Lab2 report

基本算法

K-means是一种常见的无监督学习算法，用于将数据集划分为K个簇（cluster）。其基本思想是通过迭代寻找簇的质心（centroid），并将数据点分配给距离其最近的质心所属的簇。

假设有一个包含N个数据点的数据集$[X = \{x_1, x_2, \ldots, x_N\}]$，每个数据点($x_i$)是一个d维向量，表示一个样本。

1. **初始化**：选择K个初始质心[$C = \{c_1, c_2, \ldots, c_K\}$]，通常可以随机选择数据集中的K个点作为初始质心。
2. **分配数据点到最近的质心**：对于每个数据点(x_i)，计算其与所有质心的距离，将其分配到距离最近的质心所属的簇中，形成K个簇[$S = \{S_1, S_2, \ldots, S_K\}$]，其中($S_k$)表示第k个簇中的数据点集合。

[$S_k = \{x_i : \|x_i - c_k\|^2 \leq \|x_i - c_j\|^2, \forall j, 1 \leq j \leq K\}$]

3. **更新质心**：对每个簇(S_k)，重新计算其质心为其中所有数据点的平均值。

[$c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i$]

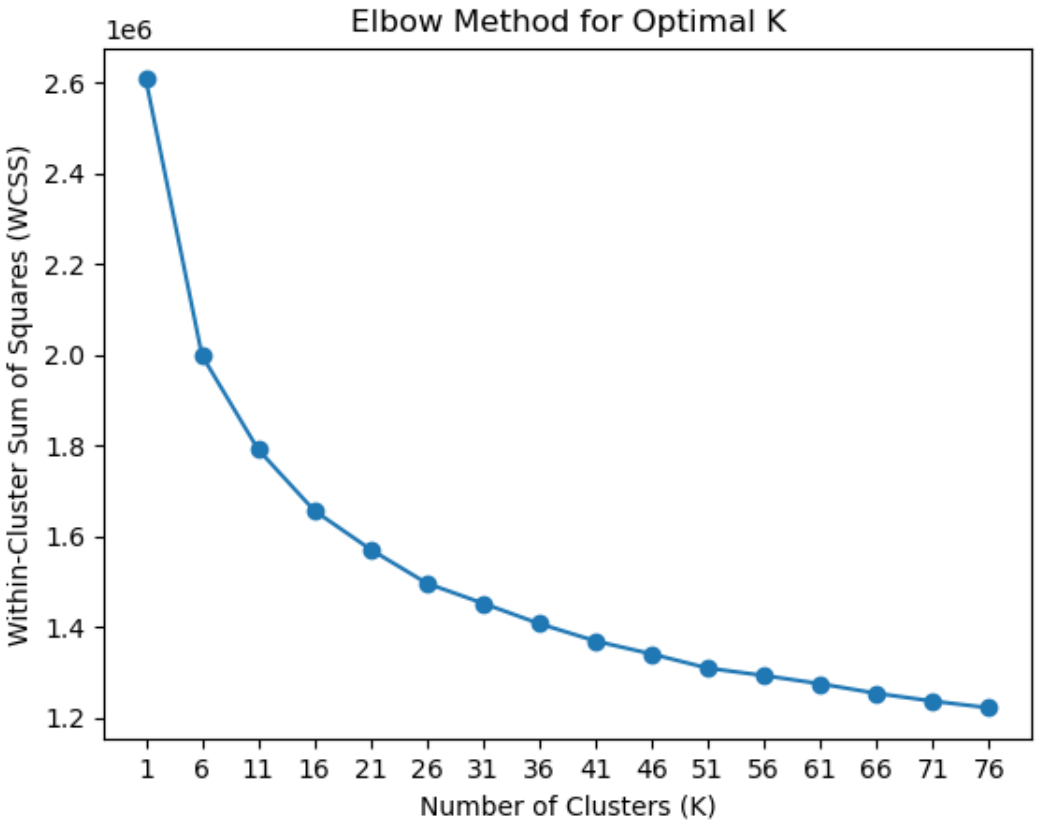
4. **重复步骤2和步骤3**，直到质心不再发生显著变化，或达到预先设定的迭代次数。
5. **收敛**：算法收敛后，得到了K个簇，每个簇由一组数据点和一个质心表示。

K均值算法的优化目标是最小化所有数据点与其所属簇质心之间的距离的平方和，即最小化下面的损失函数：

[$J = \sum_{k=1}^{K} \sum_{x_i \in S_k} \|x_i - c_k\|^2$]

确定聚类数量

通过实验发现，尽管 cifar10 有十个类，但把 K 调大对于提升最后的准确率有一定的帮助，这可能是因为增加簇的个数有利于识别每个类别内部的差异。我采用了课上提到的 elbow 方法，即绘制K值和簇内距离 (WCSS) 的折线图。



通过对 elbow 图的观察，选取作为K=30作为初始值

准确率衡量

我采用了多数投票方法来对Kmeans算法分类结果进行衡量。形式化定义如下：

假设有一个数据集 D，包含 n 个样本，每个样本 x_i 属于一个类别 C_i，共有 k 个不同的类别。令 T 为分类结果，其中 T_i 表示第 i 个样本的分类结果。

1. 对于每个样本 x_i，进行 m 次分类运行，得到 m 个分类结果向量 $T^{(1)}, T^{(2)}, \ldots, T^{(m)}$，其中 $T_i^{(j)}$ 表示第 j 次运行中第 i 个样本的分类结果。
2. 对于每个样本 x_i，统计其 m 个分类结果中出现次数最多的类别，即 $T_i = \arg\max_C \sum_{j=1}^{m} \mathbb{I}(T_i^{(j)} = C)$ 其中 $\mathbb{I}$ 是示性函数
3. 最终的多数投票分类结果为 T，其准确率可以通过与真实类别比较来进行衡量。

初始化聚类中心

我对比了随机初始化聚类中心和最远初始化两种初始化方法对聚类效果的影响，随机初始化方法初始化聚类中心，即随机选取 K 个样本作为初始中心，所谓最远初始化，是指每次选取所有点中距离当前所有距离中心距离和最远的样本作为下一个引入聚类中心，即假设初始化中心
$S = \{s_1 \cdots s_k\}, s_{i+1} = max_{x_j}(\sum_{u=0}^{i} distance(x_j, x_u))$

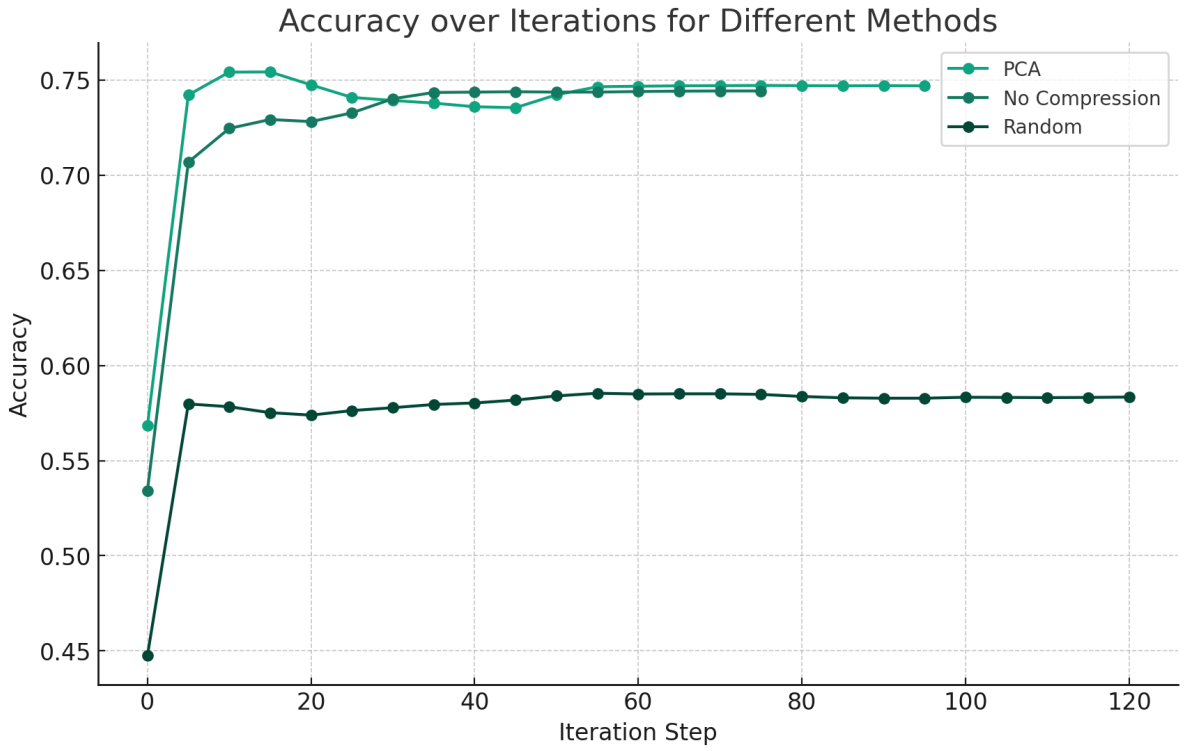
	随机初始化	**最远初始化**
准确率（%）	74.70	74.58
收敛所需迭代次数	127	117

尽管理论上随机初始化聚类中心会导致收敛速度变慢，但在实践中我发现对于大部分实验设定，随机初始化在10-30次后准确率基本就不再变化，而且收敛所需的迭代步数也不会高太多，结合之后的早停策略，这种迭代步数上的差距并不会对算法运行效率产生太大的影响。而最远优先原则方法在初始化时会面临计算复杂度高，效率低下的问题，且在初步的尝试中发现与随机初始化差异不大，因此最终选择完全随机化方法初始化聚类中心。

特征向量表示图片

本节我尝试对比了直接使用原始像素特征，随机采样和PCA三种压缩方法对最终结果的影响。其中随机采样是指从原有图像特征中随机抽取 D 维组成特征向量，实验中使用原始像素特征的特征维数为768，其余均为50。

压缩策略	不压缩	随机采样	PCA
收敛时准确率（%）	74.43	58.34	74.70
收敛所需迭代次数	77	127	97



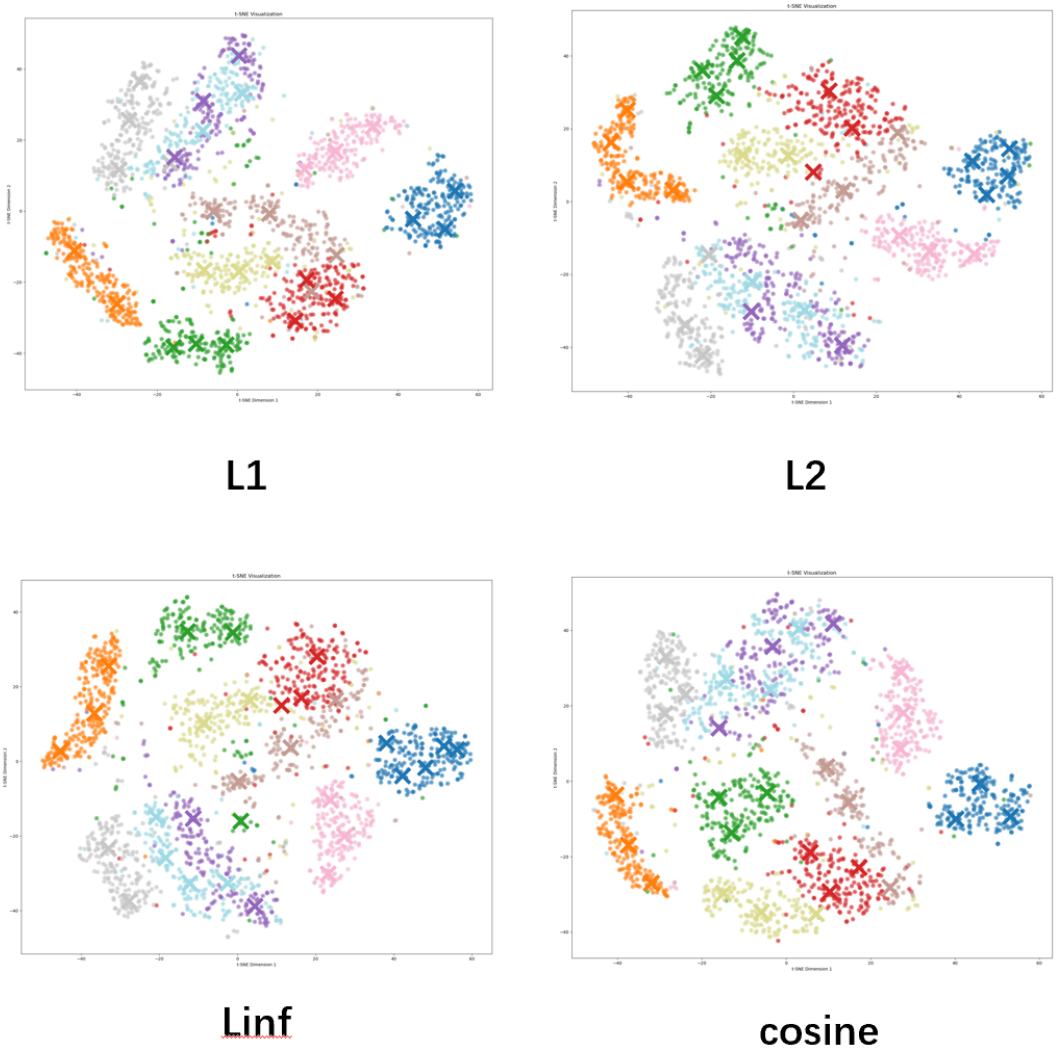
上图绘制了三组实验Kmeans迭代过程准确率的变化，可以看到，PCA和不压缩方法提取的特征向量在收敛时准确率相差不大，而且不压缩维度的方法收敛所需的步数略少，而随机提取的特征则在准确率和收敛步数上都逊于前两者。**这说明，PCA压缩技术可以从原向量中提取适合于Kmeans聚类的特征，其效果与不压缩基本一致，而如果考虑到准确率最大值和前期的准确率增长速度，PCA更胜一筹。**

衡量样本距离

本节中我选取了一范数、二范数、无穷范数和余弦相似度作为距离的衡量标准

	L1	L2	L∞	cosine
准确率(%)	74.39	74.77	68.39	76.84

tsne可视化结果



为了进一步说明四种距离函数对聚类效果的影响，我进一步绘制了它们在30次迭代时的tsne图，其中不同的颜色代表真实标签，·表示向量经过tsne可视化降维后在二维空间中的位置（从样本中随机采样2000个点），**X**表示聚类中心的位置，其类别匹配的策略和准确率衡量策略一致，都是采用投票法取多数label。

- 可以看到，对于比较简单的类别（如深蓝色、深红色），其特征向量和其它类别的特征向量距离较远，四种方法聚类中心基本都和其特征向量中心一致。说明对于较简单的类别，四种距离函数聚类效果均较好
- 对于无穷范数，其准确率明显低于其它距离函数，从tsne图中也可以反映出这一点：和其它实验相比，使用无穷范数会将为一些密度并不大的深绿色类别的特征向量生成聚类中心（tsne图的中下方深绿色锚点），而忽视了真正的深绿色类别集中的区域（图中上方深大量绿色圆点）
- 余弦函数相比其它距离函数准确率有轻微的提升，从tsne可视化来看，灰色的锚点不像其它几张图那样有一部分落入了浅蓝色样本的区域，同时紫色和浅蓝色这两组重合度比较高的样本锚点之间也距离较远，这对于提高准确率很有帮助。

设置终止条件

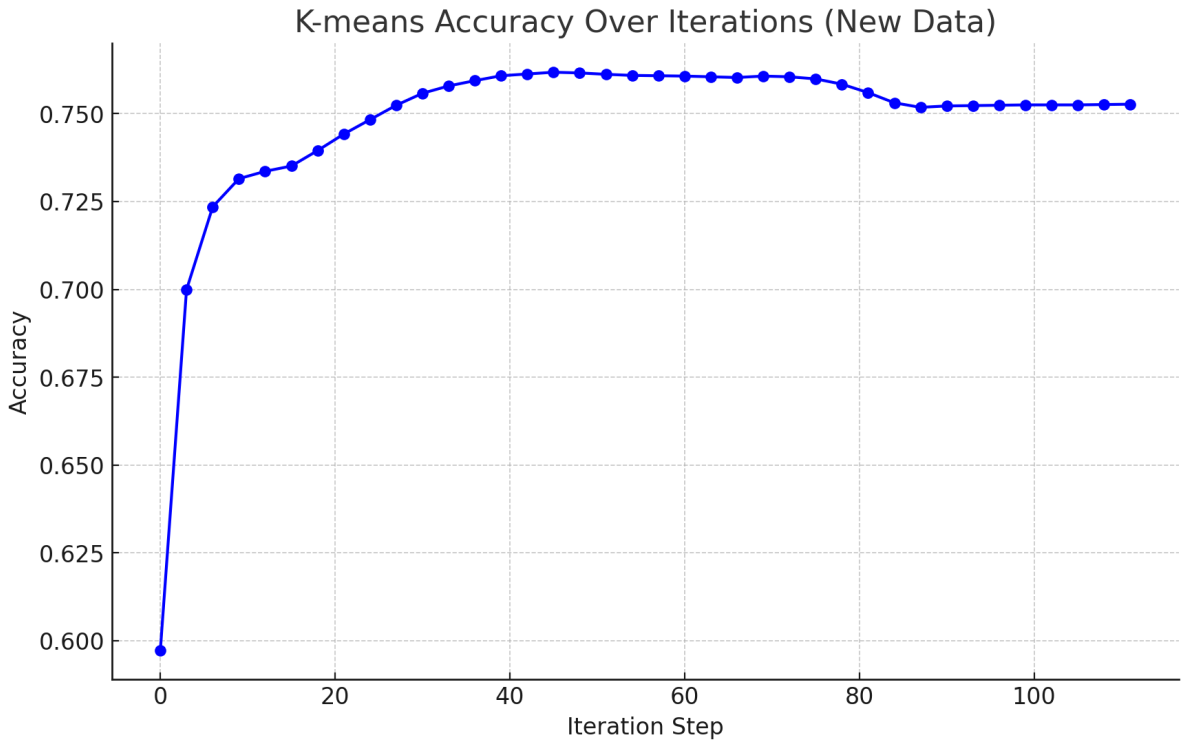
K-means 算法终止条件为当两次迭代不会改变簇中心的位置时，算法收敛终止，通过实验观察，我发现尽管K-means迭代到后面还存在小幅更新，但准确率已经不会出现太大的变化，因此在一些对比实验中，我设置了最大迭代次数来实现早停，而且这一操作本身并不涉及标签的知识，因此可以认为符合无监督学习要求。

最终实验设定及可视化分析

最终我选取的参数设定如下

	聚类数量	聚类中心初始化策略	特征向量压缩方式及维度	样本距离衡量方式
参数	30	random	pca, dim=50	cosine

最后实验在 114步收敛，收敛准确率为75.26%，训练中最高准确率76.18%



实验分析:

- 在kmeans算法前期，准确率上升十分迅速，而在30次迭代后，kmeans算法准确率已经基本趋于稳定，这与之前的实验结论一致
- **80次迭代后，准确率有所下滑，但根据无监督学习的定义，此时不能终止训练**，因为理论上准确率数据只有在训练结束后才能被观测到，这说明适当设定最大迭代步数或其它不基于label的方式以早停有一定的必要性。

  同时我绘制了训练过程中的tsne图，可视化策略与"衡量样本距离"一节中说明一致，下图展示了训练第1，12，和最后一次迭代后的tsne可视化结果

实验分析:

- 通过可视化结果可以看到，**总体上来说，随着迭代步数增加，距离中心距离真正的样本中心逐渐靠近**，例如，第一次迭代后的橙色样本存在一个离群锚点，而在后两张图中锚点基本出于橙色样本中心。

- 通过可视化结果可以看出**kmeans对简单样本和难样本学习差异，总的来说，简单样本收敛速度快于难样本**，对于简单样本（如绿色、橙色、蓝色样本）在第一次迭代后基本就能正确找到其中心所在，锚点在之后的训练过程中和该类样本的相对关系也不会发生太大的变化。而对于难样本（红圈圈出的灰、紫、蓝交融区域），一开始得到的锚点存在大量偏移样本中心的情况。例如第一张图陷入浅蓝色区的灰色锚点和缺失的紫色锚点，而在第二张图中离群的灰色锚点消失，紫色锚点有所增加，到最后收敛时灰色锚点再次恢复为3个并且基本回到灰色样本的中心，而紫色锚点也从一开始的1个到最后的三个，三类样本的锚点数量基本相等并且都比较靠近样本点中心。这说明难样本的学习是影响kmeans算法后期效率和准确率的关键制约因素。

其它实验分析

压缩维度对准确率和算法效率的影响

在使用PCA压缩技术下，探究压缩维度对准确率和算法效率的影响，选取15次迭代后测量准确率

在压缩维度小于50时，随着压缩维度增加，准确率稳步上升，超过50后，维度增加不会带来准确率的明显提升。时间成本上，相同迭代次数下，压缩维度增加会稍稍降低训练效率。

压缩维度	**2**	**5**	**10**	**20**	**50**	**100**	**no compress**
准确率（%）	44.31	59.74	69.77	72.51	**76.83**	75.60	75.28
用时（s）	435	451	343	375	391	414	422