数据挖掘常用的基础算法

1. 分类
    - k-近邻算法
    - 决策树
    - 朴素贝叶斯算法
    - Logistic回归
    - 支持向量机
    - 集成方法
        - Bagging
        - Boosting
2. 回归
    - 线性回归
3. 无监督
    - K-均值聚类

---

k-近邻算法

k-近邻算法采用测量不同特征值之间的距离方法进行分类。

优点：精度高、对异常值不敏感、无数据输入假定。
缺点：计算复杂度高、空间复杂度高。
适用数据范围：数值型和标称型。

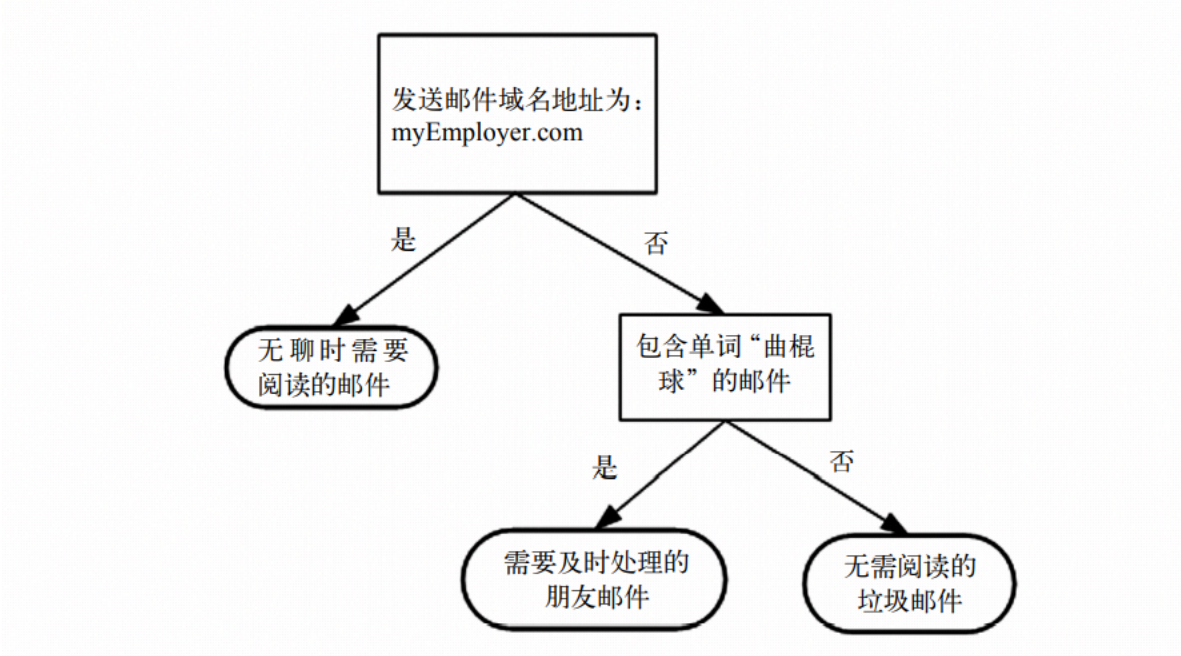
k-近邻算法（kNN），它的工作原理是：存在一个样本数据集合，也称作训练样本集，并且样本集中每个数据都存在标签，即我们知道样本集中每一数据与所属分类的对应关系。输入没有标签的新数据后，将新数据的每个特征与样本集中数据对应的特征进行比较，然后算法提取样本集中特征最相似数据（最近邻）的分类标签。一般来说，我们只选择样本数据集中前k个最相似的数据，这就是k-近邻算法中k的出处，通常k是不大于20的整数。最后，选择k个最相似数据中出现次数最多的分类，作为新数据的分类。

决策树

决策树的原理跟问答游戏类似: 游戏的规则很简单：参与游戏的一方在脑海里想某个事物，其他参与者向他提问题，只允许提20个问题，问题的答案也只能用对或错回答。问问题的人通过推断分解，逐步缩小待猜测事物的范围。决策树的工作原理与20个问题类似，用户输入一系列数据，然后给出游戏的答案。

优点：计算复杂度不高，输出结果易于理解，对中间值的缺失不敏感，可以处理不相关特征数据。
缺点：可能会产生过度匹配问题。
适用数据类型：数值型和标称型。

基于决策树的邮件分类系统，它首先检测发送邮件域名地址。如果地址为myEmployer.com，则将其放在分类“无聊时需要阅读的邮件”中。如果邮件不是来自这个域名，则检查邮件内容里是否包含单词曲棍球，如果包含则将邮件归类到“需要及时处理的朋友邮件”，如果不包含则将邮件归类到“无需阅读的垃圾邮件”。



决策树决策路径

朴素贝叶斯

概率论是许多机器学习算法的基础,朴素贝叶斯就是基于概率论进行分类的方法.

贝叶斯:

贝叶斯概率理论非常流行且效果良好。贝叶斯概率以18世纪的一位神学家托马斯·贝叶斯（Thomas Bayes）的名字命名。贝叶斯概率引入先验知识和逻辑推理来处理不确定命题。另一种概率解释称为频数概率（frequency probability），它只从数据本身获得结论，并不考虑逻辑推理及先验知识.

优点：在数据较少的情况下仍然有效，可以处理多类别问题。
缺点：对于输入数据的准备方式较为敏感。
适用数据类型：标称型数据。

Logistic回归

用Logistic回归进行分类的主要思想是：根据现有数据对分类边界线建立回归公式，以此进行分类。这里的“回归”一词源于最佳拟合，表示要找到最佳拟合参数集.

支持向量机

优点：泛化错误率低，计算开销不大，结果易解释。
缺点：对参数调节和核函数的选择敏感，原始分类器不加修改仅适用于处理二类问题。
适用数据类型：数值型和标称型数据。

集成方法

当做重要决定时，大家可能都会考虑吸取多个专家而不只是一个人的意见。机器学习处理问题时又何尝不是如此？这就是元算法（meta-algorithm）背后的思路。元算法是对其他算法进行组合的一种方式。接下来我们将集中关注一个称作AdaBoost的最流行的元算法。由于某些人认为AdaBoost是最好的监督学习的方法，所以该方法是机器学习工具箱中最强有力的工具之一。

优点：泛化错误率低，易编码，可以应用在大部分分类器上，无参数调整。
缺点：对离群点敏感。
适用数据类型：数值型和标称型数据。

Bagging

自举汇聚法（bootstrap aggregating），也称为bagging方法，是在从原始数据集选择S次后得到S个新数据集的一种技术。新数据集和原数据集的大小相等。每个数据集都是通过在原始数据集中随机选择一个样本来进行替换而得到的。这里的替换就意味着可以多次地选择同一样本。这一性质就允许新数据集中可以有重复的值，而原始数据集的某些值在新集合中则不再出现。

在S个数据集建好之后，将某个学习算法分别作用于每个数据集就得到了S个分类器。当我们要对新数据进行分类时，就可以应用这S个分类器进行分类。与此同时，选择分类器投票结果中最多的类别作为最后的分类结果。当然，还有一些更先进的bagging方法，比如随机森林（random forest）。

Boosting

boosting是一种与bagging很类似的技术。不论是在boosting还是bagging当中，所使用的多个分类器的类型都是一致的。但是在前者当中，不同的分类器是通过串行训练而获得的，每个新分类器都根据已训练出的分类器的性能来进行训练。boosting是通过集中关注被已有分类器错分的那些数据来获得新的分类器。

由于boosting分类的结果是基于所有分类器的加权求和结果的，因此boosting与bagging不太一样。bagging中的分类器权重是相等的，而boosting中的分类器权重并不相等，每个权重代表的是其对应分类器在上一轮迭代中的成功度.一些比较流行的Boosting算法,Xgboost, LightGBM

线性回归

回归的目的是预测数值型的目标值。

优点：结果易于理解，计算上不复杂。
缺点：对非线性的数据拟合不好。
适用数据类型：数值型和标称型数据。

K-均值聚类

聚类是一种无监督的学习，它将相似的对象归到同一个簇中。它有点像全自动分类。聚类方法几乎可以应用于所有对象，簇内的对象越相似，聚类的效果越好。K均值（K-means）之所以称之为K-均值是因为它可以发现k个不同的簇，且每个簇的中心采用簇中所含值的均值计算而成。

聚类与分类的最大不同在于，分类的目标事先已知，而聚类则不一样。因为其产生的结果与分类相同，而只是类别没有预先定义，聚类有时也被称为无监督分类（unsupervisedclassification）。

优点：容易实现。
缺点：可能收敛到局部最小值，在大规模数据集上收敛较慢。
适用数据类型：数值型数据。