报告

一、问题描述

1. 问题背景

本项目研究的是通过基因表达量的数据矩阵，来确定细胞的类别。

本问题的数据集包括各个细胞基因表达量的数据矩阵（deng-logcounts.csv）和细胞类型数据（deng-celltype6.csv）。

在数据矩阵中，共22431行和268列，每一行代表一个基因特征，每一列代表一个（待确定分类的）细胞。

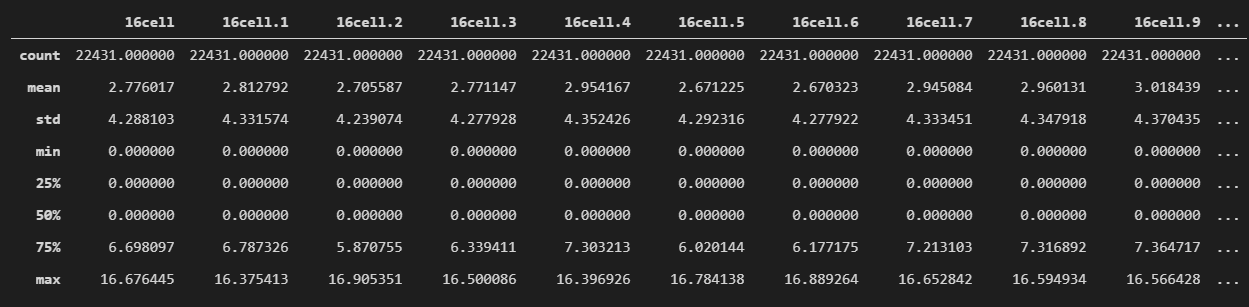
*1．Background*

*The study of this project is to determine the type of cells through the data matrix of gene expression.*

*The dataset of this question includes the data matrix (deng-logcounts.csv) and cell type data (deng-celltype6.csv) of gene expression in each cell.*

*In the data matrix, there are 22431 rows and 268 columns, each row represents a gene feature, and each column represents a cell (to be determined to be classified).*

使用python中DataFrame的describe函数可以观察这一基因特征的各元素值的分布：



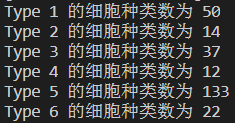
数据矩阵的概览，使用python中DataFrame的describe函数

可以看到，在数据矩阵中，所有元素均大于等于0，且各个基因在所有细胞中的均值（图中mean行）、标准差（图中std行）和最大值（图中max行）都十分接近。这种**各个特征间分布大体相同**的数据模式是**十分适合**进行数据分析的。

*In the data matrix, all elements are greater than or equal to 0, and the mean (‘mean’ row in the figure), standard deviation (‘std’ row in the figure) and maximum value (max row in the figure) of each gene in all cells are all very close. This data pattern with roughly the same distribution among various features is very suitable for data analysis.*

而在细胞类型数据中，所有细胞共有六类，我们分别用Type 1到Type6来表示。每种类型的细胞数量分布在下图给出：

*In the cell type data, there are six types of all cells, which are represented by Type 1 to Type 6 respectively. The distribution of the number of cells of each type is given in the figure below：*



每种类型细胞的种类数

可以看到，Type 5占了其中的一半左右，其他五种类型的细胞共占了剩下的一半。

*It can be seen that cell Type 5 accounted for about half of them, and the other five types of cells accounted for the remaining half.*

2. 求解方案设计

*2. Solution design*

根据问题要求，并结合课程中最大似然估计和EM算法等内容，我分别尝试了**有监督学习**和**无监督学习**的算法对细胞所述的分类进行预测。

*According to the requirements of the question, combined with the content of maximum likelihood estimation and EM algorithm in the course, I tried supervised learning and unsupervised learning algorithms to predict the classification of cells.*

对于有监督学习，我选择使用含有**最大似然估计思想**的**逻辑回归模型**对于数据集进行分类，并使用**交叉验证**（Cross Validation）的办法，将整个数据集划分为训练集和测试集，使用训练集来拟合逻辑回归模型，而使用测试集来验证这一模型的有效性。我一方面使用了**准确率**（Accuracy）来评价模型的有效性，另一方面也用了题目中要求的**AUC**来度量预测结果。

*For supervised learning, I choose to use a logistic regression model with the idea of ​​maximum likelihood estimation to classify the data set and use the method of cross validation to divide the entire data set into a training set and a test set, using the training set to fit a logistic regression model and use the test set to verify the validity of this model. On the one hand, I used the accuracy rate (Accuracy) to evaluate the effectiveness of the model, and on the other hand, I also used the AUC required in the title to measure the prediction results.*

对于无监督学习，我使用了**EM算法**的经典实例**k-means算法**来对细胞分类进行聚类。这一部分首先**定义了细胞之间的欧式空间距离**，并且使用k-means算法寻找聚类中心。最后，使用**ARI**和**NMI**来评价聚类的结果。

*For unsupervised learning, I used the k-means algorithm, a classic instance of the EM algorithm, to cluster the cell classification. This part first defines the Euclidean space distance between cells, and uses the k-means algorithm to find cluster centers. Finally, ARI and NMI are used to evaluate the clustering results.*

同时，我也使用了**t-SNE**算法对于数据进行降维，将数据降至二维之后在平面内标出数据点，并**可视化**分类情况。这一结果可以清楚地看出各个类型的细胞的分类情况与有监督学习/无监督学习下对细胞的分类/聚类情况。

*At the same time, I also used the t-SNE algorithm to reduce the dimensionality of the data. After reducing the data to two dimensions, I marked the data points in the plane and visualized the classification. This result can clearly see the classification of various types of cells and the classification/clustering of cells under supervised learning/unsupervised learning.*

3. 实验环境

本实验使用Python 3.7.0完成。主要用到的依赖库包括：

- pandas：用于处理数据文件

- numpy：提供基础的数学计算

- scipy：用于验证部分完成AUC，ARI和NMI的接口

-matplotlib：用于回归

算法的核心部分，包括有监督学习中的逻辑回归和无监督学习的Kmeans算法都使用完全纯手动实现。有监督学习的算法在交叉验证中达到了100%正确率的良好效果，证明了实现的正确性。两个算法的结果验证部分，包括求准确率、AUC、ARI和NMI由于不是本项目的核心，大都直接使用了scipy库中的函数。

本报告结构如下：第二部分将介绍使用逻辑回归算法的有监督学习的过程与实验结果；第三部分介绍使用k-means算法的无监督学习过程与实验结果；第四部分将进行总结。

二、有监督学习：使用逻辑回归模型对细胞种类进行分类

1. 逻辑回归模型：方法与实现

逻辑回归（Logistic Regression）模型是一种用于解决二分类问题的广义的线性模型。它假设因变量服从伯努利分布，通过Sigmoid函数引入了非线性因素，因此可以轻松处理二分类问题。**逻辑回归是一种使用最大似然估计的模型。这和我们上课的内容紧密相关。**

逻辑回归的假设一个模型存在，并且在已知试验结果的情况下对参数进行估计。这一参数可以求得样本属于正样本或负样本的概率。这一概率的函数形式如下：

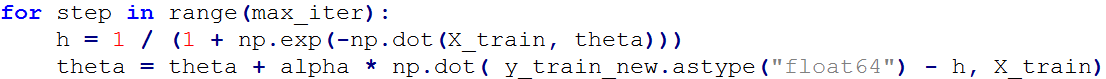
而逻辑回归模型所做的假设是：

相应的决策函数为：

而在模型训练时，逻辑回归一般使用梯度下降法进行迭代求解，这里涉及到求梯度。注意到Sigmoid的形式求梯度十分方便：

而梯度下降法的形式为：

其中，是学习率。使用上述式子，不断迭代逻辑回归的参数就能得到这一分类问题的解。

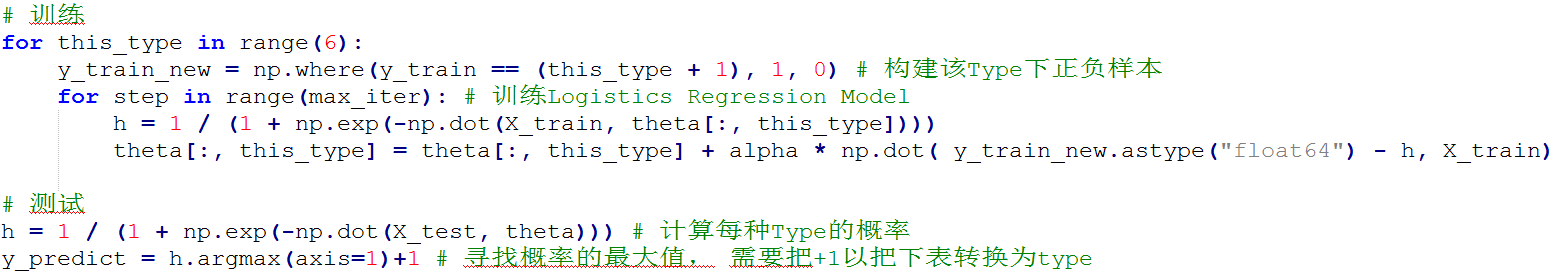


迭代部分的核心代码

上述是基本的逻辑回归算法只适用于二分类问题，而题目需要在6个分类中确定细胞的分类，是一个多分类问题。为此，需要使用“一对多”（one-versus-rest）算法对于逻辑回归进行修改。

“一对多”算法的基本想法是将**N分类问题转变为N个二分类问题**。对于每一种分类，进行一次“属于该分类/不属于该分类”这样的判断。将所有属于这一分类的标记为“正样本”，而将不属于这一分类的标记为“负样本”，这样可以训练得到一个判断某一样本“属于该种类”概率的模型。而对于N个这样的二分类模型都计算这样一个概率，寻找其中最大概率的一个，认为样本属于该种类型。

就本问题的代码实现而言，就是在实现了基本的逻辑回归之后，对于6中类型的细胞中每一个都进行一次判断，按上述要求标注正样本和负样本，并学习其样本的特性。



多分类的逻辑模型

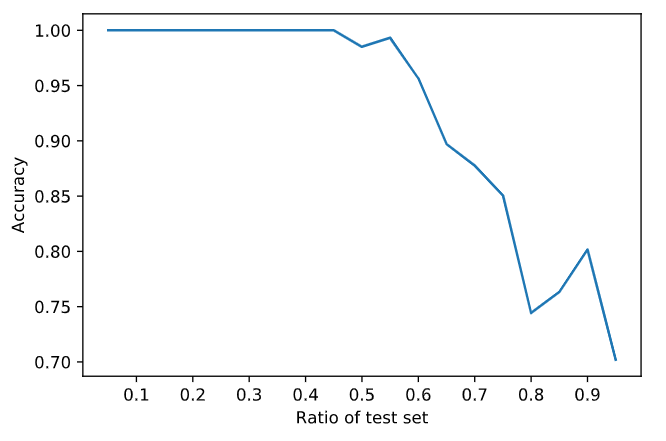
2. 实验结果

在完成了多分类逻辑回归的实现之后，使用本问题的数据集来实验验证其结果。为使结果具有更强说服力，使用**交叉验证**的方法来设计实验，将整个可用数据集随机的划分为“训练集”和“测试集”两个部分。训练集用于训练和拟合模型，而测试集用于对于拟合好的模型进行验证，观察模型的通用性。为此，我们首先将268个细胞中70%的数据（188个细胞）划分为训练集，30%（80个细胞）划分为测试集。之后，将70%的这数据与其对应的分类使用逻辑回归模型进行训练，拟合得到了一个逻辑回归的模型。

测试结果表明，首先将这一模型在原**训练集**上进行拟合，发现其**达到了100%的正确率**。

之后，我们将这一模型用在**剩余30%**的模型进行测试，观察其训练的准确率，发现其**同样达到了100%**。之后，通过把多分类转换为二分类，可以进一步计算器AUC值，发现其**AUC=1**。这一结果不仅说明了我们**实现的模型是正确**的，同时也说明了这一给定**数据集具有线性可分**的特点。

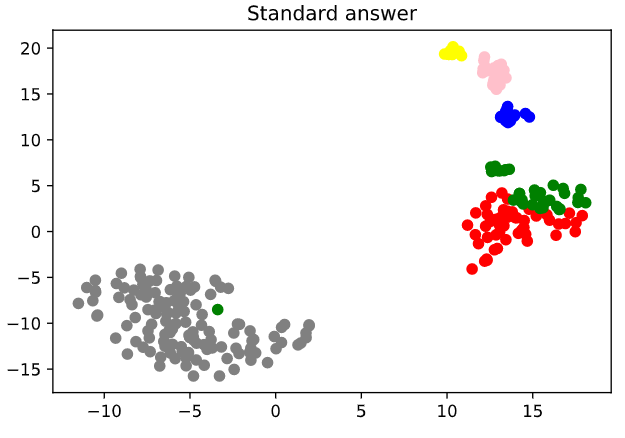
为了研究这一模型的线性可分性到底有多强（即，这一问题能够在多少小的训练集中保有目前的功能），我们进一步做实验，调低了训练集比例，提高了测试集的比例，并且观察了交叉验证时在验证集上的数据特点。实验结果如下图所示：



可以发现，在**训练集规模大于50%**（即测试集规模小于50%）时，该逻辑回归模型在训练集上可以完全达到**100%的正确率**，且**AUC=1.**

而当训练集规模在30-50%，该逻辑回归模型也表现良好，训练集正确率基本保持在90%以上。有趣的是，当我们只使用5%（268条中的13条）的数据作为训练集，这一粗糙的逻辑回归模型在测试集上**仍然能有约70%的正确率**。出现这一现象的原因与**该数据集的特性有关**：前面说到，该数据集中Type 5的细胞占整个数据集的一半，所以只要能够在小规模的数据中学到Type 5细胞的特点，就可以保证一半以上的正确率，即使没能把其他Type的特征完全学到，最终的正确率也能够达到70%左右。

虽然逻辑回归算法只是一个线性模型，但它在这一问题上的良好表现出乎了我们的想象。为了解释这一原因，我通过可视化的方法，深入研究了这一表现背后反映的数据性质。为了可视化，需要对于原来的两万多维特征进行降维，使其能够表达在二维平面上。因此我使用了降维操作中最常用的t-SNE方法对数据进行降维。在降至二维后，绘制在图表中，每种类型使用不同的颜色表示：



细胞类型数据的可视化结果，每种不同颜色代表一类不同的细胞

图中，每一个点代表一个细胞。细胞的两万多维特征经过运算，从中提取出了最有关的两个维度的特征，并且绘制在了图中。每一种颜色代表一个Type的细胞。从图中，我们发现，即使在二维空间汇总，各个类型的细胞之间也有着明显的分界线。而可以想见，在高维中，这些细胞在空间中的划分将更加明显。这一可视化结果说明了该数据集**具有线性可分性**，即各个Type之间可以用超平面来进行分割。线性模型有效的条件，就是数据集具有线性可分性。因此，即使是使用逻辑回归模型这样比较简单的线性分类模型，也可以在细胞分类这一任务上达到很好的效果。

3. 总结

这一部分使用了逻辑回归模型对于细胞分类问题进行了有监督学习，在交叉验证达到了100%正确率，AUC=1的优秀表现。同时，我们也研究了不同训练集大小下交叉验证的正确性变化曲线。结合数据曲线，并使用t-SNE降维后进行了可视化，并说明了逻辑回归模型的良好表现的原因是在于该数据集具有线性可分的特点。

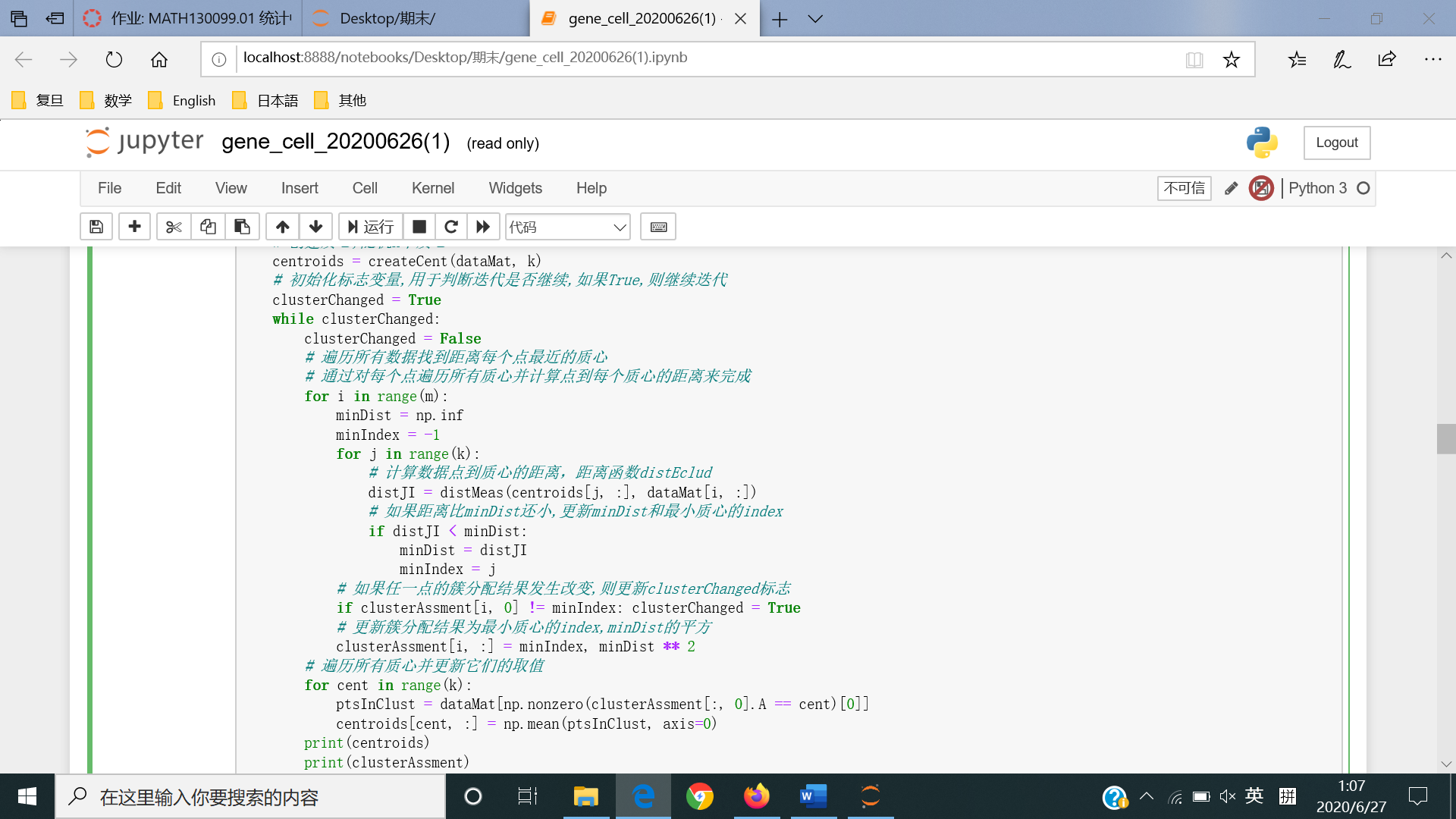
三、无监督学习：使用 k-means来对细胞种类进行聚类

1. 方法与实现

EM的算法核心就是E-step（期望步）和M-step(最大化步），而k-means也可以看作是两阶段的，总体思想都是迭代优化的过程，其中簇中心可看作参数，样本所属类别可看作隐变量。

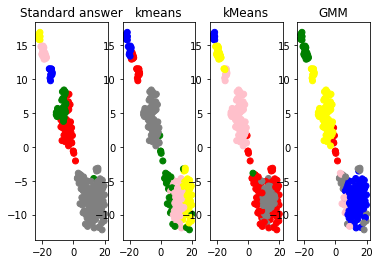
第一阶段（对应EM算法的E-step）：给定当前的簇中心，将每一个样本都指派到距离最近的簇中心所属的类别（即期望每个对象都属于最近的类别）。在这个过程中，聚类的中心保持不变。

第二部分（对应EM算法的M-step）：对每个聚类重新确定其中心，使得指派到该类别的对象到新中心的距离之和最小（即将指派到同一个类别的对象的相似度最大化）。在这个过程中，每一个样本所属的类别保持不变。



2. 实验结果与分析

调用函数计算得到NMI=0.659617，ARI=0.405131，结果并不理想。这里还尝试调库实现了GMM算法，得到NMI= 0.687105，ARI=0.551146，同样不很理想。为了直观的观察分类的情况，探究无监督学习方法在此问题上普遍表现较差的原因，我继续使用了t-SNE进行降维：



(a)实际分类情况；(b)调库实现的k-means聚类结果；

(c) 手动实现的k-means聚类结果；(d) 调库实现的GMM聚类结果

图中的（a）与上一节中的图其实是相同的，可以看到调库和手动实现都不很理想，k-means和GMM两种方法都不很理想。虽然在上一节中分析了其“线性可分”的优秀性质，但是这一算法并不适合聚类，原因在于其“同种细胞类型内”和“不同细胞类型间”的距离是不一样的，这会导致聚类算法的维度不同，这说明了使用无监督学习算法几乎无法很好的完成这一问题。

3. 总结

这一部分使用了KMeans算法对细胞进行了无监督学习中的聚类，但其效果并不理想，只有NMI=0.659617， ARI=0.405131。我们同样利用t-SNE降维技术进行可视化，发现了其数据在聚类上的特点，并说明了为何无监督学习的效果并不理想的原因所在。

四、项目总结

在本项目中，为了解决使用基因特征对细胞进行分类这一问题，我分别使用了有监督学习和无监督学习的方法进行了尝试。在有监督学习中使用了逻辑回归算法，达到了100%的正确率，且AUC=1，并且使用可视化方法探索了其数据本身的线性可分特性。在无监督学习中，使用了k-means算法，NMI=0.659617， ARI=0.405131。

该结果表明，有监督学习的方法更适合本问题，并且有潜力能够处理更大规模的细胞分类问题。

附件：gene\_cell.ipnb

(另保存为.py文件一份，以便万一没有jupyter环境的情况下可用来查看)