模式识别复习

1 模式识别系统构成

1.1 监督模式识别

- 信息获取与预处理
- 特征提取与选择
- 分类器设计（训练）
- 分类决策

1.2 非监督模式识别

- 信息获取与预处理
- 特征提取与选择
- 聚类（自学习）
- 结果解释

1.3 实例

OCR（光学字符识别）是通过扫描仪把印刷或手写的文字稿件输入到计算机中，由计算机自动识别出其中的文字内容。

1.3.1 信息获取与预处理

对印刷或手写的文字稿件扫描输入，并将内容图像进行二值化等处理，分割单字。

1.3.2 特征提取与选择

将每个单字向各个方向投影，得到像素密度分布；提取笔画分解信息等。

1.3.3 分类器设计

OCR 问题即多类分类问题，利用大量样本数据，训练多类分类器。并结合已有的对文字结构的认知，提高其准确性。

1.3.4 分类决策

根据多类分类器的输出可能结果，结合上下文的联系，得出最终的估计字符识别结果。

2 贝叶斯决策

2.1 最小风险贝叶斯决策

2.1.1 定义

样本$\mathbf{x} = [x_1 \quad x_2 \quad \dots \quad x_d]^\mathrm{T}$

状态空间$\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$

决策空间$\mathrm{A} = \{\alpha_1, \alpha_2, \dots, \alpha_k\}$

对实际状态为ω_j的向量$\mathbf{x}$，采取决策α_i所带来的损失为$\lambda(\alpha_\mathrm{i}, \omega_j)$

对某个样本$\mathbf{x}$，属于各个状态的后验概率是$\mathrm{P}(\omega_\mathrm{j}|\boldsymbol{x}), \mathrm{j} = 1, \dots, \mathrm{c}$，则对他采取决策$\alpha_\mathrm{i}, i = 1, \dots, k$的期望损失为：

$$\mathrm{R}(\alpha_\mathrm{i}|\boldsymbol{x}) = E[\lambda(\alpha_i, \omega_j)|\boldsymbol{x}] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|\boldsymbol{x}), \quad i = 1, \dots, k$$

设有一决策规则$\alpha(\mathbf{x})$，他对所有样本决策造成的期望损失为：

$$\mathrm{R}(\alpha) = \int R(\alpha(\boldsymbol{x})|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

最小风险贝叶斯决策即：

$$\min_{\alpha} R(\alpha)$$

由于$R(\alpha(\boldsymbol{x})|\boldsymbol{x})$和$p(\boldsymbol{x})$非负，且$p(\boldsymbol{x})$已知，要使积分最小，就要使对所有$\mathbf{x}$使$R(\alpha(\boldsymbol{x})|\boldsymbol{x})$最小。

2.1.2 步骤

1. 利用贝叶斯公式计算后验概率：

$$P(\omega_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_j) P(\omega_j)}{\sum_{i=1}^{c} p(\boldsymbol{x}|\omega_i) P(\omega_i)}, \quad j = 1, \dots, c$$

2. 利用决策表，计算条件风险：

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

3. 选择风险最小的决策：

$$\alpha = \arg \min_{i=1,\dots,k} R(\alpha_i|\boldsymbol{x})$$

2.1.3 两类问题

$$l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \lesseqgtr \frac{P(\omega_2)}{P(\omega_1)} \cdot \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}, \quad 则\ x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

2.1.4 例题

1. 对两类问题，若损失函数$\lambda_{11}=\lambda_{22}=0,\lambda_{12}\neq 0,\lambda_{21}\neq 0,$，试求基于最小风险贝叶斯决策分界面处的两类错误率与$\lambda_{12},\lambda_{21}$的关系。

分界面处

$$\mathrm{P}(\omega_1|\mathrm{x})\cdot\lambda_{21}=\mathrm{P}(\omega_2|\mathrm{x})\cdot\lambda_{12}$$

因此

$$P_1(e)=\mathrm{P}(\omega_2|\mathrm{x})$$

$$P_2(e)=\mathrm{P}(\omega_1|\mathrm{x})$$

$$\frac{P_1(e)}{P_2(e)}=\frac{P(\omega_2|x)}{\mathrm{P}(\omega_1|\mathrm{x})}=\frac{\lambda_{21}}{\lambda_{12}}$$

2.2 最小错误率贝叶斯决策

即最小风险贝叶斯决策的特殊情况：

$$\lambda_{11}=\lambda_{22}=0$$
$$\lambda_{12}=\lambda_{21}=1$$

2.2.1 两类问题

$$l(x)=\frac{p(x|\omega_1)}{p(x|\omega_2)}\lessgtr\frac{P(\omega_2)}{P(\omega_1)},\qquad 则\ x\in\begin{cases}\omega_1\\\omega_2\end{cases}$$

2.3 正态分布时的决策

2.3.1 一元正态分布

$$\mathrm{p(x)}=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

2.3.2 多元正态分布

$$\mathrm{p(x)}=\frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

2.3.3 最小错误率贝叶斯决策

判别函数：

$$g_i(\boldsymbol{x}) = \ln[p(\boldsymbol{x}|\omega_i)P(\omega_i)] = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) - \ln\left[(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right] + lnP(\omega_i)$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) - \frac{d}{2}ln2\pi - \frac{1}{2}ln|\boldsymbol{\Sigma_i}| + lnP(\omega_i)$$

决策面方程为：

$$g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$$

2.3.4 例题

2. 设一个二维空间中的两类样本服从正态分布，其参数分别为 $\mu_1 = (-1,0)^T$，$\sum_1 = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$， $\mu_2 = (1,0)^T$， $\sum_2 = \begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}$，先验概率 $P(\omega_1) = P(\omega_2)$，试求基于最小错误率的贝叶斯决策分界面方程。

因为两类样本服从正态分布，其比为

$$\ln\left[\frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)}\right]$$

分界面处

$$\frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)} = \frac{P(\omega_2)}{P(\omega_1)} = 1$$

所以

$$\ln\left[\frac{P(x|\omega_1)}{P(x|\omega_2)}\right] = 0$$

$$\ln[P(x|\omega_1)] = \ln[P(x|\omega_2)]$$

$$(x_1+3)^2 + {x_2}^2 = 8 + 4ln2$$

为圆的方程

3 概率密度函数的估计

3.1 最大似然估计

3.1.1 定义

每类的样本集：$\chi = \{\mathbf{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N}\}$

其中的样本都是从密度为$p(\mathbf{x}|\omega_i)$中总体中独立抽取出来的。（独立同分布条件）

因此，获得样本集的概率即出现其中各个样本的联合概率：

$$l(\theta) = p(\chi|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x_i}|\theta)$$

其为参数θ相对于样本集χ的似然函数。

最大似然估计量即：

$$\hat{\theta} = arg \max_{\theta} l(\theta)$$

为了便于分析，定义对数似然估计函数：

$$H(\theta) = ln\, l(\theta) = ln \prod_{i=1}^{N} p(\boldsymbol{x_i}|\theta) = \sum_{i=1}^{N} ln\, p(\boldsymbol{x_I}|\theta)$$

3.1.2 计算

求解

$$\nabla_{\theta} H(\boldsymbol{\theta}) = 0$$

3.1.3 正态分布

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

令$\theta_1 = \mu, \theta_2 = \sigma^2$

$$ln\, p(x|\boldsymbol{\theta}) = -\frac{1}{2} ln\, 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\theta} H(\boldsymbol{\theta}) = \sum_{k=1}^{N} \nabla_{\theta} \ln p(x|\boldsymbol{\theta}) = \sum_{k=1}^{N} \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_k - \theta_1)^2 \end{bmatrix} = 0$$

解得

$$\hat{\mu} = \widehat{\theta_1} = \frac{1}{N} \sum_{k=1}^{N} x_k$$

$$\widehat{\sigma^2} = \widehat{\theta_2} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{\mu})^2$$

3.1.4 例题

3. 设$\chi = \{x_1, x_2, \dots, x_N\}$为来自点二项分布的样本集，即

$$f(x, P) = P^x Q^{1-x}, x = 0,1, 0 \le P \le 1, Q = 1 - P$$

试求参数P的最大似然估计量$\hat{P}$

对数似然估计函数为

$$H(p) = \sum_{k=1}^{N} \ln P^{x_k}(1-p)^{1-x_k} = \ln\frac{p}{1-p}\sum_{k=1}^{N} x_k + N\ln(1-p)$$

对p**求导有**

$$\frac{dH(p)}{dp} = \sum_{k=1}^{N} x_k\left(\frac{1}{p} - \frac{1}{p-1}\right) + N\frac{1}{p-1}$$

求极值有

$$\sum_{k=1}^{N} x_k\left(\frac{1}{p(p-1)}\right) = \frac{N}{p-1}$$

得最大似然估计为

$$\hat{p} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

3.2 贝叶斯估计

3.2.1 定义

损失函数：把θ估计为$\hat{\theta}$所造成的损失$\lambda(\hat{\theta}, \theta)$

期望风险：$R = \int_{E^d}\int_{\Theta} \lambda(\hat{\theta}, \theta)p(x, \theta)d\theta dx = \int_{E^d}\int_{\Theta} \lambda(\hat{\theta}, \theta)p(\theta|x)p(x)d\theta dx = \int_{E^d} R(\hat{\theta}|x)p(x)dx$，其中$x \in E^d, \theta \in \Theta$

条件风险：$R(\hat{\theta}|x)=\int_\Theta \lambda(\hat{\theta},\theta)p(\theta|x)d\theta$，其中$x\in \mathrm{E}^{\mathrm{d}}$

贝叶斯估计，即对所有的x，最小化条件风险。

常用平方误差损失函数，此时贝叶斯估计量在给定x时θ的条件期望：

$$\hat{\theta}=E[\theta|x]=\int_\Theta \theta p(\theta|x)d\theta$$

在给定样本集χ下，θ的贝叶斯估计是：

$$\hat{\theta}=E[\theta|x]=\int_\Theta \theta p(\theta|\chi)d\theta$$

3.2.2 计算

1. 确定先验分布$p(\theta)$
2. 求样本集的联合分布

$$p(\chi|\theta)=\prod_{i=1}^{N}p(x_i|\theta)$$

3. 求θ的后验概率分布

$$p(\theta|\chi)=\frac{p(\chi|\theta)p(\theta)}{\int_\Theta p(\chi|\theta)p(\theta)d\theta}$$

4. 求θ的贝叶斯估计量

$$\hat{\theta}=E[\theta|\chi]=\int_\Theta \theta p(\theta|\chi)d\theta$$

也可直接推断总体分布

$$p(x|\chi)=\int_\Theta p(x|\theta)p(\theta|\chi)d\theta$$

3.2.3 正态分布

假设均值μ是待估计参数，方差σ^2为已知参数，其分布密度为：

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

假设均值μ的先验分布也是正态分布，其均值为μ_0、方差为σ_0^2，即

$$p(\mu)=\frac{1}{\sqrt{2\pi}\sigma_0}exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}$$

利用下式估计μ

$$p(\mu|\chi)=\frac{p(\chi|\mu)p(\mu)}{\int_\Theta p(\chi|\mu)p(\mu)d\mu}$$

分母用于归一化，暂略

$$p(\chi|\mu)p(\mu)=p(\mu)\prod_{i=1}^{N}p(x_i|\mu)=\frac{1}{\sqrt{2\pi}\sigma_0}exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}\prod_{i=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right\}\right]$$

将所有与μ无关的量写入常数中

$$\begin{aligned}p(\chi|\mu)p(\mu)&=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^{N+1}}exp\left\{-\frac{1}{2}\left[\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2+\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]\right\}\\&=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^{N+1}}exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\mu^2+\mu_0^2-2\mu_0\mu)+\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i^2+\mu^2-2x_i\mu)\right]\right\}\\&=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^{N+1}}exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\mu^2-2\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right)\mu+\frac{\mu_0^2}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i^2}{\sigma^2}\right]\right\}\\&=\frac{1}{\sqrt{2\pi}\sigma_N}exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right\}\end{aligned}$$

其中

$$\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}$$

$$\mu_N=\sigma_N^2\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right)$$

所以

$$\hat{\mu}=\mu_N$$

3.2.4 例题

4. 假定损失函数为二次函数，以及 P 的先验密度为均匀分布，即$f(P)=1,0\leq P\leq 1$，在此条件下，求题 3 的贝叶斯估计量$\hat{P}$。

$$f(p|\chi)=\prod_{k=1}^{N}f(x_k|p)=p^{\sum_{k=1}^{N}x_k}\cdot(1-p)^{N-\sum_{k=1}^{N}x_k}$$

$$\sim Be(1+\sum_{k=1}^{N}x_k\,,1+N-\sum_{k=1}^{N}x_k)$$

已知 Beta 分布$\boldsymbol{Be(\alpha,\beta)}$的期望为$\frac{\boldsymbol{\alpha}}{\boldsymbol{\alpha+\beta}}$，所以

$$\hat{p}=E(p|\chi)=\frac{1+\sum_{k=1}^{N}x_k}{1+\sum_{k=1}^{N}x_k+1+N-\sum_{k=1}^{N}x_k}=\frac{1+\sum_{k=1}^{N}x_k}{N+2}$$

5. 设总体分布密度为$N(\mu,1),-\infty<\mu<+\infty$，并设$X=\{x_1,x_2,\dots,x_N\}$，用贝叶斯估计计算$\hat{\mu}$。已知$\mu$的先验分布$p(\mu)\sim N(0,1)$。

$$p(\chi|\mu)p(\mu)=\alpha\, exp\left\{-\frac{1}{2}\left[(N+1)\mu^2-2\mu\sum_{i=1}^{N}x_i+\sum_{i=1}^{N}x_i^2\right]\right\}$$

$$=\alpha\, exp\left\{-\frac{1}{2}\left[(N+1)\mu^2-2\mu\sum_{i=1}^{N}x_i+\sum_{i=1}^{N}x_i^2\right]\right\}$$

$$=\alpha'\, exp\left\{-\frac{1}{2}\left[(N+1)\left(\mu^2-\frac{\sum_{i=1}^{N}x_i}{N+1}\right)^2\right]\right\}$$

可见

$$\hat{\mu}=\mu_N=\frac{\sum_{i=1}^{N}x_i}{N+1}$$

4 线性分类器

4.1 FISHER 线性判别分析（LDA）

4.1.1 定义

目标是找到一个投影方向w，投影后，样本变为

$$\mathrm{y_i}=\boldsymbol{w}^T\boldsymbol{x_i},i=1,2,\dots,N$$

在原样本空间中，类均值向量为

$$\boldsymbol{m}_{\mathrm{i}}=\frac{1}{N_i}\sum_{x_j\in\chi_i}\boldsymbol{x}_j\,,\mathrm{i}=1,2$$

各类的类内离散度矩阵为

$$\mathbf{S_i}=\sum_{x_j\in\chi_i}(\boldsymbol{x}_j-m_i)(\boldsymbol{x}_j-m_i)^T\,,i=1,2$$

总类内离散度矩阵为

$$\mathbf{S_w}=\boldsymbol{S_1}+\boldsymbol{S_2}$$

类间离散度矩阵为

$$\mathbf{S_b}=(\boldsymbol{m_1}-\boldsymbol{m_2})(\boldsymbol{m_1}-\boldsymbol{m_2})^T$$

在投影以后的一维空间，两类的均值分别为

$$\widetilde{m_i}=\boldsymbol{w^T}\boldsymbol{m_i}, i=1,2$$

各类的类内离散度（是一个值）为

$$\widetilde{S_i}^2=\sum_{y_j\in\mathcal{Y}_i}\left(y_j-\widetilde{m_i}\right)^2, i=1,2$$

总类内离散度为

$$\widetilde{S_w}^2=\widetilde{S_1}^2+\widetilde{S_2}^2$$

类间离散度为

$$\widetilde{S_b}=(\widetilde{m_1}-\widetilde{m_2})^2$$

我们希望最终结果使两类尽可能分开，而各类内尽可能聚集，因此可有如下准则

$$\max J_F(w)=\frac{\widetilde{S_b}}{\widetilde{S_w}}=\frac{(\widetilde{m_1}-\widetilde{m_2})^2}{\widetilde{S_1}^2+\widetilde{S_2}^2}$$

代入原样本空间的式子，可得

$$\max_{\boldsymbol{w}} J_F(\boldsymbol{w})=\frac{\boldsymbol{w^T S_b w}}{\boldsymbol{w^T S_w w}}$$

解得

$$\mathbf{w}^*=\boldsymbol{S_w^{-1}}(\boldsymbol{m_1}-\boldsymbol{m_2})$$

4.1.2 例题

6. 设两类样本的类内离散矩阵分别为：

$$S_1=\begin{bmatrix}1 & 1/2\\ 1/2 & 1\end{bmatrix},\quad S_2=\begin{bmatrix}1 & -1/2\\ -1/2 & 1\end{bmatrix},\quad m_1=(2,0)^T,\quad m_2=(2,2)^T$$

试用 fisher 准则求其决策面方程。

$$\mathrm{J}(\omega)=\frac{\omega^T S_b\omega}{\omega^T S_w\omega}$$

$$\omega^* = \arg\max_{\omega} J(\omega) = {S_w}^{-1}(m_1 - m_2) = (S_1 + S_2)^{-1}(m_1 - m_2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} (0 \quad -2)^T$$

$$= \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} (0 \quad -2)^T = (0 \quad -1)^T$$

因为m_1和m_1的中点$(2 \quad 1)^T$应位于分界面上，所以可得分界面：

$$y = (0 \quad -1)x + 1$$

4.2 感知器

4.2.1 定义

可以直接得到完整的线性判别函数。

首先将样本向量和权向量增广

$$\boldsymbol{y} = [1, x_1, x_2, \dots, x_d]^T$$

$$\boldsymbol{a} = [w_0, w_1, w_2, \dots, w_d]^T$$

线性判别函数为

$$g(\boldsymbol{y}) = \boldsymbol{a^T y}$$

定义一个新变量$\mathbf{y}'$（规范化增广样本向量）

$$y_i' = \begin{cases} y_i, & y_i \in \omega_1 \\ -y_i, & y_i \in \omega_2 \end{cases}, i = 1,2, \dots, N$$

此时，样本可分性的条件即存在$\boldsymbol{a}$（解向量）

$$\boldsymbol{a^T y_i'} > 0, i = 1,2, \dots N$$

感知器准则函数为对所有错分样本的求和惩罚

$$J_P(\boldsymbol{a}) = \sum_{\boldsymbol{a^T y_k} \le 0} (-\boldsymbol{a^T y_k})$$

解向量则为

$$\boldsymbol{a}^* = \min_{\boldsymbol{a}} J_P(\boldsymbol{a}) = 0$$

4.2.2 计算

梯度下降法

$$\boldsymbol{a}(t+1) = \boldsymbol{a}(t) - \rho_t \nabla J_P(\boldsymbol{a})$$

其中

$$\nabla J_P(\boldsymbol{a}) = \sum_{\boldsymbol{a}^T\boldsymbol{y}_k \le 0} (-\boldsymbol{y}_k)$$

因此，迭代修正公式即

$$\boldsymbol{a}(t+1) = \boldsymbol{a}(t) - \rho_t \sum_{\boldsymbol{a}^T\boldsymbol{y}_k \le 0} (-\boldsymbol{y}_k)$$

4.2.3 例题

7. 用感知器算法求下列模式分类的解向量a：

$\omega_1: \left\{(0,1,1)^T, (0,1,0)^T\right\}$， $\omega_2: \left\{(1,0,0)^T, (1,0,1)^T\right\}$

假设步长$\rho = 1$。

$$\omega_1' = \{(1 \quad 0 \quad 1 \quad 1)^T, (1 \quad 0 \quad 1 \quad 0)^T\}$$

$$\omega_2' = \{(1 \quad -1 \quad 0 \quad 0)^T, (1 \quad -1 \quad 0 \quad -1)^T\}$$

$$\mathrm{J(a)} = \sum\nolimits_{\boldsymbol{a}^T\boldsymbol{y}_k \le 0} (-a^T y)$$

$$\nabla\mathrm{J(a)} = \sum\nolimits_{\boldsymbol{a}^T\boldsymbol{y}_k \le 0} (-y)$$

$$a^{(k+1)} = a^{(k)} - \rho\nabla\mathrm{J(a)} = a^{(k)} + \sum\nolimits_{\boldsymbol{a}^T\boldsymbol{y}_k \le 0} (y)$$

设 $a^{(1)} = (1 \quad 1 \quad 1 \quad 1)^T$

$$a^{(2)} = (1 \quad 1 \quad 1 \quad 1)^T + (2 \quad -2 \quad 0 \quad -1)^T = (3 \quad -1 \quad 1 \quad 0)^T$$

此时，可以保证所有样本都被正确分类。

所以，

$$a = (3 \quad -1 \quad 1 \quad 0)^T$$

5 非线性分类器

5.1 人工神经网络

5.1.1 定义

神经元接受信号，当信号的加权和大于阈值，则神经元激活，输出信号。

激活函数可以是阶跃函数，但其数学性质不够好，难以建立模型，故使用 Sigmoid 函数（S 形函数）：

$$f(\alpha)=\frac{1}{1+e^{-\alpha}}$$

其中

$$\alpha=\sum_{i=1}^{n} w_i x_i$$

即上层神经元的输入加权和。

5.1.2 反向传播算法（BP 算法）

5.1.2.1 确定结构

总共$L+1$层（0层为输入层，L层为输出层，$1\sim(L-1)$层为隐层）。

第l层有n_l个节点。

5.1.2.2 选取样本

$$x=\left(x_1,x_2,\ldots,x_{n_0}\right)^T$$

$$y=\left(y_1,y_2,\ldots,y_{n_L}\right)^T$$

5.1.2.3 初始化权值

可以随机取较小的数。

5.1.2.4 前馈阶段

计算估计计算值

$$z_i^{(l)}=\sum_{j=0}^{n_{l-1}} a_j^{(l-1)} w_{ji}^{(l)}, l=1,2,\ldots,L$$

$$z^{(0)}=x$$

计算估计激活值

$$a_i^{(l)}=Sigmoid\left(z_i^{(l)}\right), l=0,1,\ldots,L$$

前馈计算每一层的估计值。

5.1.2.5 计算输出层梯度

代价函数

$$J(w)=\sum_{i=1}^{n_L}\frac{1}{2}\left(y_i-a_i^{(L)}\right)^2$$

偏导

$$\frac{\partial J(w)}{\partial w_{ji}^{(L)}}=\frac{\partial J(w)}{\partial a_i^{(L)}}\cdot\frac{\partial a_i^{(L)}}{\partial z_i^{(L)}}\cdot\frac{\partial z_i^{(L)}}{\partial w_{ji}^{(L)}}$$

其中

$$\frac{\partial J(w)}{\partial a_i^{(L)}}=a_i^{(L)}-y_i$$

$$\frac{\partial a_i^{(L)}}{\partial z_i^{(L)}}=\frac{\partial}{\partial z_i^{(L)}}\left[\frac{1}{1+e^{-z_i^{(L)}}}\right]=\left(1-a_i^{(L)}\right)a_i^{(L)}$$

$$\frac{\partial z_i^{(L)}}{\partial w_{ji}^{(L)}}=\frac{\partial}{\partial w_{ji}^{(L)}}\left[\sum_{k=0}^{n_{L-1}}a_k^{(L-1)}w_{ki}^{(L)}\right]=a_j^{(L-1)}$$

所以

$$\frac{\partial J(w)}{\partial w_{ji}^{(L)}}=\left(a_i^{(L)}-y_i\right)\left(1-a_i^{(L)}\right)a_i^{(L)}a_j^{(L-1)}$$

5.1.2.6 反向传播

$$\frac{\partial J(w)}{\partial w_{ji}^{(l)}}=\frac{\partial J(w)}{\partial a_i^{(l)}}\cdot\frac{\partial a_i^{(l)}}{\partial z_i^{(l)}}\cdot\frac{\partial z_i^{(l)}}{\partial w_{ji}^{(l)}},l=0,1,\dots,L-1$$

其中

$$\frac{\partial a_i^{(l)}}{\partial z_i^{(l)}}=\frac{\partial}{\partial z_i^{(l)}}\left[\frac{1}{1+e^{-z_i^{(l)}}}\right]=\left(1-a_i^{(l)}\right)a_i^{(l)}$$

$$\frac{\partial z_i^{(l)}}{\partial w_{ji}^{(l)}}=\frac{\partial}{\partial w_{ji}^{(l)}}\left[\sum_{k=0}^{n_{l-1}}a_k^{(l-1)}w_{ki}^{(l)}\right]=a_j^{(l-1)}$$

其中的第一项$\frac{\partial J(w)}{\partial a_i^{(l)}}$，由下式通过反向传播的$l+1$层的$\frac{\partial J(w)}{\partial a_j^{(l+1)}}$计算

$$\frac{\partial J(w)}{\partial a_i^{(l)}}=\sum_{j=1}^{n_{l+1}}\frac{\partial J(w)}{\partial a_j^{(l+1)}}\cdot\frac{\partial a_j^{(l+1)}}{\partial z_j^{(l+1)}}\cdot\frac{\partial z_j^{(l+1)}}{\partial a_i^{(l)}}=\sum_{j=1}^{n_{l+1}}\frac{\partial J(w)}{\partial a_j^{(l+1)}}\cdot\left(1-a_j^{(l+1)}\right)a_j^{(l+1)}w_{ij}^{(l+1)}$$

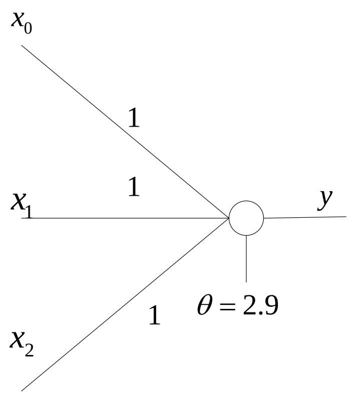
5.1.2.7 修正参数

$$\mathrm{w}_{\mathrm{ji}}^{(l)}(t+1)=\mathrm{w}_{\mathrm{ji}}^{(l)}(t)-\rho\cdot\nabla w_{ji}^{(l)}(t)=\mathrm{w}_{\mathrm{ji}}^{(l)}(t)-\rho\cdot\frac{\partial J(w)}{\partial w_{ji}^{(l)}(t)}$$

修正后，重复前馈、反向传播和修正参数步骤，直至误差小于设定值，或迭代次数过多。

5.1.3 例题

8. 由 M-P 模型组成的神经元网络的结构与参数如图所示，已知 $x_0,x_1,x_2\in\{0,1\}$，试问该网络与什么逻辑运算等价。M-C 使用的模型参数为：$\mathrm{y(h)}=\begin{cases}1 & \mathrm{h}\geq 0\\ 0 & \mathrm{h}<0\end{cases}$。



显然，只有$x_0=x_1=x_2=1$时，神经元激活。故该网络等价于与运算：AND。

6 特征选择与提取

6.1 基于类别可分性判据的特征提取

6.1.1 判据

$$\boldsymbol{\mu_i}=\mathrm{E_i}[\boldsymbol{x}]$$

$$\boldsymbol{\mu}=E[\boldsymbol{x}]$$

$$S_b=\sum_{i=1}^{c}P_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T$$

$$S_w=\sum_{i=1}^{c}P_iE_i[(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^T]$$

判据

$$J_1=tr(S_w+S_b)$$

$$J_2=tr(S_w^{-1}S_b)$$

…

6.1.2 提取

求矩阵$S_w^{-1}S_b$的特征值：$\lambda_1, \lambda_2, \ldots, \lambda_D$，从大到小排序。

选取其最大的d个特征值对应的特征向量作为最优变换矩阵$\mathbf{W}$。

6.1.3 例题

9. 已知有两类数据，分别为：

$$\omega_1 : (1,0), (2,0), (1,1)$$
$$\omega_2 : (-1,0), (0,1), (-1,1)$$

试求该数据的类内及类间离散矩阵S_w和S_b，并求使J_2达到最大的特征提取。

$$m_1 = \left(\frac{4}{3}, \frac{1}{3}\right), m_2 = \left(-\frac{2}{3}, \frac{2}{3}\right)$$

$$S_1 = \sum_{x \in \omega_1} (x - m_1)(x - m_1)^T = \frac{1}{9}\begin{pmatrix} 6 & -3 \\ -3 & 6 \end{pmatrix}$$

$$S_2 = \sum_{x \in \omega_2} (x - m_2)(x - m_2)^T = \frac{1}{9}\begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$$

$$\mathrm{S_w} = S_1 + S_2 = \frac{1}{9}\begin{pmatrix} 12 & 0 \\ 0 & 12 \end{pmatrix}$$

$$\mathrm{S_b} = (m_1 - m_2)(m_1 - m_2)^T = \frac{1}{9}\begin{pmatrix} 36 & -6 \\ -6 & 1 \end{pmatrix}$$

$$\mathrm{S_w^{-1}} = 9\begin{pmatrix} \frac{1}{12} & 0 \\ 0 & \frac{1}{12} \end{pmatrix}$$

$$\mathrm{S_w^{-1}} S_b = \begin{pmatrix} 3 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{12} \end{pmatrix}$$

特征值求得

$$\lambda_1 = \frac{37}{12}, \lambda_2 = 0$$

因此，选取λ_2对应的特征向量$(-6,1)^{\mathrm{T}}$

6.2 K-L 变换

6.2.1 简述

通过选择产生矩阵的前 d 大的特征值对应的特征向量作为变换矩阵。

当去掉均值信息时，可采用协方差矩阵作为产生矩阵，此时 K-L 变换等价于主成分分析（PCA）。

6.2.2 从类均值提取判别信息

计算总类内离散度S_w

将其作为产生矩阵进行 K-L 变换，求解特征值和特征向量。

性能指标

$$J(y_i) = \frac{\boldsymbol{u}_i^T \boldsymbol{S}_b \boldsymbol{u}_i}{\lambda_i}$$

利用性能指标计算，从大到小排序，选取前 d 个特征的特征向量作为变换矩阵。

6.2.3 例题

10. 设有一个两类问题，先验概率相等，特征为二维向量，类均值向量分别为

$$\boldsymbol{\mu}_1 = [4,2]^T$$

$$\boldsymbol{\mu}_2 = [-4,-2]^T$$

协方差矩阵分别是

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$
$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

利用 K-L 变换计算变换矩阵。

$$S_w = \frac{1}{2}(\Sigma_1 + \Sigma_2) = \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 3.5 \end{bmatrix}$$

$$\lambda_1 = 5, \lambda_2 = 2$$
$$U = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}$$
$$S_b = \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix}$$

$$J(x_1) = 3.6$$

$$J(x_2) = 1$$

所以选

$$u_1 = [0.707, 0.707]^T$$

作为变换矩阵。