商业数据分析大作业

何倩怡
3200104587

摘要—本研究对企鹅数据集和高斯混合噪声数据集进行了多项实验，包括逻辑斯蒂回归、层次聚类、数据分布可视化、PCA 降维、t-SNE 降维、KNN 分类和 K-means 聚类。通过对实验结果的分析和评估，我们得出了以下结论：逻辑斯蒂回归和 KNN 分类在这两个数据集上表现良好，具有较高的预测准确率；层次聚类的效果不佳，可能是由于数据集的特点导致的。这些实验结果为进一步研究和分析这两个数据集提供了指导。

***Index Terms*—企鹅数据集、高斯混合噪声数据集、逻辑斯蒂回归、层次聚类、数据分布可视化、PCA 降维、t-SNE 降维、KNN 分类、K-means 聚类、预测准确率**

I. 引言

本文介绍了两个实验设置，分别针对企鹅数据集和高斯混合噪声数据集。在实验设置中，我们详细描述了数据预处理、训练集和测试集拆分、特征标准化以及模型训练的步骤和参数设置。

在企鹅数据集的逻辑斯蒂回归分析实验中，我们首先进行了数据清洗，删除了具有缺失值的记录，并将物种类别映射为数值型。然后，通过计算相关系数选择了与物种类别高相关性的特征。接着，我们使用分层抽样方法将数据集拆分为训练集和测试集，并对特征进行标准化处理。针对逻辑斯蒂回归模型和层次聚类模型，我们分别设置了相应的参数，并使用训练集对模型进行训练。最后，通过计算预测准确率评估了模型在测试集上的性能。

在高斯混合噪声数据集的实验中，我们首先使用 t-SNE 算法对数据进行降维，以便可视化数据集。然后，使用 KNN 算法对降维后的数据集进行分类，并计算准确率。接着，应用 K-means 算法对降维后的数据集进行聚类，并评估不同聚类结果的性能。最后，我们使用轮廓系数评估和可视化不同聚类中心数量的聚类性能，并选择具有最高轮廓系数的聚类中心数量作为最佳聚类结果。

通过这两个实验设置，我们可以深入了解和比较不同任务的数据处理步骤、模型选择和性能评估方法。这些实验为进一步的数据分析和模型开发提供了可靠的基础，并为研究人员和从业者在实际问题中应用相应方法提供了参考。

II. 方法介绍

以下是本工作应用到的方法简介

A. 逻辑斯蒂回归 (Logistic Regression)

逻辑斯蒂回归是一种用于分类问题的统计学习方法。它通过建立一个逻辑斯蒂函数来估计两个或多个可能结果之一的概率。逻辑斯蒂函数将输入特征与对应的概率联系起来。该方法常用于二分类问题，但也可以扩展到多分类任务。逻辑斯蒂回归是一种线性模型，它可以通过最大似然估计或梯度下降等优化方法来拟合参数。

B. 层次聚类 (Hierarchical Clustering)

层次聚类是一种无监督学习方法，用于将数据集中的样本分成不同的聚类或群组。该方法通过计算样本之间的相似度或距离来构建一个层次结构，其中每个样本最初被认为是一个单独的聚类，然后通过逐步合并最相似的聚类来形成更大的聚类。层次聚类可以基于聚类的连接方式分为凝聚型（自底向上）和分裂型（自顶向下）两种方法。

C. PCA 降维 (Principal Component Analysis)

主成分分析 (PCA) 是一种常用的降维技术，用于减少高维数据的维数，并保留数据中最重要的信息。PCA 通过线性变换将原始特征投影到一组新的低维特征空间，称为主成分。这些主成分是原始特征的线性组合，被排序以便按重要性降序排列。通过选择前 n 个主成分，可以选择保留的维数。PCA 的目标是最大化投

影方差，以确保保留的主成分尽可能多地解释原始数据的方差。

D. TSNE 降维 (t-Distributed Stochastic Neighbor Embedding)

t 分布随机邻域嵌入 (t-Distributed Stochastic Neighbor Embedding, t-SNE) 是一种非线性降维技术，用于将高维数据映射到二维或三维空间以进行可视化。t-SNE 通过保留数据点之间的局部相似性来构建降维后的表示，同时试图最小化全局的 Kullback-Leibler 散度。相似的样本在降维后的空间中更接近，而不相似的样本则更远离。t-SNE 特别适用于可视化复杂数据集的聚类结构。

E. KNN 分类 (K-Nearest Neighbors)

K 最近邻分类器 (K-Nearest Neighbors, KNN) 是一种基于实例的学习方法，用于进行分类任务。KNN 算法根据样本之间的距离来进行分类决策。对于给定的未标记样本，KNN 查找其最近的 K 个邻居，并根据这些邻居的标签进行投票来预测样本的类别。KNN 的性能高度依赖于距离度量的选择和 K 值的设定。

F. K-means 聚类

K 均值聚类 (K-means Clustering) 是一种常用的聚类算法，用于将数据集分成 K 个不同的聚类。该算法通过将样本分配到距离最近的聚类中心，并使用聚类中心的均值更新聚类中心来迭代优化聚类结果。K-means 聚类是一种迭代的、局部优化的方法，它试图最小化样本与其所属聚类中心之间的平方距离的总和。K-means 聚类适用于连续型特征的数据，并假设聚类的形状是球形的。然而，该算法对初始聚类中心的选择敏感，因此常常使用多次随机初始化来获得更稳定的结果。

III. 数据集说明

本工作主要运用了两个数据集。

A. 企鹅数据集

这个数据集是关于企鹅的数据集，用于综合课程作业。数据集来源于位于南极洲的帕尔默群岛，包含了 344 只企鹅的数据（经过清洗后为 333 条数据）。数据集涵盖了帕尔默群岛上三个不同岛屿上的三种不同种类的企鹅：Adelie、Chinstrap 和 Gentoo。数据集提供了每个物种的多个特征，包括 culmen（鸟喙的上脊）的长度和深度、flipper_length（脚蹼长度）、body_mass（体重）以及性别信息。

这个数据集非常适用于进行物种分类、特征分析和可视化等任务。它可以帮助我们了解不同企鹅物种之间的差异，并探索它们与特征之间的关系。通过对这个数据集进行分析，我们可以获得关于企鹅物种分类和特征相关性的深入洞见，并从中发现有趣的结论

B. 高斯混合噪声数据集

这个数据集是一个高斯混合噪声数据集，用于进行聚类分析。数据集包含 500 个样本，每个样本具有五个维度。聚类中心的个数为 5，数据集的标准差为 0.2。X.npy 文件的维度为（500，5）表示具体的数据点，y.npy 文件的维度为（500，），表示该数据点属于哪一类。

IV. 实验设置 1

在本节中，我们将介绍用于企鹅数据集逻辑斯蒂回归分析的实验设置。

A. 数据预处理

在进行逻辑斯蒂回归分析之前，我们执行了以下数据预处理步骤：

- **数据清洗：**我们删除了具有缺失值的记录。
- **类别映射：**为了将物种类别转换为数值型，我们使用了一个映射将'Adelie'、'Chinstrap' 和'Gentoo' 分别映射为 0、1 和 2。
- **特征选择：**我们选择了与物种类别相关系数较高的特征。通过计算 Cramer's V 相关系数，我们确定了'flipper_length_mm' 和'culmen_depth_mm' 这两个特征与物种类别之间的相关性较高。

B. 训练集和测试集拆分

我们将数据集拆分为训练集和测试集，以便进行模型的训练和评估。拆分比例为 70% 的数据用于训练，30% 的数据用于测试。为了确保训练集和测试集中各个类别的比例相符，我们使用了分层抽样方法。

C. 特征标准化

为了提高模型的性能并确保各个特征之间的可比较性，我们使用了标准化方法对特征进行处理。具体而言，我们使用了均值为 0，标准差为 1 的标准化器对训练集和测试集的特征进行标准化。

D. 模型训练

在这个实验中，我们使用了逻辑斯蒂回归（Logistic Regression）模型和层次聚类（Agglomerative Clustering）模型进行分类任务和无监督聚类任务。

对于逻辑斯蒂回归模型的设置，我们采用了 sklearn 库中的 LogisticRegression 类。具体参数设置为：$C = 100.0$（正则化强度的倒数）和随机种子 1。我们使用标准化后的训练集（X_train_std）对模型进行训练，并使用标准化后的测试集（X_test_std）进行模型预测。通过计算预测结果与真实标签的准确率来评估模型在测试集上的性能。

对于层次聚类模型的设置，我们采用了 sklearn 库中的 AgglomerativeClustering 类。具体参数设置为：n_clusters=3（聚类簇的数量）、affinity='euclidean'（使用欧氏距离计算样本间的相似度）和 linkage='complete'（使用完全连接法来确定簇间的距离）。我们使用原始数据集（X）进行聚类，并得到每个样本的聚类标签。

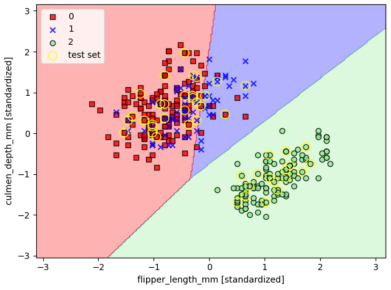
在层次聚类中，由于原始标签信息不可用，我们通过比较聚类结果与原始标签的不匹配程度来评估聚类的性能。通过定义一个比较函数，我们计算聚类标签与原始标签的误分类比例。然后，我们尝试不同的标签映射方式，将聚类标签重新映射为与原始标签相对应的标签，计算误分类比例的最小值作为最终结果。

在 PCA 降维部分的实验设置中，我们选择了'culmen_length_mm'、'culmen_depth_mm'、'flipper_length_mm' 和'body_mass_g' 作为特征列，并使用 StandardScaler 进行特征标准化，然后使用 PCA 进行降维，保留了 2 个主成分。

V. 实验结果 1

A. 逻辑斯蒂回归

逻辑斯蒂回归分类的结果如下图所示，可以看出，Gentoo 的类别区分较为优秀，但 Adelie 及 Chinstra 类别区别较为模糊。整体正确率为 0.84，说明模型在整体上有较好的预测准确性。然而，各个类别的正确率存在差异。类别 0（Adelie）的正确率为 0.977，类别 1（Chinstra）的正确率为 0.25，类别 2（Gentoo）的正确率为 1.0。这表明模型在类别 1 的预测上存在较大的误差，可能是由于类别 1 的样本量较小（20）导致支持度不够，影响了模型的表现。



通过分类报告（classification report）可以看到，模型的精确度（precision）、召回率（recall）和 F1-score 等指标在不同类别上有所差异。总体来说，宏平均（macro avg）的精确度为 0.86，召回率为 0.74，F1-score 为 0.74，加权平均（weighted avg）的精确度为 0.85，召回率为 0.84，F1-score 为 0.81。

综上所述，模型在整体上有较高的预测准确率，但在某些类别上存在较低的召回率，可能是由于样本量不足造成的。

B. 层次聚类

聚类结果显示，每个样本被分配了一个聚类标签，表示其所属的聚类簇。然而，我们发现误分类比例非常高，达到了 64.5%。这表明层次聚类在这个数据集上的效果非常不佳。

为了尝试改善聚类结果，我们尝试了多种标签映射方式，并计算了每种映射方式下的误分类比例。然而，即使在尝试不同的映射方式后，最低的误分类比例仍然较高，为 42.34%。

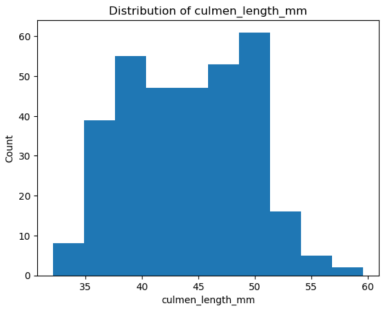
综上所述，层次聚类在这个数据集上的效果不佳。可能的原因是层次聚类忽略了标签信息，并且数据样本之间的内在结构关联性较弱。因此，选择无监督学习模型并不适用于这个数据集。

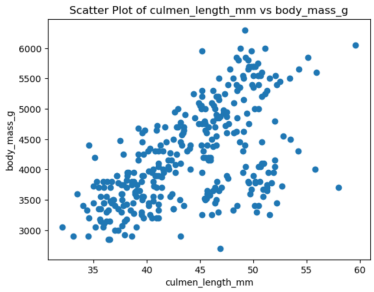
C. 数据分布可视化

我们使用 matplotlib 库绘制了两个图形。第一个图形是关于'culmen_length_mm'（喙长）的直方图。该直方图可以帮助我们了解'culmen_length_mm' 的分布情况，即不同取值范围内的样本数量。

第二个图形是关于'culmen_length_mm' 和'body_mass_g' 之间的散点图。散点图展示了这两个特征之间的关系，其中'culmen_length_mm' 表示 x 轴，'body_mass_g' 表示 y 轴。通过这个散点图，

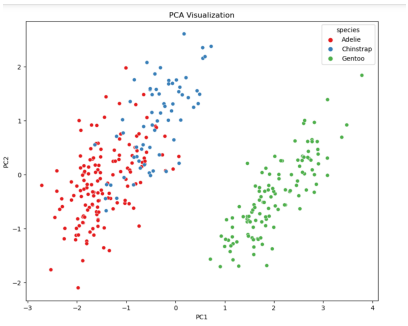
我们可以观察到两个特征之间的趋势或者是否存在相关性。





D. *PCA* 降维

通过 PCA 降维后，我们使用降维后的主成分数据进行了可视化。观察图形可以得出类似之前的结论：Gentoo Penguin 的类别区分较好，但 Adelie Penguin 和 Chinstrap Penguin 的类别区分相对较模糊。



VI. 实验设置 2

由于数据本身比较理想化，故高斯混合噪声数据集不需要太多的预处理操作，但为了后续步骤的方便进行，我调整了四个步骤的顺序，大致设置如下：

A. *t-SNE* 降维

为了可视化数据集，我们使用 t-SNE 算法将数据降低到二维。首先，我们使用 numpy 库加载数据集。然后，创建一个 t-SNE 对象并将数据集进行降维，得到降维后的数据集表示为 X_{tsne}。最后，我们使用 matplotlib 库将降维后的数据集可视化。

B. *KNN* 分类

我们使用 KNN 算法对降维后的数据集进行分类。首先，我们将数据集划分为训练集和测试集。然后，创建一个 KNeighborsClassifier 对象，并在训练集上进行训练。接下来，我们使用训练好的分类器对测试集进行预测，并计算准确率。最后，我们使用 matplotlib 库将预测结果可视化。

C. *K-means* 聚类

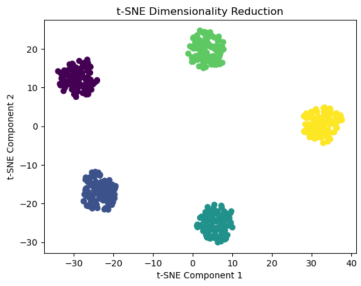
我们使用 K-means 算法对降维后的数据集进行聚类。我们尝试不同的聚类中心数量，并评估每种聚类结果的性能。对于每个聚类中心数量，我们计算轮廓系数来衡量聚类的紧密度和分离度。最后，我们选择具有最佳轮廓系数的聚类结果，并使用 matplotlib 库将最佳聚类结果和聚类中心可视化。

D. 轮廓系数评估和可视化

我们绘制了轮廓系数随聚类中心数量变化的折线图，以评估不同聚类中心数量的聚类性能。在图中，横轴表示聚类中心数量，纵轴表示轮廓系数。我们可以观察到轮廓系数随着聚类中心数量的增加而变化的趋势，并选择具有最高轮廓系数的聚类中心数量作为最佳聚类结果。

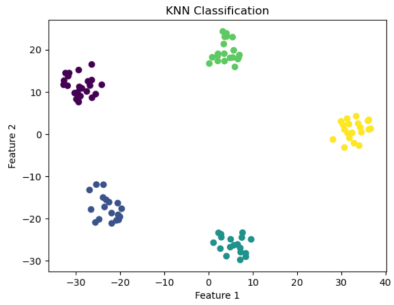
VII. 实验结果 2

A. *t-SNE* 降维



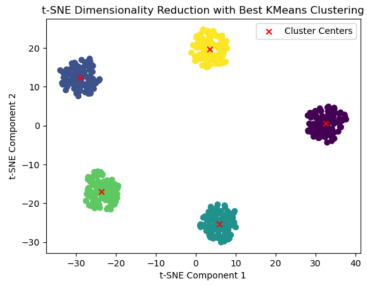
降维成二维的结果如图所示，可以发现呈现了明显的聚类特征，和理论符合。

B. KNN 分类



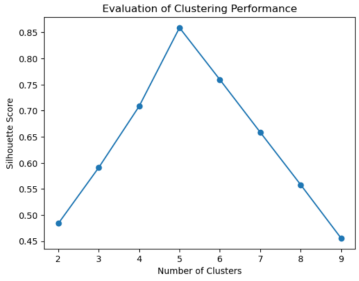
分类结果如图所示，可以发现接近的点颜色相同，说明被分到了相同的一类，同时计算了正确率，发现为1，说明分类结果很好。

C. K-means 聚类



我们尝试了多个不同的聚类中心数量方案，通过对比轮廓系数，发现聚类数量为 5 时效果最好，结果如图所示，红色的 X 表示 Kmeans 方法计算得到的聚类中心，和理论结果也是符合的。

D. 轮廓系数评估和可视化



通过折线图不难看出，选择 5 作为聚类中心数量明显优于其他值。

VIII. 结论

本研究通过对企鹅数据集和高斯混合噪声数据集进行了逻辑斯蒂回归、层次聚类、数据分布可视化、PCA 降维、t-SNE 降维、KNN 分类和 K-means 聚类等实验，并对实验结果进行了分析和评估。

在逻辑斯蒂回归实验中，整体上模型表现良好，具有较高的预测准确率。然而，针对不同类别的预测结果存在差异，其中类别 1（Chinstra）的预测准确率较低，可能是由于样本量不足导致的。因此，在未来的研究中，应该增加类别 1 的样本量，以提高模型的性能。

在层次聚类实验中，结果显示层次聚类在该数据集上效果不佳，误分类比例较高。尝试不同的标签映射方式后，效果仍然不理想。这可能是因为层次聚类忽略了标签信息，并且数据样本之间的内在结构关联性较弱。因此，无监督学习模型不适用于这个数据集。

数据分布可视化实验通过直方图和散点图展示了数据的分布情况和特征之间的关系。这些可视化结果有助于我们对数据集的特征有更深入的理解。

PCA 降维实验结果与之前的实验结论相似，Gentoo 企鹅的类别区分较好，而 Adelie 企鹅和 Chinstrap 企鹅的类别区分相对模糊。

t-SNE 降维实验结果呈现了明显的聚类特征，符合理论预期。

KNN 分类实验结果表明分类结果良好，正确率为1，说明分类效果很好。

K-means 聚类实验通过对比轮廓系数选择了最佳的聚类中心数量，结果与理论预期一致。

综上所述，通过多种实验方法和评估指标对企鹅数据集和高斯混合噪声数据集进行分析，我们得出结论：逻辑斯蒂回归和 KNN 分类在这些数据集上表现较好，而层次聚类效果不佳。这些实验结果为进一步研究和分析企鹅数据集提供了参考和指导。未来的研究可以进一步改进模型和算法，提高预测和分类的准确性。