引言

2020年9月15日

13:56



【机器学习】——数据分析技术

【数据库】———数据管理技术

【数据挖掘】



一、典型机器学习过程：

1、通过‘训练数据(包含标签~***数据集***)’

2、训练数据，获得模型。获得在这个上获得的***模型***

3、当有新的数据来的时候，通过已获得的数据信息，来判断‘新样本’的好坏。

案例：西瓜挑选问题——启示：当相应的‘指数’并没有‘全覆盖’。只从现有的有限的数据来对无线的区间进行判断。

二、基本术语：

\*pre:

* ***记录***：关于一个事件or实例的描述
* ***属性/特征***：事物上某方面的‘表现’
* ***属性值***：‘属性’上的取值
* ***属性空间/样本空间/张成空间***：属性值的取值范围，比如3个属性，3个维度，3个维度的取值。
* ***特征向量***：属性空间中的‘某一点’。
* ***维数***：维度。
* ***学习/训练***：从数据中‘学得模型的过程’
* ***训练数据***：训练用到的数据。***训练样本***：训练数据中每一个样本。***训练集***：训练样本组成的集合。
* ***假设***：学得模型 对应 有关数据集 的‘某种规律’。
* ***真相***：假设要逼近的对象or某种‘潜在规律’本身。
* ***学习机***：学习算法在给定数据和参数空间上的实例化。
* ***标记***：关于示例‘结果’的信息。（就是‘好瓜’‘坏瓜’等 ‘label’）
* ***样例***：拥有‘label’的示例。
* ***标记空间/输出空间***：所有标记的集合

1、任务：

* ***分类***：离散——预测离散值的结果，如，好瓜，坏瓜
  + ***二分类任务***：预测结果有‘两种’
  + ***多分类任务***：预测结果有‘多种’
* ***回归***：连续——预测连续值的结果，如，瓜的成熟程度
  + 输出空间是“R”
* ***聚类***：无标记信息————大概意思是，在无分类信息中进行‘标记’。

2、测试：

* ***测试***：学得模型后，使用其进行预测的过程。
* ***测试样本***：被预测的样本。
* 任务分类：根据训练数据有无‘label’信息，分成两类学习任务——
  + ***监督学习***：有label，分类、回归，均是这个代表。
  + ***无监督学习***：无label，产生新的分类，代表，‘聚类’。

3、***泛化能力***：我们总是希望‘学得的模型’能很好的适应‘新样本’。

* 模型适用于新样本的能力：泛化能力。
* 根据空间分布，数量越多，越能得到高泛化的模型

4、假设空间：

* ***归纳学习***：“从样例中学习”——模仿‘归纳-演绎’中的‘归纳过程’
  + ***广义归纳学习***：从样例中学习
  + ***狭义归纳学习/概念学习/概念形成***：从训练数据中学得‘概念’。（有利于理解机器学习中的一些基础思想）。
    - ***布尔概念学习***：（0，1）布尔学习。
* 每一个‘特征’可能的取值~‘连乘积’

PS：①一个‘学习过程’可以看成一个‘在所有假设组成的空间中进行搜索的过程’。‘搜索目标’是找到与训练集‘匹配’的假设，即与训练集中对瓜判断正确的假设。②‘假设’一旦确定，对应确定‘假设空间’及其规模大小。③例如：假设a——（色泽=亮黑，根蒂=蜷缩，敲声=雄亮）对应假设空间A。假设空间A中有某个样例；假设b——（色泽=\*，根蒂=蜷缩，敲声=雄亮）对应假设空间B。假设空间B中有多个‘样例’。④‘搜索过程’可以对假设空间进行搜索，可以在假设空间中进行 不断删除和正例不匹配的假设、和反例匹配的假设。最终获得和训练集一致的假设，这就是我们的学习结果。⑤注意：假设空间很大，训练数据的空间很小，而我们也只在现有的小空间中进行‘学习’。

* ***假设集合/版本空间***：多个假设和训练集一致，

5、归纳偏好：

* 当新的样本用不同的模型会出现不同的结果，那判断好坏就会引入‘偏好性’等‘人为因素’——如果不能容忍坏瓜，就判断成坏的，如果偶尔有个坏瓜没关系，就判断成好的。
* 机器学习算法在学习过程中对某种类型假设的偏好称作***‘归纳偏好***’。
* 比如：曲线问题。——没有免费午餐。
* ***奥卡姆剃刀***——假如多个模型和数据吻合，就选择最简单的那个。作用：用一个一般性原则，引导算法建立‘正确的’偏好。
* 注意：偏好选择也决定了算法的好坏
* 注意：所有问题出现可能性都一样，则和学习算法无关。实际问题中，并非所有问题出现的可能性相同，所以**‘离开具体问题，谈什么算法最好，毫无意义。**

5、发展历程

PS：？

PS：就是‘机器学习’也是在‘已有知识’之上，还是无法替代人工。

‘创新’就是人类的优势。