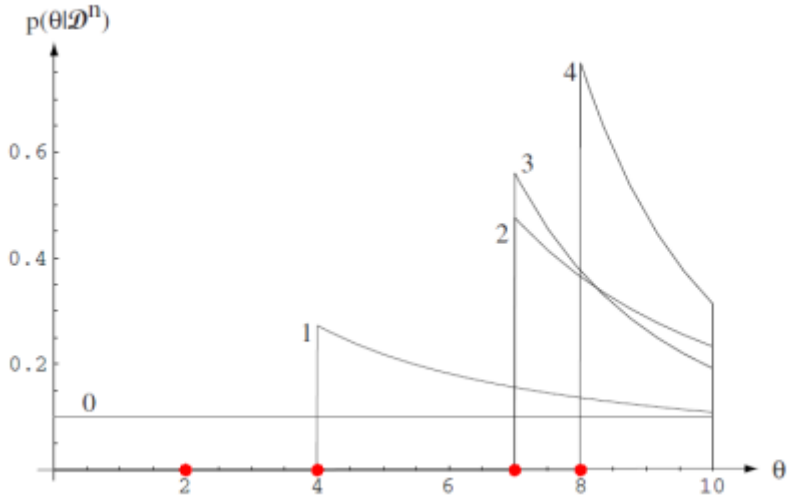
$$p(\theta\mid D^1)\propto p(x_1\mid\theta)p(\theta\mid D^0)=\begin{cases}1/\theta & \text{对于 } 4\le\theta\le10\\ 0 & \text{其他}\end{cases}$$

$$p(\theta\mid D^2)\propto p(x_2\mid\theta)p(\theta\mid D^1)=\begin{cases}1/\theta^2 & \text{对于 } 7\le\theta\le10\\ 0 & \text{其他}\end{cases}$$

$$p(\theta\mid D^n)\propto 1/\theta^n\quad \text{对于 } \max_x\left[D^n\right]\le\theta\le10$$

这里根据采样得到的最大值做出假设参数的上限为10，可以得到θ的范围是因为它一定是大于X的，小于x的时候不可能会采样得到x。

如下图所示：相当于随着采样得到的点增加，比如采样了8，则8以下的参数可能值就一下变成了0，而8的概率最大，继续递增时又继续递减。直观上理解就是没有采样到大于8的值，说明最有可能x的上限也就是参数就是8。

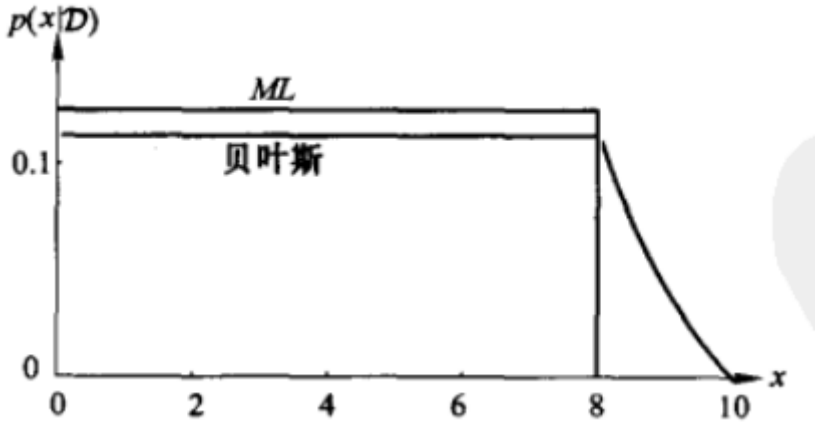


$$\begin{cases}\int_8^{10}\frac{1}{\theta}\frac{1}{\theta^4}d\theta & ,0\le x<8\\ \int_x^{10}\frac{1}{\theta}\frac{1}{\theta^4}d\theta & ,8\le x\le10\\ 0 & ,otherwise\end{cases}$$

CMP: ML & BPE

上面的例子中，贝叶斯估计得出的记过是，计算$p(x|D)=\int p(x|\theta)p(\theta|D)d\theta$

而ML会直接选择8作为参数的值，得到$p(\theta)$为0.125。如图：



从算法复杂度上看，ML避免了多重积分。对于先验知识的信任程度，贝叶斯估计得到的结果的形式与原始假设的形式不同。

	最大似然估计	贝叶斯参数估计
计算复杂度	微分	多重积分
可理解性	确定易理解	不确定不易理解
先验信息的信任程度	不准确	准确
例如 $p(x\|\theta)$	与初始假设一致	与初始假设不一致

对两者如何做出选择?

1. 对于先验概率能保证问题有解的条件下，最大似然估计和贝叶斯估计在训练样本趋近于无穷的时候效果一样
2. 最大似然估计只需要计算一些简单微分时间复杂度比贝叶斯估计要低
3. 最大似然法的可理解性比贝叶斯参数估计要强。因为它得到的结果是基于训练样本的最优解，而**贝叶斯估计的方法得到的结果则是许多可行解答的加权平均值**
4. 对概率密度函数$p(x|\theta)$，最大似然估计得到的结果$p(x|\hat{\theta})$和初始假设的形式一致，而贝叶斯估计是$p(x|D)=\int p(x|\theta)p(\theta|D)d\theta$相当于对所有可能的参数，做了一个加权求和，这样得到的概率函数和原来假设的不同。贝叶斯方法比最大似然估计方法利用到更多信息

Gibbs algorithm

贝叶斯分类器能否达到最优的分类效果。但$p(x|D)=\int p(x|\theta)p(\theta|D)d\theta$的结果可能会非常复杂，因此一个变通的结果是依据$p(\theta|D)$值选择一个参数向量$\theta$作为真实值。在较弱的假设条件下，吉布斯算法的误差概率至多是贝叶斯最优分类器的两倍。

select $\theta=\theta_0$ to make $p(x|D)\approx p(x|\theta_0)$

$P_{error}(Glibbs)\approx 2P_{error}(BPE)$

充分统计量

在实际应用中，很多模型存在很多的参数和训练样本，这么大的数据量是无法用上面的方法求解的。考虑在多元高斯密度中（如上面提到的单变量高斯情况），我们发现我们要求的目标$p(x|D)$和模式x的的具体值无关，只和样本的均值和协方差矩阵有关，因此我们可以只用均值和协方差就求出结果。但这种方法只适用于正态分布中。

不过确实存在一些分布，也能获得类似的可行解法，这个问题的关键就是**充分统计量**

充分统计量就是一个关于样本集D的函数s，其中包含了能用来估计某种参数的所有相关的信息。

常规定义如下：一个统计量s对参数θ是充分的，如果$P(D|s,\theta)$与参数θ无关，也就是s完全可以代替θ。可用于降低数据量：

$$p(\theta|s,D)=p(\theta|s)$$

因式分解定理

一个统计量s是关于参数θ的充分统计量，当且仅当$P(D|\theta)$能被分解为两个函数的积，其中一个只依赖于s,θ，另一个只依赖于数据D

$$P(D|\theta)=g(s,\theta)h(D)$$

核密度消除二义性

$$\bar{g}(s,\theta)=\frac{g(s,\theta)}{\int g(s,\theta)d\theta}$$

维数问题

从经验上看，最好满足 sample : dimension > 10 : 1

两类均值间的距离大于标准差的特征较为有用。

如果问题的概率结构完全已知，增加新的特征不会增加贝叶斯风险，并提高分类器的精确度。然而，在实际应用中，特征个数增加到某一个临界点后，继续增加反而会导致分类器的性能变差。问题的核心通常是假设的概率模型与实际情况不匹配。

解决训练样本不足的问题

- 减低问题维度：重新设计提取特征模块、只选取特征子集、将特征组合
- 假设各类协方差矩阵相同：将数据归到一起
- 寻找协方差矩阵更好的估计： $\lambda\Sigma_0+(1-\lambda)\hat{\Sigma}$

  其中Σ_0是一个合理的先验估计。

  我们可以合理的假设任何不接近1的协方差实际为0。更极端的假设是假设各个特征之间统计独立，使非对角元素为0。这种假设显然不会，但是有时候结果比MLE得到的结果要好。

PCA

目的是寻找在最小均方意义下最能够代表原始数据的投影方法。

1. 假设有n个d维的样本$x_1,x_2,\ldots,x_n$，以均方误差作为标准，如何用一个d维的向量x_0表示这n个样本?

记样本均值

$$m=\frac{1}{n}\sum_{k=1}^{n}x_k$$

则有：

$$
\begin{aligned}
J_0(x_0) &= \sum_{k=1}^{n} ||x_0 - x_k||^2 \\
&= \sum_{k=1}^{n} ||(x_0 - m) - (x_k - m)||^2 \\
&= \sum_{k=1}^{n} ||x_0 - m||^2 + \sum_{k=1}^{n} ||x_k - m||^2
\end{aligned}
$$

故在$x_0 = m$处取得最小值。

2. 样本均值是样本数据集的零维表达，不能反映样本间的不同，通过**把全部样本通过样本均值的一条直线作投影，能得到代表全部样本的一维向量。**让e表示这条通过样本均值直线上的单位向量，那么这条直线的方程可以表示为

$$
x = m + \alpha e
$$

其中α为标量，表示直线上某个点离开m的距离。e为单位向量，代表了这条直线的方向。**相当于是以样本均值 m为原点出发的一条直线**，当$\alpha = 0$的时候，取值就是样本均值。

$$
\begin{aligned}
J_1(a_1, \cdots, a_n, e) &= \sum_{k=1}^{n} ||(m + a_k e) - x_k||^2 = \sum_{k=1}^{n} ||a_k e - (x_k - m)||^2 \\
&= \sum_{k=1}^{n} a_k^2 ||e||^2 - 2\sum_{k=1}^{n} a_k e^t (x_k - m) + \sum_{k=1}^{n} ||x_k - m||^2
\end{aligned}
$$

对a_k求偏导并令结果为0，得

$$
a_k = e^t (x_k - m)
$$

将上式带入$\sum_{k=1}^{n} ||(m + a_k e) - x_k||^2$可以看出几何意义为，将向量$X_k$向通过样本均值直线e的垂直投影。

这就引入新的问题：如何找到直线e的最优方向。

记**离散度矩阵**(scatter matrix)

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

可以类比于求样本集X每两个属性之间关系的协方差矩阵，只不过是乘了$n-1$倍。$X_k - m$是一个$m \times 1$的向量，顾上述结果为一个$m \times m$的矩阵。

假设m为样本数，n为属性数（注意和上述相反）。此时散度矩阵如下图：

$$
\begin{bmatrix}
\sum_{j=1}^{m} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1) & \sum_{j=1}^{m} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) & \cdots & \sum_{j=1}^{m} (x_{1j} - \bar{x}_1)(x_{nj} - \bar{x}_n) \\
\sum_{j=1}^{m} (x_{2j} - \bar{x}_2)(x_{1j} - \bar{x}_1) & \sum_{j=1}^{m} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2) & \cdots & \sum_{j=1}^{m} (x_{2j} - \bar{x}_2)(x_{nj} - \bar{x}_n) \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{m} (x_{nj} - \bar{x}_n)(x_{1j} - \bar{x}_1) & \sum_{j=1}^{m} (x_{nj} - \bar{x}_n)(x_{2j} - \bar{x}_2) & \cdots & \sum_{j=1}^{m} (x_{nj} - \bar{x}_n)(x_{nj} - \bar{x}_n)
\end{bmatrix}
$$

带入$a_k = e^t (x_k - m)$和散度矩阵替换得：

$$
\begin{aligned}
J_1(e) &= \sum_{k=1}^{n} ||(m + a_k e) - x_k||^2 \\
&= -\sum_{k=1}^{n} e^t (x_k - m)(x_k - m)^t e + \sum_{k=1}^{n} ||x_k - m||^2 \\
&= -e^t S e + \sum_{k=1}^{n} ||x_k - m||^2
\end{aligned}
$$

显然，要让J_1最小的那个向量e。能够让$e^t Se$最大，则利用拉格朗日乘子法，约束为$||e|| = 1$可以得到拉格朗日函数：

$$
u = e^t Se - \lambda(e^t e - 1)
$$

对e求偏导，并令梯度为零。我们看到，e必须为散布矩阵的特征向量：

$$
\frac{\partial u}{\partial e} = 2Se - 2\lambda e = 0
$$

$$
u = e^t Se - \lambda(e^t e - 1)
$$

$$
\begin{aligned}
Se &= \lambda e \\
e^t Se &= \lambda
\end{aligned}
$$

为了得到最小的$J_1(e)$，即最大化$e^t Se$，**需要选取散布矩阵最大的特征值λ**

这样得到e后就可以把所有样本点映射到直线e上，这样他们对应的α_k就可以作为这样样本的属性了，这种是一维的情况。如果是多维的就需要多个单位向量e。

3. 拓展到d'维（将样本表示成多个向量，即映射到多个过m的直线上），由于离散度矩阵是实对称的，特征向量相互正交，可以选取d'个最大特征值λ，构成新维度的基向量。此时每个样本点可以表示为：

$$
x = m + \sum_{i=1}^{d'} \alpha_i e_i
$$

新的平方误差准则函数：

$$
J_{d'} = \sum_{k=1}^{n} ||(m + \sum_{i=1}^{d'} a_{ki} e_i) - x_k||^2
$$

在向量$e_1, e_2, \ldots, e_{d'}$分别为散布矩阵的d'个最大特征向量的时候最小。因为散布矩阵是实对称矩阵，因此特征向量是互相正交的。他们构成了代表任一样本x的基，而α_i就是对应基上的系数。

PCL的变换是协方差矩阵，K-L的变换矩阵有二阶矩阵、协方差矩阵、总类内离散度矩阵等。故PCA可看做K-L变换的特殊形式。

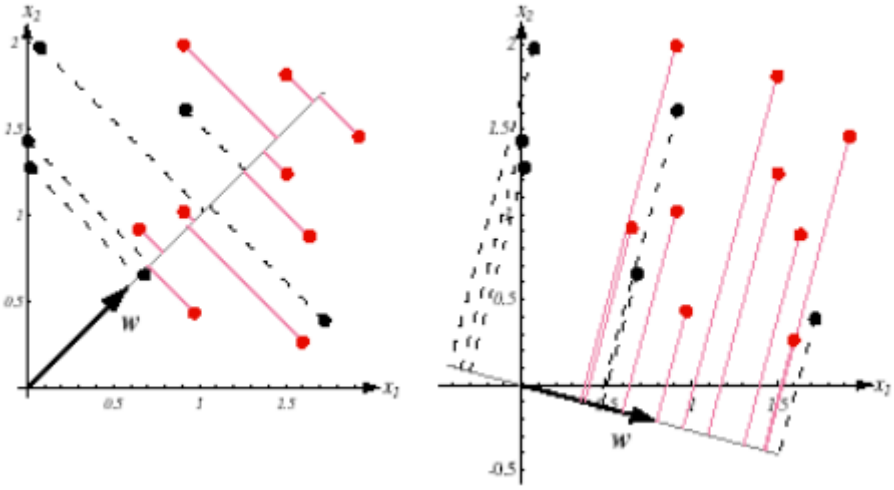
Fisher线性判别分析

PCA在数据样本表示非常有效，但是**没有理由表明PCA在分类**中有作用。

PCA寻找用来有效表示的主轴方向，而**判别分析方法(discriminant analysis)寻找有效分类**的方向。

我们考虑把d维空间中的数据点投影到一条直线上去。有可能找到能够最大限度区分各类数据点的投影方向。

e.g.



假设有n个d维的样本$x_1, \cdots, x_n$，他们分别属于两个类。即大小为n_1的样本子集D_1属于类别w_1，大小为n_2的样本样本子集D_2属于类别w_2。对x中的各个成分作线性组合，就得到点积$y = w^t x$，结果是一个标量。

设d维的样本均值为

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

投影后的点的样本均值为

$$\tilde{m}_i = \frac{1}{n_i} \sum_{x \in D_i} w^t x = w^t m_i$$

对待分开的两组样本，投影点的均值差为

$$\tilde{m}_1 - \tilde{m}_2 = |w^t (m_1 - m_2)|$$

我们可以**通过增加w的幅值来得到任意大小的投影样本均值之差**（因为$w^t x$相当于x在w方向上的投影长度乘w的长度）。但投影样本均值之差的大小总是相对而言的。

投影之后的点为$y = w^t x$，定义类内散布（方差）：

$$\tilde{s_i}^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$$

$\frac{1}{n}(\tilde{s_1}^2 + \tilde{s_2}^2)$为**全部数据的总体方差的估计。**$\tilde{s_1}^2 + \tilde{s_2}^2$称为类内总散布。**Fisher的准则函数**为：

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s_1}^2 + \tilde{s_2}^2}$$

为了让准则函数（标准化后的两组样本的均值距离）最大，我们需要求解让准则函数取最优（大）情况下，投影样本的矩阵 w ，然后再利用一个阈值进行分类。我们先将准则函数写成 w 的表达式。

定义类内散布矩阵S_i、总类内散布矩阵S_W和总类间散布矩阵S_B

$$\begin{aligned} S_i &= \sum_{x \in D_i} (x - m_i)(x - m_i)^t \\ S_W &= S_1 + S_2 \\ S_B &= (m_1 - m_2)(m_1 - m_2)^t \end{aligned}$$

然后我们有：

$$\begin{aligned}\tilde{s}_i^2 &= \sum_{x\in D_i}(w^t x - w^t m_i)^2 \\ &= \sum_{x\in D_i} w^t(x-m_i)(x-m_i)^t w \\ &= w^t S_i w\end{aligned}$$

各离散度之和可以写成：

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^t S_w w$$

投影样本均值差可以展开为：

$$\begin{aligned}(\tilde{m}_1 - \tilde{m}_2)^2 &= w^t S_B w \\ \tilde{s}_1^2 + \tilde{s}_2^2 &= w^t S_W w \\ J(w) &= \frac{w^t S_B w}{w^t S_W w}\end{aligned}$$

得到的表达式通常被称为**广义的瑞利商**。可解得最大值满足的w必须满足以下式子：

$$S_B w = \lambda S_W w$$

于是我们就可以得到一个类似求特征向量的问题：

$$S_W^{-1} S_B w = \lambda w$$

不过我们并不需要求出它的特征向量值。由于S_B是两个向量的外积，秩最多为1（每一行都是第一行的倍数），$S_B w$的方向是在$m_1 - m_2$上（用矩阵的乘法写出来就可以发现），而由于**w的模长对问题本身无关紧要**，由此可以得出线性判别器中的**典范变量w的值**(canonical variate)：

$$w = S_W^{-1}(m_1 - m_2)$$

当样本降到一维之后，剩下的问题就只是找一个点将二类分开。

当条件概率密度函数$p(x|w_i)$是多元正态函数，并且各个类别的协方差矩阵相同时，我们能够直接计算这个阈值。一般来说，如果我们对投影后的数据用一维高斯函数进行拟合，w_0就可以选择为使两个类的后验概率相同的位置。（即为两个正态分布函数的交点处）

多重判别分析

对于c类问题，需要$c-1$个判别函数。把Fisher线性判别准则作推广就需要c-1个判别函数。投影问题实际上是从d维空间向$c-1$维空间作投影。

类内散布矩阵的推广

$$S_W = \sum_{i=1}^{c} S_i$$

总体散布矩阵

$$\begin{aligned}S_T &= \sum_x (x-m)(x-m)^t \\ &= \sum_{i=1}^{c}\sum_{x\in D_i}(x - m_i + m_i - m)(x - m_i + m_i - m)^t \\ &= \sum_{i=1}^{c}\sum_{x\in D_i}(x-m_i)(x-m_i)^t + \sum_{i=1}^{c}\sum_{x\in D_i}(m_i - m)(m_i - m)^t \\ &= S_W + \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t\end{aligned}$$

得类间散布矩阵的推广

$$S_B = S_T - S_W = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t$$

利用行列式度量准则函数，可得

$$J(W) = \frac{|\tilde{S_B}|}{|\tilde{S_W}|} = \frac{|W^t S_B W|}{|W^t S_W W|}$$

由广义瑞利商知，求解

$$|S_B - \lambda_i S_W| = 0$$

的非零特征值，并进行Gram-Schmidt正交化可以得到降维后的基向量。

EM

核心思想：根据已有的数据递归似然函数

$$Q(\theta; \theta^i) = E_{D_b}[lnp(D_g, D_b; \theta)|D_g; \theta^i]$$

$Q(\theta; \theta^i)$表示一个关于θ的函数，其中θ^i被假设已经取固定值，是对当前分布最好的估计，θ是在当前估计的基础上，进一步改善估计的一个候选参数向量。该宣发表征了好数据的对数似然函数是单调递增的。

begin initialize $\boldsymbol{\theta}^0$, T, $i \leftarrow 0$
  do $i \leftarrow i + 1$
    E step: Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^i)$
    M step: $\boldsymbol{\theta}^{i+1} \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^i)$
  until $Q(\boldsymbol{\theta}^{i+1}; \boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^i; \boldsymbol{\theta}^{i-1}) \leq T$
  return $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{i+1}$
end

e.g.P103-105