INT104 CW2 report 高分模板

- by 张北海

声明：去年高分模板，用了必高分（80+）（除非文笔确实不太好，好好想想 EAP 的 essay 怎么写）；一个问题是大体的 structure 是按照去年 task sheet 里给的，学弟学妹可能需要根据今年的进行一些调整，但是方法和内容基本都是一样的；这里主要告诉你们要做什么，用哪些方法，你们只需要不会的去查，然后往里面写就行。 要说明这个 report 应该是不需要代码的，同时也不会按照你的准确率给分（数据本来就比较烂），主要是看你做了哪些工作，report 里的方法够不够 rich，所以要疯狂炫技，但是要有意义的炫技，不要做无用功。

去年的 report 包括代码可以私信问我要，随缘给（避免大家纯抄然后被老师发现判作弊）；INT104 虽然抽象但是内容还是蛮有意思，建议自己有一个主动学习的过程；有问题欢迎指正。

注意：无论你在 report 里写了你用某种方法，进行某种操作，一定要阐明他的合理性，体现 critical thinking（针对本数据的优点），切忌自嗨。

1. intro：

IEEE 会议模板，不需要 abstract。把 task sheet 里的题目要求 paraphrase 一遍，大体说自己用的方法（PCA 等），不要写太长。

1.1 数据分析

把 task sheet 里的数据描述 paraphrase 一遍，再声明第“2”类数据比较特殊，要不要作为噪音删除（个人认为都行，言之有理即可，不删掉的话在降维的时候可能展示的图更丰富，但是最后准确率不好说有可能会变低），如果删掉 2 了这里要说种类只有 0,1，这里变成了**双分类问题**，后面用到的损失函数等方法都要做调整（具体百度），如果没有删掉 2 这里就是多分类问题。

1.2 分类器选择

这里要当做 **literature review**，介绍自己用了哪些分类器（也包括 task3 中的聚类算法，也就是有监督无监督都要说，具体的建议后面会提）要把引用都集中放在这里，然后要前后文对比哪个分类器和谁比起来更有优势或者哪里有不足，做一个简单说明。

1.3 分类结果

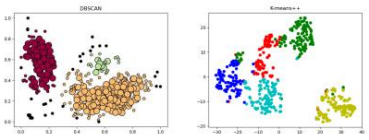
放一个**三线表**（注意一定要三线，百度 latex 格式怎么写）列出分类器，训练集准确率，测试集准确率；放两张图，作为两种聚类方法的结果；再来一个三线表写聚类结果↓

Table 1: Cross-Validation score

Classifier	Score on Train Set	Score on Test Set
DNN	96.15%	84.62%
XGB	100.00%	84.62%
SVM	100.00%	84.62%
LR	100.00%	84.62%
KNN	100.00%	92.31%

Figure 1 shows the grouping result of two clustering algorithms.

Figure 1: Clustering result of DBSCAN and K-Means



(a) DBSCAN (b) K-Means

The silihouette score of two algorithms above is shown below:

Table 2: Silihouette Score

Algorithm	Silihouette Score	Noise Point
DBSCAN	0.377	34
K-Means++	0.338	None

2. 数据观察

首先要说数据预处理（清洁），例如：

如果两行数据他具有相同的数据，但是却有不同的标签，或者有行是空数据，这样的数据是噪声，没有意义，要删掉；删除无意义没有用的列（如 ID 列）

然后是分析数据是否符合**高斯分布**，网上找个代码生成个图就行，如果是类似正态分布的样子那就是高斯分布，针对是否是高斯分布要不同处理。

然后分析 common feature 包括每一列的均值方差标准差最大值最小值，简单解释一下（谁的什么大谁的什么小，就说废话就行，别说太多）

相关性分析[重点] 分析各列（feature）的相关性，如果是双分类那么就使用 pearson 系数，如果是多分类就使用 spearman 系数，具体实现网上都有，可以直接 copy python 代码。（然后放一张相关性的热力图）解释各列是否相关，根据相关要进行不同处理（我估计列是不相关的）

然后写降维策略[重点]这里建议使用的有 PCA，NMF 和 ISOmap，在写的时候一定要写每一个方法的优劣，去网上查，然后要展示结果（可视化一张图），如果有你喜欢或者结果比较好的方法就多写几个公式放上去，显得高大上。

(1) PCA： k 值（目标维度）在这次作业里大概率就是 1 了（展示出来就是一条横线上的点），要看图像降维后点的分布情况，那条线如果太长太分散就说明效果不好。 可以展示一下 variable distribution（百度），但是这里如果只有 2 列就意义不大。描述一下图的内容（说废话）。这里还有一个 PCA score 可以展示，也可以用于 bias detection，但是我懒得说了，有兴趣可以百度。

(2) NMF： 去年我最好的降维方法，原理非常简单，要写优劣、结果如何

(3) ISOmap：适用流行（manifold）的数据[纯炫技]，说明优劣、放图

最后，选一种以上三种算法得到的降维结果（看分布好坏来选择一种的结果），对结果进行 bias removal，这里用随机森林，如果有空间就放张图，原理百度，说它为什么好。

3. 有监督学习训练

3.1 **分类器选择和原理：** 有超级超级多选择，最后选了结果较好的 SVM（支持向量机）、逻辑回归、DNN（神经网络）。XGboost，KNN（你们不用写这么多，当然想多写也行，我就写了这 5 个已经很多了，建议 3-4 个，要看篇幅删减），这一切一切方法都可以 sklearn 的包，不必要手搓代码，都很简单，CSDN 教程一堆。这里主要是解释分类器工作原理，多写点公式，显得 nb。

3.2 数据特征挑选：根据数据观察的结果，在使用[降维方法]降维并通过随机森林进行 bias removal 后的数据作为训练数据输入分类器，这里展示一张各个类别个数的柱形图，展示每个类有多少，然后讨论是否要进行**欠采样**（如果一个类比其他类多很多，为了防止分类器“作弊”，要随机从多的类里面挑取一定量数据，让各个类别的数据个数相近或者相等）这里也要有 **critical thinking（尽量全文都要有点体现，前面忘记提了），**说删除数据可能导致**欠拟合**，最好展示一张某个分类方法不同数据量的的学习曲线（折线图，很好做）。

3.3 分类器训练与交叉验证：（忘记今年有没有要求交叉验证了，去年要强制 K 折交叉验证，这里记得要去解释验证原理过程）

这里每一个分类器写一段，分别写各自的**参数设置，超参数设置【**使用**超参数网格搜索**（百度）， 其实就是穷举法去寻找超参数**】，**可以酌情写点儿 **score 计算的公式（交叉验证得分），**然后放一张表格（不要三线表）**，写这个分类器分别在原始数据**（没有欠采样或者处理之前的能用的数据）**和欠采样以后的数据进行训练后的训练集和验证集准确率（也不用全写）。**最后，**找一个最好的交叉验证的结果用一个混淆矩阵（confusion matrix - 百度）展示出来**

4. 无监督学习训练

与上相同，我用的 k-means++和 DBSCAN（如果两个原理解释不完可以略写一个详写一个，尽量详写 k-means 因为最好说）先解释工作原理，再说超参数（k）的取值，放一个轮廓系数（Silhouette coefficient）在不同 k 值下的折线图来决定一个合理的 k，跟上面一样放几个公式，能写多少写多少，不再赘述。

5. 总结

5.1 General 总结，把你写的 intro 再 paraphrase 一下，粘贴过来。

5.2 优化/future work 写你做的工作有哪些不足（可以写你因为文章字数限制没来得及炫的技）要调理清晰，在这里**是你最后的机会充分弥补前文中 critical thinking 的不足（你带英就喜欢这个）**，分 3 个 task 看看哪个能说。这里我写的 task 1 要考虑**数据集开放性（openness）**的问题，要考虑开放性系数，降低经验风险和开放空间风险，防止过度泛化（听不懂没关系，我也不懂，就摁写）；task2 的准确率过高**（去百度原因往上写）**，考虑是数据集过少等各种问题。

6. Appendix：可有可无，有的话就加上一些你喜欢的算法的代码，怎么放可以百度，这里不算字数，没有也行。