特征选择

~~问题~~：有一串 N 维特征向量 $X_1, X_2, \cdots X_p$，其中，$X_i = \{ x_{i1} \quad x_{i2} \cdots \quad x_{iN} \}$，每一个 X_i 属于 C_1 或 C_2。用 C_i 表示 X_i 的类别。

问题：要从 N 个维度中选出 M 维，($M \ll N$) 使之能尽量多的保留 C_1 和 C_2 信息，达到正确分类。

一个直观做法：遍历所有 C_N^M 个组合，每个组合构建分类器，选识别率高的那个（NP-hard 问题！）

既然全部遍历不实际，那么还有什么方法呢？

近似方法：

① 递增法：先选一个特征 x_1，使识别率最大。在 x_1 基础上，再选 x_2，构成向量 $\{x_1, x_2\}$，使识别率最高。

重复，直到识别率下降为止。

② 递减法。

③ 合成方法。

问题：每次特征选择都要构建识别器，这是一个耗时耗力的过程。

改进：用可分性判别函数。最常用的是互信息。

~~因~~ $I(X;C) = \cancel{\int \frac{p(x,c)}{p(x)p(c)} dx} \iint p(x,c)\log\frac{p(x,c)}{p(x)p(c)}dxdc$

这个值越大，说明X与C的交叉越多，可分性就越好。

~~$I(S_m, C$~~

设 S_m 为 m 个特征的组合，定义

$$I(S_m, C) = \iint p(S_m, c)\log\frac{p(S_m, c)}{p(S_m)p(c)}dS_m dc$$

$$= \iint\cdots\int p(x_1 x_2 x_3 \cdots x_n, c)\log\frac{p(x_1 x_2 \cdots x_m, c)}{p(x_1 x_2 \cdots x_n)p(c)}$$

假设：特征之间独立

~~定义：$D(S,C) = I(\{x_1 x_2 \cdots x_m\}; C)$~~

~~$\max D(S,C)$~~

定义： $D(S,C) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c)$

因此，我们要最大化 $D(S,C)$

但另一方面，两个特征间存在相关性，在特征选择中叫做 Correlation redundancy（相关性冗余）

可以用 $I(x_1, x_2)$ 来表示特征间的相关性冗余。

$$I(x_1, x_2) = \int p(x_1 x_2)\log\frac{p(x_1, x_2)}{p(x_1)p(x_2)}dx_1 dx_2.$$

特征选择

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_i, x_j) \right]$$

依次增加特征直到这个值变小为止。

概率密度估计：

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} f(x - x^{(i)}; h)$$

$$f(z, h) = \exp\left(\frac{z^T \Sigma^{-1} z}{2h^2} \right) \Big/ (2\pi)^{\frac{d}{2}} h^d |\Sigma|^{\frac{1}{2}}$$

提升算法

假定一个二类分类问题

数据集 $T=\{(x_1,y_1),(x_2,y_2)\cdots(x_N,y_N)\}$

$y_i=\{-1,1\}$

AdaBoost 算法

输入 $T=\{(x_1,y_1),(x_2,y_2),\cdots(x_N,y_N)\}$

输出：最终分类器 $G(x)$

① 初始化权值分布

$$D_1=(w_{11}\ w_{12}\cdots w_{1i},\cdots w_{1N}),\quad w_{1i}=\frac{1}{N}$$

② 对 $m=1,2,\cdots,M$

ⓐ 使用具有权值分布 D_m 的训练数据集 D_m 学习，得到基本分类器

$G_m(x)$。

ⓑ 计算 $G_m(x)$ 在训练集上误差

$$e_m=P(G_m(x_i)\neq y_i)=\sum_{i=1}^{N} w_{mi} I(G_m(x_i)\neq y_i)$$

ⓒ 计算系数

$$a_m=\frac{1}{2}\log\frac{1-e_m}{e_m}$$

ⓓ 更新训练集权值分布

$$D_{m+1}=(w_{m+1,1}\ w_{m+1,2}\cdots w_{m+1,N})$$

$$w_{m+1,i}=\frac{w_{mi}}{Z_m}\exp(-a_m y_i G_m(x))$$

$$Z_m=\sum_{i=1}^{N} w_{mi}\exp\{-a_m y_i G_m(x_i)\}$$

它使 D_{m+1} 成为一个概率分布

③ 构建基本分类面

$$f(x)=\sum_{m=1}^{M} a_m G_m(x)$$

得到最终分类器

$$G(x)=\text{sign}(f(x))=\text{sign}\left[\sum_{m=1}^{M} a_m G_m(x)\right]$$

定理：随着 M 增加，AdaBoost 得到的分类器在训练集上错误率会减小。

定理：AdaBoost 算法的训练误差界为

$$\frac{1}{N}\sum_{i=1}^{N} I(G(x_i)\neq y_i)\leq\frac{1}{N}\sum_{i=1}^{N}\exp(-y_i f(x_i))$$

$$=\prod_{m} Z_m$$

$$Z_m=\sum_{i=1}^{N} w_{mi}\exp(-a_m y_i G_m(x_i))$$

$$\frac{1}{N}\sum_{i=1}^{N}\exp(-y_i f(x_i))$$

$$=\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\sum_{m=1}^{M} a_m G_m(x_i)\right)$$

$$=\sum_{i=1}^{N} w_{1i}\prod_{m=1}^{M}\exp(-a_m y_i G_m(x_i))$$

$$=\prod_{i=1}^{M} Z_m$$

定理：二分类问题 AdaBoost 训练误差界：

$$\prod_{m=1}^{M} Z_m = \prod_{m=1}^{M} 2\sqrt{e_m(1-e_m)} = \prod_{m=1}^{M} \sqrt{(1-4\gamma_m^2)}$$

$$\leq \exp\left\{-2\sum_{m=1}^{M} \gamma_m^2\right\}$$

$$\gamma_m = \frac{1}{2} - e_m$$

证明：

$$Z_m = \sum_{i=1}^{N} w_{mi} \exp(-\alpha_m y_i G_m(x_i))$$

$$= \sum_{\substack{i=1 \\ y_i = G_m(x_i)}}^{N} w_{mi} e^{-\alpha_m} + \sum_{y \neq G_m(x_i)} w_{mi} e^{\alpha_m}$$

$$= (1-e_m)e^{-\alpha_m} + e_m e^{\alpha_m}$$

将 $\alpha_m = \frac{1}{2}\log\frac{1-e_m}{e_m}$ 代入

$$= 2\sqrt{e_m(1-e_m)} = \sqrt{1-4\gamma_m^2}$$

$$\leq \exp\left(-2\sum_{m=1}^{M} \gamma_m^2\right)$$