1基于距离的分类方法

距离的度量：欧式距离，马氏距离，明氏距离，切比雪夫距离

KNN: K最近邻算法

选出k个最近的邻居中的多数票的类编号

可以根据距离为每一个投票增加权重w = 1/d2

算法 K-近邻分类算法

输入： 训练数据T；近邻数目K；待分类的元组t。

输出： 输出类别c。

（1）N=Φ；

（2）FOR each d ∈T DO BEGIN

（3） IF |N|≤K THEN

（4） N=N∪{d}；

（5） ELSE

（6） IF ∃u∈N such that sim(t，u)<sim(t，d) THEN BEGIN

（7） N=N-{u}；

（8） N=N∪{d}；

（9） END

（10）END

（11）c=class to which the most u∈N.

**贝叶斯分类**

基于贝叶斯定理，通过计算给定样本属于一个特定类的概率来对给定样本进行分类。

公式：条件概率（事件B发生后事件A发生的概率）



乘法定理

P(AB)=P(B|A)P(A)=P(A|B)P(B)

推广形式



全概率公式



先验概率：由以往的数据分析得到的概率；

后验概率：在得到信息之后再重新加以修正的概率；

P(H|X)=P(X|H)P(H)/P(X)

P(H)、P(X)就是先验概率，P(X|H)代表假设H成立的情况下，观察到X的规律，而P(H|X)就是后验概率。

但是由于P(X)是相同的，所以只需要比较P(X|H)P(H)的大小就可以根据最大后验概率准则进行分类。如果是二分类的话，还可以根据P(yes|X)+P(no|X)=1求解出具体的P(yes|X)、P(X)

朴素贝叶斯分类的优点：容易实现；多数情况下能获得很好的结果

缺点：假设类条件独立，缺少准确性；实践中，变量之间可能存在依赖，这些依赖不能使用朴素贝叶斯算法进行分类；

**决策树**

从一组无次序、无规则的元组中推理出决策树表示形式的分类规则，其中每个内部结点表示在一个属性上的测试，每个分支代表一个测试输出，每个叶结点表示类或者类分布

Hunt算法

* 设 Dt 是与结点t相关联的训练集
* 算法步骤：
  + 如果Dt中所有记录都属于同一个类 yt，则t是叶结点，用yt标记
  + 如果Dt中包含属于多个类的记录，则选择一个属性测试条件，将记录划分成较小的子集。对于测试条件的每个输出，创建一个子结点，并根据测试结果将Dt中的记录分布到子结点中。然后，对于每个子结点，递归地调用该算法。

ID3算法

标称属性的划分

* 多路划分: 划分数（输出数）取决于该属性不同属性值的个数.
* 二元划分: 划分数为2，这种划分要考虑创建k个属性值的二元划分的所有2k-1-1种方法.

序数属性

* 多路划分: 划分数（输出数）取决于该属性不同属性值的个数.
* 二元划分: 划分数为2，需要保持序数属性值的有序性.

连续属性的划分：

* 多路划分: vi≤A＜vi+1（i=1,…,k)
* 二元划分: (A < v) or (A ≥ v)

考虑所有的划分点，选择一个最佳划分点v

分裂属性的选择

* 通过描述属性可以减少类标属性的不确定性。
* 不同描述属性对减少类标属性不确定性的贡献不同。
* **选择具有最大信息增益的属性作为当前节点的测试属性。**
  + ：正数，越大信息量越大（熵越大，越混乱），当类分布均衡时， Info值达到最大值

ID3使用**信息增益**作为属性选择度量

* ID3算法的基本思想是，以信息熵为度量，用于决策树节点的属性选择，每次优先选取信息量最多的属性，亦即能使熵值变为最小的属性，以构造一颗熵值下降最快的决策树，到叶子节点处的熵值为0。此时，每个叶子节点对应的实例集中的实例属于同一类。

Gain = Info(D)-InfoA(D):通过A上的划分我们得到了一个最纯的结果。即每一次选择一个属性，使得获得的信息增益最大（不确定消除的最多）

信息增益的度量偏向具有许多输出的测试（分的最散，比如把每个ID分成一类，这样每个分区都是纯的，但是没有意义）

因此引入信息增益率：