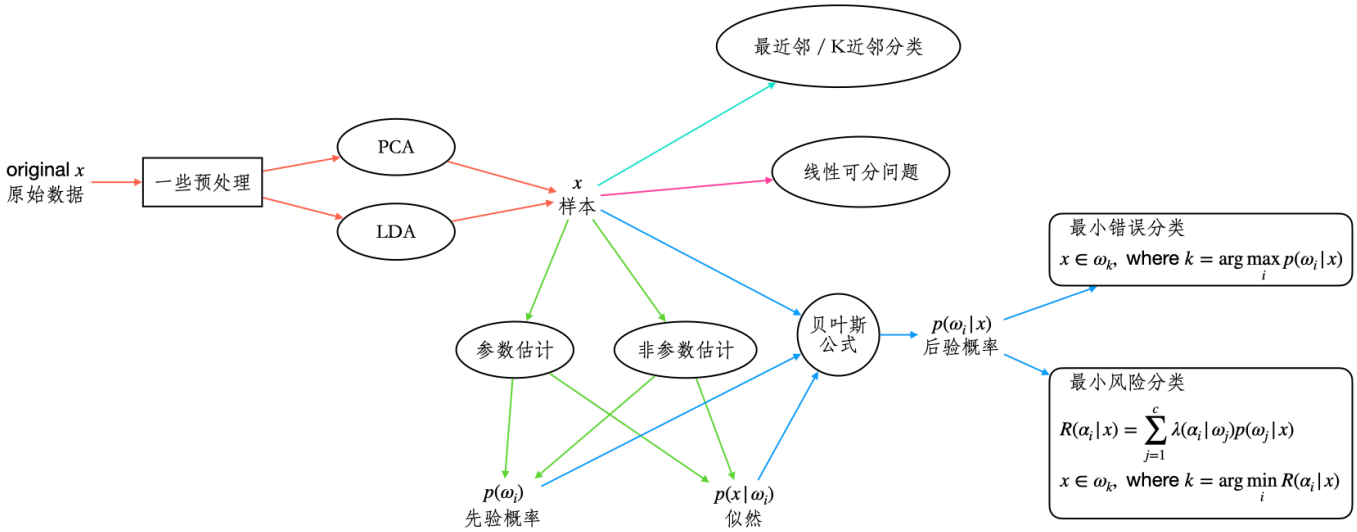
模式识别知识点总结

xyfJASON



1 最近邻 & K近邻

1.1 最近邻法

将测试样本分类为与其距离最近的训练样本的类别。

性质：渐进错误率（训练样本足够大）小于最小错误率（贝叶斯）的两倍。

不足：计算量大，存储量大，对样本敏感（多一个/少一个对分类结果有较大的影响）

1.2 K 近邻法

选择与待分类样本距离最小的 K 个样本，以这 K 个样本中大多数所属类别作为 X 的类别。

优点：

1. 思路简单，实现简单
2. 当有新样本要加入训练集中时，无需重新训练（即重新训练的代价低）
3. 计算时间和空间线性于训练集的规模（在一些场合，不算太大）

缺点：

1. 分类速度慢：时间复杂度和空间复杂度随训练集规模和特征维数的增大而快速增加。设 m 是特征维数，n 是训练集样本个数，则复杂度为 $O(nm)$
2. 各属性（特征维度）的权重相同，影响了准确率
3. 当样本不平衡时，如某类很多、其他类很少，容易导致把测试样本分到该类下，即便它并不接近这类样本
4. K 值不好确定——K 过小，近邻少，放大噪声干扰，准确率低；K 过大，对于少类来说，多类的数据也被包含，从而导致分类效果不好

---

2 贝叶斯决策，贝叶斯最小错误分类

记样本空间为 X，其中任一为 x；设类别为 $\omega_1, \ldots, \omega_c$，则所谓分类模型其实就是求给定输入样本 x 的条件下它属于类 ω_i 的概率：

$$p(\omega_i|x)$$

由贝叶斯公式知：

$$p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{p(x)} = \frac{p(x|\omega_i)p(\omega_i)}{\sum_{i=1}^{c} p(x|\omega_j)p(\omega_j)}$$

其中 $p(\omega_i)$ 为先验概率，$p(x|\omega_i)$ 称为似然，表示类别 ω_i 下样本的概率分布。也就是说，后验概率由先验概率和似然共同决定。

显然，采用贝叶斯决策进行分类，就是选择使得 $p(\omega_i|x)$ 最大的那个 ω_i 作为 x 的类别。又由于分母 $p(x)$ 与类别无关，对所有类都相同，所以我们只需要比较 $p(x|\omega_i)p(\omega_i)$ 的大小，即：

$$x \in \omega_k, \quad \text{where } k = \arg\max_i p(\omega_i|x) = \arg\max_i p(x|\omega_i)p(\omega_i)$$

贝叶斯分类的错误率为：

$$p(error|x) = \min_i\{1 - p(\omega_i|x)\} = 1 - \max_i\{p(\omega_i|x)\}$$

特别的，对于两类问题，上式可写作：

$$p(error|x) = \min(p(\omega_1|x), p(\omega_2|x))$$

贝叶斯分类器的错误率是理论上最小的错误率。

---

3 贝叶斯最小风险分类

考虑到每一个决策带来的代价是不同的，我们希望最小化风险。

设决策 α_i 表示把样本分类为 ω_i 这个动作，$\lambda(\alpha_i, \omega_j)$（或写作 $\lambda(\alpha_i|\omega_j), \lambda_{ij}$）表示把第 j 类的某样本分类为第 i 类的风险。

对于某输入 x，可以根据贝叶斯公式计算出后验概率 $p(\omega_i|x),\ i = 1, 2, \ldots, c$。在最小错误分类中，我们直接看哪个后验概率最大，但现在我们要看哪个的风险最小。故而定义**条件风险**：

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) p(\omega_j|x)$$

即给定输入 x，将其分为第 i 类的风险值。于是我们的决策就是：

$$x \in \omega_k, \quad \text{where } k = \arg\min_i R(\alpha_i|x)$$

最小风险分类特别适用于类似「宁愿把正常细胞错判为癌细胞（然后进一步人工诊断），也不要放过一个癌细胞」的情形等。

---

4 参数估计 & 非参数估计

基于贝叶斯决策理论的分类器需要两个已知信息：$p(x|\omega_i)$ 和 $p(\omega_i)$，可是它们是从哪里来的？我们手上只有一个训练集，因此我们只能从训练集样本中**估计**它们。

估计方法分为：

- 参数估计：概率密度函数的形式已知（或直接假设，例如正态分布），推断其参数
  - 最大似然估计
  - 贝叶斯估计
- 非参数估计：直接推断概率密度
  - Parzen 窗法
  - k_N 近邻法

4.1 参数估计

4.1.1 最大似然估计

记概率分布的参数为向量 θ，样本集为 $\{x_1,\dots,x_N\}$ 且**独立同分布**，定义似然函数：

$$L(\theta)=p(x_1,\dots,x_N|\theta)=p(x_1|\theta)\cdots p(x_N|\theta)=\prod_{i=1}^{N}p(x_i|\theta)$$

最大似然估计的思想是选取参数 θ 使得似然函数 $L(\theta)$ 最大，为方便计算，求对数似然：

$$\ln L(\theta)=\sum_{i=1}^{N}\ln p(x_i|\theta)$$

由于对数函数单调，故对数似然最大时似然函数也最大，因此只需求对数似然的最大值：

$$\text{let:}\quad \nabla_\theta \ln L(\theta)=\sum_{i=1}^{N}\nabla_\theta \ln p(x_i|\theta)=0$$

特别的，考虑概率分布是未知参数的**正态分布**情形。

首先考虑特征维数为 1，即样本服从 $N(\mu,\sigma^2)$：

$$p(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$\theta=\begin{bmatrix}\mu\\ \sigma^2\end{bmatrix}$$

经过计算（计算过程略），可以求出估计值：

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

即均值估计值为样本均值，方差估计值为有偏样本方差。

如果特征维数大于 1，则样本服从 $N(\mu, \Sigma)$：

$$p(x|\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

同理可以求出估计值：

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

即均值估计值为样本均值，协方差矩阵估计值为样本协方差矩阵。

4.2 非参数估计

基础思想：选取包含 x 的一个小区域 R，设其体积为 V，落入其中的样本数为 k，样本总数为 N，则估计点 x 处的概率密度值为：

$$\hat{p}(x) = \frac{k/N}{V}$$

观察上式，有一些问题值得讨论：

- 当体积 V 固定时，N 越多，$\hat{p}(x)$ 能收敛到 $p(x)$ 在区域 R 内的平均：

$$p(x) \approx \frac{k/N}{V} \approx \frac{\int_R p(x)\mathrm{d}x}{\int_R \mathrm{d}x}$$

- 如果我们希望得到 $p(x)$ 而不是平均平滑后的版本，就需要令 $V \to 0$。然而，当样本数 N 固定时，缩小体积 V，可能导致区域内不含任何样本，从而使 $\hat{p}(x) = 0$；或者如果 x 处正好有一两个样本，那么导致 $\hat{p}(x) \to \infty$；这都是没有意义的
- 实际情况下，样本数是有限的，选取的体积不可能是无限小；所以我么那不得不接受以下事实：k/N 总是有一定变动的，并且 $\hat{p}(x)$ 总是存在一定程度的平滑效果的。

在理论上，如果我们能够获取无限多的样本，那么上述局限性可以被克服。假设我们选择一系列包含 x 的区域 $R_1, R_2, \ldots, R_N$，其中 R_1 使用一个样本，R_2 使用两个样本，以此类推。记 V_N 为 R_N 的体积，k_N 为落入 R_N 的样本数，$\hat{p}_N(x)$ 为 $p(x)$ 的第 N 次估计，则：

$$\hat{p}_N(x) = \frac{k_N/N}{V_N}$$

要使之收敛于 $p(x)$，下述条件必须得到满足：

1. $$\lim_{N\to\infty} V_N = 0$$

   若区域平滑地缩小，$p(x)$ 在 x 处连续，则这一条件使得区域 R 内的平均概率 $\hat{p}(x)$ 收敛于 $p(x)$ 。

2. $$\lim_{N\to\infty} k_N = \infty$$

   对于 $p(x) \neq 0$ 的点，这一条件使得频率 k_N/N 收敛于概率（否则频率就趋向 0 了）。

3. $$\lim_{N\to\infty} \frac{k_N}{N} = 0$$

   根据第 1 个条件分母收敛于零，显然分子也必须收敛于零整个值才不会发散。

   直观而言，这一条件是说，尽管在一个小区域 R_N 内可以落入大量样本，但是与总样本数相比仍然非常小。

有两种经常采用的方法来获取区域序列：

- Parzen 窗法：根据某一个确定的体积函数，例如 $V_N = \frac{1}{\sqrt{N}}$，逐渐收缩一个给定的初始区域。这要求随机变量 k_N 和 k_N/N 能满足前文所述的条件 2 和条件 3。
- k_N 近邻法：确定 k_N 为 N 的某个函数，例如 $k_N = \sqrt{N}$。这样，我们增长体积 V_N 使得区域 R_N 内正好有 k_N 个样本。

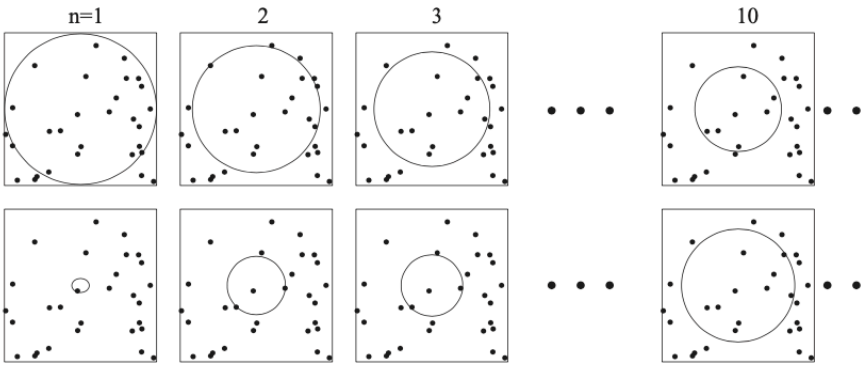


Figure 4.2: Two methods for estimating the density at a point **x** (at the center of each square) are to xxx.

这两种方法都能收敛，但是很难预测它们在有限样本情况下的效果。

4.2.1 Parzen 窗法

暂时假设区域 R_N 是一个棱长为 h_N 的 d 维超立方体，则：

$$V_N = h_N^d$$

定义窗函数（方窗函数）：

$$\varphi(u) = \begin{cases} 1, & |u_j| \leqslant 1/2,\ j = 1, \dots, d \\ 0, & \text{otherwise} \end{cases}$$

即一个中心在原点的单位超立方体。将这个超立方体（窗函数）移动到以 x 为中心、体积放缩为 V_N，则有：

$$\varphi\left(\frac{x - x_i}{h_N}\right) = \begin{cases} 1 & \text{if } x_i \text{ falls into hypercube } R_N \\ 0 & \text{otherwise} \end{cases}$$

因此，落入该超立方体的样本总数为：

$$k_N = \sum_{i=1}^{N} \varphi\left(\frac{x - x_i}{h_N}\right)$$

故：

$$\hat{p}(x) = \frac{k_N / N}{V_N} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{V_N} \varphi\left(\frac{x - x_i}{h_N}\right)$$

这就是 Parzen 窗法估计的基本公式。这个公式不必规定区域必须是超立方体，可以是更加一般化的形式。

我们知道，一个合法的概率密度函数必须满足：非负性和归一性，那我们估计的 $\hat{p}(x)$ 满足这两个性质吗？非负性是显然的，只需要验证归一性：

$$\begin{aligned} \int \hat{p}(x) \mathrm{d}x &= \int \frac{1}{N} \sum_{i=1}^{N} \frac{1}{V_N} \varphi\left(\frac{x - x_i}{h_N}\right) \mathrm{d}x \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{V_N} \int \varphi\left(\frac{x - x_i}{h_N}\right) \mathrm{d}x \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{h_N^d}{V_N} \int \varphi(u) \mathrm{d}u && \text{换元} \\ &= \frac{1}{N} \sum_{i=1}^{N} 1 && V_N = h_N^d \\ &= 1 \end{aligned}$$

因此 $\hat{p}(x)$ 是一个合法的概率密度函数。

注意上述推理过程只要求窗函数满足：

- $\varphi(u) \geqslant 0$
- $\int \varphi(u) \mathrm{d}u = 1$
- $V_N = h_N^d$

因此我们还可以定义很多其他类型的窗函数。常用的窗函数有：

1. 方窗函数

$$\varphi(u) = \begin{cases} 1, & |u_j| \leqslant 1/2,\ j = 1, \ldots, d \\ 0, & \text{otherwise} \end{cases}$$

2. 正态窗函数

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^T u\right)$$

3. 指数窗函数

$$\varphi(u) = \exp\left(-\frac{1}{2}|u|\right)$$

在样本数有限的实际情形下，窗宽对结果有非常大的影响。

- 如果窗宽 h_N 很大，只有离 x 很远的样本对概率密度才无贡献。因此，密度估计值是 N 个宽度较大、变化缓慢的函数的叠加，这是一个概率密度平均的估计，其分辨率很低——平均效应较大。
- 如果窗宽 h_N 很小，密度估计值是 N 个以样本为中心的尖峰函数的叠加，变化剧烈——不平滑效应明显。

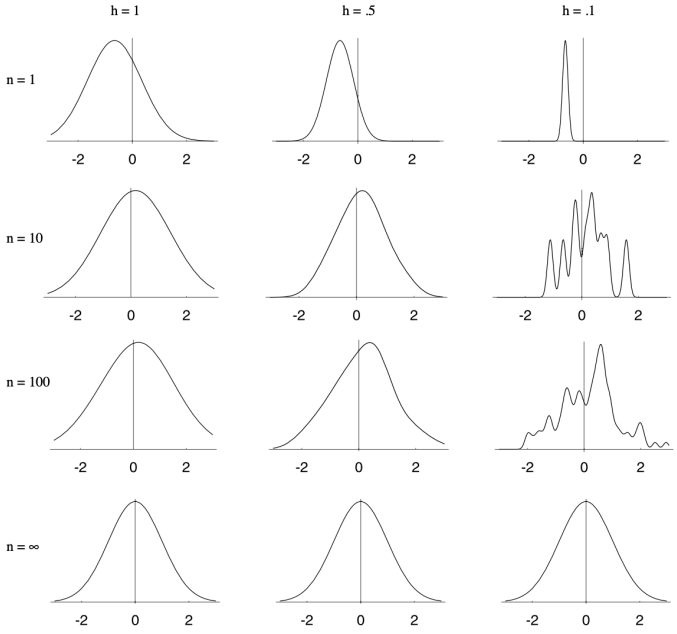


Figure 4.5: Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true generating function), regardless of window width h.

因此对于实际问题，我们只能折中考虑。一般而言，当样本数目较多时，窗宽可以取小一些。

Parzen 窗法的优缺点：

- 优点：普遍适应。对规则分布、非规则分布、单峰或多峰分布都可用此法进行密度估计
- 缺点：要求样本足够多，才能有较好的估计，计算量大，存储量大

主要困难：窗宽的选择。

4.2.2 k_N 近邻法

在 Parzen 窗法中，最佳的窗函数的选择总是一个问题，所以我们不妨转换一个角度，在估计 x 时，以 x 为中心让体积向外扩张，直到包含进 k_N 个样本为止。一般而言，我们取 k_N 为某个固定的 N 的函数，例如 $k_N = \sqrt{n}$。

k_N 近邻法的优缺点：

- 优点：
    - 如果点 x 附近的密度比较高，则包含 k_N 个样本的体积就比较小，可以提高分辨率；
    - 反之，如果点 x 附近的密度比较低，则体积就比较大，分辨率较低
- 缺点：需要样本多，计算量大，存储量大

---

5 主成分分析

主成分分析是一种无监督数据降维方法，人们希望用较小的维度替代原来较大的数据维度，并且尽可能保留原有信息。其基本思想是将高维数据通过线性组合投影到某低维空间中。

设我们的样本集为：$\{x_1, x_2, \dots, x_n\}$, $x_i \in \mathbb{R}^m$，**中心化**后得到：$\{d_1, d_2, \dots, d_n\}$, $d_i = x_i - \bar{x}$。

构建协方差矩阵（无偏还是有偏无所谓，因为特征值/奇异值大小关系不变）：

$$C = \frac{1}{n} \sum_{i=1}^{n} d_i d_i^T = \frac{1}{n} A A^T \in \mathbb{R}^{m \times m}$$

其中，$A = (d_1, d_2, \dots, d_n) \in \mathbb{R}^{m \times n}$。

对协方差矩阵进行特征值分解，求出特征值和特征向量。选取最大的 p 个特征值对应的特征向量 $v_1, v_2, \dots, v_p$，则我们要找的低维空间就是由它们作为正交基组成的特征空间：

$$W = (v_1, v_2, \dots, v_p) \in \mathbb{R}^{m \times p}$$

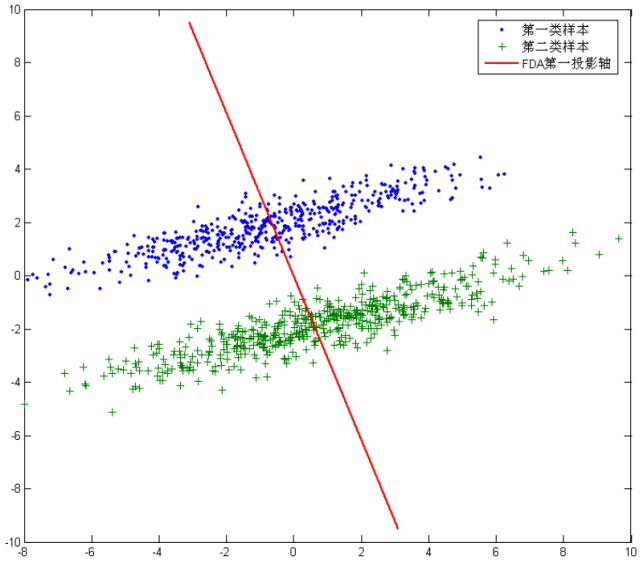
要将一个原始数据 $x \in \mathbb{R}^m$ 投影到该空间中，只需要左乘 W^T 即可：

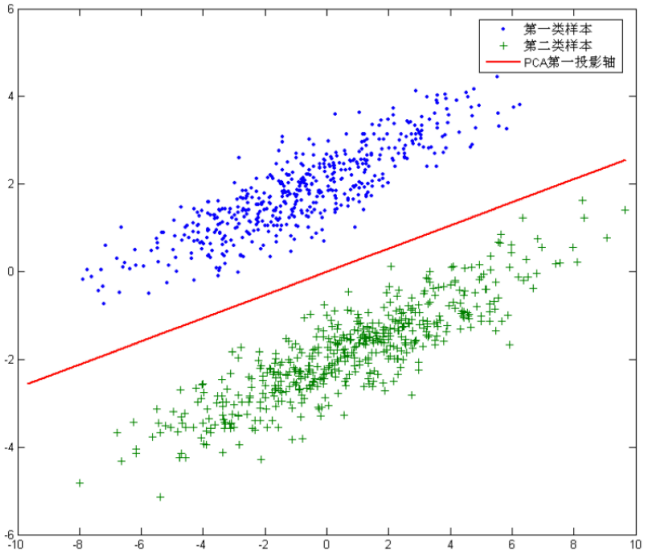
$$W^T x \in \mathbb{R}^p$$

---

6 线性判别分析

线性判别分析是一种有监督的数据降维方法，其基本思想是希望投影后不同类别之间的数据点距离更大，同一类别的数据点更紧凑。因为它保证在投影空间上有最大类间距离和最小类内距离，所以它也是一种有效的特征提取方法。





上面左图是 LDA 的投影轴，右图是 PCA 的投影轴，可以看见它们的区别。

考虑把 d 维空间的二分类数据点投影到一条直线上去。设样本为 $\{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$，它们分别属于两类，大小 n_1 的子集 $\mathcal{D}_1$ 属于类别 ω_1，大小 n_2 的子集 $\mathcal{D}_2$ 属于类别 ω_2。

投影直线是一条过原点的直线，方向向量为 w，如果 $||w|| = 1$，那么根据点积的几何意义，样本点 x 与 w 的点积 $y = w^T x$ 就是 x 在投影直线上的投影。事实上，w 的幅值并不重要，因为所有点都会被放缩同样的系数，重要的是 w 的方向。记投影后的子集 $\mathcal{D}_1, \mathcal{D}_2$ 分别为 $\mathcal{Y}_1, \mathcal{Y}_2$。

设第 i 类的样本均值为：

$$m_i = \frac{1}{n_i} \sum_{x \in \mathcal{D}_i} x$$

则投影后的样本均值为：

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y = \frac{1}{n_i} \sum_{x \in \mathcal{D}_i} w^T x = w^T m_i$$

投影后样本均值之差为：

$$|\tilde{m}_1 - \tilde{m}_2| = w^T |m_1 - m_2|$$

定义类别 ω_i 的类内散布如下：

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2$$

为了达到类内散布小、类间差异大的效果，我们最大化下列目标函数：

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

为了求解这个优化问题，首先把 $J(w)$ 写成 w 的表达式。分子部分：

$$\begin{aligned} |\tilde{m}_1 - \tilde{m}_2|^2 &= (w^T(m_1 - m_2))^2 \\ &= w^T(m_1 - m_2)(m_1 - m_2)^T w \qquad \text{因为 } w^T(m_1 - m_2) \text{ 是一个标量} \\ &= w^T S_B w \end{aligned}$$

其中 $S_B = (m_1 - m_2)(m_1 - m_2)^T$ **是类间散布矩阵**。又：

$$\begin{aligned} \tilde{s}_i^2 &= \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2 \\ &= \sum_{x \in \mathcal{D}_i} (w^T(x - m_i))^2 \\ &= \sum_{x \in \mathcal{D}_i} w^T(x - m_i)(x - m_i)^T w \qquad \text{因为 } w^T(x - m_i) \text{ 是一个标量} \\ &= w^T S_i w \end{aligned}$$

其中 $S_i = \sum_{x \in \mathcal{D}_i} (x - m_i)(x - m_i)^T$ **是第 i 类的类内散布矩阵**。**定义总类内散布矩阵**：

$$S_W = S_1 + S_2$$

为各类别的类内散布矩阵之和。

因此，优化问题可以写作：

$$\max_w \quad J(w) = \frac{w^T S_B w}{w^T S_W w}$$

这通常被称为广义瑞利商。

怎么解上述优化问题呢？注意到 w 的幅值对目标函数的大小没有影响（因为上下都是二次型，幅值变化可以约掉），所以我们可以令分母为 1，把问题转化成带约束优化问题：

$$\begin{aligned} \max_w \quad & J(w) = w^T S_B w \\ \text{s.t.} \quad & w^T S_W w = 1 \end{aligned}$$

应用拉格朗日乘数法，定义拉格朗日函数：

$$L(w) = w^T S_B w - \lambda(w^T S_W w - 1)$$

求偏导并令为零：

$$\frac{\partial L(w)}{\partial w} = 2S_B w - 2\lambda S_W w = 0$$

得到**广义特征方程**：

$$S_B w = \lambda S_W w$$

变换一下（先假设 S_W 可逆）：

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

这是什么？特征方程啊！所以我们要求的 w 就是矩阵 $S_W^{-1} S_B$ 的特征向量！

但是！在我们的问题中，其实没有必要去计算特征值和特征向量，因为我们可以直接瞧出一个解——注意到 $S_B w$ 和 $(m_1 - m_2)$ 共线，即存在这么一个 k 使得：$S_B w = k(m_1 - m_2)$，于是乎，很容易知道上述特征方程有这么一组特征值和特征向量：

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

特征方程是在 S_W 可逆的前提下得到的，如果 S_W 不可逆，可以用奇异值分解代替。

注意我们一直到现在都只在考虑二分类问题且投影到一条直线的情况，更多类问题和更高维投影的多重判别分析参看教材。

---

7 线性可分问题

对于两类问题，设有 n 个样本 $\{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$，如果存在一个线性判别函数 $g(x) = a^T x$ 能够完全正确地分类，则这些样本是线性可分的，权重向量 a 称为**分离向量或解向量**。

对 x_i 根据其类别指定标签 b_i，则我们可以将分类问题视作一个回归问题。我们希望 $a^T x_i = b_i$ 对所有样本都成立，即：

$$\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} a = b \iff Xa = b$$

其中 $X \in \mathbb{R}^{n \times d}$。但显然事实上，我们一般无法找到这样的一个完美的 a，而是具有一个误差向量：

$$e = Xa - b$$

欲最小化该误差，最小化平方误差损失函数：

$$\min_a \quad J(a) = ||Xa - b||^2 = \sum_{i=1}^{n} ||a^T x_i - b_i||^2$$

求梯度并令为零：

$$\nabla_a J(a) = \sum_{i=1}^{n} 2(a^T x_i - b_i) x_i = 2X^T(Xa - b) = 0$$

得到：

$$X^T X a = X^T b$$

若 $X^T X$ 可逆，则解为：

$$a = (X^T X)^{-1} X^T b = X^+ b$$

其中 X^+ 为 X 的伪逆。

分类问题被转化成了回归问题看起来很奇怪，但是对于二分类是自然的，例如设定：若 $a^T x > 0$ 则分为第一类，$a^T x < 0$ 分为第二类。特别地，如果把所有属于 ω_2 的样本乘以 -1（标准化），那么只有一个不等式 $a^T x > 0$。

通过上文可知，找线性判别函数被转化成了最小化损失函数问题。对于平方误差损失函数，我们可以用解析的方法求出解，但更一般的，我们可以用数值方法求解：

- 梯度下降法
- 牛顿法，拟牛顿法

- 共轭梯度法

选择适当的 b，则平方误差判别函数 a^Tx 和 Fisher 线性判别可以等价。首先我们对每个样本的第一个维度添加一个分量 $x^{(0)}=1$ 得到“增广模式向量”，然后对所有属于 ω_2 的样本的特征向量乘以 -1。不失一般性地，假设前 n_1 个样本属于 ω_1，后 n_2 个样本属于 ω_2，那么增广后的矩阵可写作：

$$Y=\begin{bmatrix}\mathbf{1}_1 & X_1\\ -\mathbf{1}_2 & -X_2\end{bmatrix}$$

其中 $\mathbf{1}_i$ 表示由 n_i 个 1 组成的列向量，$X_i\in\mathbb{R}^{n_i\times d}$。

设我们要求的向量 a 为：

$$a=\begin{bmatrix}w_0\\ w\end{bmatrix}$$

设定 b 为：

$$b=\begin{bmatrix}\frac{n}{n_1}\mathbf{1}_1\\ \frac{n}{n_2}\mathbf{1}_2\end{bmatrix}$$

那么这样的特定选法和 Fisher 线性判别是相关的，证明参看教材。

对于非线性可分问题，可以通过**核方法**转换成线性可分问题，具体参看教材。