KNN知识点

1. 一般，knn最近邻方法在( )的情况下效果较好。

A. 样本较多但典型性不好

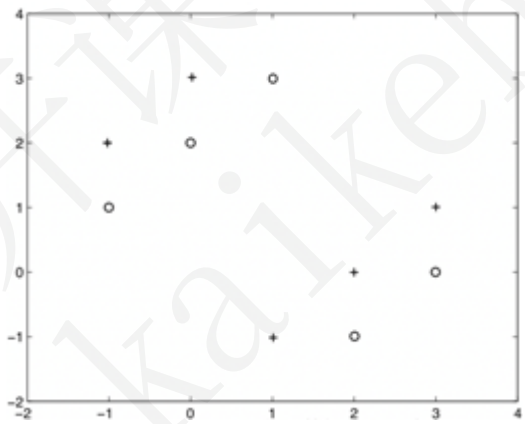
B. 样本较少但典型性好

C. 样本呈团状分布

D. 样本呈链状分布

---

2. 使用k=1的KNN算法, 下图二类分类问题, “+” 和 “o” 分别代表两个类, 那么, 用仅拿出一个测试样本的交叉验证方法, 交叉验证的错误率是多少：( )



A. 0%

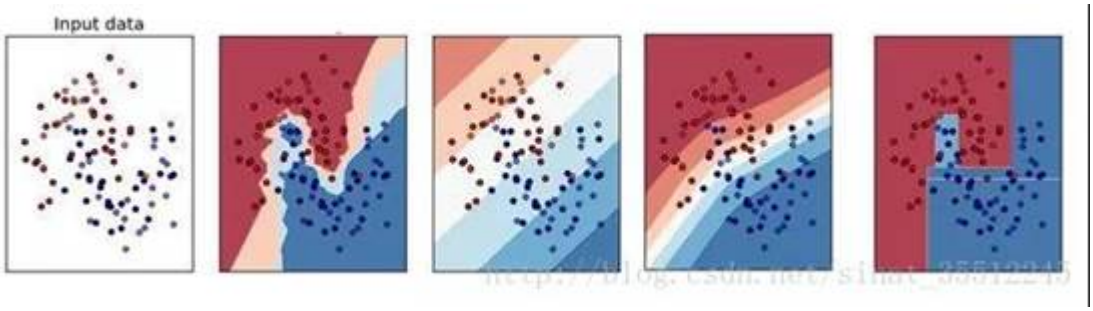
B. 100%

C. 0% 到 100%

D. 以上都不是

---

3. 以下哪个图是KNN算法的训练边界？( )



A. B

B. A

C. D

D. C

E. 都不是

---

4. 简述KNN最近邻分类算法的过程？

---

5. 在knn，我们是用欧氏距离来计算最近的邻居之间的距离。为什么不用曼哈顿距离？

---

6. knn算法是否需要做归一化处理？****

---

7. 关于knn，以下说法正确的是？

决策树——DT

8. 假设我们有一个数据集，在一个深度为 6 的决策树的帮助下，它可以使用 100% 的精确度被训练。现在考虑一下两点，并基于这两点选择正确的选项。( )

注意：所有其他超参数是相同的，所有其他因子不受影响。

1.深度为 4 时将有高偏差和低方差

2.深度为 4 时将有低偏差和低方差

A. 只有 1

B. 只有 2

C. 1 和 2

D. 没有一个

正确答案: (A)

---

9. (判断题) 决策树中，随着树中结点数变得太大，即使模型的训练误差还在继续降低，但是检验误差开始增大，这是出现了模型拟合不足的原因 ( )

---

10. (判断题) 决策树算法不需要做归一化处理。

---

11. 试析使用“最小训练误差”作为决策树划分选择的缺陷。

---

12. 以下关于决策树说法正确的事是 ( )

A. 决策树是一种自上而下，对样本数据进行树形分类的过程，由结点和有向边组成。

B. 结点分为内部结点和叶结点，其中每个内部结点表示一个特征或属性，叶结点表示类别。

C. 从顶部根结点开始，所有样本聚在一起。经过根结点的划分，样本被分到不同的子结点中。再根据子结点的特征进一步划分，直至所有样本都被归到某一个类别 (即叶结点) 中。

D. 决策树作为最基础、最常见的有监督学习模型，常被用于分类问题和回归问题，在市场营销和生物医药等领域尤其受欢迎，主要因为树形结构与销售、诊断等场景下的决策过程十分相似。

---

13. 假设共有5个人追求的一位女孩，年龄有两个属性 (老，年轻)，长相有三个属性 (帅，一般，丑)，工资有三个属性 (高，中等，低)，会写代码有两个属性 (会，不会)，最终分类结果有两类 (见，不见)。我们根据女孩有监督的主观意愿可以得到表如下所示：

	年龄	长相	工资	写代码	类别
小A	老	帅	高	不会	不见
小B	年轻	一般	中等	会	见
小C	年轻	丑	高	不会	不见
小D	年轻	一般	高	会	见
小L	年轻	一般	低	不会	不见

求出每一个特征的信息增益。

---

14. 如上题所示数据，求出每个特征的信息增益比。

---

15. 如13题所示数据，求出每个特征的Gini指数是多少。

---

16. 分析上面三题的计算结果，关于ID3、C4.5、CART，下面说法正确的是：（ ）

A. ID3是采用信息增益作为评价标准，除了“会写代码”这一逆天特征外，会倾向于取值较多的特征。因为，信息增益反映的是给定条件以后不确定性减少的程度，特征取值越多就意味着确定性更高，也就是条件熵越小，信息增益越大。比如，我们引入特征“DNA”，每个人的DNA都不同，如果ID3按照“DNA”特征进行划分一定是最优的（条件熵为0），但这种分类的泛化能力是非常弱的。因此，C4.5实际上是对ID3进行优化，通过引入信息增益比，一定程度上对取值比较多的特征进行惩罚，避免ID3出现过拟合的特性，提升决策树的泛化能力。

B. 从样本类型的角度，ID3只能处理离散型变量，而C4.5和CART都可以处理连续型变量。C4.5处理连续型变量时，通过对数据排序之后找到类别不同的分割线作为切分点，根据切分点把连续属性转换为布尔型，从而将连续型变量转换多个取值区间的离散型变量。而对于CART，由于其构建时每次都会对特征进行二值划分，因此可以很好地适用于连续性变量。

C. 从应用角度，ID3和C4.5只能用于分类任务，而CART（ClassificationandRegressionTree，分类回归树）从名字就可以看出其不仅可以用于分类，也可以应用于回归任务（回归树使用最小平方误差准则）。

D. 从实现细节、优化过程等角度，这三种决策树还有一些不同。比如，ID3对样本特征缺失值比较敏感，而C4.5和CART可以对缺失值进行不同方式的处理；ID3和C4.5可以在每个结点上产生出多叉分支，且每个特征在层级之间不会复用，而CART每个结点只会产生两个分支，因此最后会形成一颗二叉树，且每个特征可以被重复使用；ID3和C4.5通过剪枝来权衡树的准确性与泛化能力，而CART直接利用全部数据发现所有可能的树结构进行对比。

---

17. 决策树的剪枝通常有哪儿两种方法？

18. 决策树的预剪枝对于何时停止决策树的生长有哪几种方法？

朴素贝叶斯部分

19.Nave Bayes是一种特殊的Bayes分类器,特征变量是X,类别标签是C,它的一个假定是:( )

A. 各类别的先验概率P(C)是相等的

B. 以0为均值，sqr(2)/2为标准差的正态分布

C. 特征变量X的各个维度是类别条件独立随机变量

D. P(X|C)是高斯分布

---

20. 假定某同学使用Naive Bayesian（NB）分类模型时，不小心将训练数据的两个维度搞重复了，那么关于NB的说法中正确的是：( )

A. 这个被重复的特征在模型中的决定作用会被加强

B. 模型效果相比无重复特征的情况下精确度会降低

C. 如果所有特征都被重复一遍，得到的模型预测结果相对于不重复的情况下的模型预测结果一样。

D. 当两列特征高度相关时，无法用两列特征相同时所得到的结论来分析问题

E. NB可以用来做最小二乘回归

---

21. 为什么朴素贝叶斯如此“朴素”？

22、以下公式说法正确的是：

$$P(A_k|B) = \frac{P(A_k B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

A、P(Ak) 是先验概率

B、P(Ak | B) 是后验概率

C、P(B | Ak) 是似然函数

D、P(Ak | B) 是似然函数

E、P(B | Ak) 是后验概率

23、朴素贝叶斯的主要优点有：()

A、朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

B、对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。

C、对缺失数据不太敏感，算法也比较简单，常用于文本分类。

D、对输入数据表达形式敏感。

24、朴素贝叶斯的主要缺点有：()

A 、 理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

B 、需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。

C、由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

D、对输入数据的表达形式很敏感。

25、scikit-learn 中的使用sklearn.naive_bayes模块中只有三个分类器，分别为（ ）

A、BernoulliNB()

B、GaussianNB()

C、MultinomialNB()

D、KNeighborsClassifier()

#