Pattern Recognition

University of Chinese Academy of Sciences

Fall 2023

Shiming Xiang, Gaofeng Meng

Homework 5

Chenkai GUO

2023.12.20

1. 简述 *PCA* 的原理、学习模型和算法步骤

   PCA (Principal Components Analysis，主成分分析*)*:

   (*1*) 模型原理:

   ①从可区分性角度出发，即为了使投影后的数据具有最大的方差

   若给定投影矩阵 $\mathbf{W}$，则有投影后的数据点 $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i, i = 1, 2, \cdots, n$，以及投影后的数据均值 $\overline{\mathbf{y}} = \mathbf{W}^T\overline{\mathbf{x}}$，其中 $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$，此时可以计算投影后的数据点方差:

$$
\begin{aligned}
Var(\mathbf{y}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\overline{\mathbf{x}}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\overline{\mathbf{x}}\right)\left(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\overline{\mathbf{x}}\right)^T \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{W}^T\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T\mathbf{W} \\
&= \mathbf{W}^T\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T\right)}_{\mathbf{C}\ (Covariance\ Matrix)}\mathbf{W} \\
&= \mathbf{W}^T\mathbf{C}\mathbf{W}
\end{aligned}
$$

   则优化的目标为: $\max Var(\mathbf{y})$, $s.t.\ \mathbf{W}^T\mathbf{W} = 1$，即 $\max \mathbf{W}^T\mathbf{C}\mathbf{W}$, $s.t.\ \mathbf{W}^T\mathbf{W} = 1$，引入拉格朗日乘子 λ，可得以下目标优化函数:

$$
\begin{aligned}
obj &= \mathbf{W}^T\mathbf{C}\mathbf{W} - \lambda(\mathbf{W}^T\mathbf{W} - 1) \\
\frac{\partial obj}{\partial \mathbf{W}} &= 2\mathbf{C}\mathbf{W} - 2\lambda\mathbf{W}
\end{aligned}
$$

   对 $\mathbf{W}$ 求导并令导数为零求极值得:

$$
\mathbf{C}\mathbf{W} = \lambda\mathbf{W}
$$

即为特征值方程，故投影矩阵 $\mathbf{W}$ 必由协方差矩阵 $\mathbf{C}$ 的特征值组成

②从可重构角度出发，即样本到这个超平面的距离都足够近

由投影矩阵 W 定义新坐标系：假定投影变换是正交变换，即新坐标系由 $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d]^T \in \mathbb{R}^{m\times d}$ 来表示 $(d < m)$，$\mathbf{w}_i$ 模等于 1，且 $\mathbf{w}_i$ 与 $\mathbf{w}_j$ 两两正交。设样本点 $\mathbf{x}_i$ 在新坐标系下的坐标为 $\mathbf{y}_i = [y_{i1}, y_{i2}, ..., y_{id}]^T \in \mathbb{R}^d$，在正交坐标系下，对样本点 $\mathbf{x}_i$，有：$y_{ij} = \mathbf{w}_j^T\mathbf{x}_i, \mathbf{w}_j \in \mathbb{R}^d, j = 1, 2, \cdots, d$，此时可推导出 $\mathbf{x}_i$ 在新坐标系下的新表示：

$$\hat{\mathbf{x}}_i = \sum_{j=1}^{d} y_{ij}\mathbf{w}_j, \quad i = 1, 2, ..., n$$

则样本点的重构误差为：

$$\begin{aligned}
\sum_{i=1}^{n}\|\mathbf{x} - \widehat{\mathbf{x}}_i\|_2^2 &= \sum_{i=1}^{n}\left\|\mathbf{x}_i - \sum_{j=1}^{d}\mathbf{y}_{ij}\mathbf{w}_j\right\|_2^2 = \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{W}\mathbf{y}_i\|_2^2 \\
&= \sum_{i=1}^{n}\left(\left(\mathbf{W}\mathbf{y}_i\right)^T\mathbf{W}\mathbf{y}_i - 2\mathbf{x}_i^T\mathbf{W}\mathbf{y}_i + \mathbf{x}_i^T\mathbf{x}_i\right) \\
\left(\because \mathbf{W}^T\mathbf{W} = \mathbf{I}\right) &= \sum_{i=1}^{n}\left(\mathbf{y}_i^T\mathbf{y}_i - 2\mathbf{y}_i^T\mathbf{y}_i + \mathbf{x}_i^T\mathbf{x}_i\right) \\
\left(\because \mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i\right) &= -\sum_{i=1}^{n}\mathbf{y}_i^T\mathbf{y}_i + const^2 = -\sum_{i=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i\right)^T\left(\mathbf{W}^T\mathbf{x}_i\right) + const \\
\left(\because tr(A_{1\times 1}) = A_{1\times 1}\right) &= -tr\left(\sum_{i=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i\right)^T\left(\mathbf{W}^T\mathbf{x}_i\right)\right) + const \\
\left(\because tr(AB) = tr(BA)\right) &= tr\left(\mathbf{W}^T\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T\right) + const
\end{aligned}$$

进一步假设数据进行了零均值化，且令 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_d] \in \mathbb{R}^{m\times n}$，则有：

$$\begin{aligned}
\sum_{i=1}^{n}\|\mathbf{x} - \widehat{\mathbf{x}}_i\|_2^2 &= tr\left(\mathbf{W}^T\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T\right) + const \\
&= tr\left(\mathbf{W}^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{W}^T\right) + const
\end{aligned}$$

即得到 PCA 的最优化模型：$\max_{\mathbf{W}\in R^{m\times d}} tr\left(\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}\right),\ s.t.\ \mathbf{W}^T\mathbf{W} = \mathbf{I}$

(3) 算法步骤：

输入：n 个样本数据 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ ① 计算数据的均值：

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

② 计算数据的协方差矩阵：

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

③ 对矩阵 $\mathbf{C}$ 进行特征值分解，并取最大的 m 个特征值（$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$）对应的特征向量 $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d$，组成投影矩阵 $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d] \in \mathbb{R}^{m\times d}$
④ 将每一个数据进行投影：$\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, \cdots, n$

2. 简述 *LDA* 的原理和学习模型

LDA (Linear Discriminant Analysis，线性判别分析*)*：

(1) 模型原理：

LDA 意在寻找一组投影方向，使样本在投影之后（即在新的坐标系下），类内样本点尽可能靠近，类间样本点尽可能远离，并保证投影的维度（方向数）小于原始数据的维度，从而达到降维的目的。

(2) 学习模型：

考虑样本集 $D = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$ $yi \in 0, 1$，令 $\mathbf{X}i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ 分别表示第 $i \in 0, 1$ 类的示例集合、均值向量、协方差矩阵。则两类样本的中心在直线上的投影分别为 $\mathbf{w}^T\boldsymbol{\mu}_0$ 和 $\mathbf{w}^T\boldsymbol{\mu}_1$，应使得两类的中心距离尽可能远，保证类别区分性；两类样本的协方差分别为 $\mathbf{w}^T\boldsymbol{\mu}_0\mathbf{w}$ 和 $\mathbf{w}^T\boldsymbol{\mu}_1\mathbf{w}$，应使得同类样本的协方差尽可能小，保证类内聚集性，因此考虑最大化下述损失函数：

$$J = \frac{\left\|\mathbf{w}^T\boldsymbol{\mu}_0 - \mathbf{w}^T\boldsymbol{\mu}_1\right\|_2^2}{\mathbf{w}^T\boldsymbol{\Sigma}_0\mathbf{w} + \mathbf{w}^T\boldsymbol{\Sigma}_1\mathbf{w}} == \frac{\mathbf{w}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\mathbf{w}}{\mathbf{w}^T(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\mathbf{w}}$$

或广义瑞利（*Rayleigh*）商：

$$\mathbf{S}_W = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 = \sum_{\mathbf{x}\in X_0}(\mathbf{x} - \boldsymbol{\mu}_0)(\mathbf{x} - \boldsymbol{\mu}_0)^T + \sum_{\mathbf{x}\in X_1}(\mathbf{x} - \boldsymbol{\mu}_1)(\mathbf{x} - \boldsymbol{\mu}_1)^T$$

$$\mathbf{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T, \quad J = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}$$

则学习目标为：$\max \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}$，$s.t.\ \mathbf{w}^T\mathbf{w} = 1$
采用更直观的方式：$\max \mathbf{w}^T\mathbf{S}_b\mathbf{w}$，$s.t.\ \mathbf{w}^T\mathbf{S}_w\mathbf{w} = 1$，根据拉格朗日乘子法得：$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{w} = \lambda\mathbf{w}$，因此 $\mathbf{w}$ 是矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征向量，又因为：

$$\mathbf{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T, \mathbf{S}_b\mathbf{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\mathbf{w} = s\cdot(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1), (s\ is\ a\ scalar)$$

$$let\ \mathbf{S}_b\mathbf{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1),\ thus\ \mathbf{w} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

由此可以得到所需的投影向量 $\mathbf{w}$ 和投影矩阵 $\mathbf{W}$

3. 作为一类非线性降维方法，简述流形学习的基本思想

流形学习（*Manifold Learning*）的基本思想为高维空间中相似的数据点，映射到低维空间后的距离也是相似的；处理好局部是构造新的流形学习算法的关键。以下为一些经典的流形学习方法及其基本思想：

① ***LLE***（*Locally Linear Embedding*）：给定数据集，通过最近邻等方式构造一个数据图（*data graph*）。然后在每一个局部区域，高维空间中的样本线性重构关系在低维空间中均得以保持。

② ***Isomap***（*Isometric Feature Mapping*）：给定数据集，通过最近邻等方式构造一个数据图（*data graph*）。然后，计算任意两个点之间的最短路径（即测地距离），对于所有的任意两个点对，期望在低维空间中保持其测地距离。

③ ***LE***（*Laplacian Eigenmapping*）：给定数据集，通过最近邻等方式构造一个数据图（*data graph*）。然后，在每一个局部区域，计算点与点之间的亲和度（相似度），期望点对亲和度在低维空间中也得到保持。

④ ***LTSA***（*Local Tangent Space Alignment*，局部切空间对齐）：对每一个数据，在局部引入一个线性变换，将其近邻点映射到低维坐标系中的对应近邻点

⑤ ***LSE***（*Local Spline Embedding*，局部样条嵌入）：对每一个数据，在局部引入一个非线性变换，将其近邻点映射到低维坐标系中的对应近邻点

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

根据特征选择与分类器的结合程度，可分为以下三种特征选择方法：

(*1*) **过滤式特征选择方法**（“选择”和“学习”独立）：首先定义一个评价函数，并用它来度量某个给定特征与类别标签之间的相关度；最后选取具有最大相关度的 m 个特征作为选择结果。(核心任务：如何定义特征的评价函数)

算法特点：

① 过滤式方法先对数据集进行特征选择，然后再训练学习器。特征选择过程与后续学习器无关

② 启发式特征选择方法，无法获得最优子集

③ 与包裹式选择方法相比，计算量降低了很多

常用算法：单独特征选择法、顺序前进特征选择法、顺序后退特征选择法、前向-后向特征选择法（增 l 减 r 特征选择法）、*Relief* 方法

(*2*) **包裹式特征选择方法**（“选择”依赖“学习”）：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，通常采用交叉验证，如 K 折交叉验证 *(k-fold cross validation)*，留一法 *(Leave-one-out)* 来评价选取的特征子集的好坏。

算法特点：

① 分类器能够处理高维特征向量

② 在特征维度很高、样本个数较少时，分类器依然可以取得较好的效果

常用算法：直观方法、替代方法 *(递归策略)*（*Recursive SVM, SVM-RFE, Adaboost*）

(3) **嵌入式特征选择方法**（“选择”和“学习”同时进行）：将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成，即在学习器训练过程中自动地进行了特征选择。即在学习特征的权重矩阵 w 的时候，对 w 进行限制，使其不仅能满足训练样本的误差要求，同时使得 w 中非零元素尽可能少（即只对少数特征进行加权）。

常用算法：

① *LASSO*（基于 *L1* 范数的特征选择）：对目标函数才有 *L1* 范数限制

② 基于稀疏学习的特征选择：矩阵行稀疏性度量（结构化稀疏，权重矩阵 w 的一行平方和为 0，采用矩阵的 $L_{p,r}$ 范数）：

$$\|W\|_{p,r} = \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{m} |w_{ij}|^p \right)^{\frac{r}{p}} \right)^{\frac{1}{r}}$$

5. 简述最优特征选择的基本思想

最优特征选择包括以下两种方法：

① 穷举法：从给定的 d 个特征中，挑选出最优特征子集，若采用穷举法，需要遍历 2^d 个子集。该方法的缺点是当 d 很大时，该方法计算量巨大：$O(2^d)$。

② 分支定界法（*Branch and Bound*）：将所有可能的特征选择组合以树的形式进行表示，采用分枝定界方法对树进行搜索，使得搜索过程尽早达到最优解，而不必搜索整个树（基本前提为特征评价准则所使用的判据对特征具有单调性，即特征增多时，判据值不会减少）。

6. 计算编程题

(*1*) 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。

(*2*) 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。

本题使用的数据如下所示：

(a) 所用数据集 *1*: *AT&T* 40 个人脸数据集（即著名的 *ORL* 数据集）。样本个数：400，样本维数：256，类别总数：40

提示：降维时可以以 5 为间隔，比如，$10, 15, 20$ ...

读取数据和类标签信息的 *Matlab* 代码如下：

```
load ORLData_25;
X = ORLData';
X = double(X);
[n, d] = size(X);

labels = X(:, dim);              %获取各样本的类别标签
labels = floor(double(labels));
c = max(labels);                 % c = 40

X(:, dim) = [];                  % 获取样本数据
clear ORLData;
```

(b) 所用数据集 *1*: *AT&T* 40 个人脸数据集（即著名的 *ORL* 数据集）。样本个数：400，样本维数：256，类别总数：40

提示：降维时可以以 5 为间隔，比如，$10, 15, 20$ ...

读取数据和类标签信息的 *Matlab* 代码如下：

```
load ORLData_25;
X = ORLData';
X = double(X);
[n, d] = size(X);

labels = X(:, dim);              %获取各样本的类别标签
labels = floor(double(labels));
c = max(labels);                 % c = 40

X(:, dim) = [];                  % 获取样本数据
clear ORLData;
```

训练结果如下：

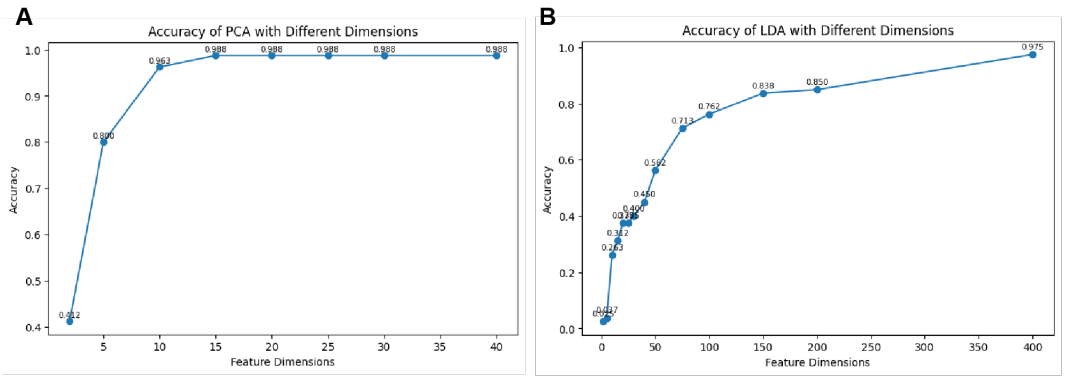

Fig. 1: PCA 和 LDA 保留不同维度的 1-NN 模型训练精度（ORL 数据集）

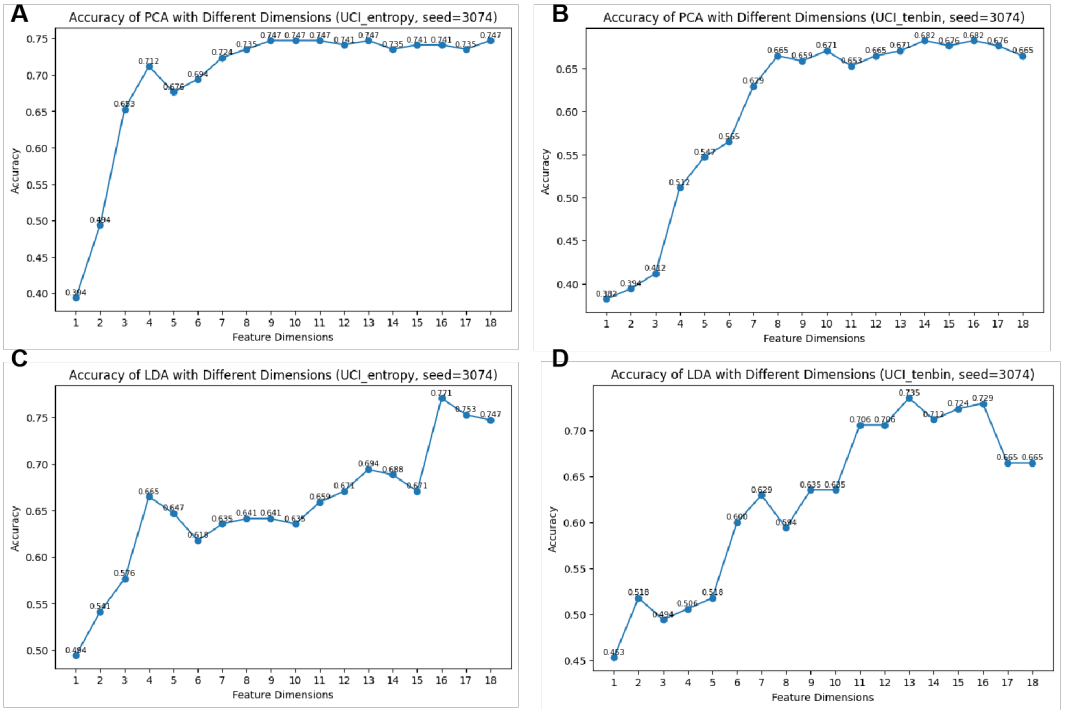

Fig. 2: PCA 和 LDA 保留不同维度的 1-NN 模型训练时间（Vehicle 数据集）

对 *PCA* 降维后而言，即使保留较少维度（*ORL* 数据集中 654 个维度仅保留 40 甚至 10，*Vehicle* 数据集中仅保留一半），其 1-*NN* 模型的分类精度仍能保持和全维度下分类精度相当，说明 *PCA* 降维的特征保留效果很好，基本保留了数据的所有有效特征信息；而对 *LDA* 降维后而言，其需要保证较多的维度保留下（*ORL* 数据集中保留维度越多精度越大，*Vehicle* 数据集中仅保留一半的维度），其 1-*NN* 模型的分类精度才能保持和全维度下分类精度相当，说明 *LDA* 降维会对数据的有效特征信息产生一定的损失。

另外，对于 *Vehicle* 数据集而言，其保留不同维度下的分类精度差异较大的可能原因在于其数据本身特征较少，难以使模型在训练中抓取类别的全部信息。

Supplementary Figures:

代码见附件(*homework5.py*)

一个很有意思的现象：保留维度大反而 *KNN* 模型的拟合时间越小（*PCA* 和 *LDA* 均有出现该现象，且更换随机种子也是这样），有点难以理解 *emmm*

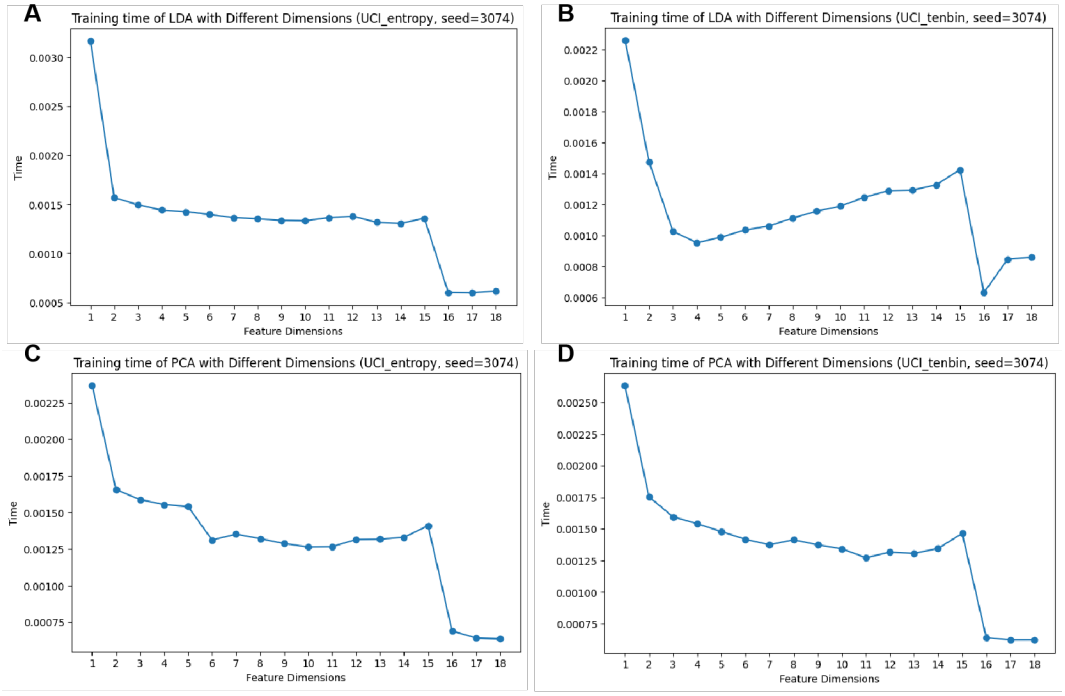

Fig. 3: PCA 和 LDA 保留不同维度的 1-NN 模型训练时间