第八章 特征选择

苏智勇
可视计算研究组
南京理工大学
suzhiyong@njust.edu.cn
https://zhiyongsu.github.io

主要内容

8.1 引言

8.2 用于分类的特征评价标准

8.3 特征选择的最优算法

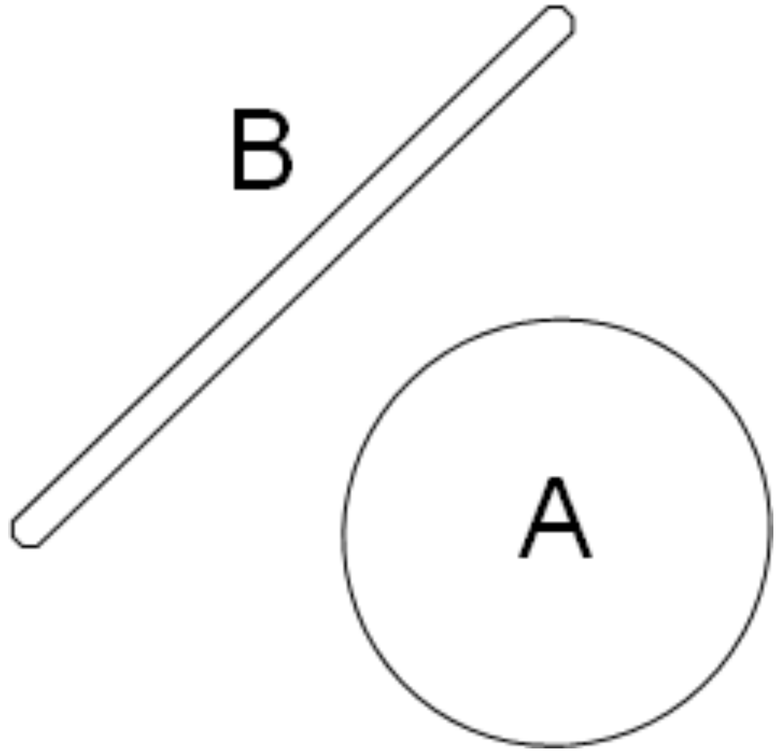
8.4 特征选择的次优算法

8.5 遗传算法

8.6 包裹法

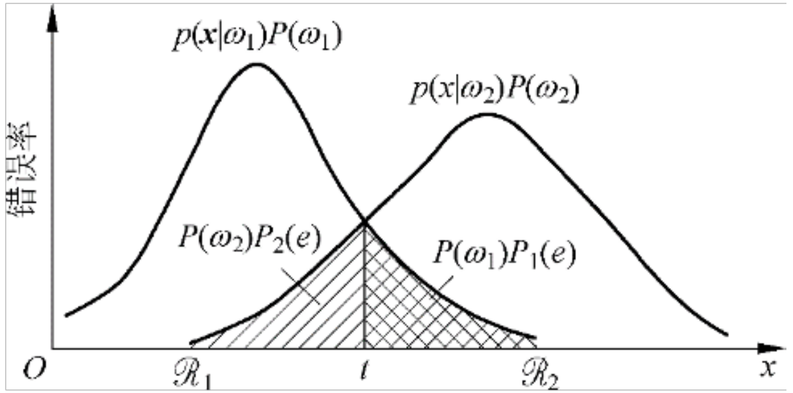
8.1 引言

- **什么是特征选择问题**
  - 用计算方法从一组给定的特征中选择一部分特征进行分类
    - 例：周长、面积、两个互相垂直的内径比
- **为什么要进行特征选择**
  - 计算上的考虑
    - 时间、空间复杂度
  - 性能上的考虑
    - 泛化性：人脸识别（肤色）



8.1 引言

- 特征选择的任务是找出一组分类最好的特征 → 评价准则
  - 从D各特征中选择使准则函数最优的d个特征（$d < D$）
- **概念**
  - 数学上定义的用以衡量特征对分类的效果的准则
    - 错误率：类条件概率密度函数

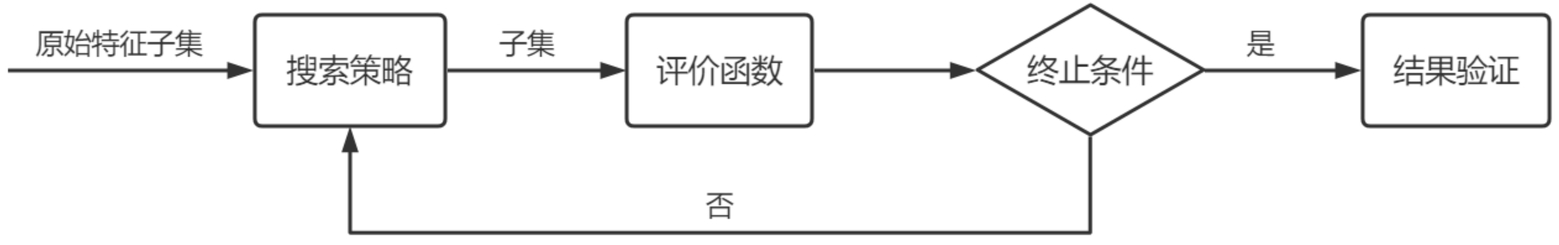


$$\min P(\text{error}) = \min \int P(\text{error} \mid x)P(x)dx \Rightarrow \min P(\text{error} \mid x) \quad \forall x$$

  - 实际问题中需要根据实际情况人为确定

8.1 引言

- **过程**
  - 搜索策略
  - 评价函数
  - 终止条件
  - 结果验证



- 特征选择
  - 搜索策略
    - 全局最优搜索策略
      - 穷举法
      - 分支界定法
    - 启发式搜索策略
      - 前向序列选择方法
      - 后向序列删除方法
      - 增 l 减 r 方法
      - 序列浮动搜索方法
    - 随机搜索策略
      - 遗传算法
      - 模拟退火算法
  - 评价函数
    - 距离度量
      - 几何距离度量
      - 概率距离度量
    - 一致性度量：不一致率
    - 依赖性度量
      - t − test
      - F − measure
      - 皮尔逊相关系数
      - Fisher 分数
    - 信息度量
      - 信息熵
      - 信息增益
      - 互信息
    - 分类误差度量：调用分类器
  - 终止条件：常用四种终止条件
  - 结果验证：不同的测试集和分类器来验证

8.2 用于分类的特征评价标准

- **可分性准则** J_{ij}：衡量第i类和第j类的可分程度
  - 与错误率有单调关系：J_{ij} 大⇒ P_e 小
  - 度量特性：$J_{ij} > 0 \ (i \neq j)$ ， $J_{ij} = 0 \ (i = j)$ ， $J_{ii} = 0,\ J_{ij} = J_{ji}$
  - 对独立的特征有可加性：$J_{ij}\left(x_1, x_1, \cdots, x_d\right) = \sum_{k=1}^{d} J_{ij}\left(x_k\right)$
  - 增加特征时判据不减小：$J_{ij}(x_1, x_2, \cdots, x_d) \leqslant J_{ij}(x_1, x_2, \cdots, x_d, x_{d+1})$

8.2.1 基于类内类间距离的可分性判据

- **类间平均距离**：各类特征向量之间的平均距离

$$J_d(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \delta(\boldsymbol{x}_k^{(i)}, \boldsymbol{x}_l^{(j)})$$

  - c为类别数；d为特征向量维度
  - $\boldsymbol{x}_k^{(i)} \in \omega_i,\ k = 1,\cdots,n_i;\ \boldsymbol{x}_l^{(j)} \in \omega_j,\ l = 1,\cdots,n_j$
  - P_i，P_j为相应类别的先验概率，n_i，n_i分别为ω_i和ω_j类中的样本数
  - $\delta(\boldsymbol{x}_k^{(i)}, \boldsymbol{x}_l^{(j)})$：$\boldsymbol{x}_k^{(i)}$和$\boldsymbol{x}_l^{(j)}$的距离度量，通常采用**欧式距离**：

$$\delta\left(\boldsymbol{x}_k^{(i)}, \boldsymbol{x}_l^{(j)}\right) = \left(\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_l^{(j)}\right)^T \left(\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_l^{(j)}\right)$$

8.2.1 基于类内类间距离的可分性判据

- **类间平均距离**

$$J_d(\boldsymbol{x}) = \mathrm{tr}(\widetilde{S}_w + \widetilde{S}_b)$$

_ 类均值向量：$\boldsymbol{m}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_k^{(i)}$；总均值向量：$\boldsymbol{m} = \sum_{i=1}^{c} P_i \boldsymbol{m}_i$

_ 类间离散度矩阵S_b的估计：$\widetilde{S}_b = \sum_{i=1}^{c} P_i\left(\boldsymbol{m}_i - \boldsymbol{m}\right)\left(\boldsymbol{m}_i - \boldsymbol{m}\right)^T$

_ 类内离散度矩阵S_w的估计：$\widetilde{S}_w = \sum_{i=1}^{c} P_i \frac{1}{n_i}\sum_{k=1}^{n_i}\left(\boldsymbol{x_k}^{(i)} - \boldsymbol{m}_i\right)\left(\boldsymbol{x_k}^{(i)} - \boldsymbol{m}_i\right)^T$

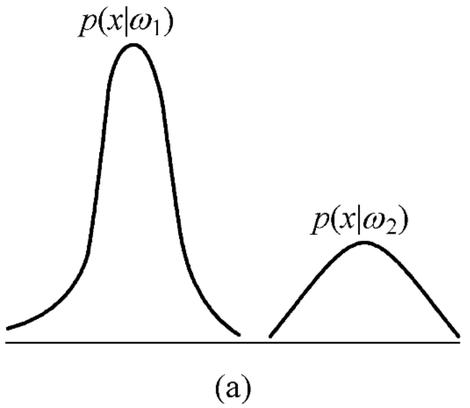
8.2.1 基于类内类间距离的可分性判据

- 常用的基于类内类间距离的可分性判据
  - $J_1 = \mathrm{tr}(S_w + S_b)$
  - $J_2 = \mathrm{tr}(S_w^{-1} S_b)$
  - $J_3 = \ln \dfrac{|S_b|}{|S_w|}$
  - $J_4 = \dfrac{\mathrm{tr}(S_b)}{\mathrm{tr}(S_w)}$
  - $J_5 = \dfrac{|S_b - S_w|}{|S_w|}$

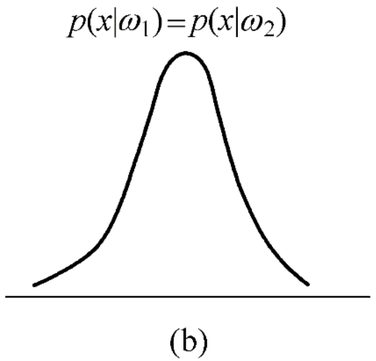
8.2.2 基于概率分布的可分性判据

- 考察两类分布密度间的交叠程度
- 定义：两个密度函数间距离：

$$J_P(\cdot) = \int g\left[p(\boldsymbol{x}\,|\,\omega_1), p(\boldsymbol{x}\,|\,\omega_2), P_1, P_2\right] d\boldsymbol{x}$$



(a)



(b)

必须满足三个条件：

- $J_P \geq 0$
- 若$p(\boldsymbol{x}\,|\,\omega_1)p(\boldsymbol{x}\,|\,\omega_2) = 0$，$\forall \boldsymbol{x}$，则$J_P = \mathrm{Max}$ （完全不重叠，a图）
- 若$p(\boldsymbol{x}\,|\,\omega_1) = p(\boldsymbol{x}\,|\,\omega_2)$，$\forall \boldsymbol{x}$，则$J_P$=0 （完全重叠，b图）

8.2.2 基于概率分布的可分性判据

- **常用的概率距离度量**
  - **Bhattacharyya距离**

$$J_B = -\ln \int \left[p(\boldsymbol{x}\,|\,\omega_1)p(\boldsymbol{x}\,|\,\omega_2)\right]^{\frac{1}{2}} d\boldsymbol{x}$$

两类完全重合时，$J_B = 0$；两类完全不交叠时，$J_B = \infty$

  - **Chernoff界**

$$J_C = -\ln \int p^s\left(\boldsymbol{x}\,|\,\omega_1\right)p^{1-s}\left(\boldsymbol{x}\,|\,\omega_2\right)d\boldsymbol{x}$$

当$s = 0.5$时，Chernoff界限与Bhattacharyya距离相同

8.2.2 基于概率分布的可分性判据

- **常用的概率距离度量**
  - **散度**

$$J_D = \int_{\boldsymbol{x}} \left[p(\boldsymbol{x} \mid \omega_1) - p(\boldsymbol{x} \mid \omega_2) \right] \ln \frac{p(\boldsymbol{x} \mid \omega_1)}{p(\boldsymbol{x} \mid \omega_2)} d\boldsymbol{x}$$

在两类样本都是正态分布情况下，散度为：

$$J_D = \frac{1}{2} \mathrm{tr} \left[\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1 - 2\boldsymbol{I} \right] + \frac{1}{2} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \right)^T (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

当两类协方差矩阵相等时，散度和Bhattacharyya距离有如下关系：

$$J_D = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \right)^T \Sigma^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \right) = 8 J_B$$

也等于两类均值之间的Mahalanobis距离

8.2.3 基于熵的可分性判据

- **熵：**事件的不确定性度量
  - A事件的不确定性越大（熵大），则对A的事件的观察所提供的信息量大。
- **思路**
  - 把各类ω_i看作一系列事件
  - 把后验概率$P(\omega_i \mid \boldsymbol{x})$看作特征$\boldsymbol{x}$上出现$\omega_i$的频率
  - 如从$\boldsymbol{x}$能确定ω_i，则对ω_i的观察不提供信息量，熵为0。（特征$\boldsymbol{x}$有利于分类）
  - 如$\boldsymbol{x}$完全不能确定ω_i，则对ω_i的观察信息量大，熵最大。（特征$\boldsymbol{x}$无助于分类）

8.2.3 基于熵的可分性判据

- **常用熵度量**

  _ Shannon熵：$H = -\sum_{i=1}^{c} P(\omega_i \mid \boldsymbol{x}) \log_2 P(\omega_i \mid \boldsymbol{x})$

  _ 平方熵：$H = 2\left[1 - \sum_{i=1}^{t} P^2(\omega_i \mid \boldsymbol{x})\right]$

  _ 基于熵的可分性判据：$J_{\mathrm{E}} = \int H(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$，$J_{\mathrm{E}}$越小可分性越好

8.2.4 利用统计检验作为可分性判据

- **假设检验：**检验某一变量在两类样本间是否存在显著差异
- **参数化检验方法：t-检验（t-test）**
  - 用一个统计量来反映两类样本间的差别
  - 基本假设：两类样本均服从正态分布：$x_i \sim N(\mu_x, \sigma^2)$，$y_i \sim N(\mu_y, \sigma^2)$。总体样本方差为：

$$s_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{m+n-2}$$

其中，S_x^2，S_y^2为两类样本各自的估计方差，并记两类样本的均值为$\bar{x}$和$\bar{y}$，则t-检验的统计量为：

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

-局限：对数据分布有一定的假设

8.2.4 利用统计检验作为可分性判据

- **非参数化检验方法：秩和检验**
  - Wilcoxon秩和检验（rank-sum test），亦称Mann-Whitney U 检验
  - 基本做法
    - 把两类样本混合在一起，按照所考察的特征从小到大排序
    - 如果一类样本的排序序号之和（秩和）显著的比另一类样本小（或大），则两类样本在所考察的特征上有显著差异。
  - 优点：不对数据分布作特殊假设

8.3 特征选择的最优算法

- **目标**
  - 从D维特征中选取d维（$d < D$）使分类性能最佳（J 最大）
- **两个问题**
  - 标准：根据实际情况选择某种判据
  - 算法

搜索问题

组合数：$C_D^d = \dfrac{D!}{(D-d)!d!}$

e.g.

$D = 100, d = 2, C = 4950$

$D = 100, d = 10, C = 1.73e+13$

$D = 100, d = 50, C = 1.01e+29$

$D = 1000, d = 2, C = 499500$

$D = 10000, d = 2, C = 5.00e+7$

穷举搜索：最优

非穷举搜索：次优

搜索方向：从底向上：$\chi_0 = \varnothing$

从顶向下：$\chi_0 = \chi$

8.3 特征选择的最优算法

- **穷举算法**
  - 计算每一可能的组合，逐一比较准则函数。
  - 适用于：d 或 $D-d$ 很小（组合数较少）的情况。
- **分支界定法**
  - 从顶向下，有回溯
  - 应用条件：准则函数有单调性，即

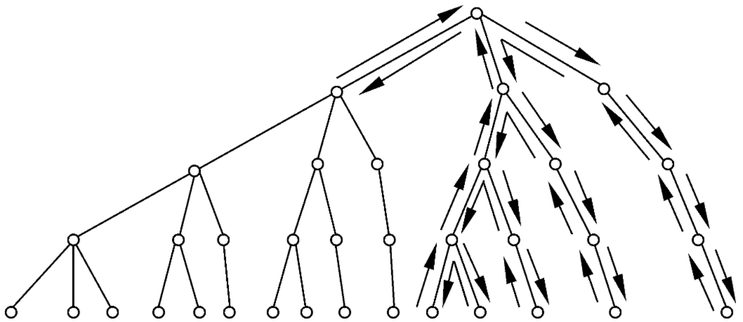
    对特征组 $\bar{\chi}_1 \supset \bar{\chi}_2 \supset \ldots \supset \bar{\chi}_i$，有 $J(\bar{\chi}_1) \geq J(\bar{\chi}_2) \geq \cdots \geq J(\bar{\chi}_i)$

8.3 特征选择的最优算法

- **特征选择的分支定界法**
  - **基本思想**

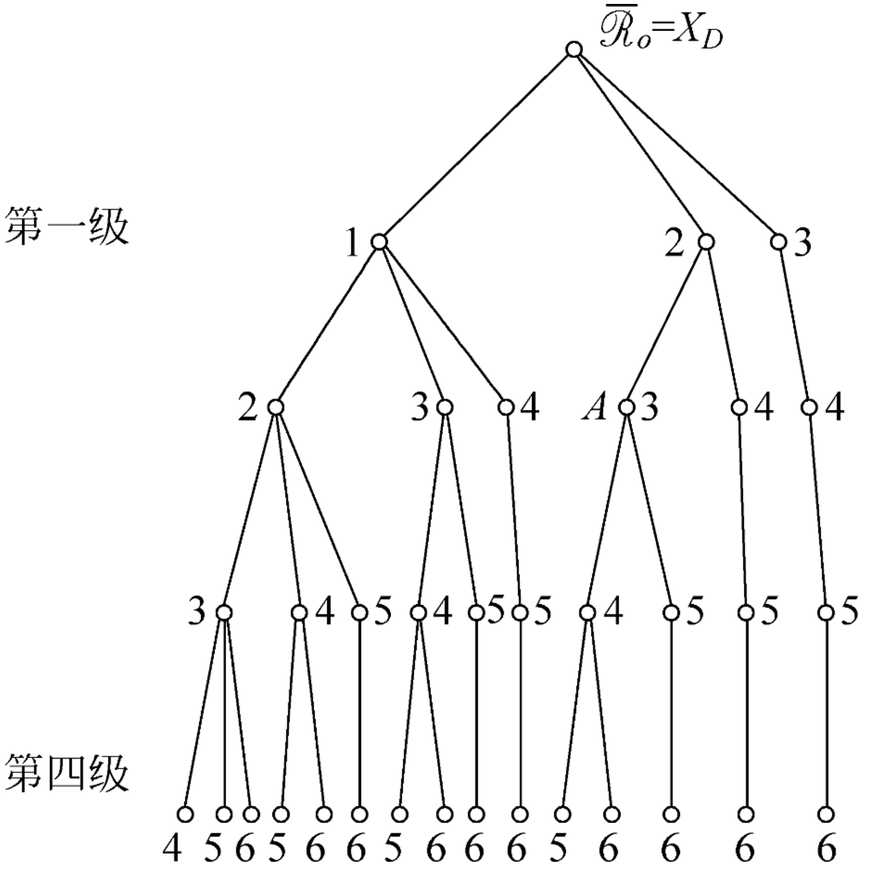
    按照一定顺序将所有可能的组合排列成一棵树，沿途搜索，避免一些不必要的计算，使找到最优解的机会更早。

  - **特点**
    - 最优搜索算法，所有可能的组合都被考虑到
    - 建立搜索树的过程就是特征选择的过程
    - 前提：**准则函数单调性** （注：实际中可能不满足，因 J 是估计值。）
    - 节约计算与存储
    - $d \approx D/2$时最经济



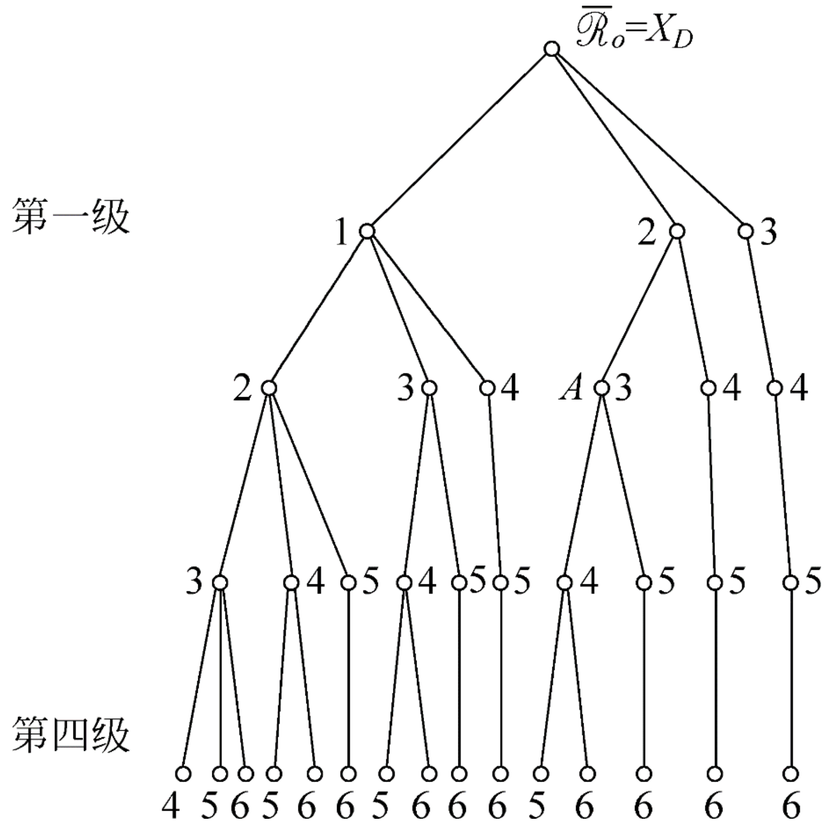
8.3 特征选择的最优算法

- **特征选择的分支定界法**
  - **算法要点**：$D=6$，$d=2$；广度优先搜索
    - 根节点为第0级，包含全体特征
    - 每个节点舍弃一个特征，各叶结点代表选择的各种组合
    - 每级中将最不可能被舍弃（即舍弃后 J 最小）的特征放在最左侧：**按照重要度从左到右排列**
    - 左侧同级中，已经在左侧节点上的特征，在本结点之下不再进行舍弃
    - 避免在整个树中出现相同组合的树枝和叶结点



8.3 特征选择的最优算法

- **特征选择的分支定界法**
  - **算法要点**：$D=6$，$d=2$；广度优先搜索
    - 从**右侧**开始搜索
    - 记录当前搜索到的**叶结点**的**最大准则函数**（界限B），初值置零
    - 搜索到叶结点后，**更新B值**，然后回溯到上一分支处
    - 如果结点上 $J<B$，则不向下搜索，向上回溯
    - 如已回溯到顶（根）而不能再向下搜索，则 $J=B$的叶结点即为解



8.4 特征选择的次优算法

- **单独最优组合**
  - 选前d个单独最佳的特征
- **顺序前进法（sequential forward selection, SFS）**：做加法
  - 第一个特征找最优的
  - 从底向上，从剩余特征中选择一个跟已有组合最优的特征，使加入该特征后所得组合最大
  - 特点：考虑了特征间的组合，但某一特征一经入选，则无法淘汰
  - 广义SFS法：每次增加l个特征
- **顺序后退法（sequential backward selection, SBS）**：做减法
  - 从顶向下，每次删减一个特征寻优一次，使删除该特征后所得组合最大
  - 特点：考虑了特征间的组合，但某一特征一经剔除，则无法入选
  - 广义SBS法：每次删减r个特征

8.4 特征选择的次优算法

- **增l减r法（$l-r$法）**
  - 从底向上，每次增l个特征再减r个特征（$l>r$）
  - 或从顶向下，每次减r个特征再增l个特征（$l<r$）
  - 特点：带有局部回溯过程
  - 广义l-r法：每次选择或提出多个特征

8.5 遗传算法

- **基本思想**
  - 随机搜索算法
  - 模拟生物进化的现象
  - 把优化问题比喻成在无数可能的重组和突变组合中发现适应性最强的组合的问题
  - 开创新的领域：进化计算（Evolutionary Computing）
- **用遗传算法进行特征选择（D个特征里面选择d个）**
  - 染色体（chromosome)编码：二进制字符串m
  - 适应度（fitness）函数：每条染色体对应一个适应度值$f(m)$
  - 选择概率模型 $p(f(m))$

8.5 遗传算法

- **遗传算法基本步骤**

  1.初始化，$t = 0$，随机产生一个包含L个染色体的种群$M(0)$

  2.计算当前种群$M(t)$中每一条染色体的适应度$f(m)$

  3.按照选择概率$p(f(m))$对种群中的染色体进行采样，由采样出的染色体经过一定的操作繁殖出下一代染色体，组成下一代种群$M(t+1)$

  4.回到 2，直到到达终止条件，输出适应度最大的染色体作为找到的最优解。终止条件是某条染色体的适应度达到设定的阈值

- **遗传与变异**

  – 重组（recombination），也称交叉（crossover）

  – 突变（mutation）

  – 基因组反转（inversion）、转座（transposition）

8.6 包裹法

- **包裹法（wrapper法）**
  - 任务相关的选择策略：集成分类器与特征选择、利用分类器进行特征选择的方法
- **例如：R-SVM（递归SVM)和SVM-RFE（SVM递归特征剔除）**

1.用当前所有**候选特征**训练支持向量机：每次选择一定比例特征

2.评估当前所有特征对支持向量机的相对贡献，按贡献大小排序

3.根据事先确定的递归选择特征的数目选择出排序在前面的特征，用这组特征构成新的候选特征，转1，直到达到所规定的特征选择数目

8.6 包裹法

- **区别：评估特征在分类器中的贡献**
  - 线性核

R-SVM：$s_j = \omega_j\left(m_j^+ - m_j^-\right), j = 1, \cdots, d$

SVM-RFE：$s_j^{RFE} = \omega_j^2$

  - 非线性核

$$Q = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$DQ(k) = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \left[K\left(x_i, x_j\right) - K(x_i^{(-k)}, x_j^{(-k)}) \right]$$