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

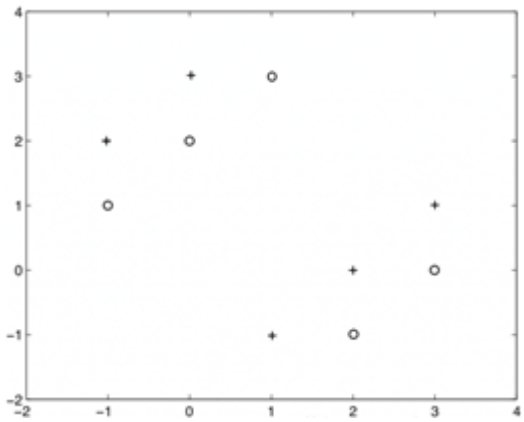
D. 样本呈链状分布

正确答案：(B)

解析： K近邻算法主要依靠的是周围的点，因此如果样本过多，那肯定是区分不出来的。因此应当选择B。样本呈团状颇有迷惑性，这里应该指的是整个样本都是呈团状分布，这样knn就发挥不出其求近邻的优势了，整体样本应该具有典型性好，样本较少，比较适宜。

---

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

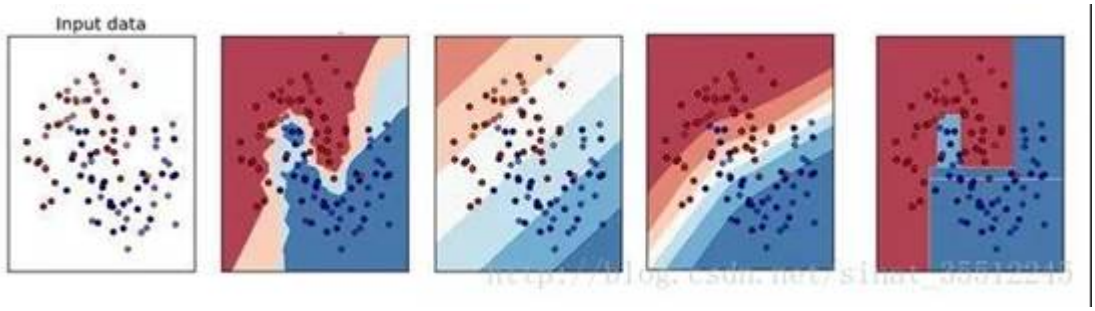
正确答案：(B)

解析：

KNN算法就是, 在样本周围看k个样本, 其中大多数样本的分类是A类, 我们就把这个样本分成A类. 显然, k=1 的KNN在上图不是一个好选择, 分类的错误率始终是100%。

---

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

正确答案：(B)

解析：

KNN算法肯定不是线性的边界，所以直的边界就不用考虑了。另外这个算法是看周围最近的k个样本的分类用以确定分类，所以边界一定是坑坑洼洼的。

---

4. 简述KNN最近邻分类算法的过程？

解析：（注意对算法时间空间复杂度的了解）

1.计算训练样本和测试样本中每个样本点的距离（常见的距离度量有欧式距离，马氏距离等）；

2.对上面所有的距离值进行排序；

3.选前k个最小距离的样本；

4.根据这k个样本的标签进行投票，得到最后的分类类别；

---

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？

解析：（**k-means**该问题也适用，提前可以了解下**）

曼哈顿距离只计算水平或垂直距离，有维度的限制。另一方面，欧氏距离可用于任何空间的距离计算问题。因为，数据点可以存在于任何空间，欧氏距离是更可行的选择。例如：想象一下国际象棋棋盘，象或车所做的移动是由曼哈顿距离计算的，因为它们是在各自的水平和垂直方向做的运动。

---

6. knn算法是否需要做归一化处理？**

解析：（对其它算法也做了些说明）

概率模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率，如决策树、RF。而像Adaboost、GBDT、XGBoost、SVM、LR、KNN、KMeans之类的最优化问题就需要归一化。

---

7. 关于knn，以下说法正确的是？

A. K值较小，则模型复杂度较高，容易发生过拟合，学习的估计误差会增大，预测结果对近邻的实例点非常敏感。

B. K值较大可以减少学习的估计误差，但是学习的近似误差会增大，与输入实例较远的训练实例也会对预测起作用，使预测发生错误，k值增大模型的复杂度会下降。

C. 在应用中，k值一般取一个比较小的值，通常采用交叉验证法来来选取最优的K值。

D. 以上说法都不正确

正确答案：(B)

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)

解析：（熟悉了解过拟合和欠拟合问题，偏差和方差问题）

如果在这样的数据中你拟合深度为 4 的决策树，这意味着其更有可能与数据欠拟合。因此，在欠拟合的情况下，你将获得高偏差和低方差。

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（ ）

正确答案: (错误)

解析：这是过拟合的原因，决策树越复杂越容易过拟合，这时候我们考虑剪枝

10.（判断题）决策树算法不需要做归一化处理。

正确答案:（正确)

解析：

概率模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率

---

11. 试析使用“最小训练误差”作为决策树划分选择的缺陷。

解析：

答：若以最小训练误差作为决策树划分的依据，由于训练集和真是情况总是会存在一定偏差，这使得这样得到的决策树会存在过拟合的情况，对于未知的数据的泛化能力较差。因此最小训练误差不适合用来作为决策树划分的依据。

---

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。

正确答案:（ABCD)

---

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。

解析：

$$H(D)=-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}=0.971$$

计算出4个分支结点的信息熵为：

$$\begin{aligned}H(D\mid 年龄)&=\frac{1}{5}H(老)+\frac{4}{5}H(年轻)\\&=\frac{1}{5}(-0)+\frac{4}{5}\left(-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}\right)=0.8\end{aligned}$$

$$\begin{aligned}H(D\mid 长相)&=\frac{1}{5}H(帅)+\frac{3}{5}H(一般)+\frac{1}{5}H(丑)\\&=0+\frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+0=0.551\end{aligned}$$

$$\begin{aligned}H(D\mid 工资)&=\frac{3}{5}H(高)+\frac{1}{5}H(中等)+\frac{1}{5}H(低)\\&=\frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+0+0=0.551\end{aligned}$$

$$\begin{aligned}H(D\mid 写代码)&=\frac{3}{5}H(不会)+\frac{2}{5}H(会)\\&=\frac{3}{5}(0)+\frac{2}{5}(0)=0\end{aligned}$$

进而可计算出各个特征的信息增益为：

$$g(D,年龄)=0.171,\ g(D,长相)=0.42,$$
$$g(D,工资)=0.42,\ \ g(D,写代码)=0.971$$

显然，特征“写代码”的信息增益最大，所有的样本根据此特征，可以直接被分到叶结点（即见或不见）中，完成决策树生长。当然，在实际应用中，决策树往往不能通过一个特征就完成构建，需要在经验熵非0的类别中继续生长。

---

14. 如上题所示数据，求出每个特征的信息增益比。

解析：

特征A对于数据集D的信息增益比定义为

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$$

$H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$ 称为数据集D关于A的取值熵

求出数据集关于每个特征的取值熵为

$$H_{年龄}(D)=-\frac{1}{5}\log_2\frac{1}{5}-\frac{4}{5}\log_2\frac{4}{5}=0.722$$

$$H_{长相}(D)=-\frac{1}{5}\log_2\frac{1}{5}-\frac{3}{5}\log_2\frac{3}{5}-\frac{1}{5}\log_2\frac{1}{5}=1.371$$

$$H_{工资}(D)=-\frac{3}{5}\log_2\frac{3}{5}-\frac{1}{5}\log_2\frac{1}{5}-\frac{1}{5}\log_2\frac{1}{5}=1.371$$

$$H_{写代码}(D)=-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}=0.971$$

根据公式可计算出各个特征的信息增益比为

$$g_R(D,\text{年龄})=0.236,\ g_R(D,\text{长相})=0.402,$$
$$g_R(D,\text{工资})=0.402,\ g_R(D,\text{写代码})=1.$$

信息增益比最大的仍是特征“写代码”，但通过信息增益比，特征“年龄”对应的指标上升了，而特征“长相”和特征“工资”却有所下降。

---

15. 如13题所示数据，求出每个特征的Gini指数是多少。

解析：

Gini描述的是数据的纯度，与信息熵含义类似。

$$\text{Gini}(D)=1-\sum_{k=1}^{n}\left(\frac{|C_k|}{|D|}\right)^2.$$

CART在每一次迭代中选择基尼指数最小的特征及其对应的切分点进行分类。但与ID3、C4.5不同的是，CART是一颗二叉树，采用二元切割法，每一步将数据按特征A的取值切成两份，分别进入左右子树。特征A的Gini指数定义为

$$\text{Gini}(D\mid A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}\text{Gini}(D_i).$$

可计算出各个特征的Gini指数为：

$\text{Gini}(D\mid\text{年龄=老})=0.4$，$\text{Gini}(D\mid\text{年龄=年轻})=0.4$，

$\text{Gini}(D\mid\text{长相=帅})=0.4$，$\text{Gini}(D\mid\text{长相=丑})=0.4$，

$\text{Gini}(D\mid\text{写代码=会})=0$，$\text{Gini}(D\mid\text{写代码=不会})=0$，

$\text{Gini}(D\mid\text{工资=高})=0.47$，$\text{Gini}(D\mid\text{工资=中等})=0.3$，

$\text{Gini}(D\mid\text{工资=低})=0.4$.

在“年龄”“长相”“工资”“写代码”四个特征中，我们可以很快地发现特征“写代码”的Gini指数最小为0，因此选择特征“写代码”作为最优特征，“写代码=会”为最优切分点。按照这种切分，从根结点会直接产生两个叶结点，基尼指数降为0，完成决策树生长。

---

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。

正确答案:（ABCD)

---

17. 决策树的剪枝通常有哪儿两种方法？

解析：

预剪枝（PrePruning）和后剪枝（PostPruning）

预剪枝，即在生成决策树的过程中提前停止树的增长。而后剪枝，是在已生成的过拟合决策树上进行剪枝，得到简化版的剪枝决策树。

预剪枝的核心思想是在树中结点进行扩展之前，先计算当前的划分是否能带来模型泛化能力的提升，如果不能，则不再继续生长子树。

后剪枝的核心思想是让算法生成一棵完全生长的决策树，然后从最底层向上计算是否剪枝。剪枝过程将子树删除，用一个叶子结点替代，该结点的类别同样按照多数投票的原则进行判断。同样地，后剪枝也可以通过在测试集上的准确率进行判断，如果剪枝过后准确率有所提升，则进行剪枝。相比于预剪枝，后剪枝方法通常可以得到泛化能力更强的决策树，但时间开销会更大。

---

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？

解析：

（1）当树到达一定深度的时候，停止树的生长。（2）当到达当前结点的样本数量小于某个阈值的时候，停止树的生长。（3）计算每次分裂对测试集的准确度提升，当小于某个阈值的时候，不再继续扩展。

预剪枝具有思想直接、算法简单、效率高等特点，适合解决大规模问题。但如何准确地估计何时停止树的生长（即上述方法中的深度或阈值），针对不同问题会有很大差别，需要一定经验判断。且预剪枝存在一定局限性，有欠拟合的风险，虽然当前的划分会导致测试集准确率降低，但在之后的划分中，准确率可能会有显著上升。

#

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

正确答案：C

解析：

朴素贝叶斯的条件就是每个变量相互独立。

---

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

正确答案：BD

解析：

NB的核心在于它假设向量的所有分量之间是独立的。在贝叶斯理论系统中，都有一个重要的条件独立性假设：假设所有特征之间相互独立，这样才能将联合概率拆分。

---

21. 为什么朴素贝叶斯如此“朴素”？

解析：

因为它假定所有的特征在数据集中的作用是同样重要和独立的。正如我们所知，这个假设在现实世界中是很不真实的，因此，说朴素贝叶斯真的很“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

正确答案：：ABC

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。

正确答案：：ABC

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。

正确答案：：ABCD

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()

正确答案：：ABC

#