1. Part 1
   1. 评估课程作业问题 1

你好，

对于评估课程中的问题 1，什么时候应该删除属性“fnlwgt”？我们应该在函数 read\_csv\_1 中这样做吗？我可以假设在 1b 中，属性名称列表不应包含属性“fnlwgt”吗？

我对函数 read\_csv\_1 的用途也有些困惑，据我了解，它以 data\_file（此处为成人数据集）作为输入，但也输出相同的数据文件。此外，由于我计算机上数据集的路径肯定与您的不同，我应该如何编写代码使其也可以在您的计算机上运行？

亲切的问候，

迈克尔

亲爱的 Michael，  
  
您需要回答以下问题：  
- 根据课程作业描述，您什么时候应该删除“fnlwgt”属性？为什么？  
- read\_csv\_1 函数的输入是什么，它的输出是什么？  
  
亲切的问候，  
迪米特里斯

1. 2**关于课程作业任务 1 中的属性 fnlwgt**

亲爱的迪米特里奥斯，

在任务 1 的说明中，它说“对于这部分，应删除属性 fnlwgt，并应考虑以下属性”。

您的意思是我们在读取 csv 时需要删除此列吗？因为它会影响“num\_rows”等其他一些函数的输出结果。

最好的祝福，

Yunyue chen

之前有人回答过，但我仍然不确定，希望得到这个答案。我认为我们将其删除，然后使用删除的属性计算练习的其余部分。我的问题是，当单元测试正在测试 read\_csv\_1 时，它是否期望“fnlwgt”属性出现在数据中？  
  
如果不是，这意味着对于每个其他函数，我们必须在继续之前删除“fnlwgt”，因为原始读取没有。

亲爱的 Yunyue 和 Jagdeep，  
  
由于要求删除“fnlwgt”，您可以假设它与您想要执行的任务无关。因此，没有理由将其存储在内存中。因此，它不应出现在 read\_csv\_1 函数的输出中。我的期望是你自己可以做出这种推理。这同样适用于课程作业的所有问题。  
  
亲切的问候，  
迪米特里斯

### 1.3 连续波不一致

你好，

在“bachelors\_masters\_percentage”函数的adult.py 文件中，它表示将百分比保留到小数点后3 位。但是，在课程作业说明中，它要求保留 1 位小数。

我们应该遵循课程作业说明还是模板中的评论？

谢谢你。

亲爱的 U Ri，  
  
感谢您发现这一点。请使用 3 位小数精度。  
  
亲切的问候，  
迪米特里斯

### 1.4课程数据集成人的一些问题

你好

我在adults.cvs中发现了一些问题，每个实例的age属性都是一位数，有的甚至是0。但是在原网站的数据集中都是正常的年龄。我想知道这个课程作业是否应该根据adult.cvs中的数据完成？

来自成人.cvs 的年龄数据

年龄数据形成 UCI 机器学习存储库

亲爱的微宝，  
  
请按照课程作业说明：  
  
“本课程作业所需的数据集在模块的 KEATS 页面中提供。请勿从其原始来源下载它们以用于您的课程作业实施。提供指向其来源的链接仅供参考。”  
  
亲切的问候，  
迪米特里斯

### 1.5 CW one\_hot\_encoding

嗨，只是想知道我们应该如何命名新列，因为我们将有 3 列或更多列，而不是一列。例如。年龄列将变为age\_1,age\_2,age\_3.....另外，我们应该删除原始列吗？

我也有同样的问题。我们是否需要“仅使用属性值”以便我们将获得年龄或 get\_dummies() 中的默认方式的“1, 2, 3, ...”？（这样我们就会得到“age\_1，age\_2，age\_3，...”）  
  
我不认为“仅使用属性值”会起作用，因为我们有多个属性的值为 1 或 0。例如。年龄和资本收益

嗨易遥和云月，  
  
在执行一次热编码后，一个分类属性确实被多个属性替换，每个原始属性值一个，例如 get\_dummies()。因此，如果您明确新属性代表什么，例如age\_1 等，这将很有用。  
  
创建新属性后，是否需要旧列？  
  
亲切的问候，  
迪米特里斯  
  
以我的理解，以“age”为例，age\_1表示age属性为1的原始数据。因此如果数据集中有4种数据，“age”分别为1、2、3、4，那么使用一种后-hot 编码，它将有 4 个新列，名称分别为 age\_1、age\_2、age\_3 和 age\_4。此外，由于我们使用 one-hot 编码代替了原始编码，因此将删除名为“age”的原始列。  
  
我认为如果我们将新数据添加到数据集中，并且这些数据与所有现有数据相比具有不同的属性值，那么 one-hot encoding 将需要一个新列。  
  
至于列名，我觉得只用get\_dummies()的默认设置就完美了，因为易遥说的问题。

### 1.6 CW 几项澄清

* 第一季度

1. missing\_values(df)：我们应该返回 pandas 如何显示每列缺失值的默认设置，还是只返回一个数字，其中包含所有列中缺失实例的总和？
2. one\_hot\_encoding(df)：我们是否应该删除男性和女性等二元属性之一？我们还应该保留education\_num，因为它与教育100％相关吗？

不确定我是否理解您关于缺失值的问题。missing\_values(df) 计算数据中所有缺失值的数量。正如我们在讲座/教程中提到的，如果给定实例的给定属性没有值，那么这应该为计数贡献 1。

- one\_hot\_encoding(df)：执行一次热编码后，保留您创建的所有新属性。不再需要原来的分类属性。对于某些人来说，性别属性不是二元的，所以请保留男性和女性。不要求您执行特征选择，因此请保留所有应该创建的新属性。

1.7 关于第 3 部分的问题。是否应返回 df 副本

Part3 大部分功能是直接在数据框上编辑的。我们应该创建一个数据框的副本并修改副本值并返回它，还是我们应该只修改原始数据框？

例如，

def remove\_non\_alphabetic\_chars(df):

            df['xxx'] = df['xxx'].str.replace('[^a-zA-Z]', ' ', 正则表达式=True)

要么

def remove\_non\_alphabetic\_chars(df):

            dfc=df.copy()

            dfc['xxx'] = dfc['xxx'].str.replace('[^a-zA-Z]', ' ', 正则表达式=True)

             返回 dfc

亲爱的易遥，  
  
感谢您的留言。原则上，这里给您的建议是不要在不需要时创建数据副本，因为这可能会占用大量内存。修改原始数据框更合适。这也是您希望在您作为数据科学家的可能职业中采用的建议。在这个特定的练习中，您还可以通过构建新的数据框对象获得一些运行时间。  
  
亲切的问候，  
迪米特里斯

亲爱的 Dimitris，  
  
我们不应该对每个问题都使用副本吗？在问题1中，当我们进行一次热编码和标签编码时，如果在标签编码之前进行了一次热编码，则从数据帧中删除了属性“类”，因此无法进行标签编码。我们应该在 one hot encoding 函数中进行复制，还是我们可以假设在测试文件中，标签编码将在 one hot encoding 之前执行以避免这个问题？  
  
亲切的问候，  
  
迈克尔

亲爱的 Michael，  
  
one\_hot\_encoding() 函数可以返回仅包含输入属性的 X\_train，而 label\_encoding() 函数可以返回仅包含目标属性的 y\_train。然后可以由 dt\_predict() 使用这些。  
  
亲切的问候，  
迪米特里斯

### 1.8CW 第 1 部分 q1 和 q2

亲爱的 Letsios 博士，

从 CW 第 1 部分开始，我有 4 个问题：

1. CW 中的 q1 提到“例如，如果百分比是 0.21547%，那么函数应该返回 0.216” - 这是否意味着预期结果必须是 0.216？

2. 虽然指令 pdf 中的 q2 要求我们“使用 one-hot 编码将****所有属性（类****除外）转换为数字”，但 CW 程序中的指令说这应该通过“使用 one-热编码”。扫描提供的成人.csv，所有属性似乎都不是分类的。我们应该遵循哪一个？

3. Q2 还要求我们“****下一步****，将类值转换为带有标签编码的数字”。我们是否希望在 label\_encoding() 中调用one\_hot\_encoding *()* *以便*可以按顺序完成？

4. 根据 CW 程序中的指令，label\_encoding() 必须返回一个序列“仅包含一列，其中 df 实例的标签使用标签编码转换为数字”。虽然指令 pdf 要求我们“将类值转换为带有标签编码的数字”，就好像必须返回整个数据帧一样，但是否应该只返回编码的类列？

你能澄清这些吗？

谢谢你。

亲爱的 Soyoung，  
  
感谢您的留言。  
1. 是的。  
2. 在课程作业的第 1 部分中，所有属性都被视为分类属性。甚至教育年数也被认为是绝对的。当然，教育年限也可以被认为是名义上的，甚至是数字上的。但是，在这个特定的课程中，请假设它是分类的，并使用 one-hot 编码将其转换为数字。  
3. 有一个名为“one\_hot\_encoding”的特定函数应该将分类属性转换为数字。我\*不\*期望你会在另一个函数中使用这个函数。如说明中所述，评估函数工作情况的正确方法是创建一个单独的文件，您将在其中通过导入和调用函数实现来执行所有操作。  
4. 函数 label\_encoding 应该得到一个带有数据集的 data\_frame 并创建一个带有 target 属性的新向量（熊猫系列），其中所有标签都将被标签编码。  
  
希望以上对您有所帮助。  
  
亲切的问候，  
迪米特里斯

### 1.9 Q1 - dt\_predict() & dt\_error\_rate()

你好教授，

所以我的理解是我们需要在 ****dt\_predict(X\_train,y\_train)****中拟合树，因此我们将使用训练样本进行拟合，然后我们被要求预测“ ****X\_train**** ”而不是通常的“ ****X\_test**** ”，但这很好部分。

但是，当我们移动到 ****dt\_error\_rate(y\_pred,y\_true)****时，我们需要计算错误率，我会假设“ ****y\_pred**** ”将从****dt\_predict(X\_test,y\_test)****生成，但这会改变拟合。

决策树的传统伪代码如下所示：

 - 适合（X\_train，y\_train）

 -****预测（X\_test）****

 - 错误（ ****预测（X\_test）****，y\_test）

但是根据当前的说明，它将是：

 -****适合（X\_test，y\_test）****

 - 预测（X\_test）

 - 错误（预测（X\_test），y\_test）

或假设我对如何生成“ ****y\_pred**** ”的假设是错误的：

 - 适合（X\_train，y\_train）

****- 预测（X\_train）****

 - 错误（ ****预测（X\_train）****，  ****y\_train****）

我不确定我想说的是否清楚，但我认为由此产生的错误率不是真实的。

谢谢你的时间

亲爱的 Fadel，  
  
感谢您的留言。请在下面找到您的评论/查询的一些答案：  
  
- 正确，使用训练集进行拟合是课程作业第一部分的期望。  
  
- Q1 的必修课标记部分不要求拆分为培训和测试。在这里，您被要求仅计算训练误差。目标是使用整个提供的数据集作为您的训练集。使用此数据构建模型后，您可以使用对训练集中实例的预测来计算 y\_pred 向量。计算训练误差相当于比较 y\_pred 和 y\_true，其中 y\_true 是训练数据中的真实标签。  
  
- 计算测试错误是您课程作业的一个可选的、未标记的部分。很高兴看到您也对这部分感兴趣。在这种情况下，您确实需要将原始数据分成两部分（X\_train，y\_train）和（X\_test，y\_test）。我们在讲座和教程中讨论了如何进行这种拆分的各种方法。向量 y\_train 和 y\_test 分别包含 X\_train 和 X\_test 中实例的真实标签。然后，您可以使用您的训练集 X\_train 构建模型。完成后，您可以使用您的模型为您的训练集进行预测 y\_pred\_train = predict(X\_train) 并为您的测试集进行预测 y\_pred\_test = predict(X\_test)。通过将 y\_pred\_train 与 y\_train 进行比较，您可以计算出训练误差。通过比较 y\_pred\_test 和 y\_test，  
  
- 鉴于上述情况，在提供的模板文件中正确计算了训练误差。  
  
- 您确实是对的，执行测试的全部目的是评估模型对未见实例的泛化程度。训练误差通常不能很好地估计我们在未见过的实例上所期望的真实误差。原则上，训练误差应该较低，因为它评估模型对已用于训练的实例的预测质量。这些预测往往是好的，因此是有偏见的。所以，我们保留一些实例作为我们的测试集。这些实例不用于训练，因此可以更好地估计我们的模型的泛化程度。

- 执行训练和测试的伪代码可能如下所示：

(X,y)：原始数据

(X\_train, y\_train, X\_test, y\_test) = split(X,y) 分成训练集和测试集。

模型=火车（X\_train）

y\_pred\_train = model.predict(X\_train)

y\_pred\_test = model.predict(X\_test)

train\_error = (y\_pred\_train, y\_train)

test\_error = (y\_pred\_test, y\_test)

- 您还可以使用交叉验证来计算对未见实例的真实错误的更好估计，方法是将不同的方式拆分为训练和测试。

亲切的问候，  
迪米特里斯

1.10CW 问题 1

嗨迪米特里奥斯，

希望你一切安好，保持安全。只是想问一个关于课程作业中第一个问题的快速问题；dt\_predict 函数的注释说应该对 X\_train 的标签进行预测。但是，我们所做的所有示例都使用 X\_test 进行预测。课程作业中是否存在拼写错误，或者要求是否存在差异？

非常感谢！

亲爱的萨普，  
  
感谢您的提问。事实上，课程作业描述和模板文件都要求评估训练错误。  
  
评估测试错误是可选的，没有标记，并且可以让您更好地评估结果模型的泛化程度。  
  
亲切的问候，  
迪米特里斯

2.Q2

2.1CW 几项澄清

1. KMeans 和 Kmeans++：sklearn 中 KMeans() 的默认设置是 init = "k-means++" 那么我们是否应该为 KMeans() 设置为 init = "random"？
2. 在“可视化上一个问题中计算的****最佳集群集****”的描述中 - 我假设 - 从 ****best\_clustering\_score(rdf) 生成。****虽然在评论中它说“****使用 k=3 运行您选择的****一些聚类算法并生成散点图”，这不一定对应于最好的分数？

- 在您的 kmeans 函数实现中，您可以根据需要调整它，只要该函数产生请求的输入和输出。但我建议您使用最基本的实现，即使没有随机化。在填写性能评估矩阵时，您有机会研究在随机化和不随机化的情况下 kmeans 的行为方式。  
  
- 当您可视化您的集群时，为此运行任何集群算法就足够了。但是，我建议您使用具有最佳轮廓分数的算法运行，以便手动检查它分离数据点的程度。

### 2.2Coursework\_Q2\_standardize

亲爱的迪米特里奥斯博士，

关于问题 2 ****def standardize(df)****，我们是希望从****summary\_statistics(df)****中获取先前的函数结果，还是从原始数据帧中获取？如果换新的情况属实，是否还期望在标准化过程中遵循“将均值和标准差四舍五入到最接近的整数”的指导？

最好的祝福，

KHC

亲爱的宽浩，  
  
标准化应该应用于原始数据。在对具有数字属性的一组数据点进行聚类时，如果属性以不同的尺度测量，则生成的聚类集可能会很弱。例如，如果您的实例是体重以公斤为单位，身高以毫米为单位的人，则这两个属性的值将显着不同。标准化允许以相同的比例表示输入属性，因此有助于不过度重视大规模测量的属性。  
  
亲切的问候，  
迪米特里斯

2.3 关于cw的一些问题

你好，教授

我对 cw 中的一些澄清有一些疑问。

在第二季度，

对于函数 summary\_statistics，我们被要求“将平均值和标准差四舍五入到最接近的整数”。我们应该使用 ''round( , 0)'' 还是 ''int''，这将导致不同的结果（例如 2 或 2.0）。

### 2.4 课程 Q2：K-Means 和凝聚函数输入类型

嗨迪米特里奥斯，

希望你安然无恙。对于课程的第二部分，您能否解释一下如何将 k 数赋予函数“kmeans”和“agglomerative”？课程作业 pdf 文件说它将是 3、5 和 10，但函数描述似乎只显示一个输入值。换句话说，函数（kmeans 和凝聚）中的输入“k”是单个整数（即 3、5 或 10），然后将其放入分类器，还是像一个列表：k= [3,5,10]，因此我们需要遍历列表中的每个数字？非常感谢您的帮助。

亲切的问候，

萨普

嗨 Dimitris，  
同样对于 Q2，它在说明中说我们应该“使用剪影分数作为评估指标来确定哪个运行导致了最佳集群集”，而在 python 脚本中它说“返回最佳计算剪影分数'。我们是否应该返回最佳运行的轮廓分数，或者哪个运行最好（例如，kmeans 标准化 k = 5）。如果是后者，我们应该怎么写呢？也许是“Kmeans\_Standardized\_5”或类似的东西？  
亲切的问候，  
迈克尔

亲爱的 Sarp 和 Michael， @Sarp  
  
函数 kmeans(df,k) 和 agglomerative(df,k) 接收带有要聚类的实例的数据帧，并使用欧几里德距离度量对它们进行聚类。如果将其作为值传递给参数 k，它们可以产生任何合理数量的集群，例如 3、5 或 10。这些函数接收一个 k 值作为输入参数，例如 k=3。函数 cluster\_evaluation(df) 采用数据帧并使用不同的 k 参数值执行数值评估。实际上，您可以创建一个列表 k=[3,5,10] 来执行此函数内请求的 k 值的分析，但不能在其他两个函数中执行分析。  
  
@Michael 函数 best\_clustering\_score(rdf) 应该返回评论中指定的最佳计算轮廓分数。这是要求的。您当然可以按照与您指示的方式类似的方式，将计算出最佳分数的方法计算为可选的、未标记的任务。通过这种方式，您可以反思我们在讲座中讨论的聚类方法的性能。  
  
亲切的问候，  
迪米特里斯

2.4 约 10 次 k-means 的执行，CW 任务 2

亲爱的迪米特里奥斯，

在课程作业的说明（任务 2 问题 2）中，我们被要求“对每个 k 值重复 10 次 k-means 执行”。

据我了解，在运行 10 次 k-means 后，我们将得到 10 个聚类结果。我们是否需要只保留这些执行中的最佳剪影分数（此设置仅 1 行），还是需要将它们全部保存在函数返回的数据框中（相同设置为 10 行）？

最好的祝福，

云月

亲爱的云月和银瑶，  
  
感谢您的留言。  
  
@Yunyue 当您实现 kmeans(df,k) 函数时，您只需要执行一次 kmeans。当您实现 cluster\_evaluation(df) 函数时，您需要对集合 {3,5,10} 中的每个 k 值执行 10 次 kmeans 执行。函数 cluster\_evaluation(df) 要求创建一个包含性能评估结果的新数据框。在此数据框中，模板文件要求为您的每次运行填充一行，从而报告每次运行的剪影分数。  
  
@Yiyao 感谢您指出如何使用 sklearn 运行具有多个初始质心的 kmeans。这可以在 cluster\_evaluation(df) 函数中使用。但是，在 cluster\_evaluation(df) 中手动执行每次运行并且不使用 sklearn 的批处理执行功能以便能够填充性能评估矩阵可能很有用。  
  
希望这会有所帮助。  
  
亲切的问候，  
迪米特里斯

2.5   
关于课程作业 Q2 和 Q3 的问题

嗨，Letsios 博士，

希望你一切顺利。

我对课程作业有一些疑问。问题如下。

1. 在挖掘文本数据中，我们需要词干。我想知道是否有标准的词干分析器供我们使用，因为据我所知，不同的词干分析器会给出不同的结果。

例如，如果使用 PorterStemmer，“冠状病毒”将“冠状病毒”进入航空公司，而如果使用 SnowballStemmer，则将“冠状病毒”转化为“冠状病毒”。

因此，它会对结果产生影响。

2. 在聚类分析中，我们应该为每个 K 值重复 10 次 kmeans。

我们要在函数内重复 10 次吗？或者 main 函数会调用我们的 kmeans 函数 10 次。

如果我们需要在我们的函数中重复 kmeans 过程，我们是否应该调用 best\_clustering\_score 函数来选择最好的并返回它？

提前感谢您的回答！

此致，

金宏

亲爱的 Jinhong，  
  
1. 事实上，在预处理文本数据时，不同的词干分析器可能会给出不同的结果。很高兴看到您的推文预处理达到了这种详细程度。一般来说，有不同的词干提取方法。建议您使用 nltk PorterStemmer。对于您的 DM1 项目，它应该足够了。在这里，您将根据您是否执行了某种程度的词干进行评估。您选择最佳词干分析器的能力将不会被评估。最后一点可能是研究工作的主题。  
  
如果您有兴趣深入了解该主题，您可以研究往往会执行相当好的词干提取水平的后缀剥离方法。这是最基本的方法之一。但是，如果您可以访问适当的训练语料库，则可以使用更复杂的方法，例如基于 n-gram 和隐藏马尔可夫模型。这里有一个很好的权衡：一方面，您希望减小术语文档矩阵的大小。另一方面，您不希望丢失有价值的信息（通过用相同的术语表示一些单词，当用不同的术语表示它们可能会提供更多信息）。最终目标是通过训练方法的高效运行时间来实现良好的分类准确性。  
  
2. 您的数据集包含 3 个旨在将您的数据划分为集群的函数：kmeans()、kmeans\_plus() 和 agglomerative()。这些函数中的每一个都应该执行一个聚类算法的一次运行（具有一个固定的 k 值）。  
  
功能集群评估是执行多次集群算法运行以用您的性能评估结果填充数据框的功能。此功能的原因是研究不同聚类方法的性能差异，例如凝聚和kmeans 聚类。此外，您还有机会研究这些算法的性能如何受 (a) 不同的 k 值、(b) 不同的初始化（在 kmeans 的情况下）和标准化的影响。

best\_clustering\_score() 函数的目的是接收带有性能评估结果的数据帧，并确定具有最佳聚类结果的运行。这应该反映您对我们如何根据剪影分数在多个选项中选择一组好的集群的理解。

亲切的问候，  
迪米特里斯

### 2.6关于CW task2函数cluster\_evaluation的问题

亲爱的迪米特里奥斯，

Q1：在课程作业的说明中（任务 2 问题 2），我们被要求对每个 K 值重复执行 10 次 kmeans。那么凝聚层次聚类呢？我们是否需要为每个 K 值重复执行 10 次凝聚层次聚类？

Q2：函数cluster\_evaluation的输出格式是什么？根据我的理解，我在下面画了一个草稿。我的理解对吗？

亲爱的 Jiahang，  
  
本课程包含 3 个旨在将数据点划分为集群的函数：kmeans()、kmeans\_plus() 和 aglomerative()。其中每一个都是针对一个固定的 k 值执行一次聚类算法。它接收包含您的数据的数据帧，并将它们分成 k 个集群。  
  
功能 clustering\_evaluation 是执行多次运行的功能。在这个函数内部，期望是执行不同算法的多次运行。实现此特定功能的目的是让您有机会将不同聚类算法（kmeans，凝聚）的性能与特定数据进行比较，同时评估聚类算法的性能如何受各种参数的影响，例如（a ) k 值，(b) 初始化（在 k-means 的情况下），以及 (c) 基于数据是否已标准化。  
  
结果将存储在新的数据框中。这个新数据框的结构在模板文件中描述。您对 cluster\_evaluation() 函数输出的说明似乎朝着好的方向发展，但不执行 10 次凝聚层次聚类可能很有用。您可以研究不同的 k 值和标准化如何影响凝聚层次聚类的性能。因此，对于 {3,5,10} 中的 k，您可以运行该算法 6 次，每次运行都使用原始数据或标准化数据。  
  
您不需要执行 10 次凝聚层次聚类。对于给定的集群内和集群间距离度量以及给定的 k 值，凝聚层次聚类应该返回相同的结果，因为它是一种确定性方法。您被要求执行 kmeans 聚类算法的多次运行，因为它的性能受初始化（通常是随机的）的影响。因此，希望能够让您了解不同初始化如何影响 kmeans 的性能。  
  
亲切的问候，  
迪米特里斯

2.7 CW 问题 2 cluster\_evaluation(df)

亲爱的迪米特里斯，

在课程作业问题 2 的 python 文件中，在函数 cluster\_evaluation(df) 的注释中，它说“给定带有要聚类的数据的数据帧 df，返回一个 pandas 数据帧，其中包含每个聚类算法执行的条目。每个条目应该包含：“算法”名称：“Kmeans”或“凝聚”，“数据”类型：“原始”或“标准化”，“k”：产生的聚类数，“剪影分数”：用于评估结果一组集群”。我们如何以这种方式创建数据框？我最初的想法是会有一个 3d 数据框，其中行和列名称对应于这些值，例如，一列用于凝聚，另一列用于 kmeans，并且条目对应于轮廓分数，但据我了解，我们不能拥有 3d 数据框。那么我们应该如何在这个新的数据框中呈现数据呢？

亲切的问候，

迈克尔

亲爱的迈克尔，  
  
感谢您的留言。您可以拥有一个包含 4 列的数据框：(a) 算法（kmeans 或凝聚），(b) 数据（原始或标准化），(c) k（基于已用于特定执行的 k 值），以及(d) 分数（此执行产生的集群集的轮廓分数）。此类带有执行统计信息的表可用于性能评估。  
  
亲切的问候，  
迪米特里斯

2.8CW 第 2 部分散点图

嗨迪米特里奥斯，

问题的描述在文档和模板文件中有所不同。在文档中，我们被要求生成得分最高的聚类的 sctter 图，但在文件中，我们被要求获取特定 k =3 且没有数据集类型和聚类方法的散点图。

我需要关注哪一个？

谢谢你。

亲切的问候，

世明

亲爱的世明，  
  
请按照模板文件中的说明进行操作。给定一组以 k=3 生成的聚类，生成散点图。它是使用 kmeans 还是使用凝聚以及原始数据或标准化数据生成的并不重要。这是您手动检查集群相对于各种属性分离实例的程度的机会。  
  
亲切的问候，  
迪米特里斯

亲爱的 Dimitris，  
  
很抱歉，我不知道我是否正确理解了您的话。  
  
您的意思是函数“scatter\_plots”的输入参数“df”是聚类结果而不是等待聚类的数据集？如果是这样，您介意澄清这个“df”的格式吗？例如，数据框“df”是否有 6+1 列，最后一列是类标签？

另外，我们需要使用“plt.show()”在屏幕上显示图表还是直接保存在本地路径中？  
  
最好的问候，  
云月

亲爱的云月，  
  
scatter\_plots 函数接收一个数据框 df 作为输入，并生成 15 个散点图。数据框的格式可以是您使用 read\_csv\_2 文件获得的格式。为了在函数采用这种格式时绘制散点图，您当然需要在该函数内聚集数据点。否则，您将不知道如何为数据点着色。为此，您可以调用其他函数之一，例如实现 kmeans。鉴于没有提供路径作为函数的参数，请将输出存储在与 .py 文件相同的路径中。请不要在该函数中使用 plt\_show() 。  
  
亲切的问候，  
迪米特里斯

### 2.9 关于Q2 part2散点图的问题

你好，

我想知道对于 k = [3,5,10]，文档中指定的 15 个散点图怎么可能存在。

6 for k 表示非标准化数据和标准化数据

6 用于凝聚层次聚类非标准化数据和标准化数据

和 2 用于非标准化数据和标准化数据

我最多只能得到 14 个地块

我认为他的意思是对，k 只等于 3。我们有 6 个属性，例如属性 ABCDEF  
然后我们应该绘制 [A,B] [A,C] [A,D] [A,E] [A,F] [B,C] [B,D] [B,E] [B,F] [C,D] [C,E] [C,F] [D,E] [D,F] [E,F]所以总共有15个地块

亲爱的大家，  
  
感谢您的留言。  
  
@Chi Ho 这个问题要求可视化一组集群（最好的集群集）。一组集群是将您的数据点划分为 k 个集群。对于这个特定的练习，预计您将获得导致一组良好集群的运行（例如具有最高轮廓分数的集群）。给定这样一组集群，问题要求为每对属性绘制一个散点图。这里有6个属性。因此，共有5+4+3+2+1对属性，即总共15对。  
  
备注：如果您将每次运行可视化，则会产生大量散点图。此外，您只能绘制最多 3 个维度的散点图。  
  
@Yiyao 非常感谢您的澄清。这正是所要求的。很高兴看到您已经清楚这一点。  
  
@Yunyue 我总是很乐意澄清你不清楚的任何事情。根据我与工业合作者的经验，这个课程公告比您在专业环境中可能遇到的许多情况要清晰得多。在那里，我们通常必须确定利益相关者的要求是什么（即写下说明）。  
  
@Chi Ho 确实，一次运行使用了所有 6 个属性并将点分类为集群。每次运行的数据点位置都相同。可能会改变的是不同运行中数据点的颜色。颜色应该指定每个数据点属于哪个集群。  
  
希望这会有所帮助。  
  
亲切的问候，  
迪米特里斯

### 2.10 CW Q2 函数：scatter\_plots(df)

亲爱的迪米特里奥斯教授，

问题 2 说我们需要使用剪影分数作为评估指标来确定哪个运行产生了最佳集群集。我们是否要使用相同的方法来获得来自批发客户.csv 的最佳结果？或者在函数 scatter\_plots(df) 中编写一些代码，通过调用函数 cluster\_evaluation(df) 和 best\_clustering\_score(df) 来选择 df 的最佳运行？我们在函数 scatter\_plots(df) 中使用的方法会随着 df 的变化而变化吗？

最好的祝福，

泽耀

亲爱的 Zeyao，  
  
事实上，您可以在性能评估中使用产生最佳集群集（就剪影分数而言）的运行。无需在您的 scatter\_plots 函数中重复此性能评估。  
  
原则上，您在课程作业的前面步骤中使用的任何聚类算法运行都应该足以满足本练习的目的。例如，您可以使用随机化运行 kmeans。您可以在此散点图函数中包含这些运行之一。我建议您使用产生最佳集群集的运行，以便手动评估生成的集群的质量。此运行应 (a) 由 ak 值指定，(b) 使用或不使用标准化，(c) 采用一些聚类方法（例如 kmeans 或凝聚）等。  
  
亲切的问候，  
迪米特里斯

1. Q3

### 3.1CW\_Q3

亲爱的教授，

在课程作业问题 3 的说明中，您指出“此外，您应该使用原始 Python 字符串函数进行文本处理操作。” 我现在处于标记化阶段，这是否意味着我们不应该使用 NLTK 库？

亲爱的Kuan-Hao，  
  
您可以在课程作业的文本挖掘部分使用ntlk。但是，对于某些任务，nltk 可能会很慢。因此，预计您对这些使用原始 Python 字符串操作，因为它们可能比 nltk 快得多。  
  
亲切的问候，  
迪米特里斯

### 3.2 课程作业 - 第 3 部分

你好教授，

我刚刚阅读了课程作业，并且在适用的情况下明确提到了 pandas、numpy 和 scikit-learn。尽管如此，第 3 部分似乎需要 NLTK，但没有提及，还是我们要学习其他方法？

谢谢

亲爱的 Hadel，  
  
您课程作业的文本挖掘部分\*不\*需要使用 nltk。您可以为此使用 ntlk。但是，对于某些任务，nltk 可能会很慢。因此，预计您对这些使用原始 Python 字符串操作，因为它们可能比 nltk 快得多。  
  
亲切的问候，  
迪米特里斯

### 3.3 连续波第 3 部分

下午好，迪米特里奥斯

当我们需要从推文中删除非字母/空白字符时，我有一个关于修改数据框的问题。

我们是否需要创建一个单独的列来包含修改后的推文，或者我们在 OriginalTweet 列中就地执行？

亲爱的 Alisher，  
  
您可能希望 \* 不 \* 创建一个包含修改后的推文的附加列，因为这对于实际应用程序可能会占用大量内存。  
  
亲切的问候，  
迪米特里斯

### 3.4CW 第 3 部分表演

嘿迪米特里奥斯，

对于 remove\_stop\_words 函数，我的实现很慢。我们的功能实现会在性能和速度上打上标记吗？

非常感谢，  
法希姆

亲爱的 Fahim，  
  
这是您提高 remove\_stop\_words 功能性能的机会。一个非常慢的实现表明该实现效率不高。这在专业/工业环境中可能非常关键。您的代码的运行时间确实会影响您的课程成绩。  
  
亲切的问候，  
迪米特里斯

### 3.5 关于课程作业 Q2 和 Q3 的问题

嗨，Letsios 博士，

希望你一切顺利。

我对课程作业有一些疑问。问题如下。

1. 在挖掘文本数据中，我们需要词干。我想知道是否有标准的词干分析器供我们使用，因为据我所知，不同的词干分析器会给出不同的结果。

例如，如果使用 PorterStemmer，“冠状病毒”将“冠状病毒”进入航空公司，而如果使用 SnowballStemmer，则将“冠状病毒”转化为“冠状病毒”。

因此，它会对结果产生影响。

亲爱的 Jinhong，  
  
1. 事实上，在预处理文本数据时，不同的词干分析器可能会给出不同的结果。很高兴看到您的推文预处理达到了这种详细程度。一般来说，有不同的词干提取方法。建议您使用 nltk PorterStemmer。对于您的 DM1 项目，它应该足够了。在这里，您将根据您是否执行了某种程度的词干进行评估。您选择最佳词干分析器的能力将不会被评估。最后一点可能是研究工作的主题。  
  
如果您有兴趣深入了解该主题，您可以研究往往会执行相当好的词干提取水平的后缀剥离方法。这是最基本的方法之一。但是，如果您可以访问适当的训练语料库，则可以使用更复杂的方法，例如基于 n-gram 和隐藏马尔可夫模型。这里有一个很好的权衡：一方面，您希望减小术语文档矩阵的大小。另一方面，您不希望丢失有价值的信息（通过用相同的术语表示一些单词，当用不同的术语表示它们可能会提供更多信息）。最终目标是通过训练方法的高效运行时间来实现良好的分类准确性。

### 3.6 课程第 3 部分词干

下午好，教授

对于 CW Part 3，你有推荐的词干分析器吗？我使用 PorterStem 来阻止这个词。可以吗？

最好的，

Yingchen

亲爱的映辰，  
  
感谢您的留言。事实上，我的建议是您使用 nltk PorterStemmer()。这很好。  
  
如果您有兴趣了解更多关于词干的信息，我建议您参考 3 月 15 日由蒋金宏发帖发起的论坛中的讨论。  
  
亲切的问候，  
迪米特里斯

### 3.7 CW Q3 函数：remove\_non\_alphabetic\_chars(df) 和 remove\_multiple\_consecutive\_whitespaces(df)

嗨迪米特里奥斯，

希望你会好好的。我只是想问一个关于该主题中功能的快速问题。我们应该将这些应用于仅包含推文的列还是整个数据框？非常感谢您的帮助和理解。

亲切的问候，

陡

亲爱的 Sarp，  
  
确实，请仅在推文上应用文本处理操作，例如去除非字母字符和去除多个连续空格。带有推文的列本质上是您的语料库，只有这一列应该经过预处理，这将有助于您根据情绪对推文进行分类测试。  
  
亲切的问候，  
迪米特里斯

3.8 关于任务 3 MultinomialNB 运行时间和准确率

亲爱的迪米特里斯，

在task 3的最后一个问题中，我们是否会根据最终的预测准确度进行标记？

在这个问题中，我们需要考虑如何限制term-document矩阵的大小（否则数组的内存不足），也会影响准确率和训练时间。至于训练时间，我们是否也需要减少它（例如在 15 秒内），因为我们被要求高效地完成 NLP 任务？

最好的祝福，

云月

亲爱的云月，  
  
确实，您的多项朴素贝叶斯分类器的训练精度将在标记过程中被考虑在内。这里的目标是决定要考虑哪些属性（即术语-文档矩阵的术语）以提高准确性。这确实可以通过限制术语文档矩阵的大小来实现。  
  
不确定我是否正确理解了您关于记忆的问题。内存是您在本课程中需要考虑的资源，更重要的是，在实际应用中。理想情况下，您应该避免使用超过给定任务所需的内存。然而，数据集已被选择，以便它可以轻松地放入相当新的个人计算机的主存储器中。  
  
15s 对于 Q3 来说是一个相当长的运行时间。我的建议是您尝试确定您可以改进的实现方面，以便将其减少到至少 10 秒以下。事实上，这是有效执行任务的一部分。  
  
亲切的问候，  
迪米特里斯

亲爱的 Dimitris，  
  
非常感谢您的回答！  
  
之所以说内存是因为如果不限制term-document矩阵的特征数量，它的大小很容易超过16GB。  
  
另外，关于运行时间10s的问题，你是指“coronavirus\_tweets.py”中所有函数执行的总时间还是每个函数的运行时间（这样总时间可能超过10s）？  
  
最好的问候，  
云月

顺便说一句，我已经将 PorterStemmer() 与 pandas API“apply()”一起使用，这将非常快速地操作数据帧中的元素。但是，运行时间仍然在 13s 左右，我想请问您介意如何进一步改进它吗？

亲爱  
  
的云月，coronavirus\_tweets.csv 文件大小为 10.5MB。为什么它可以超过1GB？  
  
实际上，理想情况下，运行所有的coronavirus\_py 函数应该不再需要 10 秒。如果高于此值，那么我的猜测是您可能会改进某些方面。如果你做不到，不用担心。用于区分不同课程实施质量的分数会有一些小罚分。我们可以在课程作业的评分过程完成后详细讨论这个问题。  
  
亲切的问候，  
迪米特里斯

3.9 散点图、库和数据定位课程

嘿迪米特里奥斯，

只是对课程作业有几个问题。

1. 在当前工作目录中保存散点图时，您希望我们将 15 个图保存为单独的 .png 文件，还是我们可以将所有 15 个图作为子图绘制在一个图中并简单地保存为 1 个 .png 文件？

2.我用过的库有pandas、sklearn、matplotlib、nltk、requests。您将用于测试的环境是否安装了这些库？

3. 假设 .csv 文件在当前工作目录中并且我们可以使用 pandas.read\_csv("adult.csv") 简单地读取是否安全？

谢谢，

贾格迪普

亲爱的 Jagdeep，  
  
- 请生成 15 个不同的散点图并将它们存储在本地目录中。无需使输出更紧凑。  
  
- 实际上，运行代码的环境将安装 pandas、sklearn、matplotlib 和 nltk 库。  
  
- 实际上，您可以假设这些文件与您的 python 文件位于同一目录中。但是，您的函数实现原则上应该接受“path/adult.csv”形式的任何路径。  
  
亲切的问候，  
迪米特里斯

### 3.10关于停用词列表

亲爱的迪米特里奥斯教授，

我注意到，在第三季度，删除停用词的部分。你放了一个关于english-stop-words-large.txt的链接。我很困惑我们是否需要下载文本文档，或者只是使用请求库来获取它。你能给我一些建议吗？

最良好的祝愿，

HONGLIN CHEN

亲爱的洪林，  
  
在实用的7解决方案中，有一种方法可以通过使用url来获取停用词列表。您可以使用这种方法在您的实现中访问它。  
  
亲切的问候，  
迪米特里斯

### 3.11 关于第 3 部分的问题。是否应返回 df 副本

Part3 大部分功能是直接在数据框上编辑的。我们应该创建一个数据框的副本并修改副本值并返回它，还是我们应该只修改原始数据框？

eg,

def remove\_non\_alphabetic\_chars(df):

            df['xxx'] = df['xxx'].str.replace('[^a-zA-Z]', ' ', regex=True)

or

def remove\_non\_alphabetic\_chars(df):

            dfc= df.copy()

            dfc['xxx'] = dfc['xxx'].str.replace('[^a-zA-Z]', ' ', regex=True)

             return dfc

亲爱的易遥，  
  
感谢您的留言。原则上，这里给您的建议是不要在不需要时创建数据副本，因为这可能会占用大量内存。修改原始数据框更合适。这也是您希望在您作为数据科学家的可能职业中采用的建议。在这个特定的练习中，您可能还会从构建新的数据框对象中获得一些运行时间。  
  
亲切的问候，  
迪米特里斯

亲爱的 Dimitris，  
  
我们不应该对每个问题都使用副本吗？在问题1中，当我们进行一次热编码和标签编码时，如果在标签编码之前进行了一次热编码，则从数据帧中删除了属性“类”，因此无法进行标签编码。我们应该在 one hot encoding 函数中进行复制，还是我们可以假设在测试文件中，标签编码将在 one hot encoding 之前执行以避免这个问题？  
  
亲切的问候，  
  
迈克尔

亲爱的 Michael，  
  
one\_hot\_encoding() 函数可以返回仅包含输入属性的 X\_train，而 label\_encoding() 函数可以返回仅包含目标属性的 y\_train。然后可以由 dt\_predict() 使用这些。  
  
亲切的问候，  
迪米特里斯

### 3.12CW 任务：冠状病毒推文，问题 3

嗨迪米特里奥斯，

希望你会好好的。我似乎对最终任务中的最后一个问题有些困惑，如果可能的话，我想要求澄清一下（当然我的课程伙伴也可以贡献）。

当我们谈论将“语料库”存储在 numpy 数组中时，“语料库”是推文和情绪，还是整个数据，还是其他什么？

另外，在使用 CountVectorizer 时，我们应该在整个数据帧上使用它还是只在推文和情绪上使用它？

非常感谢您的帮助。

亲切的问候，

陡

亲爱的 Sarp，  
  
语料库是原始数据的一个属性，即包含推文的属性。CountVectorizer 可用于文本预处理操作：  
https ://scikit-learn.org/stable/tutorial/text\_analytics/working\_with\_text\_data.html  
  
当然，您还需要使用情绪来执行感兴趣的分类任务，例如，通过对标签使用额外的 numpy 向量。  
  
亲切的问候，  
迪米特里斯

### 3.13关于cw的一些问题

在第三季度，

1.对于函数tokenize，你说我们可以使用原始python字符串函数。但是，原始 python 字符串函数和 nltk 的结果是不同的。

（例如，单词 'cannot' 将导致 nltk 中的 'can' 和 'not'，但只有使用字符串相关函数 ''split'' 的 'cannot'）哪一个是正确的？

2. 我注意到您说我们应该修改原始列“OriginalTweet”的语料库，而不是添加新列。它在代币化方面也有效吗？总之，我们是否需要另一列“tokenizedtweet”？

3. 函数count\_words\_without\_repetitions，不知道怎么算？

例如。第一列：[a，b，a，c]；第二列：[a, b, c] 那么函数的结果是 3 还是 6？

导致使用 textblob(unique word) 的结果包含一个单词的多次。

3.14

* 第三季度

1. frequent\_words(tdf,k)：我们应该只返回单词列表还是元组列表 [(word1, count1), (word2, count2),...]？
2. mnb\_predict(df)：我使用的是 TfidfVectorizer，它基本上是一步结合的 CountVectorizer 和 TfidfTransformer。这个可以吗？

-frequent\_words(tdf,k) 应该只返回 k 个最常用词的列表。他们的频率\*不\*要求。  
  
- 只要实现 mnb\_predict 以便它接收请求的输入并产生请求的输出，它应该没问题。此外，运行时间应该是合理的。