

## pro4

结合问题2的关联矩阵和UMAP模型，接下来我们使用配对的t检验，在与上面相同的样本数据上构建并测试分类器朴素贝叶斯，确定了训练和测试集的准确性结果的均值，方差和置信区间（CI）。换句话说，我做了一个保持数据分区，并创建和测试了分类器。然后，该过程重复约150次。

现假设事件 $A_1$ 和事件 $A_2$ 关于事件B独立, 根据事件的独立性, 则有 $P[(A_1|B)(A_2|B)] = P(A_1|B)P(A_2|B)$ , 又由于只有在事件B这个共同条件下事件 $A_1$ 才会与事件 $A_2$ 独立, 我们不妨将式子改写为 $P(A_1, A_2|B) = P(A_1|B)P(A_2|B)$ , 从而可以将我们的贝叶斯公式改写为:

$$P(B_i | A_1, A_2) = \frac{P(B_i) P(A_1 | B_i) P(A_2 | B_i)}{\sum_{j=1}^n P(B_j) P(A_1 | B_j) P(A_2 | B_j)} \quad (1)$$

假设事件组 $A_m$ 中的所有事件关于完备事件组 $B_n$ 中任意一个事件两两相互独立, 进而公式可以推广为:

$$P(B_i | A_m) = \frac{P(B_i) \prod_{t=1}^m P(A_t | B_i)}{\sum_{j=1}^n P(B_j) \prod_{t=1}^m P(A_t | B_j)} \quad (2)$$

其中A为音乐特征, B为艺术家。

UMAP模型和朴素贝叶斯模型都建立在相同的训练集上，并且针对150个采样路径的每个路径，在相同的测试集中测量了性能。然后，我可以使用配对t检验来衡量两个模型的性能是否存在显著差异。

t检验的**原假设**是音乐特征与艺术家的音乐方面相似。

## UMAP反向缩放, 找出每个集群的平均值(Reverse the scaling, find the mean for each cluster ) (表格)

	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalness	liveness	speechiness	explicit	duration_ms	popularity	year
UMAP_KM_ClusterID															
0	0.670447	0.638378	0.767477	115.415646	-9.265074	0.734177	5.135624	0.258715	0.016363	0.181110	0.087266	0.054250	222073.685353	40.003617	198
1	0.460088	0.408589	0.420035	114.043362	-15.164071	0.629921	4.937008	0.584785	0.768913	0.147766	0.050605	0.019685	255816.732283	28.523622	197
2	0.491136	0.300995	0.422750	110.470121	-13.253623	0.752094	5.078727	0.755892	0.015855	0.168683	0.044052	0.003350	210633.157454	30.164154	197
3	0.465752	0.717841	0.462007	130.928748	-7.761762	0.741611	5.189597	0.140462	0.029264	0.291630	0.064888	0.072148	257987.055369	43.065436	199

在计算t统计量和自由度之后，结合UMAP的值，我们可得到，在90%置信度下，p值约为0.98；UMAP模型表明各参数与艺术家高度关联；因此，我们推翻原假设，也就是说，我们有足够的证据证明有影响力的人“真的会影响追随者创作的音乐,音乐特征在影响某个特定艺术家的音乐方面都扮演着相似的角色。

## 模型检验

升力图和增益图均用于评估模型的性能。他们都测量了与没有模型的情况相比，使用预测模型可以期望做的更好的事情。

对于提升图，我们可以了解到，使用给定模型时，对于训练数据集，我们能够在属于0类的前10%百分率观测值中捕获约14.43%属于0类的观测值，即1.4倍就像我们不应用模型一样 在测试集上，由于我们不应用模型，因此该模型能够在最高的10%百分位数上捕获1.2倍的阳性分类。

对于增益图，要使我们的模型在训练和测试集上都达到0级的准确度达到80%，就必须在训练集和测试集上至少获取70%的样本。对于测试数据，通过最大化（灵敏度+特异性-1）截止值，我能够找到最佳截止阈值为0.242。故我们的模型灵敏度较高，具有很强的鲁棒性。

