判别与分类

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

简介

1.1 简介

- **判别** (Discrimination): 使用具有类别信息的观测数据 (Training Set, Learning set) 建立一个分类器 (classifier) 或者分类法则 (classification rule)，其可以最大可能的区分事先定义的类。(**Separation**)
- **分类** (Classification)：给定一组新的未知类别信息的观测数据集，使用分类器将其分配到一些已知的类中.(**Allocation**)
- 实际应用中, 判别与分类常常混在一起:
  - 一个作为判别的 p 元函数也可以用于对新的观测进行分类.
  - 一个分类准则常常作为判别法则使用

- 机器学习领域中，判别与分类称为**有指导 (监督) 的学习**(Supervised learning)
- 判别与分类的研究是一个交叉领域, 常用方法有
  - Bayes 判别
  - 最大似然判别方法
  - Fisher 线性判别分析 (FLDA)
  - 最近邻分类 (NNC)
  - 支持向量机 (SVM), C4.5, 神经网络, 等等

1.2 Bayes 判别分析

- 假设有两个总体 (类), $G = 1, 2$ 表示类别. X 为取值 Ω 上的多元观测, 且 $X|G = g \sim f_g(x)$, $g = 1, 2$. f 为概率函数.

- 记两个类的先验概率为 $P(G = g) = p_g, g = 1, 2$.

- 对任意给定的观测 X, 必须把 X 归到两个类中的某个. 等价地, (**分类规则, 决策法则**)

$$\delta(X) = \begin{cases} 1, & X \in R_1 \\ 2, & X \in R_2 = \Omega - R_1 \end{cases}$$

- 决策的损失:

$$L(\delta(X), G) = \begin{cases} c(2|1) > 0, & X \in R_2, G = 1 \\ c(1|2) > 0, & X \in R_1, G = 2 \\ 0, & otherwise \end{cases}$$

- 从而错误分类带来的平均损失为 (ECM, Expected Cost of Misclassification)(Bayes 风险):

$$\begin{aligned} ECM &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \\ &= E_G E_{X|G} L(\delta(X), G) = R(\delta, G) \end{aligned}$$

其中分类 $P(2|1), P(1|2)$ 为错误分类概率:

  - $P(2|1) = P(X \in R_2 | G = 1) = \int_{R_2} f_1(\mathbf{x}) d\mathbf{x}$
  - $P(1|2) = P(X \in R_1 | G = 2) = \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$

- 因此最优决策 (**Bayes 解**) 即为

$$\delta^*(X) = \arg\min_{R_1, R_2} ECM = \arg\min_{\delta} R(\delta, G)$$

从而得到最优分类法则 (练习 11.3)

$$\delta^*(X) = \begin{cases} 1, & X \in \{R_1^* : \frac{f_1(\mathbf{X})}{f_2(\mathbf{X})} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}\} \\ 2, & X \in \{R_2^* : \frac{f_1(\mathbf{X})}{f_2(\mathbf{X})} < \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}\} \end{cases}$$

- 分类法则中等式成立时候, 可以采取进一步的措施 (比如随机化) 来确定分类法则的值.
- $p_2/p_1 = 1$, 两个类的先验概率相同. 常常用于对两个类没有先验信息时候.
- $c(1|2)/c(2|1) = 1$, 错误分类的损失相同. 常常用于没有明确分类损失时候.
- $(c(1|2)/c(2|1))(p_2/p_1) = 1$, 此时为似然法则.
- 对一个观测 x_0, 由于

$$P(G = 1|X = \mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)p_1}{f_1(\mathbf{x}_0)p_1 + f_2(\mathbf{x}_0)p_2}$$
$$P(G = 2|X = \mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)p_2}{f_1(\mathbf{x}_0)p_1 + f_2(\mathbf{x}_0)p_2}$$

▷ 因此按照**后验概率原则**: 当 $P(G = 1|X = \mathbf{x}_0) > P(G = 2|X = \mathbf{x}_0)$ 时候将 x_0 分到第一类, 否则分到第二类.

▷ 按照**Bayes 因子原则**: 当 $f_1(\mathbf{x}_0) > f_2(\mathbf{x}_0)$ 时候将 $\mathbf{x}_0$ 分到第一类, 否则分到第二类.

两个多元正态总体场合:$X|G = g \sim N_p(\mu_g, \Sigma), g = 1, 2.$

↑Example
↓Example

此时,

$$f_1(\mathbf{x})/f_2(\mathbf{x}) = exp\Big[-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)\Big]$$
$$= exp\Big[(\mu_1-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\Big]$$
$$= exp\Big[(\mu_1-\mu_2)'\Sigma^{-1}(\mathbf{x}-\frac{\mu_1+\mu_2}{2})\Big]$$

因此第一个类的决策域为

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$
$$\Longleftrightarrow (\mu_1-\mu_2)'\Sigma^{-1}(\mathbf{x}-\frac{\mu_1+\mu_2}{2}) \geq \text{常数} = log\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

由此, 对新的观测点 $\mathbf{x}_0$, 可以得到分类法则. 实际中, μ_g, Σ 往往未知, 在观测到 (训练) 样本 (第一个总体中得到 n_1 个观测, 第二个总体中有 n_2 个观测) 后, 得到它们的估计

$$\hat{\mu}_g = \bar{\mathbf{x}}_g (g = 1, 2);$$
$$\hat{\Sigma} = S_{pool} = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_1 + (n_2 - 1)S_2]$$

记 $W(\mathbf{x}) = (\hat{\mu}_1 - \hat{\mu}_2)'\hat{\Sigma}^{-1}(\mathbf{x} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2})$, 从而经过训练后的分类器为

$$\delta^*(X) = \begin{cases} 1, & W(\mathbf{x}) \geq \text{常数} \\ 2, & W(\mathbf{x}) < \text{常数} \end{cases}$$

- 当 $\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} = 1$ 时候, 上述分类器中的 常数 $= 0$ (**Fisher 线性判别器**)

↑Example

两个多元正态总体场合:$X|G=g \sim N_p(\mu_g,\Sigma_g), g=1,2.$ 其中 $\Sigma_1 \neq \Sigma_2$.

↓Example

容易得到,

$$f_1(\mathbf{x})/f_2(\mathbf{x}) = exp\Big[-\frac{1}{2}\mathbf{x}'(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{x}+(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})\mathbf{x}-d\Big]$$

其中 $d=\frac{1}{2}log(\frac{|\Sigma_1|}{|\Sigma_2|})+\frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1-\mu_2'\Sigma_2^{-1}\mu_2)$ 因此第一个总体的分类域为

$$R_1: -\frac{1}{2}\mathbf{x}'(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{x}+(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})\mathbf{x}-d \geq log\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

当 μ_g,Σ_g 未知时候, 使用训练样本得到其估计 $\hat{\mu}_g=\bar{\mathbf{x}}_g, \hat{\Sigma}_g=S_g$ 代入, 得到训练后的分类器 (**二次判别法则**).

$$R_1: -\frac{1}{2}\mathbf{x}'(\hat{\Sigma}_1^{-1}-\hat{\Sigma}_2^{-1})\mathbf{x}+(\hat{\mu}_1'\hat{\Sigma}_1^{-1}-\hat{\mu}_2'\hat{\Sigma}_2^{-1})\mathbf{x}-\hat{d} \geq log\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

多个类场合

- 假设有 $G = 1, \ldots, k$ 个类, 观测 X 来自第 $G = g$ 个类时有概率函数, 即 $X|G = g \sim f_g(x)$, 先验概率为 $P(G = g) = p_g$, $g = 1, \ldots, k$.

- 决策函数为 $\delta(X) = i$, 当$X \in R_i$ 时 $(i = 1, \ldots, k.)$. 其中 $R_1, \ldots, R_k$ 为 Ω 的划分.

- 分类的损失函数为 $L(\delta(X), G) = c(i|g) > 0$, 当 $\delta(X) = i, G = g$ 时候, 其中 $c(i|i) = 0$.

- 平均损失为

$$ECM = EL(\delta(X), G) = \sum_{g=1}^{k} p_g \sum_{j=1}^{k} c(j|g) P(j|g)$$

- 最优决策为

$$\delta^*(X) = \arg\min_{R_1,\ldots,R_k} ECM$$

得到

$$\delta^*(X) = i, \text{当} X \in \{R_i^* : h_i(x) < h_j(x), j \neq i, j = 1, \ldots, k\}$$

其中

$$h_i(x) = \sum_{g \neq i, g=1}^{k} p_g f_g(x) c(i|g)$$

1.3 最大似然方法

- 最大似然分类器 (MLC) 选择使观测机会最大的类
- 假设每个类的条件概率函数 (密度或者分布律) 为

$$p_g(\mathbf{x}) = Pr(\mathbf{x}|G = g), g = 1, \dots, k$$

- 最大似然判别法则通过确定 X 的最大似然来预测观测 $\mathbf{x}$ 的类:

$$\delta(\mathbf{x}) = \arg\max_g p_g(\mathbf{x})$$

  - **QDA**(二次型法则) 若 $X|g \sim N_p(\mu_g, \Sigma_g)$, 则最大似然判别法则为

$$\delta(\mathbf{x}) = \arg\min_g \left[(\mathbf{x} - \mu_g)'\Sigma_g^{-1}(\mathbf{x} - \mu_g) + log|\Sigma_g| \right]$$

实际中, μ_g, Σ_g 使用训练样本估计 $\hat{\mu}_g = \bar{\mathbf{x}}_g, \hat{\Sigma}_g = S_g$. 从而判别函数为

$$\delta(\mathbf{x}) = \arg\min_g \left[(\mathbf{x} - \hat{\mu}_g)' \hat{\Sigma}_g^{-1} (\mathbf{x} - \hat{\mu}_g) + log|\hat{\Sigma}_g| \right]$$

– **LDA**(线性法则) 若 $X|g \sim N_p(\mu_g, \Sigma)$, 则最大似然判别法则为

$$\begin{aligned}
\delta(\mathbf{x}) &= \arg\min_g \left[(\mathbf{x} - \mu_g)' \Sigma^{-1} (\mathbf{x} - \mu_g) \right] \\
&= \arg\min_g \left[-2\mathbf{x}' \Sigma^{-1} \mu_g + \mu_g' \Sigma^{-1} \mu_g + \mathbf{x}' \Sigma^{-1} \mathbf{x} \right] \\
&= \arg\max_g \left[2\mathbf{x}' \Sigma^{-1} \mu_g - \mu_g' \Sigma^{-1} \mu_g \right]
\end{aligned}$$

此时 $\hat{\mu}_g = \bar{\mathbf{x}}_g$, $\hat{\Sigma} = S_{pool} = \sum_g (n_g - 1) S_g / (n - k)$, $n = n_1 + \cdots + n_k$.

– **DQDA**(对角二次法则) 若 $X|g \sim N_p(\mu_g, \Sigma_g)$, 其中 $\Sigma_g = diag(\sigma_{g1}^2, \ldots, \sigma_{gp}^2)$, 则最大似然判别法则为

$$\delta(\mathbf{x}) = \arg\min_g \sum_{i=1}^{p} \left[\frac{(x_i - \mu_{gi})^2}{\sigma_{gi}^2} + log(\sigma_{gi}^2) \right]$$

此时, $\hat{\mu}_g = \bar{\mathbf{x}}_g$, $\hat{\Sigma}_g = diag(S_g)$

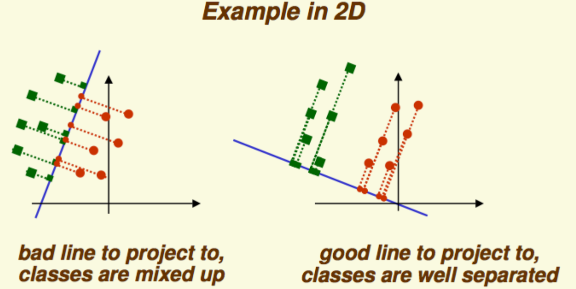
– **DLDA**(对角线性法则) 若 $X|g \sim N_p(\mu_g, \Sigma)$, 其中 $\Sigma_g = diag(\sigma_1^2, \ldots, \sigma_p^2)$, 则最大似然判别法则为

$$\delta(\mathbf{x}) = \arg\min_g \sum_{i=1}^{p} \frac{(x_i - \mu_{gi})^2}{\sigma_i^2}$$

此时, $\hat{\mu}_g = \bar{\mathbf{x}}_g$, $\hat{\Sigma} = diag(S_{pool})$

1.4 Fisher 线性判别

- Fisher 的思想：将 p 元数据投影到某个方向转化为一维数据，使得投影后的数据类和类尽可能分开.

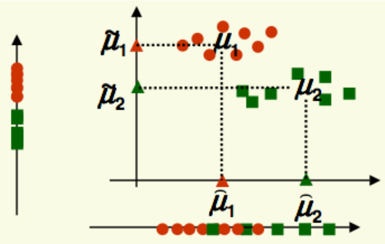


- 对两个类，假设训练样本为 $\mathbf{x}_1, \ldots, \mathbf{x}_n$，其中有 n_1 个来自第一个类 C_1，$n - n_1 = n_2$ 个来自第二个类 C_2.

- 将训练样本投影到向量 a 上得到 $a'\mathbf{x}_1, \ldots, a'\mathbf{x}_n$, 投影后两个类的中心为

$$\tilde{\mu}_1 = \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1} a'\mathbf{x}_i = a'\hat{\mu}_1, \quad \tilde{\mu}_2 = \frac{1}{n_2} \sum_{\mathbf{x}_i \in C_2} a'\mathbf{x}_i = a'\hat{\mu}_2$$

- 直接使用 $|\tilde{\mu}_1 - \tilde{\mu}_2|$ 作为选择最佳投影直线的准则可能会出问题:



原因在于没有考虑到两个类的方差没有考虑.

- 由于投影后为一维数据, 因此记

$$\tilde{s}_1^2 = \sum_{\mathbf{x}_i \in C_1} (a'\mathbf{x}_i - \tilde{\mu}_1)^2 = a'S_1 a$$

其中 (Scatter matrix)$S_1 = \sum_{\mathbf{x}_i \in C_1} (\mathbf{x}_i - \hat{\mu}_1)(\mathbf{x}_i - \hat{\mu}_1)'$. 类似定义 $\tilde{s}_2^2 = a'S_2 a$. 记

$$S_B = (\hat{\mu}_1 - \hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_2)', \quad S_W = S_1 + S_2$$

从而最大化目标函数

$$J(a) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{a'S_B a}{a'S_W a}$$

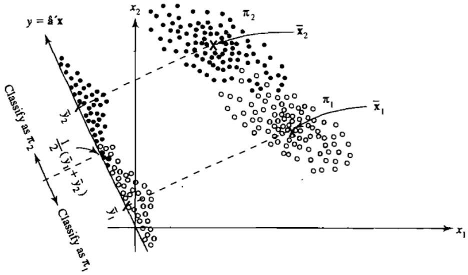
来选择最优的投影直线 a. 若 S_W 可逆, 则此问题的解为 (课本 P61 页)

$$\hat{a} = cS_W^{-1}(\hat{\mu}_1 - \hat{\mu}_2), \forall c \neq 0$$

- 常数 c 可以通过对 $\hat{a}$ 进行 "规范化" 以保证 $\hat{a}$ 的唯一性. 当 $\mathbf{x}$ 是标准化后的样本点时, 一般建议对 $\hat{a}$ 也进行规范化.

- 因此分类法则为

$$\delta^*(\mathbf{x}) = \begin{cases} 1, & \hat{a}'\mathbf{x} \geq (\tilde{\mu}_1 + \tilde{\mu}_2)/2 \\ 2, & \hat{a}'\mathbf{x} < (\tilde{\mu}_1 + \tilde{\mu}_2)/2 \end{cases}$$



- (**降维**) FLDA 将二元数据投影到直线 $\hat{a}$ 上后得到一维数据, 因此在一维空间上可以探索数据是否存在类, 异常点等等. 这一思想可以推广.

- (**方差分析想法**) 对两个类来说, 由一元方差分析知若两组均值差异显著, 则

$$
\begin{aligned}
F(a) &= \frac{SS_B/(2-1)}{SS_W/(n-2)} \\
&= \frac{n_1 n_2 (n-2)}{n_1+n_2} \frac{a' S_B a}{a' S_W a} \\
&= \frac{n_1 n_2 (n-2)}{n_1+n_2} J(a)
\end{aligned}
$$

应充分大. 其中

$$
\begin{aligned}
SS_B &= n_1(\tilde{\mu}_1 - a'\bar{\mathbf{x}})^2 + n_2(\tilde{\mu}_2 - a'\bar{\mathbf{x}})^2 \\
&= a'\Big[n_1(\hat{\mu}_1 - \bar{\mathbf{x}})(\hat{\mu}_1 - \bar{\mathbf{x}})' + n_2(\hat{\mu}_2 - \bar{\mathbf{x}})(\hat{\mu}_2 - \bar{\mathbf{x}})'\Big]a \\
&= a'\Big[\frac{n_1 n_2}{n_1 + n_2}(\hat{\mu}_1 - \hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_2)'\Big] \\
&:= \frac{n_1 n_2}{n_1 + n_2} a' S_B a
\end{aligned}
$$

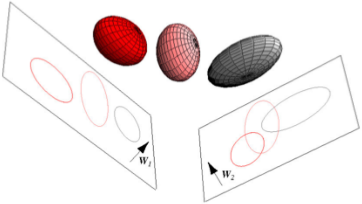
以及

$$
\begin{aligned}
SS_W &= \sum_{\mathbf{x}_i \in C_1}^{n} (a'\mathbf{x}_i - \tilde{\mu}_1)^2 + \sum_{\mathbf{x}_i \in C_2}^{n} (a'\mathbf{x}_i - \tilde{\mu}_2)^2 \\
&= a'\Big[\sum_{\mathbf{x}_i \in C_1} (\mathbf{x}_i - \hat{\mu}_1)(\mathbf{x}_i - \hat{\mu}_1)' + \sum_{\mathbf{x}_i \in C_2} (\mathbf{x}_i - \hat{\mu}_2)(\mathbf{x}_i - \hat{\mu}_2)'\Big]a \\
&:= a' S_W a
\end{aligned}
$$

因此, 最大化 $F(a)$ 等价于最大化 $J(a)$.

多个类的判别(Multiple Discriminant Analysis, MDA)

- Fisher 线性判别可以推广到多个类场合
- 当有 g 个类时, 可以将维数降低为 $1, \ldots, \min\{g-1, p\}$ 维
- 将每个样本点 $\mathbf{x}_i$ 投影到一个线性子空间 $\mathbf{y}_i = V'\mathbf{x}_i$, 其中 V 为投影矩阵



- 利用方差分析的想法, 假设有 g 个类, 第 i 个类训练样本为 $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$, $\bar{\mathbf{x}}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}$ $(i = 1, \ldots, g)$, $n = n_1 + \cdots +$

n_g, $\bar{\mathbf{x}} = \sum_{i,j} \mathbf{x}_{ij}/n$, 若假定 g 个类的总体协方差矩阵相同. 将所有训练样本投影到直线方向 a, 记投影后的各类均值为 $\tilde{\mu}_i = a'\bar{\mathbf{x}}_{i\cdot}$, 则 g 个类的投影均值差异明显时候,

$$F(a) = \frac{SS_B/(g-1)}{SS_W/(n-g)}$$

应尽可能的大. 其中

$$\begin{aligned}
SS_B &= \sum_{i=1}^{g} n_i(\tilde{\mu}_i - a'\bar{\mathbf{x}})^2 \\
&= a'\Big[\sum_{i=1}^{g}(\mathbf{x}_{i\cdot} - \bar{\mathbf{x}})(\mathbf{x}_{i\cdot} - \bar{\mathbf{x}})'\Big]a := a'Ba \\
SS_W &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(a'\mathbf{x}_{ij} - \tilde{\mu}_i)^2 \\
&= a'\Big[\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \mathbf{x}_{i\cdot})(\mathbf{x}_{ij} - \mathbf{x}_{i\cdot})'\Big]a := a'Wa
\end{aligned}$$

从而当 W 可逆时候,

$$\hat{a}_1 = \arg \sup_{\|a\|=1} \frac{a'Ba}{a'Wa} = \hat{e}_1$$

其中 $\hat{e}_1$ 为矩阵 $W^{-1}B$ 的最大特征根对应的特征向量.

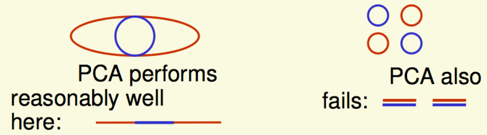
- 记 $W^{-1}B$ 的所有非零特征根分别为 $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_s > 0$, $\hat{e}_1, \ldots, \hat{e}_s$ 为对应的特征向量, $s \leq \min\{g-1, p\}$.

- $\hat{a}_1'\mathbf{x}$ 称为样本第一判别函数, 有时候一个判别函数不能很好区分各个类, 可取 $\hat{a}_2 = \hat{e}_2$, $\hat{a}_2'\mathbf{x}$ 作为样本第二判别函数, 以此类推, 最多有 $\min\{g-1, p\}$ 个样本判别函数.

- 当有 r 个判决函数时候, 这相当于将原始 p 元数据投影到 r 维空间, 因此可以使用基于上节的线性判别来进行分类. 比如

$$\delta^*(\mathbf{x}) = k \text{ 如果 } \sum_{j=1}^{r} [\hat{a}_i'(\mathbf{x} - \bar{\mathbf{x}}_k)]^2 \leq \sum_{j=1}^{r} [\hat{a}_j'(\mathbf{x} - \bar{\mathbf{x}}_i)]^2 \text{对所有} i \neq k$$

- Fisher 线性判别假设了各类同 (协) 方差
- Fisher 线性判别也可能会失效:
  - 所有类的中心 (均值) 相同. 此时 $F(a)$ 总是 (近似) 为零.



  - 当所有类向任何方向投影时候都有很大重叠时候, $F(a)$ 总是很大. FLDA 和 PCA 均会失效.

- FLDA 的变种: 非参数 LDA, 正交 LDA, 广义 LDA 等

FLDA 和回归分析

- 给定两个类的数据, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n, \mathbf{x}_i \in R^p, y_i \in \{+1, -1\}$.
- 考虑以 $\mathbf{y} = (y_1, \ldots, y_n)'$ 为响应变量, $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ 为自变量的线性回归问题

$$\min L(\beta, \beta_0) = \frac{1}{2}\|\mathbf{y} - \mathbf{x}'\beta - \beta_0\|^2$$

- 若假定 $\{\mathbf{x}_i\}$ 和 $\{y_i\}$ 均中心化, 即 $\sum_{i=1}^n \mathbf{x}_i = 0, \sum_{i=1}^n y_i = 0$. 因此

$$y_i \in \{-2n_2/n, 2n_1/n\}$$

此时, $\beta_0 = 0$, 因此最小化目标函数

$$L(\beta) = \frac{1}{2}\|\mathbf{y} - \mathbf{x}'\beta\|^2$$

- 其解为 $\hat{\beta} = (\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}\mathbf{y}$(假定 $\mathbf{x}$ 满秩, 否则使用广义逆代替). 注意到 $\mathbf{x}\mathbf{x}' = S_W$, $\mathbf{x}\mathbf{y} = \frac{2n_1n_2}{n}(\hat{\mu}_1 - \hat{\mu}_2)$, 因此

$$\hat{\beta} = \frac{2n_1n_2}{n} S_W^{-1}(\hat{\mu}_1 - \hat{\mu}_2) = \frac{2n_1n_2}{n}\hat{a}$$

从而回归函数为 $\hat{\mathbf{y}} = \hat{\beta}'\mathbf{x} \propto \hat{a}'\mathbf{x}$, 即 Fisher 线性判别分析等价于回归模型.

- 对多个类的场合, $y = 1, 2, \ldots, g.(g > 2)$, 使用线性模型时候必须对类别变量 y 进行编码, 比如使用 $y_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ 表示第 i 个类, 可以证明在这种编码下回归模型和 LDA 不等价, 但有关系 (Hastie et al., 2001)

- 那么是否存在使得线性回归模型和 LDA 等价的类别变量编码? 可以证明, 存在这种编码方式, 使得线性模型和 LDA 是等价的 (方开泰,1982).

1.5 支持向量机

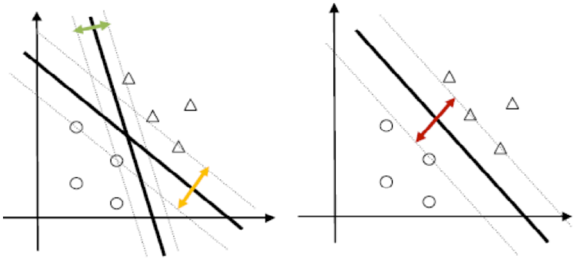
- 设有两个类, 训练集为 $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$, 其中 $\mathbf{x}_i \in R^p$, y_i**取值 $+1$ 或 -1**.
- 目的是找到一个分类器 $f : R^p \to R$, 使得分类法则为

$$\delta(\mathbf{x}) = sgn(f(\mathbf{x}))$$

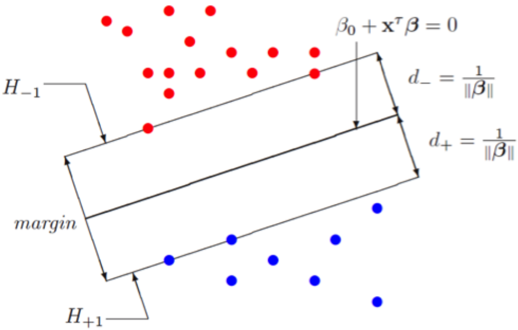
- 假设两个类的训练集可以通过一个超平面分开

$$\{\mathbf{x} : f(\mathbf{x}) = \beta_0 + \mathbf{x}'\beta = 0\}$$

- 若此超平面可以将两个类分开而没有错误, 则称其为**可分超平面**. 显然这样的超平面可能有无穷多, 那么哪个最好?



- 记可分平面分别到两类点的最近距离记为 d_- 和 d_+.
- 可分超平面的间隔 (**margin**) 定义为 $d = d_- + d_+$.
- SVM 通过最大化可分超平面到两类样本点的最近距离和间隔 (margin) 来得到最优可分超平面



- 两个类线性可分当且仅当存在 β_0, β 使得

$$\beta_0 + \mathbf{x}_i'\beta \geq +1, \text{当} y_i = +1$$
$$\beta_0 + \mathbf{x}_i'\beta \leq -1, \text{当} y_i = -1$$

使得等式成立的样本点称为**支持向量**(support vector):

$$H_{+1} : (\beta_0 - 1) + \mathbf{x}'\beta = 0$$

$$H_{-1} : (\beta_0 + 1) + \mathbf{x}'\beta = 0$$

- 因此求解最优超平面转化为下述优化问题

$$\begin{aligned} &\text{最小化} \quad \frac{1}{2}\|\beta\|^2 \\ &st. \quad y_i(\beta_0 + \mathbf{x}'\beta) \geq 1, i = 1, \ldots, n. \end{aligned}$$

- 由 Lagrange 乘子法上述优化问题等价于

$$\begin{aligned} &\text{最大化} \quad \mathbf{1}_n'\alpha - \frac{1}{2}\alpha' H\alpha \\ &st. \quad \alpha \geq 0, \alpha'\mathbf{y} = 0. \end{aligned}$$

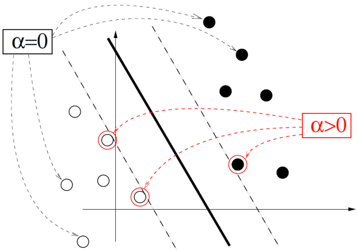
其中 $\mathbf{y} = (y_1, \ldots, y_n)'$, $H = (H_{ij})$ 为 n 阶方阵且 $H_{ij} = y_i y_j(\mathbf{x}_i'\mathbf{x}_j)$.

- 若记 $\hat{\alpha}$ 为上述优化问题的解, 则

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i = \sum_{i \in sv} \hat{\alpha}_i y_i \mathbf{x}_i$$

其中 sv 表示支持向量集.

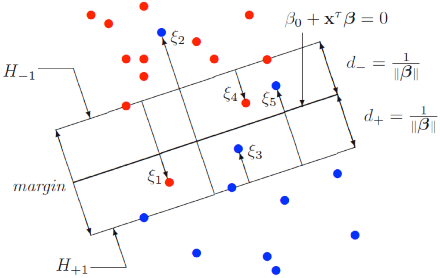
$$\hat{\beta}_0 = \frac{1}{|sv|} \sum_{i \in sv} \left(\frac{1 - y_i \mathbf{x}_i' \hat{\beta}}{y_i} \right)$$



- 从而决策函数为

$$\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \mathbf{x}'\hat{\beta} = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i y_i < \mathbf{x}_i, \mathbf{x} >$$

- 此时 $\|\hat{\beta}\|^2 = \sum_{i \in sv} \hat{\alpha}_i$, 因此最优可分超平面有间隔 $2/\|\hat{\beta}\|^2$
- 如果两个类不是线性可分的，或者两个类之间不存在明确的线性或非线性可分性



- 引入松弛变量 (slack variable): $\xi = (\xi_1, \ldots, \xi_n)' \geq 0$

- 最优超平面同时控制间隔 (margin) 和松弛变量:

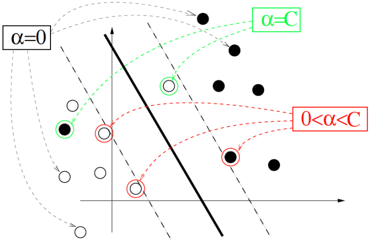
$$\text{最小化} \quad \frac{1}{2}\|\beta\|^2 + C\|\xi\|_1$$

$$st.\ \xi_i \geq 0, \quad y_i(\beta_0 + \mathbf{x}_i'\beta) \geq 1 - \xi_i, i = 1, \ldots, n.$$

这等价于

$$\text{最大化} \quad \mathbf{1}_n'\alpha - \frac{1}{2}\alpha' H\alpha$$

$$st. \quad 0 \leq \alpha \leq C\mathbf{1}_n, \alpha'\mathbf{y} = 0.$$



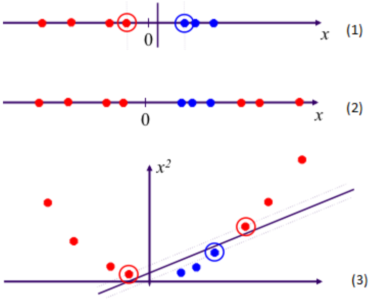
大的 C 值: 使松弛变量趋于零, 因此错误较少; 小的 C 值, 使得间隔较大; 中等的 C 值: 间隔与错误之间的权衡

非线性 SVM

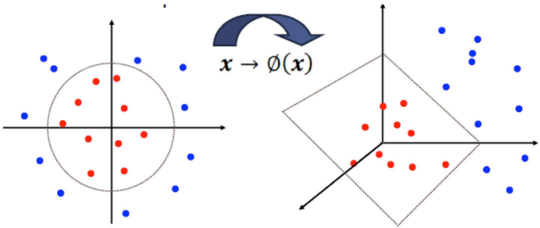
- 当数据为线性可分时候，可以很好的分开两类. 图 (1).

- 但是如果数据不是线性可分的, 图 (2).

- 将数据映射到 2 维空间里，则变成线性可分, 图 (3).



- 因此, 在高维的特征空间建立线性支持向量机:



- 将训练点 $\mathbf{x}_i$ 变换到高维空间 $\mathcal{H}$(feature space), 记变换为 $\Phi(\mathbf{x}_i) = (\phi_1(\mathbf{x}_i), \ldots, \phi_{N_{\mathcal{H}}})' \in \mathcal{H}$, $N_{\mathcal{H}}$ 为 $\mathcal{H}$ 的维数.

- 在特征空间 $\mathcal{H}$ 内考虑线性可分超平面:

$$f(\mathbf{x}) = \Phi(\mathbf{x})'\beta + \beta_0 = 0$$

- 类似前面的所述知决策函数 (decision function) 为

$$\hat{f}(\mathbf{x}) = \Phi(\mathbf{x})'\hat{\beta} + \hat{\beta}_0 = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i y_i < \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) >$$

- (Kernel Trick) 使用核函数 $K(\mathbf{x}_i, \mathbf{x}_j) =< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) >$ 来代替高维特征空间下的内积. 使用核函数的优点是只需指定核函数 K, 而不必指定 Φ.

- 部分常用核函数

Kernel	$K(x,y)$	In R
Linear	$< x, y >$	vanilladot
Gaussian RBF	$e^{-\sigma\|x-y\|^2}$	rbfdot
Laplacian	$e^{-\sigma\|x-y\|}$	laplacedot
Polynomial of degree d	$(< x, y > +c)^d$	polydot

see ?dots in kernlab

C-SVM 给定训练集 $\mathcal{L} = \{(\Phi(\mathbf{x}_i), y_i), i = 1, \ldots, n\}$

- 决策函数 $f(\mathbf{x}) = \Phi(\mathbf{x})'\beta + \beta_0$
- 引入松弛变量后最大化间隔, 即

$$\text{最小化} \quad \frac{1}{2}\|\beta\|^2 + C\|\xi\|_1$$

$$st.\ \xi_i \geq 0, \quad y_i(\beta_0 + \Phi(\mathbf{x}_i)'\beta) \geq 1 - \xi_i, i = 1, \ldots, n.$$

- 利用对偶问题等价于

$$\text{最大化} \quad \mathbf{1}_n'\alpha - \frac{1}{2}\alpha' H\alpha$$

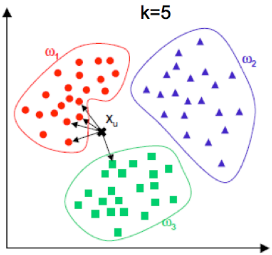
$$st. \quad 0 \leq \alpha \leq C\mathbf{1}_n, \alpha'\mathbf{y} = 0.$$

其中 $H = (H_{ij}), H_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. 类似前面讨论得决策函数为

$$\hat{f}(\mathbf{x}) = \hat{\beta} + \sum_{i \in sv} \hat{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x})$$

1.6 k-NN 方法

- 基于测量两个观测之间的距离 (例如欧式距离或者相似性度量)
- $k-NN$ 法则 (Fix & Hodges 1951) 分类一个观测 $\mathbf{x}$ 的方法如下:
  - 在训练集中找到 k 个与 $\mathbf{x}$ 最近的训练点
  - 采取最多投票制分类 $\mathbf{x}$: 即将 $\mathbf{x}$ 分类到包含其 k 个最近邻训练点最多的类里
  - 最优的 k 采用交叉验证方法选择



1.7 分类效果的评价

- 对两个类的分类决策效果进行评价时, 需要计算 $P(1|2)$ 和 $P(2|1)$. 比如总的错误分类概率为

$$TPM = p_1 P(2|1) + p_2 P(1|2) = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$$

- 当总体分布密度 $f_1(\mathbf{x})$ 和 $f_2(\mathbf{x})$ 未知时候, 则不能直接计算评价样本分类准则的性能.

- 下面我们讨论几种估计实际错误率 (Actual Error Rate, AER) 的方法

  – 表现失误率 (Apparent error rate, APER) 方法

  – 提留法 (holdout procedure)

  – 数据分割方法 (Data splitting)

– 交叉验证 (Cross-validation)

– Bootstrap 方法

- Actual error rate (AER)实际失误率：样本分类函数的效果可以使用实际失误率来衡量：比如在两个类别场合，

$$AER = p_1 \int_{\hat{R}_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x}) d\mathbf{x}$$

一般来说，AER 不能直接计算，因为 f_1, f_2 未知.

- 模糊矩阵 (confusion matrix)

		Predicted membership π_1	Predicted membership π_2	
Actual membership	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

– n_{1M}, n_{2M} 表示各类中被错误分类的个数

- n_{1c}, n_{2c} 表示各类中被正确分类的个数

- Apparent error rate (APER) 表现失误率: 由训练样本训练出分类器后, 对训练样本中的每个观测点进行分类, 其中错误比例

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

  - 样本被重复使用, 因此 $APER$ 为有偏估计, 它倾向于低估 AER, 样本量增加时候偏差会减少.
  - 需要样本量很大

- Lachenbrush's holdout“提留” 方法: 使用类 1 中 $n_1 - 1$ 个样本点和类 2 的 n_2 个样本点训练分类器, 然后对类 1 中提留出的一个样本点进行检验, 重复直至类 1 中所有点都被当为提留点, 记 n_{1M}^H 为其中错误分类的数目; 对类 2 进行类似的过程, 记 n_{2M}^H 表示其中被错误分类的数目, 则则期望的 AER 的一个

(近似无偏) 估计为

$$\hat{E}(AER) = \frac{n_{1M}^H + n_{2M}^H}{n_1 + n_2}$$

– 这个估计较 APER 要好

– 对中等规模样本量也成立

- Data Splitting 当训练集较大时候, 随机将训练集分割为 training 和 validation 两个集合, 大约 %25 – %35 的样本应该被划分为 validation 集.

  – 基于 trainning 集训练分类器, 对 validation 集使用训练好的分类器进行分类, 据此估计错误分类概率

  – 由于部分随机训练样本被用来估计错误分类概率, 因此估计偏差较小, 但方差较大. 实际中不使用

- **交叉验证** 类似于数据分割方法, 只不过将数据随机分割为 g 个组 (一般各组数据相同)
  - 使用其中 $g-1$ 个组训练分类器, 使用剩余的组估计错误分类概率
  - 重复上述过程 g 次, 每次使用一个不同的组样本估计错误分类概率
  - 总的错误分类概率估计通过平均 g 个估计来得到
  - 比数据分割方法有较小的方差, 估计是近似无偏的, 但当变量个数增加时候, 偏差会增加.
  - 当 $g = n_1 + n_2$ 时候, 称为留一 (**leave-one-out**) 交叉验证方法.
- bootstrap 通过 Bootstrap 方法得到 "新" 训练集.
  - 使用所有训练样本估计错误分类概率, 即得到 $\hat{P}(1|2)$ 和 $\hat{P}(2|1)$

- 从原始训练集中有放回的随机抽取得到同样大小的训练集 (n_1 和 n_2), 估计错误分类概率
- 重复上过程 B 次, 得到 $\hat{P}_i(1|2)$ 和 $\hat{P}_i(2|1), i = 1, \ldots, B$
- 使用 B 个估计来估计估计的偏差

$$\widehat{bias}(1|2) = \frac{1}{B}\sum \hat{P}_i(1|2) - \hat{P}(1|2)$$
$$\widehat{bias}(2|1) = \frac{1}{B}\sum \hat{P}_i(2|1) - \hat{P}(2|1)$$

- 从而 $P(1|2)$ 和 $P(2|1)$ 的 bootstrap-corrected 估计分别为

$$\hat{P}^c(1|2) = \hat{P}(1|2) - \widehat{bias}(1|2)$$
$$\hat{P}^c(2|1) = \hat{P}(2|1) - \widehat{bias}(2|1)$$