在n维空间中，以下哪一个方法最适用于异常点检测？

A.正态概率图

B.盒图

C.Mahalonobis 距离

D.散点图

答案：C

Mahalonobis 距离是基于卡方分布的多变量异常的程度的统计量。更多内容点击此处。

问题2

线性回归在以下哪些方面和多元回归不一样？

A. 它是专门用来预测一个事件的概率

B. 拟合优度指数

C. 在回归系数的估计方面

D. 以上所有

答案：D

A：线性回归用来解决分类问题，我们可以计算出事件发生的概率

B：总体来说，拟合优度测试是用来测量数据与模型的一致性，我们用逻辑回归来检验模型拟合程度。

C：拟合逻辑回归模型之后，我们可以以他们的系数为目标，观察独立特征之间的关系（正相关或负相关）。

问题3：

引导数据的意义是什么？

A.从M个总体中有放回的抽样出m个特征

B.从M个总体中无放回的抽样出m个特征

C.从N个总体中有放回的抽取n个样本

D.从N个总体中无有放回的抽取n个样本

答案：C

如果我们没有足够的数据来训练我们的算法，我们就可以从训练集中有放回的随机重复一些数据项。

问题4

“过拟合只是监督学习中的问题，对于无监督学习影响不大”这句话是正确还是错误

A.正确

B.错误

答案：B

我们可以使用无监督矩阵来评估一个无监督机器学习算法。举个例子，我们可以用“调整兰德系数”来评估聚类模型。

问题5：

关于选择k层交叉检验中“k”的值，以下说法正确的是？

A.k并不是越大越好，更大的k会减慢检验结果的过程

B.选择更大的k会导致降低向真实期望错误的倾斜

C.选择总是能最小化交叉验证中的方差的k

D.以上所有

答案：D

更大的k会减少过高估计了真正的预期误差的情况（因为训练层更接近总体数据集），但是会带来更长的运行时间（因为这样会接近留一交叉的极限情况），当选择k的时候，我们也要考虑k层精度之间的方差。

问题6：

回归模型具有多重共线性效应，在不损失太多信息的情况下如何应对这种情况？

去除所有共线变量1.去除所有共线变量

去除一个变量而不是都去掉

我们可以计算VIF（方差膨胀因子）来检验多重共线性效应，然后根据情况处理

去除相关的变量可能会导致信息的丢失。为了保证数据的完整性，我们应该选取比如岭回归和套索回归等惩罚回归模型。

以上那些是正确的？

A. 1

B. 2

C. 2 和3

D. 2,3 和 4

答案：D

为了检查多重共线性，我们可以创建一个相关矩阵来识别和删除具有75%相关性的变量（阈值的选择是主观的）。此外，我们使用VIF（方差膨胀因子）来检查多重共线性，如果VIF小于4表示没有多重共线性，如果大于10则表示严重的多重共线性。我们也可以使用一个宽容量作为多重共线性的指数。

但是，移除相关变量会导致信息的损失。为了保证数据的完整性，我们应该选取比如岭回归和套索回归等惩罚回归模型。我们也可以在变量中增加随机噪声，这样数据会变得不一样。但是这种方法会降低预测的准确性，所以要慎用。

问题7：

评估完模型后，我们发现模型中有很高的偏差。我们怎样来减少这个偏差？

A.减少模型中特征的数量

B.增加模型中特征的数量

C.增加模型中的数据点

D.B和C

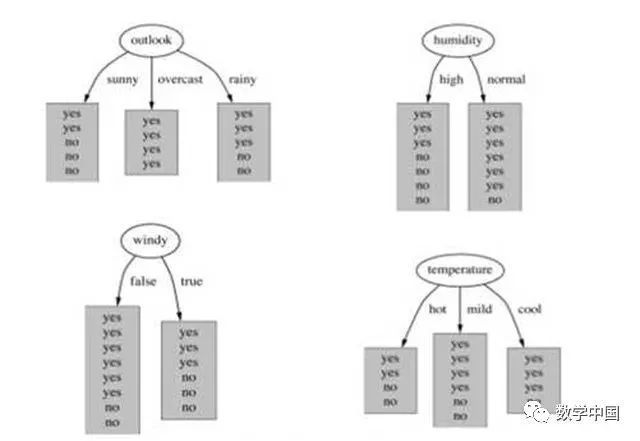
E.以上所有

答案：B

如果模型偏差大，说明模型相对过于简单。我们可以在特征空间中增加更多的特征来提高模型的鲁棒性。增加数据点也会减少方差。

问题8：

当我们建立基于决策树的模型时，我们将有最高信息增益的节点分离出来作为属性， 在下图中，哪一个属性有最高的信息增益？



A. Outlook

B. Humidity

C. Windy

D. Temperature

答案：A

信息增益随子集平均纯度的增加而增加。要了解信息增益的计算，点这里阅读。你也可以查看这张幻灯片。

问题9：在决策树中，当一个节点分叉的时候，以下关于“信息增益”正确的是？

不纯的节点越少，越需要更多的信息来描述种群

信息增益可以用熵作为“1-Entropy”来推导

信息增益偏向于数值大的属性

A. 1

B. 2

C. 2 和3

D. 都正确

答案：C

想了解详情，请阅读这篇文章和这个幻灯片。

问题10：使用SVM模型遇到了欠拟合的问题，以下哪个选项能提高模型性能？

A.增加惩罚参数“C”

B.减少惩罚参数

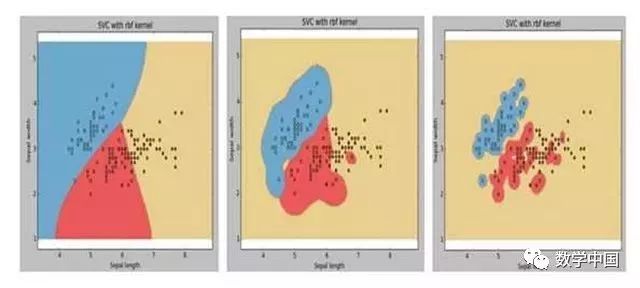
C.减少核系数（gamma的值）

答案：A

如果是欠拟合情况，我们需要增加模型的复杂性，如果我们增大C，则意味着决策边界变复杂，所以A是正确答案。

问题11:

假如我们已经画出SVM算法中的不同点的gamma值（Kernel coefficient）。但由于一些原因，我们没有在可视化界面中显示出来。在这种情况下，以下哪个选项最好的解释了三张图的gamma值关系（图中从左向右分别是图1,、图2、图3，对应的gamma值分别是g1,、g2、g3）。



A. g1 > g2 > g3

B. g1 = g2 = g3

C. g1 < g2 < g3

D. g1 >= g2 >= g3

E. g1 <= g2 <= g3

答案：C

如果gamma值高，则会尽全力去拟合数据集中的每一条数据，会带来过拟合问题。所以最合适的选项是C。

问题12：

我们在解决一个分类问题（二值分类预测）。然而，我们并不是要得到0或1的真实结果，而是要获取每一个类的概率。现在假设我有一个概率模型，并且使用一个0.5的阈值来预测结果，如果概率大于等于0.5，则认为是1，如果小于0.5，我们则认为是0。如果我们使用一个比0.5高的阈值，一下哪条最合适？

增加阈值那么分类器会有相同或者更低的查对率

增加阈值分类器会有更高的查对率

增加阈值会有相同或者更高的准确率

增加阈值会有更低的准确率

A. 1

B. 2

C. 1和 3

D. 2和 4

E. 无

答案：C

想了解调整阈值对查对率和准确率的影响，请参考这篇文章。

问题13：

当使用比例失调的数据（数据集中99%的negative class和1%的positive class）进行“点击率”预测的时候，假如我们的准确率是99%，那么我们的结论是？

准确率很高，我们不用再做任何工作。

B.准确率不够好，我们需要尝试构建一个更好的模型

C.无法判断这个模型

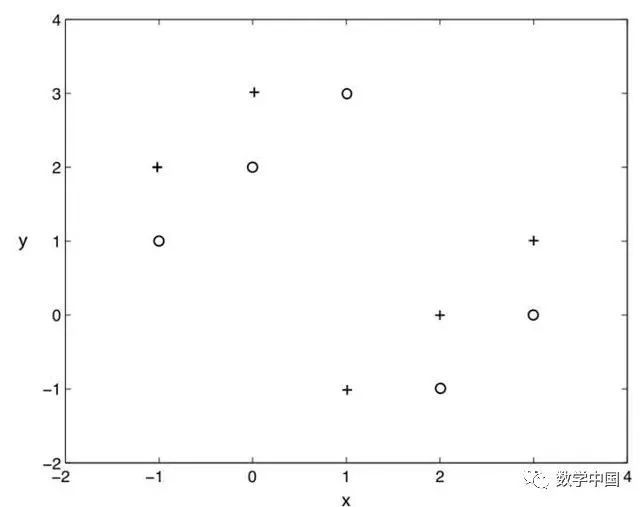
D.以上都不正确

答案:B

当使用不平衡数据集的时候，准确率不能作为性能的指标，因为99%（正如上文提到的）可能只是预测多数类别正确，但是往往重要的是少数的类（1%）。因此，对于这样的模型，我们应该使用敏感性和特异性来描述分类器的性能。如果占少数的类别预测不准的话，我们需要采取一些必要的措施。更多不平衡分类问题，可以参考这篇文章。

问题14：

比方说，我们使用KNN训练观测数据较少的模型（以下是训练数据的快照，x和y分别表示两个属性，“＋”“o”分别表示两种标签）。已知k=1，leave one out 交叉验证的错误率会是多少。



A. 0%

B. 100%

C. 从0 到 100%

D. 以上没有

答案：B

在Leave-One-Out交叉验证中，我们选取n-1条观测量作为训练集，1条观测量用来验证。如果把每个点作为交叉验证点并找到最近的点总会得到相反的类别。所以错误率是100%。

问题15：

当我们使用大数据集进行决策树训练的时候，一下哪个选项可以用来减少训练时间？

增加树的深度

增加学习率

减少树的深度

减少树的数量

A. 2

B. 1 and 2

C. 3

D. 3 和 4

E. 2 和 3

F. 2, 3 和 4

答案：C

如果决策树的参数是固定的话，我们可以考虑以下的选项。

增加深度会导致所有节点扩张，因此增加深度会导致时间变长。

在单一决策树种，学习率是不能作为一个可调整的参数的。

使用决策树时，我们只会建一颗树。

问题16：

关于神经网络，一下那种说法是正确的？

在测试数据中增加层数可能会增加分类错误

在测试数据中减少层数总会减少分类错误

在测试数据中增加层数总会减少分类错误

A.1

B.1和3

C.1 和2

D. 2

答案：A

通常来说，增加层数会让模型更加一般化，所以它将会在训练集和测试集上都表现更优异。但这个不是真理，在这篇文章中，笔者发现深层的网络比浅层的网络有更高的错误率。所以选项2和3都是错误的，因为这个假设不总是正确的，而1只是说可能是正确的。

问题17：

假设我们使用原始的非线性可分离SVM最优化目标函数，我们做什么可以保证结果是线性可分的？

A. C = 1

B. C = 0

C. C = 无穷大

D.以上没有正确答案

答案C：

如果我们使用原始的非线性可分离SVM最优化目标函数，我们需要将C设置成无穷大来保证结果是线性可分的。因此C是正确答案。

问题18：

训练完SVM之后，我们可以丢掉所以不支持向量的样本而不影响对新的样本进行分类。

A.正确

B.错误

答案：A

这是正确的，因为只有支持向量会影响边界。

问题19：

以下哪些算法可以借助神经网络来构建？

K-NN

线性回归

逻辑回归

A.1 和2

B.2 和 3

C.1, 2 和 3

D.无

答案：B

KNN是一个机遇实例的学习方法，它没有用来训练的参数，所以它不能用神经网络来构建

神经网络的最简单形式就是最小二乘回归。

神经网络和逻辑回归有关。主要在于，我们可以把逻辑回归看成一层神经网络。

问题20：

请选择可以用来实施隐马尔可夫模型的数据集。

A.基因序列数据集

B.电影评论数据集

C.股票价格数据集

D.以上所有

答案D：

以上所有的数据集都可以用隐马尔可夫模型。

问题21：

我们想在在一个百万级的数据集上构建机器学习模型，每条数据有5000个特征。可是训练这么大量的数据集会有很多困难，一下哪些步骤可以有效地训练模型？

A.我们可以从数据集中随机选取一些样本，在样本上构建模型

B.我们可以尝试联机机器学习算法

C.我们可以用主成分分析来减少特征

D.B和C

E.A和B

F.以上所有

答案:F

在一个内存有限的机器上处理高维数据是一项非常费力的工作。以下的方法可以用来应对这样的情况。

我们可以采取随机采样的方式，这意味着我们创建一个更小的数据集。举个例子来说，我们可以抽取300000条数据，每条有1000个特征然后再进行计算。

我们可以使用展示在Vowpal Wabbit中的联机学习算法

我们可以使用主成分析来选取能反映最大方差的部分。

因此所有的都是正确的。

问题22：

我们想减少数据的特征，以下哪些做法是合适的？

使用预选的方式

使用向后消除的方式

首先使用所有特征来计算模型的准确度。我们选择一个特征，然后将测试集的该特征的数值都打乱，然后对打乱过的数据集进行预测。经过对预测模型的分析，如果模型准确率提高，则删掉这个属性

查找关联性表，关联性高的特征就可以去掉

A. 1和2

B. 2, 3和 4

C. 1, 2 和4

D. 以上所有

答案：D

l预选和向后消除是特征选择的两个常用的主要方法。

l如果不用上面的两种方法，我们也可以选择3中所说的方法，这种方法应对大数据量时非常有效。

l我们也可以使用基于特征选择的关联分析，然后去除共线性特征。

问题23：

关于随机森林和梯度提升树，请选择正确的选项。

在随机森林中，中间树互相不独立，而在梯度回归树中，中间树相互独立。

他们都使用随机特征子集来构建中间树。

在梯度提升树的情况下我们可以生成并行树，因为树互相独立。

梯度提升树在任何数据集上都比随机森林要好。

A. 2

B. 1 和 2

C. 1, 3 和 4

D. 2 和 4

答案 A：

随机森林是基于bagging而梯度提升是基于boosting

这两种算法都使用随机特征子集来构建中间树

由于随机森林的中间树互相独立，因此算法可以并行，而梯度提升树则不可能。

这个不是绝对的，不同数据结果不同。

问题24：

对于主成分析转换的特征，朴素贝叶斯的基本假设是成立的，因为主成都是正交的，因此是无关的。这句话正确么？

A.正确

B.错误

答案：B

这句话是错误的。首先，不相关不等同于独立。第二，转换过的特征也不一定是不相关的。

问题25：

以下关于主成分析哪些是正确的？

在PCA前必须将数据标准化

我们应该选择说明最高方差的主成分

我们应该选择说明最低方差的主成分

我们可以用PCA来可视化低维数据

A. 1, 2 和4

B. 2 和 4

C. 3 和 4

D. 1 和 3

E. 1, 3 和 4

答案:A

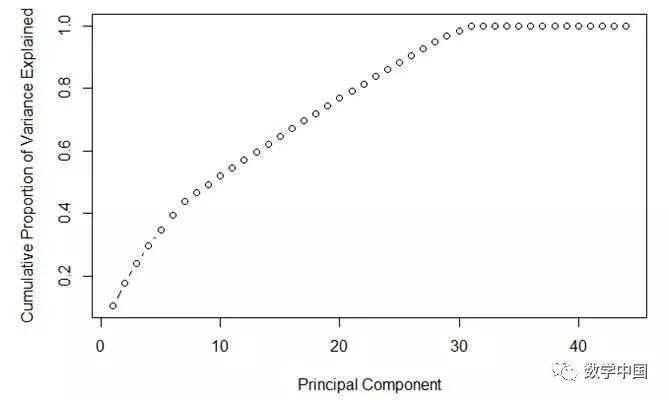
lPCA 对数据中变量的大小是敏感的，所以在PCA之前必须要对数据进行标准化。举个例子，如果我们将一个变量的单位从km改成cm，该变量可能会从影响很小一跃成为主成分。

l第二条是正确的，因为我们总是选择最大方差的主成分。

l有时候用低维来画出数据是十分有效地。我们可以选择前二的主成分，然后用散点图来描绘数据。

问题26：

在下图中的主成分的最佳数目是多少？



A. 7

B. 30

C. 35

D.不确定

答案：B

在上图中，成分数为30时候就达到了最大的方差，所以选择B

问题27：

数据科学家经常使用混合算法来做预测，然后将混合算法的结果合并（也叫集成学习）。这些混合算法的输出更加健壮且具有一般性，而且比任何一种单一模型都要准确。以下哪些选项是正确的？

A. 基础模型有更高的相关性

B. 基础模型有更低的相关性

C. 使用平均加权而不是投票的方式来集成

D. 基础模型源自相同的算法

答案B：

请参阅下面的集成指南来了解细节

Basics of Ensemble Learning Explained in Simple English

Kaggle Ensemble Guide

Easy questions on Ensemble Modeling everyone should know

问题28：

我们如何在监督的机器学习挑战使用聚类方法？

我们可以先创建簇，然后在不同簇中分别使用监督机器学习算法。

我们在使用监督机器学习算法之前可以把簇的id作为特征空间中额外的特征。

我们无法在使用监督机器学习算法之前创建簇。

我们在使用监督机器学习算法之前不能把簇的id作为特征空间中额外的特征。

A. 2和4

B. 1和2

C. 3和4

D. 1和3

答案：B

l我们可以在不同的簇中使用不同的机器学习模型，这样一来，预测的准确性可能会提高。

l增加簇的id可以提高预测的准确性，因为id是对数据很好的概括。

因此B是正确的。

问题29

以下的说法哪些是正确的？

一个机器学习模型如果能得到很高的 准确率，则说明这是个好的分类器。

如果增加一个模型的复杂度，测试错误总会增加。

如果增加一个模型的复杂度，训练错误总会增加。

A. 1

B. 2

C. 3

D. 1和3

答案C：

当类不平衡的时候，准确率不是一个很好的评价指标。而 precision 和recall是最好的评价方式。

增加一个模型的复杂度可能会导致过拟合。而过拟合会引起训练错误的减少和测试错误的增加。

问题30：

以下有关于梯度回归树算法的说法正确的是？

当我们增加用于分割的最小样本数时，我们总是试图得到不会过拟合数据的算法。

当我们增加用于分割的最小样本数时, 数据会过拟合。

当我们减少用于拟合各个基本学习者的样本的分数时，我们总是希望减少方差。

当我们减少用于拟合各个基本学习者的样本的分数时，我们总是希望减少偏差。

A. 2和4

B. 2和3

C. 1和3

D. 1和4

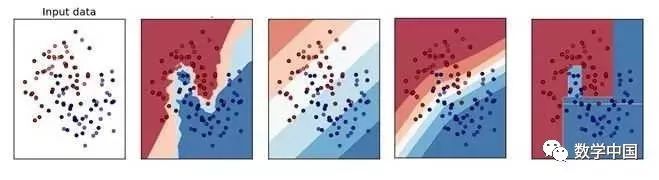
答案: C

最小化样本的数量，在分裂节点的地方用于控制过拟合, 太高的数值会导致欠拟合因此应该用CV来进行调整.

每棵树选择观测值的分数是通过随机采样的方式来做的。如果数值比1小一点点则会使模型健壮，而且方差也会减小。典型的数值是0.8，当然，也要根据实际情况微调。

问题31：

以下哪个是KNN算法的决策边界？（下图从左到右分别是A,B,C,D）



A) B

B) A

C) D

D) C

E) 不确定

答案：B

KNN算法的原理是为观测变量寻找K个最近邻居，将邻居中的多数的标签赋给观测变量。所以决策边界不会是线性的。因此，选择B。

问题32：

如果一个机器学习模型在测试集上获得的100%的准确性，是否意味着在新的测试集上也能获得100%的准确性。

A.是的，因为这个模型足够一般，可以适用于所有类型的数据

B.不是，仍然有模型不能控制的因素，比如噪声。

答案：B

答案选择B，因为实际数据不可能没有噪声，所以不可能得到100%的准确性。

问题33：

以下是交叉验证的常见方法：

i. Bootstrap with replacement.

ii. Leave one out cross validation.

iii. 5 Fold cross validation.

iv. 2 repeats of 5 Fold cross validation

如果样本的数量是1000，那么这这四种方法执行时间的排序是？

A. i > ii > iii > iv

B. ii > iv > iii > i

C. iv > i > ii > iii

D. ii > iii > iv > i

答案：B

Bootstrapping是一个统计的技术，属于广泛的重采样的范畴，所以只有1个验证集使用了随机采样。

Leave-One-Out cross validation的时间最长，因为我们要n次训练模型（n是观测值的数量）

5 Fold cross validation 会训练五个模型，而训练时间和观测值数量无关。

2 repeats of 5 Fold cross validation则是训练10个模型。

因此答案选择B。

问题34：已取消

问题35：

变量选择旨在选择预测变量的“最佳”子集。当我们选择变量的时候，考虑到系统的性能，我们需要注意些什么？

类似的多个变量

模型的可解释性

特征信息

交叉检验

A. 1和4

B. 1, 2和3

C. 1,3和4

D. 以上所有

答案：C

如果几个变量具有很高的xiang s，则会展现出共线性。

相对于模型的性能，我们不需要关注模型的可解释性。

如果特征有很高的信息，则会为模型带来价值。

我们需要使用交叉检验来验证模型的普遍性。

因此C是正确答案。

问题36：

线性回归模型中的其他变量下列哪些语句是正确的关于？

R-Squared和Adjusted R-squared 都会增长

R-Squared 是常数，Adjusted R-squared 会增长

R-Squared 和Adjusted R-squared 都会减少 4.R-Squared 减少而 Adjusted R-squared 增长

A. 1和2

B. 1和3

C. 2和4

D.以上没有正确的

答案: D

R-squared 不能确定系数估计和预测是否有偏差，这就是为什么我们要评估残差图。 Adjusted R-squared 是R-squared的增强版，该方法调整了模型中预测器的数量. 如果有新方法将模型改进的几率大于预期时，Adjusted R-squared 会增加。 当预测变量将模型改进的几率小于预期时，它减少。

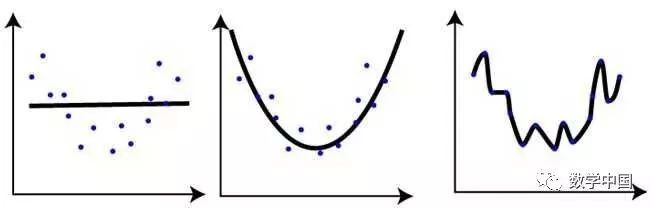
但是 R-squared 比adjusted R-squared 有更多的问题，因此predicted R-squared被提出。

如果为模型增加一个预测器，则R-squared会保持不变或者增加。

想讨论更多，请点击这里。

问题37：

下图我们画出了在同一个数据集上解决回归问题的三种不同模型，从下图中我们可以总结出什么？



和其他的相比，第一张图的训练错误最大。

最后一个是最好的模型，因为在第三张图有最小的训练错误。

第二个模型比第一个和第三个更健壮，因为它能更好的处理不可预见的数据。

第三个模型和前两个相比属于过拟合。

所有模型的性能都一样，因为我们还没有看到测试集。

A. 1和3

B. 1和4

C. 1, 3和4

D. 5

答案：C

对于自变量X来说，图中数据的趋势像一个多项式函数。最右边图中的多项式形式更复杂，准确率也最高，但是对于测试集则会表现不佳。而最左边的图很明显属于欠拟合的情况。

问题38：

在应用线性回归时我们需要遵循哪些假设？

由于线性回归对于异常值很敏感，所以检查异常值是十分重要的。

线性回归要求所有变量都遵循正态分布。

线性回归假设数据中很少或不存在多重共线性。

A. 1和2

B. 2和3

C. 1,2和3

D. 以上所有

答案：D

l异常值是数据中对最终回归线的斜率影响最高的点。所以在回归分析中去除离群值总是很重要的。

l了解自变量的分布是非常必要的。自变量的正负偏态分布可以影响模型的性能，并将高度偏态的自变量转换正态将改进模型性能

l当模型包含彼此相关的多个要素时，会出现多重共线性。换句话说就是有多余因素线性回归假设在数据中应该有很少冗余或者尽可能没有。

问题39：

当建立线性模型的时候，我们会关注数据之间的关联。假如我们在关联矩阵中找到了三对数据(Var1和Var2 , Var2和Var3 , Var3和Var1) 的关联分别是 -0.98, 0.45 and 1.23。那么从这些信息中我们可以推断出什么？

Var1和Var2具有很高的关联性。

Var1和Var2有很高的关联度，所以它们具有多重共线性。所以我们要将Var1或者Var2移除出我们的模型。

Var3和Var1之间的关联值是1.23是不可能的。

A. 1和3

B. 1和2

C. 1,2和3

D. 1

答案: C

lVar1和Var2的关联值很高，所以是具有多重共线性的，因此我们可以从中去除一个。

l一般来说，关联系数高于0.7的说明数据具有多重共线性。

l第3个是不可能的，关联系数一定会在-1和1之间。

问题40：

如果独立和不独立的变量之间有很高的非线性且复杂的关系，那么一个树模型将会比一般的经典回归有更好的效果。这个说法正确么？

A.正确

B.错误

答案：A