(1) 轮廓系数：结合了聚类的凝聚度和分离度，用于评估聚类的效果

该值处于 -1 到 1，越大表示聚类效果越好

(3, 2) 归属于簇 1　　$S(i)=[b(i)-a(i)]/\max\{a(i), b(i)\}$.

$b(i)$ 为 (3, 2) 与 2 簇之间点的平均距离.

$a(i)$ 为 (3, 2) 与 1 簇之间其它点的平均距离.

(2) MinPts：表示领域邻域最小值/阈值，若对象 O 密度大于 Minst，则为核心对象.

邻域：$N_\varepsilon(O)=\{p \mid dist(p,O)<\varepsilon, p\in D\}$.

∴ (1,2), (2,2) (2,1) (3,2) (2,3)

(4,4) (4,3) (5,3) ~~(5,4)~~ (6,4) (5,5).

分类：给定数据集 $D=\{(x_i, y_i), i=1,2,\dots,|D|, y_i \in \{+1,-1\}\}$

(1) 假设用线性函数 h 来做分类，采用最小二乘法求解，请给出该优化问题的数学描述

(2). 请利用仅有的数据集，设计实验和评估方法，用于寻找好的分类器。

要求：写出主要步骤，评估标准及计算方法。

ID3. C4.5

(1) $h(x)=\vec{w}\cdot\phi(x)+b \Rightarrow h(x)=\vec{w}\phi(x)$

令 $y=f(x)$. $\vec{w}\cdot\phi(x)-y$ 为残差 residual.

优化目标：$\min \sum_x loss_2(\vec{w}, f, x)=(f(x)-h(x))^2=(\vec{w}\phi(x)-f(x))^2$.

(2). ①. 通过对数据 D 进行处理，任取 $\vec{w}$ 值，得到初始分类器 $h(x)=\vec{w}\phi(x)$. $f(x)=y_i$

②. 选取损失函数为平方损失，~~loss TrainLoss(·)=Σ los~~

$TrainLoss(\cdot)=\sum loss_2(\vec{w}, f, x)=(f(x)-h(x))^2$

$=\sum(\vec{w}\cdot\phi(x)-f(x))^2$.

优化目标是使 TrainLoss(·) 最小，~~时二得~~

③. $\vec{w}$ 的迭代过程：采用梯度下降法.

$$\vec{w}_{s+1}=\vec{w}_s-\lambda_s\nabla_w\left(\frac{\sum_{(x,y)\in D_{train}}(\vec{w}_s\cdot\phi(x)-y)^2}{|D_{train}|}\right)=$$

$$=\vec{w}_s-\lambda_s\frac{\sum_{(x,y)\in D_{train}}2(\vec{w}_s\cdot\phi(x)-y)\phi(x)}{|D_{train}|}\Big).$$

采用不同的方法，例如梯度下降法和随机梯度下降，寻找 $\vec{w}$.

④. 最终得到的 $\vec{w}$ 即为 $h(x)=\vec{w}\phi(x)$.

(3). 评估标准：最终求得的损失函数越小，则分类器越好.

计算方法.

如图所示混淆矩阵，为一分类器在测试集上的统计结果

1、请计算分类错误率

2、计算召回率和精度，并判断该分类器是否适合该数据集的分类问题，请阐明理由

实际类/预测类	yes	no	total
yes	170	(30)	200
no	(330)	470	800
total	500	500	1000

Ture positive　False Negative

False positive　True negative

1、$\frac{30+330}{1000} \times 100\%$.

2、精度.

$$precision = \frac{true\ positive}{true\ positive + false\ positive}.$$

$$= \frac{170}{500}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative}.$$

$$= \frac{170}{200}$$

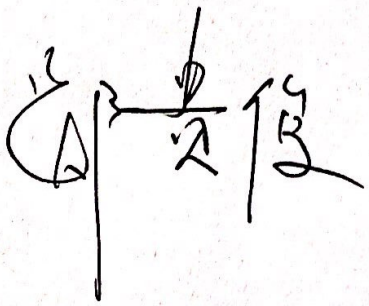
$$F1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall}$$

第二届“中金所杯”获奖总结

获得奖项：第二届中金所杯一等奖
指导老师：徐守萍
获奖时间：2015 年 1 月
竞赛简介：全国大学生金融衍生品知识竞赛
获奖总结：

我是工商管理系 2012 级金融营销专业的学生，在一次浏览金融资讯网页的时候了解到了由中国金融期货交易所举办的第二届“中金所杯”全国大学生金融衍生品知识竞赛。顾名思义，这项竞赛主要考察大学生对金融衍生品知识的掌握，知识竞赛内容范围涵盖国债期货、外汇期货、金融期权、场外衍生品、结构化产品，涉及范围十分之多，看完大赛官网的例题之后，更是发现其中考察的深度也难。本来是心生退意，但是向学校实验中心徐守萍老师表达困惑后，得到了鼓舞的力量，她鼓励我勇敢参加这样的竞赛，即便参赛的对象不乏各类名校的优秀学生，但是我们只要用心去做这件事，其实最终没有得到任何奖项，但不也是收获了一路的知识吗？她用这样的话语启迪了我，比赛第二，知识的收获才是真的很重要，是第一。

在竞赛的期间，由于考察的金融知识十分之多，并且具有较高的难度，许多题目或者问题是我翻书都没法独立解决的，咨询身边学习较好的同学也对许多题目表示很棘手，而徐守萍老师因为在从事金融建模竞赛的组织工作，每当我遇到困难找她时，她都耐心地跟我讨论解题的思路：有些题目她曾经了解过，一见到题目就能给我指出需要哪些知识点，需要找什么样的参考资料；有一些题目她之前没接触过的，也会用上不少时间跟我一起琢磨，一起寻求解题路径；当然老师也不是全知的，也会碰上她不甚了解，我们一起寻求也解不出的答案，在这种时候她就会帮我向金融系的专业老师寻求答案的求解。徐老师用一种身体力行的方式使我耳濡目染-----书山有路勤为径。
其实这个过程许多次产生自我怀疑，毕竟知识体系上面缺失的实在是很多，常常觉得自己的能力参加这个竞赛是很困难的，但总是在徐老师的鼓励下重新恢复信心，老师在我参加竞赛的过程中的持续鼓舞让我深刻感悟一句话，既然选择了远方，便只顾风雨兼程。
大赛结束后当得知我得到了一等奖的成绩时，十分欣喜激动，感谢徐守萍老师在竞赛期间对我的持续鼓舞和指导，之所以能得到这样的好成绩最重要的是她用心在不断的将希望孕育并将温暖让我所感知----风雨兼程过后，面朝大海，春暖花开。



$A > split\_point$ 的元组集合。

2. 增益率

信息增益度量偏向具有许多输出的测试。换句话说，它倾向于选择具有大量值的属性。例如，考虑充当唯一标识符的属性，如 *product_ID*。在 *product_ID* 的划分将导致大量分区（与值一样多），每个只包含一个元组。由于每个分区都是纯的，所以基于该划分对数据集 D 分类所需要的信息为 $Info_{product\_ID}(D) = 0$。因此，通过对该属性的划分得到的信息增益最大。显然，这种划分对分类没有用。

ID3 的后继 C4.5 使用一种称为增益率（gain ratio）的信息增益扩充，试图克服这种偏倚。它用“分裂信息（split information）”值将信息增益规范化。分裂信息类似于 $Info(D)$，定义如下

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{8.5}$$

该值代表由训练数据集 D 划分成对应于属性 A 测试的 v 个输出的 v 个分区产生的信息。注意，对于每个输出，它相对于 D 中元组的总数考虑具有该输出的元组数。它不同于信息增益，信息增益度量关于分类基于同样划分的所获得的信息。增益率定义为

340

$$GrianRate(A) = \frac{Grain(A)}{SplitInfo_A(D)} \tag{8.6}$$

选择具有最大增益率的属性作为分裂属性。然而需要注意的是，随着划分信息趋向于0，该比率变得不稳定。为了避免这种情况，增加一个约束：选取的测试的信息增益必须较大，至少与考察的所有测试的平均增益一样大。

例8.2 属性 *income* 的增益率的计算。属性 *income* 的测试将表8.1中的数据划分成3个分区，即 *low*、*medium* 和 *high*，分别包含4、6和4个元组。为了计算 *income* 的增益率，首先使用（8.5）式得到

$$SplitInfo_A(D) = -\frac{4}{14} \times \log_2\frac{4}{14} - \frac{6}{14} \times \log_2\frac{6}{14} - \frac{4}{14} \times \log_2\frac{4}{14} = 1.557$$

由例8.1，$Gain(income) = 0.029$。因此，$GainRatio(income) = 0.029/1.557 = 0.019$。■

3. 基尼指数

基尼指数（Gini index）在CART中使用。使用上面介绍的概念，基尼指数度量数据分区或训练元组集 D 的不纯度，定义为

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{8.7}$$

其中，p_i 是 D 中元组属于 C_i 类的概率，并用 $|C_{i,D}|/|D|$ 估计。对 m 个类计算和。

基尼指数考虑每个属性的二元划分。首先考虑 A 是离散值属性的情况，其中 A 具有 v 个不同值 $\{a_1, a_2, \cdots, a_v\}$ 出现在 D 中。为了确定 A 上最好的二元划分，考察使用 A 的已知值形成的所有可能子集。每个子集 S_A 可以看做属性 A 的一个形如“$A \in S_A$?”的二元测试。给定一个元组，如果该元组 A 的值出现在 S_A 列出的值中，则该测试满足。如果 A 具有 v 个可能的值，则存在 2^v 个可能的子集。例如，如果 *income* 具有3个可能的值 $\{low, medium, high\}$，则可能的子集是 $\{low, medium, high\}$、$\{low, medium\}$、$\{low, high\}$、$\{medium, high\}$、$\{low\}$、$\{medium\}$、$\{high\}$ 和 $\{\}$。不考虑幂集 $\{low, medium, high\}$ 和空集，因为从概念上讲，它们不代表任何分裂。因此，基于 A 的二元划分，存在 $\left(\frac{2^v - 2}{2}\right)$ 种形成数据集 D 的两个分区的可能方法。

341

当考虑二元划分裂时，计算每个结果分区的不纯度的加权和。例如，如果 A 的二元划分