第7章：特征提取和特征选择（作业）

周强 电子学院 202128019427002

简述题

1. **简述PCA的原理、学习模型和算法步骤。**

   PCA降维是一种线性降维方法，其目标将原始数据投影到一组方差较大的方向，达到数据变换和降维的同时，尽可能保留原始数据信息的目的。其中方差最大的投影方向称为第一主成分，其次为第二主成分，以此类推。

   设原始数据为n个m维空间的样本，即$x_i \in R^m, i = 1, 2, 3, \ldots, n$。我们的目标是求一个变换矩阵$W \in R^{m \times d}$,使得投影后的样本$y_i = W^T x_i \in R^d, i = 1, 2, \ldots, n$是在主成分的方向上。

   投影后的均值为

   $$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} W^T x_i = W^T \bar{x}$$

   其中$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$是原始数据的均值。

   投影后的方差为

   $$var = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T = \frac{1}{n} \sum_{i=1}^{n} W^T (x_i - \bar{x})(x_i - \bar{x})^T W = W^T C W$$

   其中$C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \in R^{m \times m}$是原始数据的协方差矩阵。因此我们的问题转换为

   $$\max W^T C W$$
   $$s.t. W^T W = I$$

   使用拉格朗日乘子法，即目标函数变成

   $$\max obj = \max W^T C W - \lambda (W^T W - I)$$

   目标函数对W求导并置零，即

   $$\frac{\partial obj}{\partial W} = 2CW - 2\lambda W = 0$$

   则有

   $$CW = \lambda W$$

   将W按列展开，则W各列是C的特征向量。

   因此，为了计算投影后的d个主成分方向，需要先计算原始数据的协方差矩阵C，求C的特征向量和特征值。取最大的d个特征值的对应的特征向量作为主成分的方向即可。

   将PCA降维的步骤总结如下

   1. 设$X \in R^{n \times m}$是原始数据矩阵，代表n个m维向量，即X的每一行是一个样本。
   2. 计算均值向量：$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$。
   3. 去均值化：X的每一行减去均值向量。

4. 计算协方差矩阵：$C=\frac{1}{n}X^TX$。
5. 计算C的特征值和特征向量。
6. 将特征向量按照对应特征值降序排列，取前d个特征向量组成变换矩阵$W\in R^{m\times d}$,W的每一列是一个特征向量。
7. 降维：计算$Y=XW\in R^{n\times m}$即为降维后的数据。

2. **简述LDA的原理和学习模型。**

LDA是一种有监督的线性降维方式，主要针对分类问题进行降维。其基本思想是寻找一组投影方向，使样本在投影之后满足

1. 类内的样本点尽可能相互靠近。
2. 类间的样本点尽可能相互远离。

满足上述约束条件的投影方向对于特定的分类任务有效。从约束条件可知，我们需要知道哪些样本属于同一类，即样本的标签。投影的维数通常小于原始数据的维数，因此投影样本相当于将样本在子空间内表示，从而达到降维的目的。

下面以二分类问题为例阐述LDA的算法思想。设样本集$D=\{(x_j,y_j)\},j=1,2,3...,n,y_j\in\{0,1\}$。设$X_i,\mu_i,\Sigma_i$分别表示第$i\in\{0,1\}$类样本的样本集合、均值向量和协方差矩阵。将数据投影到一条直线上，则样本中心在该直线上的投影分别时$w^T\mu_0,w^T\mu_1$,投影后样本的协方差矩阵分别为$w^T\Sigma_0w,w^T\Sigma_1w$。欲使同类样本更接近，即期望$w^T\Sigma_0w+w^T\Sigma_1w$尽可能小；欲使不同类样本尽可能远离，即期望$||w^T\mu_0-w^T\mu_1||^2$尽可能大。因此LDA的目标是最大化如下函数：

$$J=\frac{||w^T\mu_0-w^T\mu_1||^2}{w^T\Sigma_0w+w^T\Sigma_1w}$$

为了更方便的表示目标函数，同时将LDA推广到多分类问题的情形，我们可以定义如下矩阵。

1. 类内散度矩阵。

$$S_w=\Sigma_0+\Sigma_1$$

2. 类间散度矩阵。

$$S_b=(\mu_0-\mu_1)(\mu_0-\mu_1)^T$$

3.

因此目标函数可以写成如下形式

$$J=\frac{w^TS_bw}{w^TS_ww}$$

由于目标函数值与长度无关，仅与方向有关，不失一般性可令$w^Tw=1$。应用拉格朗日乘子法有

$$S_bw=\lambda S_ww$$

即

$$S_w^{-1}S_bw=\lambda w$$

换言之，λ是矩阵$S_w^{-1}S$的特征值，w是特征向量。

将LDA推广到高纬，并总结其步骤如下。

1. 设样本集为$D=\{(x_j,y_j)\},j=1,2,3...,n,y_j\in\{0,1,\dots,c\}$,即n个样本分属c类。

2. 计算全局散度矩阵。

$$S_t = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$$

其中$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$为全局均值。

3. 计算类内散度。

$$S_w = \sum_{j=1}^{c} S_{wj}$$

其中$S_{wj} = \sum_{x \in X_j} (x - \mu_j)(x - \mu_j)^T$是第$j$类的协方差矩阵，$\mu_j = \frac{1}{n} \sum_{x \in X_j} x$。

4. 计算类间散度矩阵。

$$S_b = S_t - S_w$$

5. 计算$S_w^{-1} S_b$，并计算其特征值和特征向量。
6. 将特征向量按照对应特征值降序排列，取前d个特征向量组成变换矩阵W。
7. 降维：$y_i = W^T x_i$即为降维后的数据。

3. **作为一类非线性降维方法，简述流形学习的基本思想。**

流形是一种具有局部欧几里得空间性质的空间，但是在全局欧氏空间不成立，通过线性投影将高维数据降维到低维空间难以展开非线性结构。因此流形学习是一种非线性降维方式，其基本思想是高维空间中相似的样本映射到低维空间以后保持相似性。

经典的流形学习算法有**局部线性嵌入（LLE)、Isomap、拉普拉斯特征映射（LE）**。

1. LLE。其基本思想是给定数据集后，通过最近邻等方式构造一个数据图，然后在每一个局部空间，高维空间中的样本的线性重构关系在低维空间中均得以保持。
2. Isomap。其基本思想是给定数据集后，通过最近邻等方式构造一个数据图，然后计算任意两个点之间的最短路径（即测地距离）。对于任意两个点，期望在低维空间中保持其测地距离。
3. LE。其基本思想是给定数据集后，通过最近邻等方式构造一个数据图，在每一个局部图内计算点与点之间的亲和度（相似度），期望亲和度在低维空间得以保持。

4. **根据特征选择和分类器的结合程度，简述特征提取的主要方法，指出各类方法的特点。**

高维数据具有计算复杂度高、特征冗余等缺点，因此需要到在低维空间处理数据，特征选择是处理高维数据的主流方式之一。其基本方法是给定学习任务后，对于给定的数据属性集，从中选择和任务相关的特征子集，从而缓解维数灾难，去除冗余特征，提高分类器性能。特征选择的总体技术路线分为如下两步：

1. **子集搜索。**从特征集合$x_1, x_2, \ldots, x_d$中选搜索最优的特征子集。
2. **子集评价。**对于某种特征的特征子集，依据某种评价准则，对该特征子集进行评价。

根据特征选择和分类器的结合程度，可以将特征选择方法分为以下三类。

1. **过滤式特征选择。**其基本思想是定义评价函数后，度量某个给定特征和类别标签之间的相关性，最后选择具有最大相关度的一些特征作为选择结果。这类方法的特点是先对数据集进行特征学习，然后再训练学习器，特征选择过程与后续学习器无关，与启发式特征选择方式相比，它无法获得最优子集；与包裹式特征选择方式相比，它大大降低了计算成本。这类方法的特点是特征选择独立于学习器。
2. **包裹式特征选择。**其基本思想是先对数据集进行特征选择，然后再训练分类器，特征选择过程与分类单独进行，特征选择评价判据间接反应分类器性能。这类方法的特点是特征选择依赖于学习器。
3. **嵌入式特征选择。**这类方法的特点是特征选择与学习同时进行。

5. **简述最优特征提取的基本思想。**

最优特征选择可以采用穷举法和分支定界法。

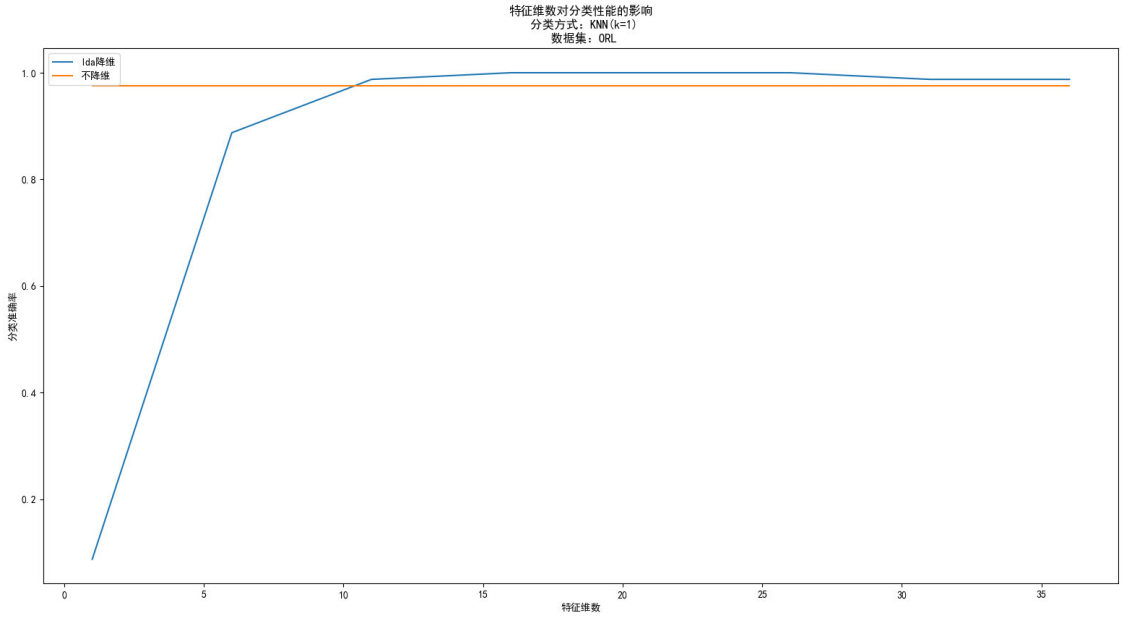
1. **穷举法**：从给定的d个特征中，遍历2^d个子集，依据评价函数找出最优特征子集。穷举法的计算复杂度巨大，子集个数是特征维数的指数。
2. **分支定界法**。将所有可能的特征选择组合以树的形式表示，采用分支定界法对树进行搜索，使得搜索尽可能到达最优解而不必搜索整个树。这种方法依赖于评价准则关于特征的单调性，即包含的特征增多时，判据值不会减少。

编程题

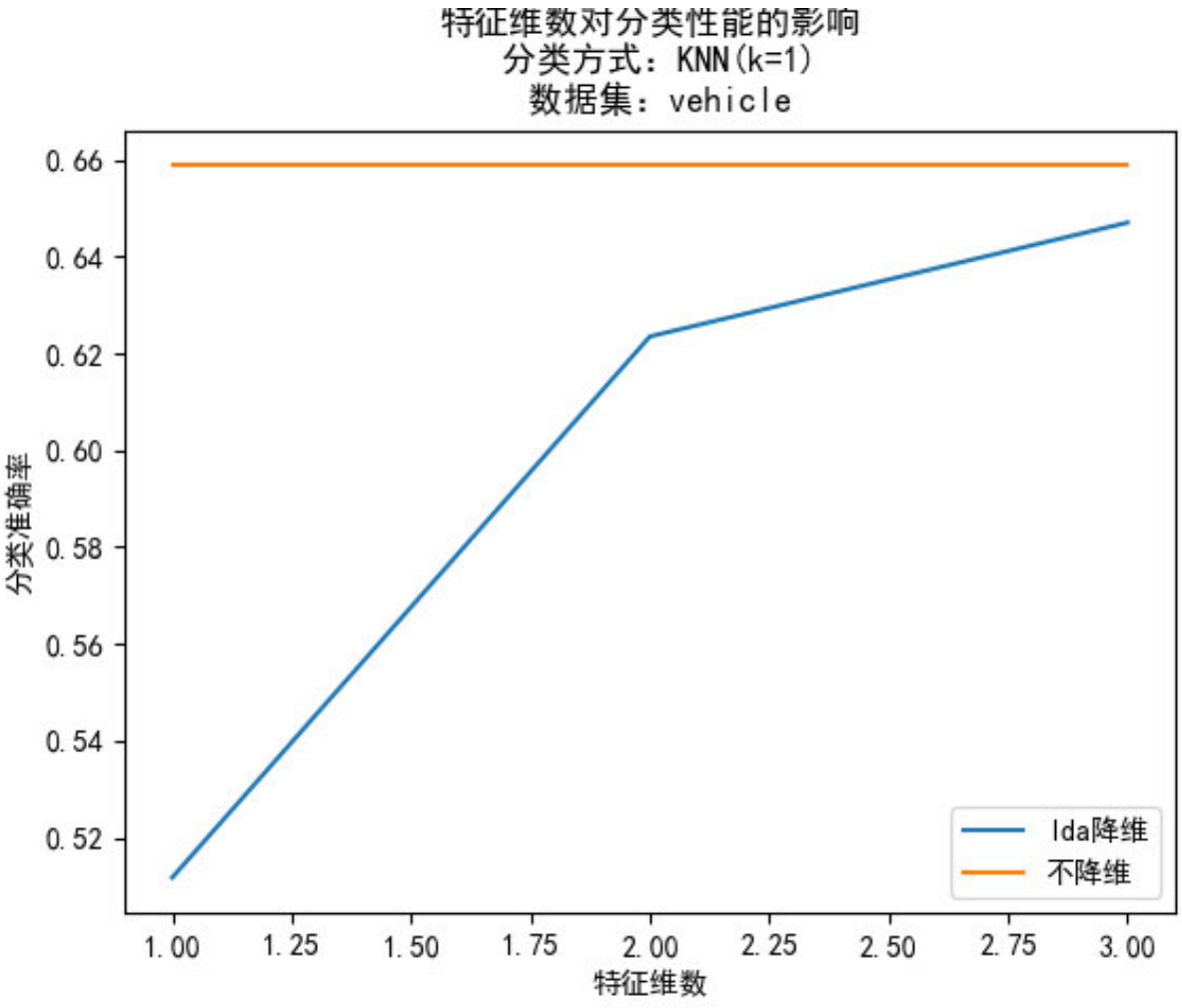
1. **编程实现：PCA+KNN，即首先进行PCA降维，然后采用最近邻分类器（1近邻分类器）作为分类器进行分类。**
2. **编程实现：LDA+KNN，即首先进行LDA降维，然后采用最近邻分类器（1近邻分类器）作为分类器进行分类。**
3. **采用80%作样本训练集，20%作样本测试集，报告降至不同维数时的分类性能。**

   在完成PCA降维和KNN分类后，采用ORL和Vehicle数据集研究不同降维方式和特征维数对分类性能的影响。

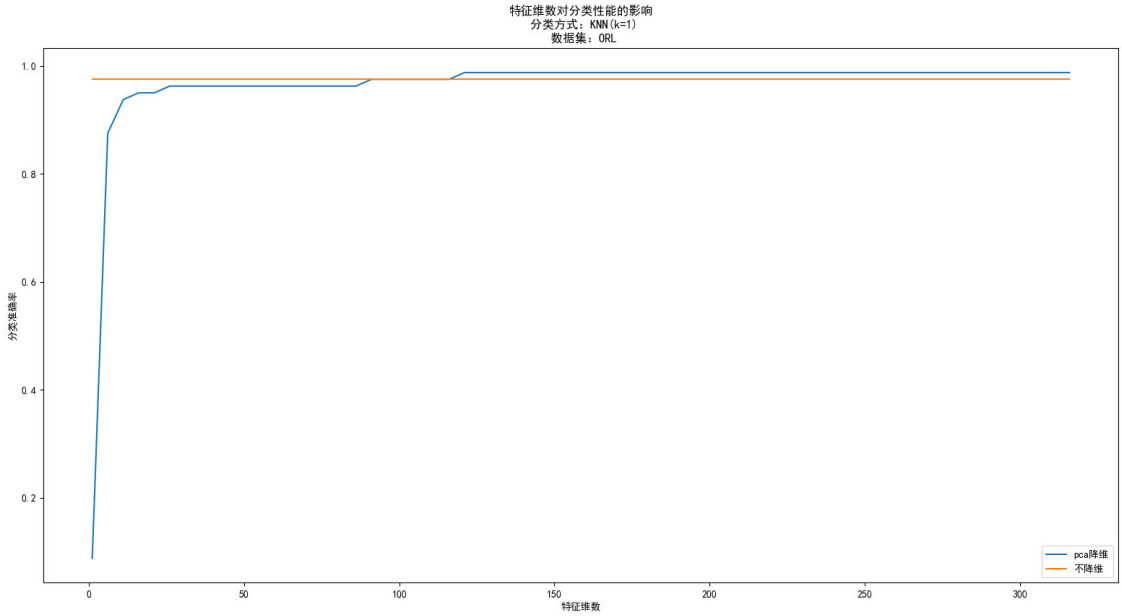
实验1：特征维数对分类性能的影响。



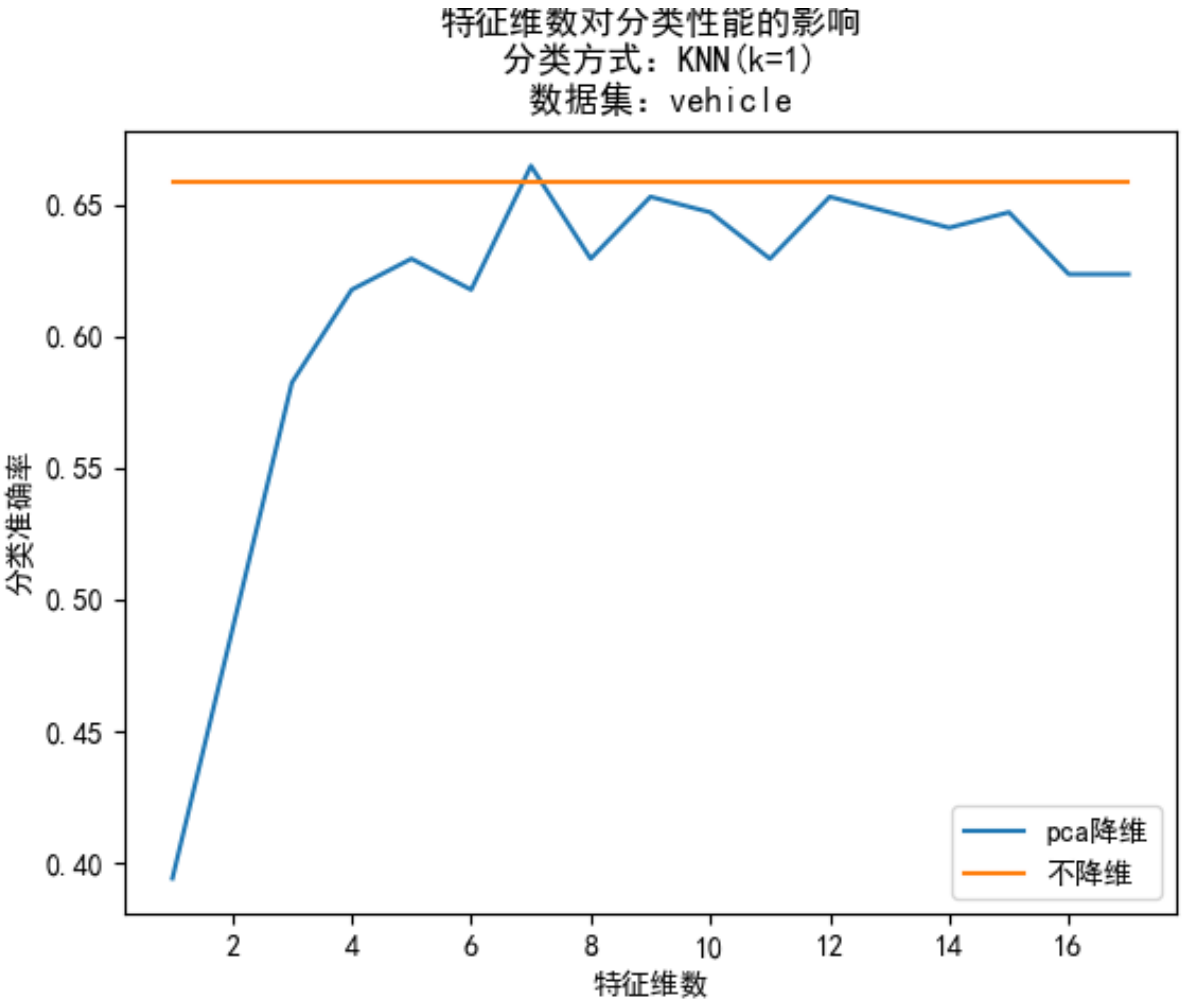
所用数据集维ORL，降维方式是LDA，降维时以5为间隔，研究特征维数对分类性能的影响。特征维数较低时，分类性能较差，即较少的特征不能充分保留原始数据集的信息。随着特征维数的增加，分类性能显著提升。当降维到11维时，分类性能与原始接近。继续增加特征维数，分类性能趋于平稳。



所用数据集是Vehicle，降维方式是LDA。此数据集仅有4类样本，因此使用LDA降维时，最大的特征维数为3。随着特征维数的增加，分类性能显著增加，但是仍与原始数据有较大差距。

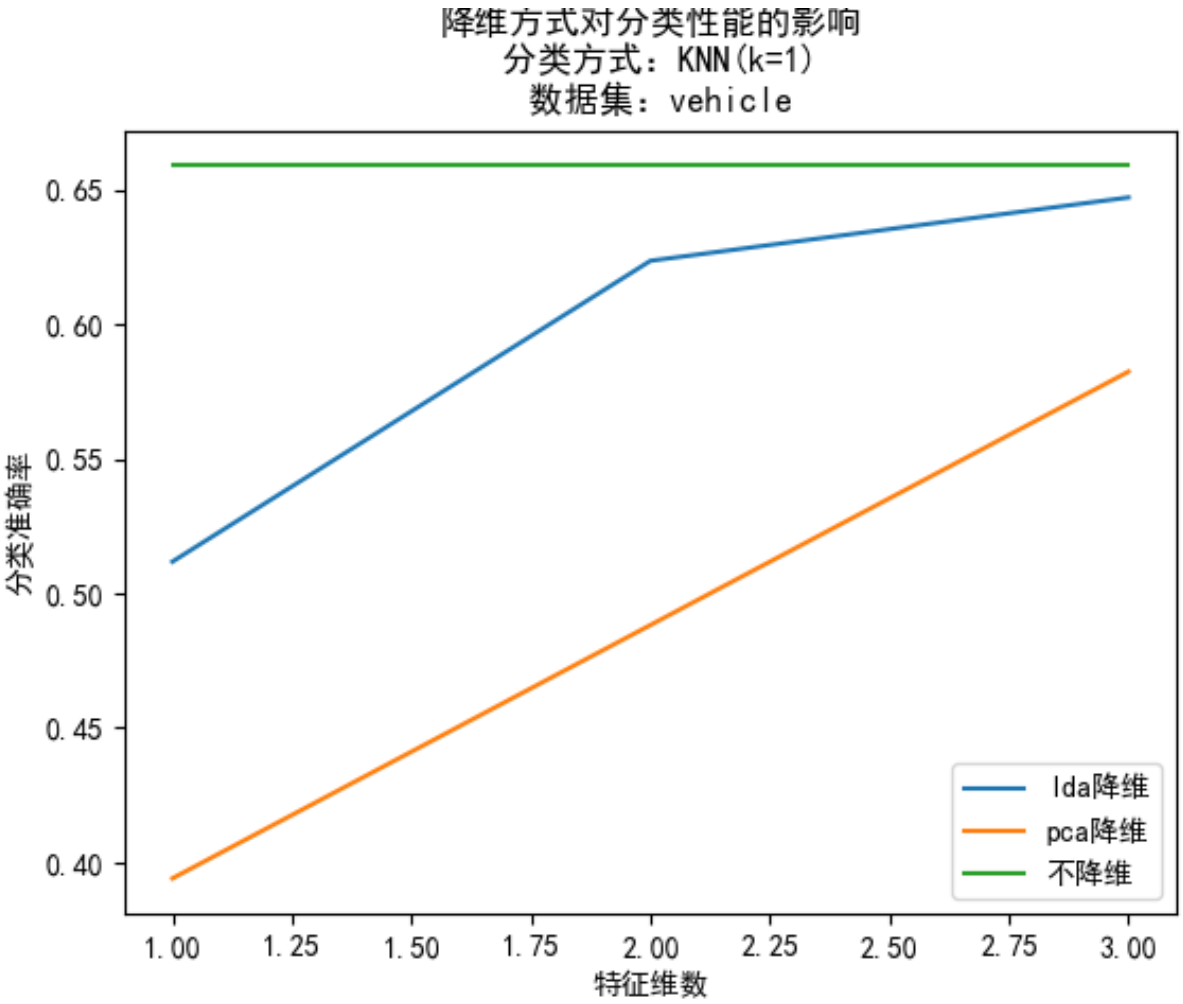


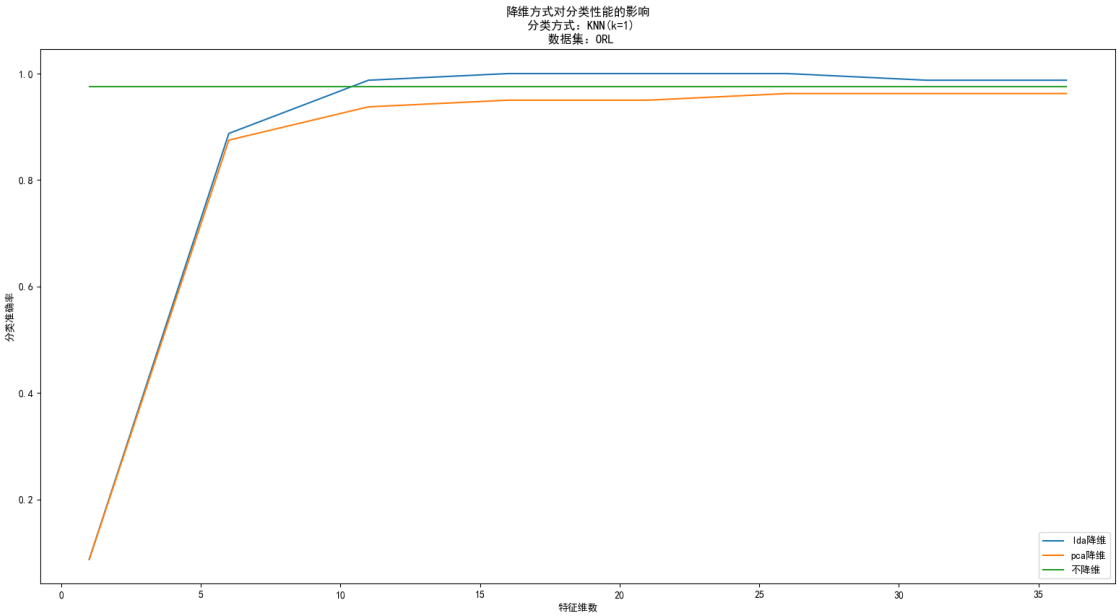
所用数据集维ORL，降维方式是PCA，降维时以5为间隔，研究特征维数对分类性能的影响。结论与LDA降维基本一致。



所用数据集维Vehicle，降维方式是PCA，研究特征维数对分类性能的影响。结论与LDA降维基本一致。但是PCA降维时可以选择更多的特征，达到更好的分类性能。

实验2:降维方式对分类性能的影响





分别使用ORL数据集和Vehicle数据集采用PCA降维和LDA降维，研究不同维度时的分类性能后发现，在特征维数相同时，LDA的性能更高。

结论

1. PCA降维和LDA降维都能在保证分类性能基本不变的前提下，大大降低特征维数，进而减少后续处理环节的复杂度。
2. PCA是一种无监督的线性降维方式，适用于无标签样本和有标签样本。
3. LDA是一种有监督的线性降维方式，适用于有标签样本。
4. 降维到相同维数时，LDA的性能更好，是因为该方法更好的利用了标签信息。因此在有标签时，可以考虑优先使用LDA降维。
5. LDA要求降维数不大于类别数，因此不适用于原始特征维数很高，而类别数较少的情况。因为此时只能选择较少的特征维数，难以充分保留原始数据的信息，导致分类性能大大下降。此时可以考虑PCA降维。

实验中需要的问题

1. **转换数据格式**

   我使用numpy完成本次实验。原始数据类型为`np.uint8`,如果不进行数据格式转换而直接计算，会导致计算结果溢出。因此应该在读取完原始数据后将其转化为`np.float`类型的数据，避免上述问题。

2. **矩阵求逆**

   实现LDA时，需要计算类内散度矩阵的逆，但是我们无法保证该矩阵可逆。实际计算过程中，该矩阵可能不可逆或者非常接近一个不可逆矩阵，导致计算结果误差很大，进而影响分类性能。在实际计算中进行矩阵求逆时，需要在带求逆矩阵后加上一个很小的单位矩阵，保证计算结果的准确性。