传统机器学习算法

2019年10月6日 20:44

blog：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12282042.0.0.775b2042AnFXZQ&postId=6239

一、传统机器学习算法的分类

- 回归：建立一个回归方程来预测目标值，用于连续型分布预测
- 分类：给定大量带标签的数据，计算出未知标签样本的标签取值
- 聚类：将不带标签的数据根据距离聚集成不同的簇，每一簇数据有共同的特征
- 关联分析：计算出数据之间的频繁项集合
- 降维：原高维空间中的数据点映射到低维度的空间中

二、算法讲解

1、**回归**--线性回归

即建立一个回归方程来预测目标值，用于连续型分布预测。

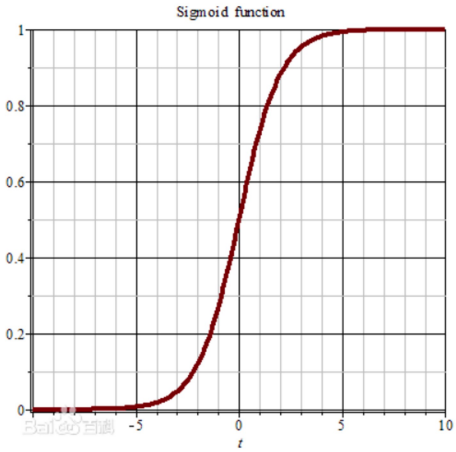
2、**分类**（给定大量带标签的数据，计算出未知标签样本的标签取值）

1）逻辑回归

通过Sigmoid函数将线性函数的结果映射到Sigmoid函数中，预估事件出现的概率并分类。Sigmoid是归一化的函数，可以把连续数值转化为0到1的范围，提供了一种将连续型的数据离散化为离散型数据的方法。

sigmoid函数表达式：

$$S(x) = \frac{1}{1+e^{-x}}$$



2）K-近邻--用距离度量最相邻的标签

工作原理如下：

- 计算样本数据中的点与当前点之间的距离
- 算法提取样本最相似数据(最近邻)的分类标签
- 确定前k个点所在类别的出现频率. 一般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
- 返回前k个点所出现频率最高的类别作为当前点的预测分类

3）朴素贝叶斯--选择后验概率最大的类作为分类标签

P(X|C): 条件概率，C中X出现的概率

P(C): 先验概率，C出现的概率

P(C|X): 后验概率，X属于C类的概率

4）决策树--构造一棵熵值下降最快的分类树

决策树自顶向下递归，使得使用某特征对数据集划分之后，各数据子集的纯度要比划分前的数据集D的纯度高，不确定性降低

熵：描述信息的不确定性。熵值越大不确定性越大，越混乱。

$$H(X) = -\sum_{i=1}^{n} p_i log_2 p_i$$

决策点的分裂特征：选择信息增益最大的特征

信息增益（ID3算法）：以某特征划分数据集前后熵的差异（自然是越大越好，越大说明熵值下降越快）

信息增益比（C4.5算法）：解决信息增益偏向取值较多的特征。（参见 https://blog.csdn.net/Tomcater321/article/details/80699044）

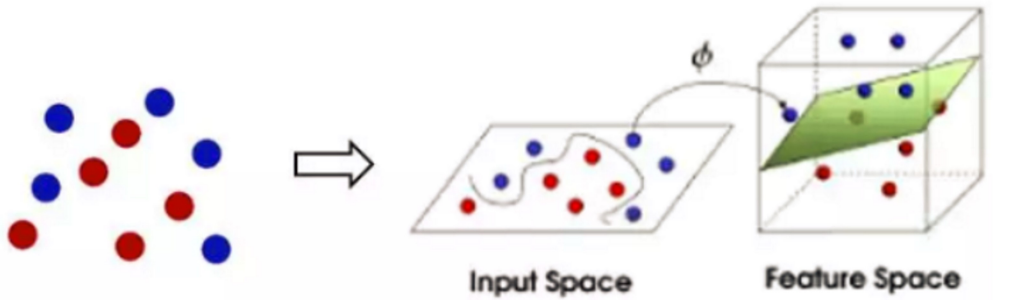
基尼指数（CART算法--分类树）：

基尼指数（基尼不纯度）：表示在样本集合中一个随机选中的样本被分错的概率。Gini指数越小表示集合中被选中的样本被分错的概率越小，也就是说集合的纯度越高，反之，集合越不纯。

5、SVM支持向量机：构造超平面，分类非线性数据集

原理过程：（1）当一个分类问题，数据是线性可分时，只要将线的位置放在让小球距离线的距离最大化的位置即可，寻找这个最大间隔的过程，就叫做最优化。（2）一般的数据是线性不可分的，可以通

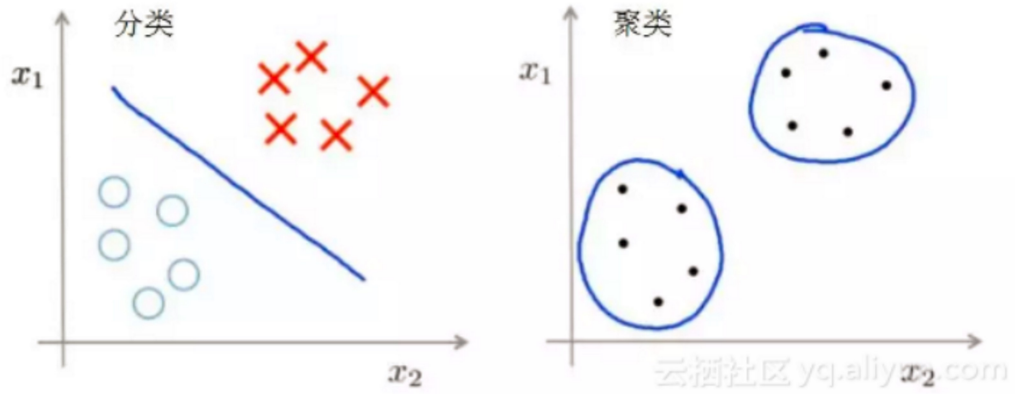
过核函数，将数据从二维映射到高位，通过超平面将数据切分。



目的是找到最大间隔的决策面的最优解

3、**聚类（无监督学习）**

需要被分类的数据集已经有标记，例如数据集已经标记为○或者×，通过学习出假设函数对这两类数据进行划分。而对于没有标记的数据集，希望能有一种算法能够自动的将相同元素分为紧密关系的子集或簇，这就是聚类算法。



聚类算法的训练数据是没有类别标签的，输出的是类别的标号

分类算法的训练数据的有类别标签的，输出的是类别

1）K-means：

原理步骤：

（1）随机生成k个初始点作为质心；（2）数据集中的数据按照距离质心的远近分到各个簇中；（3）各个簇中的数据求平均值，作为新的质心，重复上一步，直到所有的簇不再改变。 两个分类间隔越远，则聚类效果越好。

4、关联分析

1）FP-growth算法

- 频繁项集：在数据库中大量频繁出现的数据集合。例如购物单数据中{'啤酒'}、{'尿布'}、{'啤酒', '尿布'}出现的次数都比较多。
- 关联规则：由集合 A，可以在某置信度下推出集合 B。即如果 A 发生了，那么 B 也很有可能会发生。例如购买了{'尿布'}的人很可能会购买{'啤酒'}。
- 支持度：指某频繁项集在整个数据集中的比例。假设数据集有 10 条记录，包含{'啤酒', '尿布'}的有 5 条记录，那么{'啤酒', '尿布'}的支持度就是 5/10 = 0.5。
- 置信度：有关联规则如{'尿布'} -> {'啤酒'}，它的置信度为 {'尿布'} -> {'啤酒'}

假设{'尿布', '啤酒'}的支持度为 0.45，{'尿布'}的支持度为 0.5，则{'尿布'} -> {'啤酒'}的置信度为 0.45 / 0.5 = 0.9。

分析步骤为：

（1）从购物车数据中挖掘出频繁项集

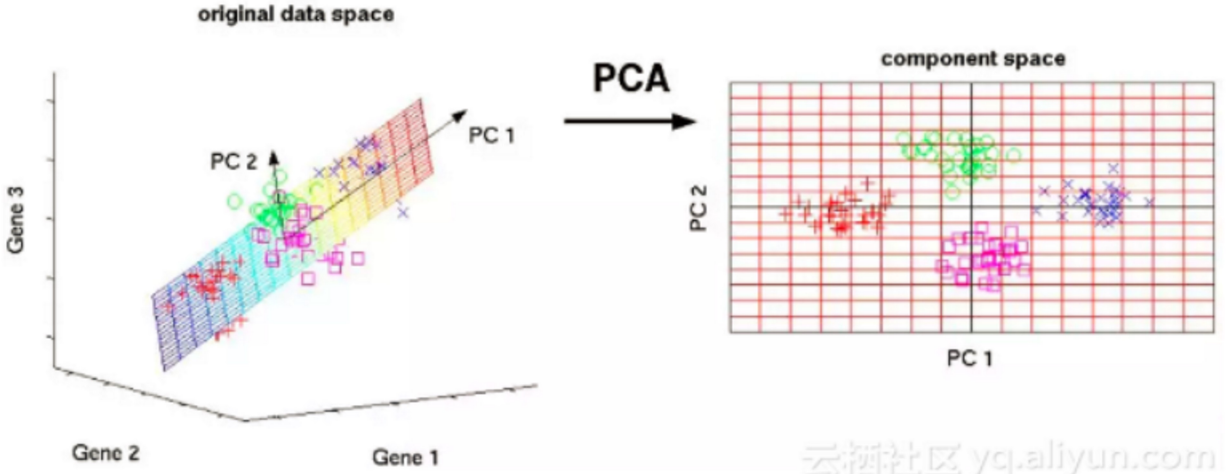
（2）从频繁项集中产生关联规则，计算支持度

（3）输出置信度

5、降维

1）PCA降维：减少数据维度，降低数据复杂度

降维是指将原高维空间中的数据点映射到低维度的空间中。因为高维特征的数目巨大，距离计算困难，分类器的性能会随着特征数的增加而下降；减少高维的冗余信息所造成的误差,可以提高识别的精度。



比较常用的是主成分分析算法（PCA）。它是通过某种线性投影，将高维的数据映射到低维的空间中表示，并期望在所投影的维度上数据的方差最大，以此使用较少的数据维度，同时保留住较多的原数据点的特性。

6、神经网络

不同节点之间的连接被赋予了不同的权重，每个权重代表了一个节点对另一个节点的影响大小。每个节点代表一种特定函数，来自其他节点的信息经过其相应的权重综合计算。是一个可学习的函数，接受不同数据的训练，不断通过调整权重而得到契合实际模型。

多层神经网络的每一层神经元学习到的是前一层神经元值的更抽象（更大更泛化）的表示，通过抽取更抽象的特征来对事物进行区分，从而获得更好的区分与分类能力。