《模式识别》作业五

姓名：谷绍伟　　学号：202418020428007

1 简述题

1. 简述 PCA 的原理、学习模型和算法步骤。

答：主成分分析的原理：

PAC 的原理是寻找一个超平面对数据进行投影，使得样本点在这个超平面上的投影能够尽可能地分开，即使方差最大化；同时使样本到这个超平面的距离都足够近，即使重构误差最小化。

学习模型：

主成分分析的模型是一个线性变换，假设输入数据 $\mathbf{x_1}, \ldots, \mathbf{x_n} \in R^d$，则 PCA 的目标是寻找一个 $d \times k$ 的矩阵 $\mathbf{W}$，使得经过线性变换 $X^T W$ 后，得到降维后的数据，即为主成分。

算法步骤：

- 计算数据均值 $\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$；
- 计算数据的协方差矩阵：$\mathbf{C} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}) (\mathbf{x}_i - \overline{\mathbf{x}})^T$；
- 对矩阵 $\mathbf{C}$ 进行特征值分解，并取最大的 m 个特征值 $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m)$ 对应的特征向量 $\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_m}$，组成投影矩阵 $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m] \in \mathbb{R}^{d \times m}$；
- 将每一个数据进行投影：$\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i \in R^m, \quad i = 1, 2, \ldots, n$，得到降维后的数据。

2. 简述 LDA 的原理和学习模型，给出多类 LDA 的计算步骤。

答：LDA 的原理：

寻找一组投影方向，使样本在投影之后（即在新坐标系下）类内样本点尽可能靠近，即类内散度最小化；类间样本点尽可能远离，即类间散度最大化，提升样本表示的分类鉴别能力。

学习模型：

LDA 的学习模型是一个线性投影模型。假设原始数据是 $x \in R^d$（d 维数据），通过学习得到一个投影矩阵 $W \in R^{d \times k}$ （$k < d$），将原始数据投影到 k 维空间，投影后的新数据 $y = W^T x$。

多类 LDA 的计算步骤：

假设类别数为 c，计算步骤如下：

- 先计算全局散度矩阵：$\mathbf{S}_t=\sum_{i=1}^{n}\left(\mathbf{x}_i-\mu\right)\left(\mathbf{x}_i-\mu\right)^T$， $\mu=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$；
- 计算类内散度矩阵：$\mathbf{S}_w=\sum_{j=1}^{c}\mathbf{S}_{wj}$，其中 $\mathbf{S}_{wj}=\sum_{\mathbf{x}\in X_j}(\mathbf{x}-\mu_j)(\mathbf{x}-\mu_j)^T$， $\mu_j=\frac{1}{n_j}\sum_{\mathbf{x}\in X_j}\mathbf{x}$；
- 计算类间散度矩阵：$\mathbf{S}_b=\mathbf{S}_t-\mathbf{S}_w=\sum_{j=1}^{c}n_j(\mu_j-\mu)(\mu_j-\mu)^T$，其中，$n_j$ 为属于第 j 类的样本个数。
- $\max \frac{|\mathbf{W}^T\mathbf{S}_b\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_w\mathbf{W}|}, \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{W}=\mathbf{I}$，可以通过求解广义特征值问题得到：$\mathbf{S}_b\mathbf{w}=\lambda\mathbf{S}_w\mathbf{w}$。

3. 作为一类非线性降维方法，简述流形学习的基本思想。

答：流形学习的基本思想是认为高维空间相似的数据点，映射到低维空间距离也是相似的。同时流形学习假设高维数据分布在一个低维流形上。流形是一种局部类似于欧几里得空间的拓扑空间。高维数据中的每个数据点都位于这个低维流形的某个位置。由于通过线性投影将高维数据降到低维将难以展开非线性结构，流形学习的目标是找到这个低维流形的内在结构，然后将高维数据投影到低维空间，同时尽可能地保留数据在原始高维空间中的局部几何结构。

4. 根据特征选择与分类器的结合程度，简述特征选择的主要方法，指出各类方法的特点。

答：

- 过滤式特征选择方法：“选择”与“学习”独立。其特点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性能。常见的方法有方差选择、互信息等。
- 包裹式特征选择方法：“选择”依赖“学习”。其特点是特征选择过程与分类性能相结合，特征评价判据为分类器性能。对给定分类方法，选择最有利于提升分类性能的特征子集。常见的方法有递归特征消除（RFE）等。
- 嵌入式特征选择方法：“选择”与“学习”同时进行。其特点是特征选择与分类器训练过程融为一体。常见的方法有 LASSO 回归、决策树等。

2 编程题

编程实现 1

PCA+KNN：即首先 PCA 进行降维，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。

编程实现 2

LDA +KNN，即首先 LDA 进行降维，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。

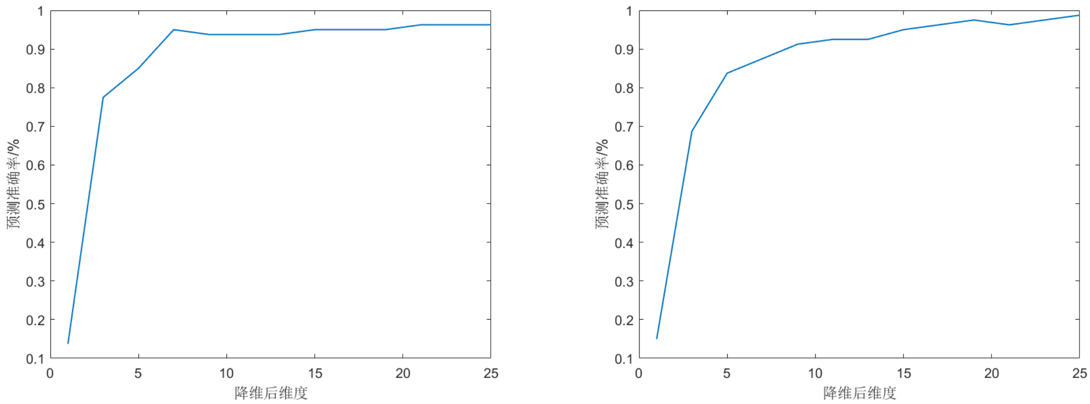
任务：采用 80% 作样本作训练集，20% 样本做测试集，报告降至不同维数时的分类性能。

实现的代码见压缩包中 main.m 和 lda_knn.m 及 pca_knn.m 文件，其中 main.m 为主程序文件，其余两个为对应的函数，再主程序中设置对应的参数，选择数据集后可以进行测试。结果如下：

ORL 数据集

对于 PCA 降维，设置降维后的维度取值范围为 1 到 25，步长为 2，得到的预测准确率结果如图 2(a)所示。

保持相同的参数，改用 LDA 降维方法，结果如图 2(b)所示。不使用数据降维时，直接使用 KNN 方法，得到的预测准确率为 96.25%，与数据降维之后的结果相差不大，但数据降维有效减小了计算量。

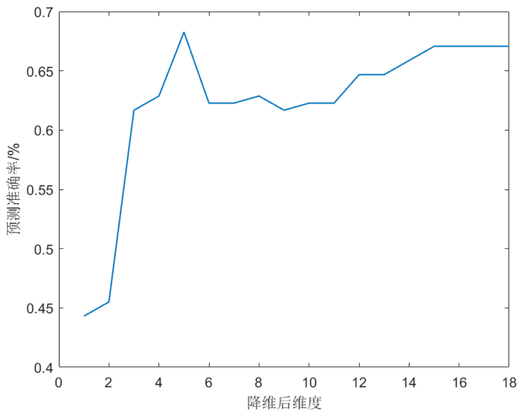


(a) PCA+KNN　　(b) LDA+KNN

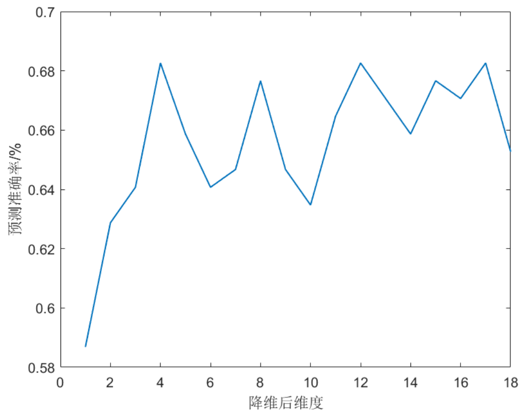
Figure 1: ORL 数据集预测准确率

Vehicle 数据集

由于原始数据维度为 18，相对较小，直接设置步长为 1，最大维度为 18，分别使用 PCA 和 LDA 进行降维，其中 PCA 降维的结果如图所示，LDA 降维预测结果如图所示。两种方法中，均出现了数据降维后预测准确率高于未降维的情况，说明数据降维在一定程度上可以减小原始数据中的噪声和干扰，提高准确率。



(a) PCA+KNN



(b) LDA+KNN

Figure 2: Vehicle 数据集预测准确率