大数据原理与技术 Lab6

20337025 崔璨明

实验过程

实现的K-means函数如下:

```
#kmeans算法

def kmeans(X, n_clusters, max_iterations=10000):
    # 初始化聚类中心
    centroids = X[np.random.choice(range(X.shape[0]), size=n_clusters,
replace=False)]

for _ in range(max_iterations):
    # 分配样本到最近的聚类中心
    labels = assign_labels(X, centroids)
    # 更新聚类中心
    new_centroids = update_centroids(X, labels, n_clusters)
    # 检查聚类中心是否收敛
    if np.allclose(centroids, new_centroids):
        break
    centroids = new_centroids
return labels
```

它接受数据集 x 、聚类数目 n_clusters 和最大迭代次数 max_iterations 作为参数。首先通过随机选择样本来初始化聚类中心 centroids。然后,使用 assign_labels 函数将样本分配到最近的聚类中心,并使用 update_centroids 函数更新聚类中心。重复以上步骤,直到聚类中心收敛或达到最大迭代次数。最后,函数返回分配给每个样本的标签 labels ,即聚类结果。

在主函数中,加载鸢尾花数据集(Iris),该数据集包含了四个特征。使用其中的数据 x 和标签 y_true。然后,指定聚类数目 n_clusters=3 ,并调用 kmeans 函数运行K-means聚类算法,将聚类的 结果存储在labels中。

最后,调用 sklearn 库的 metrics.normalized_mutual_info_score 函数计算标准化互信息 (Normalized Mutual Information, NMI) 作为聚类结果的评估指标。NMI的取值范围为0到1,值越接近1表示聚类结果与真实标签的一致性越高。

实验结果

在UCI的鸢尾花数据集上的实验结果如下,得到的NMI值为0.7581756800057784,表示聚类结果与真实标签的一致性较高:

我随即在uci的其他数据集上进行了实验,得到结果如下:

数据集	iris	breast_cancer	wine
聚类数目	3	2	3
NMI	0.758	0.464	0.428

可见聚类结果的好坏也会受到数据集特性的影响。

总结 && 感想

通过实现K-means聚类算法,并在鸢尾花数据集上进行实验,我得到了较高的NMI值,这说明聚类结果与真实标签的一致性较高。这表明K-means算法在这个数据集上表现良好,能够有效地将样本分配到不同的聚类中心。然而,需要注意的是,聚类结果的好坏也受到数据集特性的影响,对于其他数据集可能会有不同的表现。因此,在实际应用中需要根据具体情况选择适合的聚类算法和评估指标。

这次复现K-means聚类算法的实验让我对聚类算法有了更深入的理解。通过编写代码和运行实验,我能够清楚地看到算法是如何工作的,这也加深了我在课上学习到的理论知识。