机器学习白板推导系列笔记 (4.1~4.9)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 10 月 31 日

1 感知机

本节介绍最简单的线性分类器感知机，它是硬分类算法．感知机是二分类的线性分类模型，输入实例的特征向量，输出实例的类别．旨在在特征空间中求出线性划分的分离超平面，属于判别模型，感知机的数学模型如下

$$y = f(\mathbf{x}) = sign(\mathbf{w}^T\mathbf{x}) \tag{1}$$

通常来说 $x \in \mathbb{R}^{p\times 1}$, $w \in \mathbb{R}^{p\times 1}$, $y \in \mathbb{R}$, sgn 函数的定义

$$sign(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2}$$

样本集合为 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$，定义一个集合 D 为被误分类点的集合．由此可以定义出感知机的损失函数

$$\mathcal{L}(\mathbf{w}) = \sum_{\mathbf{x}_i \in D} \mathbb{I}\{y_i\mathbf{w}^T\mathbf{x} < 0\} \tag{3}$$

注. 当能被正确分类时，$y_i\mathbf{w}^T\mathbf{x} > 0$

但该损失函数存在一个问题，$\mathbb{I}(\cdot)$ 是跳变函数，因此不可导，无法求其梯度，但我们观察式 $y_i\mathbf{w}^T\mathbf{x}$，此式可导，也能很好的反应模型的损失特性，因此构造新的损失函数

$$\mathcal{L}(\mathbf{w}) = -\sum_{\mathbf{x}_i \in D} y_i\mathbf{w}^T\mathbf{x}_i \tag{4}$$

损失函数的梯度

$$\nabla_{\mathbf{w}}\mathcal{L} = -y_i\mathbf{x}_i \tag{5}$$

使用随机梯度下降法 (SGD) 进行更新

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \eta\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w}^k + \eta y_i\mathbf{x}_i \tag{6}$$

此算法要求数据必须线性可分，因此这也是感知机的缺点，可以使用此算法的变形，pocket algorithm.

2 线性判别分析

本节讲述**硬分类**算法中的线性判别分析 (LDA), 也叫 Fisher 判别分析, 先做一下简单的符号说明, 设数据集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, 规定 $y = +1$ 属于类别 c_1, $y = -1$ 属于类别 c_2.

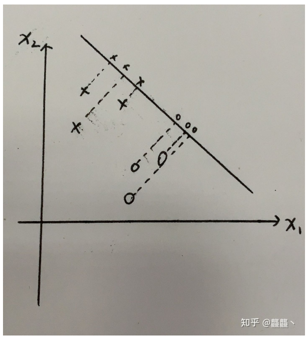
$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{pmatrix}^T = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}_{N \times p}$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

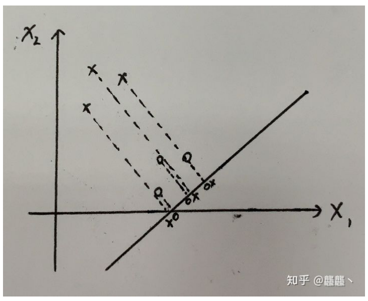
注. $\mathbf{x}_{c_1} = \{\mathbf{x}_i | y_i = +1\}, \mathbf{x}_{c_2} = \{\mathbf{x}_i | y_i = -1\} |c_1| = N_1, |c_2| = N_2, N_1 + N_2 = N$

2.1 建模

LDA 算法的主要思想是: 同类内差距小, 类间差距大. 从降维的角度出发, 以 $p = 2$ 为例, 将 p 维的数据投影到一维空间中, 在新空间中选择一个阈值 (threshold), 从而完成分类, 如下所示



但是, 如果投影的轴选的不好, 可能会无法分类, 如下所示



因此, 选择投影轴的目标为类内距离小, 类间距离大 (松耦合, 高内聚), 转化为数学描述就是类内方差足够小, 类间均值差距大. 假设投影轴为 $\mathbf{w}$, 且 $||\mathbf{w}|| = 1$, 那么数据在 $\mathbf{w}$ 轴上的投影即

$\mathbf{w}^T\mathbf{x}_i$(因为投影到直线上, 因此 $\mathbf{w}^T\mathbf{x}_i$ 为**实数**), 因此均值和方差可以表示为

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N}\mathbf{w}^T\mathbf{x}_i$$

$$S_z = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^T$$

那么对于 c_1 和 c_2 类, 分别有

$$\bar{z}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} z_i = \frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{w}^T\mathbf{x}_i, S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(z_i - \bar{z}_1)(z_i - \bar{z}_1)^T$$

$$\bar{z}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} z_i = \frac{1}{N_2}\sum_{i=1}^{N_2}\mathbf{w}^T\mathbf{x}_i, S_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(z_i - \bar{z}_2)(z_i - \bar{z}_2)^T$$

接下来定义**目标函数**, 根据上面所述, 类间 $(\bar{z}_1 - \bar{z}_2)^2$, 类内 $S_1 + S_2$, 即

$$\arg\max_{\mathbf{w}}(\bar{z}_1 - \bar{z}_2)^2$$

$$\arg\min_{\mathbf{w}} S_1 + S_2$$

把二者进行结合, 定义一个新的目标函数

$$\mathcal{J}(\mathbf{w}) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2} \tag{7}$$

对于分母 $(\bar{z}_1 - \bar{z}_2)^2$ 有

$$\begin{aligned}
(\bar{z}_1 - \bar{z}_2)^2 &= (\frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{w}^T\mathbf{x}_i - \frac{1}{N_2}\sum_{i=1}^{N_2}\mathbf{w}^T\mathbf{x}_i)^2 \\
&= (\mathbf{w}^T(\frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{x}_i - \frac{1}{N_2}\sum_{i=1}^{N_2}\mathbf{x}_i))^2 \\
&= (\mathbf{w}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2 \\
&= \mathbf{w}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{w}
\end{aligned}$$

对于分子 $S_1 + S_2$ 有

$$\begin{aligned}
S_1 + S_2 &= \frac{1}{N_1}\sum_{i=1}^{N_1}(z_i - \bar{z}_1)(z_i - \bar{z}_1)^T + \frac{1}{N_2}\sum_{i=1}^{N_2}(z_i - \bar{z}_2)(z_i - \bar{z}_2)^T \\
&= \frac{1}{N_1}\sum_{i=1}^{N_1}(\mathbf{w}^T\mathbf{x}_i - \frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{w}^T\mathbf{x}_i)(\mathbf{w}^T\mathbf{x}_i - \frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{w}^T\mathbf{x}_i)^T \\
&+ \frac{1}{N_2}\sum_{i=1}^{N_2}(\mathbf{w}^T\mathbf{x}_i - \frac{1}{N_2}\sum_{i=1}^{N_2}\mathbf{w}^T\mathbf{x}_i)(\mathbf{w}^T\mathbf{x}_i - \frac{1}{N_2}\sum_{i=1}^{N_2}\mathbf{w}^T\mathbf{x}_i)^T \\
&= \mathbf{w}^T\frac{1}{N_1}\sum_{i=1}^{N_1}[(\mathbf{x}_i - \bar{\mathbf{x}}_1)(\mathbf{x}_i - \bar{\mathbf{x}}_1)^T]\mathbf{w} + \mathbf{w}^T\frac{1}{N_2}\sum_{i=1}^{N_2}[(\mathbf{x}_i - \bar{\mathbf{x}}_2)(\mathbf{x}_i - \bar{\mathbf{x}}_2)^T]\mathbf{w} \\
&= \mathbf{w}^T S_{c_1}\mathbf{w} + \mathbf{w}^T S_{c_2}\mathbf{w} \\
&= \mathbf{w}^T(S_{c_1} + S_{c_2})\mathbf{w}
\end{aligned}$$

式 (7) 变为

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{w}}{\mathbf{w}^T(S_{c_1} + S_{c_2})\mathbf{w}} \tag{8}$$

2.2 模型求解

接下来对式 (8) 进行求解, 先做一下符号假设. 设 $S_b = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T$, $S_w = S_{c_1} + S_{c_2}$, $S_b \in \mathbb{R}^{p\times p}$ 意为类间方差 (between-class), $S_w \in \mathbb{R}^{p\times p}$ 意为类内方差 (within-class). 因此式 (8) 化简为

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} = (\mathbf{w}^T S_b \mathbf{w})(\mathbf{w}^T S_w \mathbf{w})^{-1} \tag{9}$$

对上式求导得

$$\frac{\partial}{\partial \mathbf{w}}\mathcal{J}(\mathbf{w}) = 2S_b\mathbf{w}\cdot(\mathbf{w}^T S_w \mathbf{w})^{-1} + (\mathbf{w}^T S_b \mathbf{w})\cdot(-1)(\mathbf{w}^T S_w \mathbf{w})^{-2}\cdot 2S_w\mathbf{w}$$

令 $\frac{\partial}{\partial \mathbf{w}}\mathcal{J}(\mathbf{w}) = 0$ 得

$$2S_b\mathbf{w}\cdot(\mathbf{w}^T S_w \mathbf{w})^{-1} + (\mathbf{w}^T S_b \mathbf{w})\cdot(-1)(\mathbf{w}^T S_w \mathbf{w})^{-2}\cdot 2S_w\mathbf{w} = 0$$

又因为 $\mathbf{w} \in \mathbb{R}^{p\times 1}$, $\mathbf{w}^T \in \mathbb{R}^{1\times p}$, 因此 $\mathbf{w}^T S_w \mathbf{w}$ 和 $\mathbf{w}^T S_b \mathbf{w}$均为常数, 上式子可化为

$$S_b\mathbf{w} - (\mathbf{w}^T S_b \mathbf{w})(\mathbf{w}^T S_w \mathbf{w})^{-1}\cdot S_w\mathbf{w} = 0$$

解得

$$\mathbf{w} = \frac{\mathbf{w}^T S_w \mathbf{w}}{\mathbf{w}^T S_b \mathbf{w}} S_w^{-1} S_b \mathbf{w} \tag{10}$$

注意求解的目的是找到 $\mathbf{w}$ 的方向, 大小并不是我们关心的, 因此设

$$\mathbf{w} \propto S_w^{-1} S_b \mathbf{w} = S_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{w}$$

$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \in \mathbb{R}^{1\times p}$, 因此 $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{w} \in \mathbb{R}$ 上式又可以简化为

$$\mathbf{w} \propto S_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{11}$$

因此 $S_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ 就是最终要求的 $\mathbf{w}$ 的方向. 特别的, 如果 S_w 是对角矩阵, 各向同性的话, $S_w^{-1} \propto I$, 那么

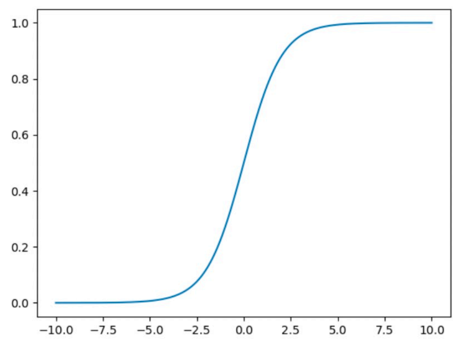
$$\mathbf{w} \propto (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{12}$$

3 逻辑回归 (logistic regression)

逻辑回归是软分类中的概率判别模型, 直接对条件概率 $p(y|x)$ 进行建模, 通过在线性回归上添加激活函数从而进行分类, 逻辑回归中使用的激活函数是Sigmoid 函数.

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$

函数图像如下所示



下面进行推导, 设数据集 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^N$, $\mathbf{x}_i\in\mathbb{R}^p$, $y_i\in\{0,1\}$. 逻辑回归模型是如下的条件概率分布

$$\begin{aligned} p(y=1|\mathbf{x}) &= \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} \\ p(y=0|\mathbf{x}) &= 1-\sigma(\mathbf{w}^T\mathbf{x}) = 1-\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{\exp(-\mathbf{w}^T\mathbf{x})}{1+\exp(-\mathbf{w}^T\mathbf{x})} \end{aligned} \tag{13}$$

将上式表达成一般形式即

$$p(y|\mathbf{x}) = \pi(\mathbf{x})^y\cdot(1-\pi(\mathbf{x}))^{1-y} \tag{14}$$

注. $\pi(\mathbf{x})=p(y=1|\mathbf{x}), 1-\pi(\mathbf{x})=p(y=0|\mathbf{x})$

使用极大似然估计 (MLE) 估计模型参数

$$\begin{aligned} \arg\max_{\mathbf{w}}\log P(\mathbf{Y}|\mathbf{X}) &= \arg\max_{\mathbf{w}}\log\prod_{i=1}^N p(y_i|\mathbf{x}_i) \\ &= \arg\max_{\mathbf{w}}\sum_{i=1}^N\log p(y_i|\mathbf{x}_i) \\ &= \arg\max_{\mathbf{w}}\sum_{i=1}^N\log[\pi(\mathbf{x}_i)^{y_i}\cdot(1-\pi(\mathbf{x}_i))^{1-y_i}] \\ &= \arg\max_{\mathbf{w}}\sum_{i=1}^N[y_i\log\pi(\mathbf{x}_i)+(1-y_i)\log(1-\pi(\mathbf{x}_i))] \end{aligned}$$

最终得似然函数

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^N[y_i\log\pi(\mathbf{x}_i)+(1-y_i)\log(1-\pi(\mathbf{x}_i))] \tag{15}$$

对 $\mathcal{L}(\mathbf{w})$ 求最大值, 得到 $\mathbf{w}$ 的估计值, 通常逻辑回归采用的是梯度下降法或者拟牛顿法.

4 高斯判别分析 (GDA)

本节介绍软分类中的概率生成模型高斯判别分析 (Gaussian Discriminant Analysis), 与逻辑回归直接对 $p(y|x)$ 建模不同, 高斯判别分析通过贝叶斯定理间接获得 $p(y|x)$ 的表示. 贝叶斯定理表述如下

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} \tag{16}$$

4.1 模型建立

做一下简单的符号说明, 设数据集 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^N$, 其中 $\mathbf{x}_i\in\mathbb{R}^p$, $y_i\in\{0,1\}$. 设集合 $c_1=\{\mathbf{x}_i|y_i=1, i=1,2,\cdots,N\}$, $c_2=\{\mathbf{x}_i|y_i=0, i=1,2,\cdots,N\}$, 且 $|c_1|=N_1$, $|c_2|=N_2$, $N_1+N_2=N$.

又因为在式 (16) 中对于不同 y, 分母 $p(x)$ 不变, 因此

$$\hat{y} = \arg\max_{y\in\{0,1\}} p(y|\mathbf{x}) \Rightarrow \arg\max_{y\in\{0,1\}} p(y)p(\mathbf{x}|y) \tag{17}$$

接下来做出一些假设, 认为随机变量 y 服从伯努力分布, 即 $p(y=1)=\phi, p(y=1)=1-\phi$,

写成一个通式即

$$p(y) = \phi^y (1-\phi)^{1-y} \tag{18}$$

假设 $\mathbf{x}|y$ 条件概率服从高斯分布, 即

$$\mathbf{x}|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$
$$\mathbf{x}|y = 0 \sim \mathcal{N}(\mu_2, \Sigma)$$

同理

$$p(\mathbf{x}|y) = \mathcal{N}(\mu_1, \Sigma)^y \mathcal{N}(\mu_2, \Sigma)^{1-y} \tag{19}$$

下面建立似然函数 $\mathcal{L}(\theta)$

$$\begin{aligned}
\mathcal{L}(\theta) &= \log \prod_{i=1}^{N} p(\mathbf{x}_i, y_i) \\
&= \sum_{i=1}^{N} \log p(\mathbf{x}_i, y_i) = \sum_{i=1}^{N} \log[p(y_i)p(\mathbf{x}_i|y_i)] \\
&= \sum_{i=1}^{N} [\log p(y_i) + \log p(\mathbf{x}_i|y_i)] \\
&= \sum_{i=1}^{N} [\log(\phi^{y_i}(1-\phi)^{1-y_i}) + \log \mathcal{N}(\mu_1, \Sigma)^{y_i} + \log \mathcal{N}(\mu_2, \Sigma)^{1-y_i}]
\end{aligned}$$

注. $\theta = (\mu_1, \mu_2, \Sigma, \phi)$

4.2 求解参数

首先求解参数 ϕ

$$\begin{aligned}
\frac{\partial}{\partial \phi}\mathcal{L}(\theta) &= \frac{\partial}{\partial \phi}(\sum_{i=1}^{N} \log(\phi^{y_i}(1-\phi)^{1-y_i})) \\
&= \frac{\partial}{\partial \phi}\sum_{i=1}^{N}(y_i \log \phi_i + (1-y_i)\log(1-\phi)) \\
&= \sum_{i=1}^{N}(\frac{y_i}{\phi} - \frac{1-y_i}{1-\phi})
\end{aligned}$$

令 $\frac{\partial}{\partial \phi}\mathcal{L}(\theta) = 0$ 解得

$$\hat{\phi} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{N_1}{N} \tag{20}$$

下面求解参数 μ_1/μ_2, 因为二者较为类似, 因此以求解 μ_1 为例

$$\begin{aligned}
\frac{\partial}{\partial\mu_1}\mathcal{L}(\theta) &= \frac{\partial}{\partial\mu_1}\sum_{i=1}^{N}\log\mathcal{N}(\mu_1,\Sigma)^{y_i} \\
&= \frac{\partial}{\partial\mu_1}\sum_{i=1}^{N}y_i\log(\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}_i-\mu_1)^T\Sigma^{-1}(\mathbf{x}_i-\mu_1))) \\
&= \frac{\partial}{\partial\mu_1}\sum_{i=1}^{N}y_i(-\frac{1}{2})(\mathbf{x}_i-\mu_1)^T\Sigma^{-1}(\mathbf{x}_i-\mu_1) \\
&= \frac{\partial}{\partial\mu_1}\sum_{i=1}^{N}y_i(-\frac{1}{2})(\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i-\mathbf{x}_i^T\Sigma^{-1}\mu_1-\mu_1^T\Sigma^{-1}\mathbf{x}_i+\mu_1^T\Sigma^{-1}\mu_1) \\
&= \frac{\partial}{\partial\mu_1}\sum_{i=1}^{N}y_i(-\frac{1}{2})(-2\mu_1^T\Sigma^{-1}\mathbf{x}_i+\mu_1^T\Sigma^{-1}\mu_1) \\
&= \sum_{i=1}^{N}y_i(-\frac{1}{2})(-2\Sigma^{-1}\mathbf{x}_i+2\Sigma^{-1}\mu_1)
\end{aligned}$$

令 $\frac{\partial}{\partial\mu_1}\mathcal{L}(\theta)=0$ 解得

$$\sum_{i=1}^{N}y_i(-\frac{1}{2})(-2\Sigma^{-1}\mathbf{x}_i+2\Sigma^{-1}\mu_1)=0\rightarrow\sum_{i=1}^{N}y_i\mathbf{x}_i=\sum_{i=1}^{N}y_i\mu_1$$

得

$$\hat{\mu}_1=\frac{\sum_{i=1}^{N}y_i\mathbf{x}_i}{\sum_{i=1}^{N}y_i}=\frac{1}{N_1}\sum_{x_i\in c_1}x_i \tag{21}$$

最后求协方差矩阵 Σ

$$\begin{aligned}
\frac{\partial}{\partial\Sigma}\mathcal{L}(\theta) &= \frac{\partial}{\partial\Sigma}\sum_{i=1}^{N}(y_i\log\mathcal{N}(\mu_1,\Sigma)+(1-y_i)\log\mathcal{N}(\mu_2,\Sigma)) \\
&= \frac{\partial}{\partial\Sigma}(\sum_{\mathbf{x}_i\in c_1}\log\mathcal{N}(\mu_1,\Sigma)+\sum_{\mathbf{x}_i\in c_2}\log\mathcal{N}(\mu_2,\Sigma))
\end{aligned} \tag{22}$$

下面对 $\sum_i^N\log\mathcal{N}(\mu,\Sigma)$ 得的一般形式做一下讨论

$$\begin{aligned}
\sum_i^N\log\mathcal{N}(\mu,\Sigma) &= \sum_i^N\log(\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu))) \\
&= \sum_i^N(\log\frac{1}{(2\pi)^{\frac{p}{2}}}-\frac{1}{2}\log|\Sigma|-\frac{1}{2}(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)) \\
&= C-\frac{1}{2}\sum_i^N\log|\Sigma|-\frac{1}{2}\sum_i^N(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)
\end{aligned} \tag{23}$$

对 $\sum_i^N (\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)$ 进一步讨论, 可以看出 $(\mathbf{x}_i-\mu)^T_{1\times p}\Sigma^{-1}_{p\times p}(\mathbf{x}_i-\mu)_{p\times 1}\in\mathbb{R}$, 因此

$$\begin{aligned}\sum_i^N(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu) &= \sum_i^N tr[(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)]\\ &= tr[\sum_i^N(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)]\\ &= tr[\sum_i^N(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^T\Sigma^{-1}](\text{迹的性质})\\ &= tr[NS\Sigma^{-1}]\\ &= N\cdot tr[S\Sigma^{-1}]\end{aligned}\tag{24}$$

注. 其中 $S=\frac{1}{N}\sum_{i=1}^N(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^T$

将式 (23), (24) 代入式 (22) 得

$$\begin{aligned}\frac{\partial}{\partial\Sigma}(\sum_{\mathbf{x}_i\in c_1}\log\mathcal{N}(\mu_1,\Sigma)+\sum_{\mathbf{x}_i\in c_2}\log\mathcal{N}(\mu_2,\Sigma)) &= \frac{\partial}{\partial\Sigma}(-\frac{1}{2}N_1\log|\Sigma|-\frac{1}{2}N_1tr(S_1\cdot\Sigma^{-1})\\ &\quad -\frac{1}{2}N_2\log|\Sigma|-\frac{1}{2}N_2tr(S_2\cdot\Sigma^{-1})+C)\\ &= \frac{\partial}{\partial\Sigma}(-\frac{1}{2}N\log|\Sigma|-\frac{1}{2}N_1tr(S_1\cdot\Sigma^{-1})-\frac{1}{2}N_2tr(S_2\cdot\Sigma^{-1}))\\ &= -\frac{1}{2}(N\frac{1}{|\Sigma|}|\Sigma|\Sigma^{-1}+N_1S_1^T(-1)\Sigma^{-2}+N_2S_2^T(-1)\Sigma^{-2})\\ &= -\frac{1}{2}(N\Sigma^{-1}-N_1S_1^T\Sigma^{-2}-N_2S_2^T\Sigma^{-2})\end{aligned}$$

令上式等 0 得

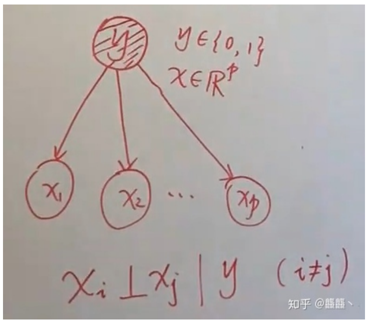
$$\hat{\Sigma}=\frac{N_1S_1^T+N_2S_2^T}{N}=\frac{N_1S_1+N_2S_2}{N}\tag{25}$$

注. 因为 S 为对称矩阵

至此完整的求解了高斯判别分析中的所有参数.

5 朴素贝叶斯

本节主要介绍软分类中概率生成模型的朴素贝叶斯分类器, 其主要思想是朴素贝叶斯假设, 即条件独立性假设, 最简单的概率图模型表示如下



设数据集 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^N, \mathbf{x}_i\in\mathcal{R}^p, y_i\in\{c_1,c_2,\cdots,c_k\}$, 由式 (17) 分析可知

$$\hat{y}=\arg\max_{y} p(y)p(\mathbf{x}|y)$$

即

$$\hat{y}=\arg\max_{c_k} p(y=c_k)p(\mathbf{x}|y=c_k) \tag{26}$$

又

$$p(\mathbf{x}|y=c_k)=\prod_{i=1}^{p} p(x^{(i)}|y=c_k)$$

式 (26) 化为

$$\hat{y}=\arg\max_{c_k} p(y=c_k)\prod_{i=1}^{p} p(x^{(i)}|y=c_k) \tag{27}$$

随机变量 y 在二分类的情况下满足伯努利分布, 在多分类的情况下满足类别分布 (Categorical Distribution). 当 x^i 为离散型随机变量时, $x^i|y$ 服从类别分布; 当 x^i 为连续型随机变量时, $x^i|y$ 服从 $\mathcal{N}(\mu_i,\sigma_i^2)$, $p(y)$ 和 $p(x^i|y)$ 的参数估计较为简单, 使用 MLE 即可, 参数估计结果如下

$$\begin{aligned}p(y=c_k)&=\frac{\sum_{i=1}^N \mathbb{I}(y_i=c_k)}{N}\\ p(x^{(j)}|y=c_k)&=\frac{\sum_{i=1}^N \mathbb{I}(x_i^{(j)},y_i=c_k)}{\sum_{i=1}^N \mathbb{I}(y_i=c_k)}\end{aligned} \tag{28}$$