数据挖掘作业一

141220132　银琦 141220132@smail.nju.edu.cn

1. LDA 基本原理

对于二分类，给定特征为 d 维的样例集，设法将样例投影到一条直线上使得同类样例的投影点尽可能接近、异类样例的投影点尽可能远离，能够使投影后的两类样本中心点尽量分离的直线是好的直线。在对新样本进行分类时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别。

2. NCA 基本原理

NCA 是一种监督学习方法，用于根据数据上的给定距离度量将多变量数据分类为不同类别。NCA 旨在通过找到输入数据的线性变换来“学习”距离度量，使得平均离开（LOO）分类性能在变换空间中最大化。算法的关键是矩阵 A 可以通过定义一个可微的目标函数来找到对应的转换 A，然后使用迭代求解器，如共轭梯度下降。该算法的优点之一是类的数量 k 可以确定为一个函数 A，直到一个标量常数。因此，该算法的使用解决了模型选择的问题。

3. PCA 基本原理

主成分分析搜索 k 个最能代表数据的 n 维正交向量，其中 k<=n，这样，原数据投影到一个小得多的空间上，导致维规约。基本过程如下：对输入数据规范化，使得每个属性都落入相同的区间；PCA 计算 k 个标准正交向量，作为规范化输入数据的基；对主成分按“重要性”或强度降序排列；去掉较弱的成分（方差较小的那些）来规约数据。

4. LDA、NCA 与 PCA 的比较

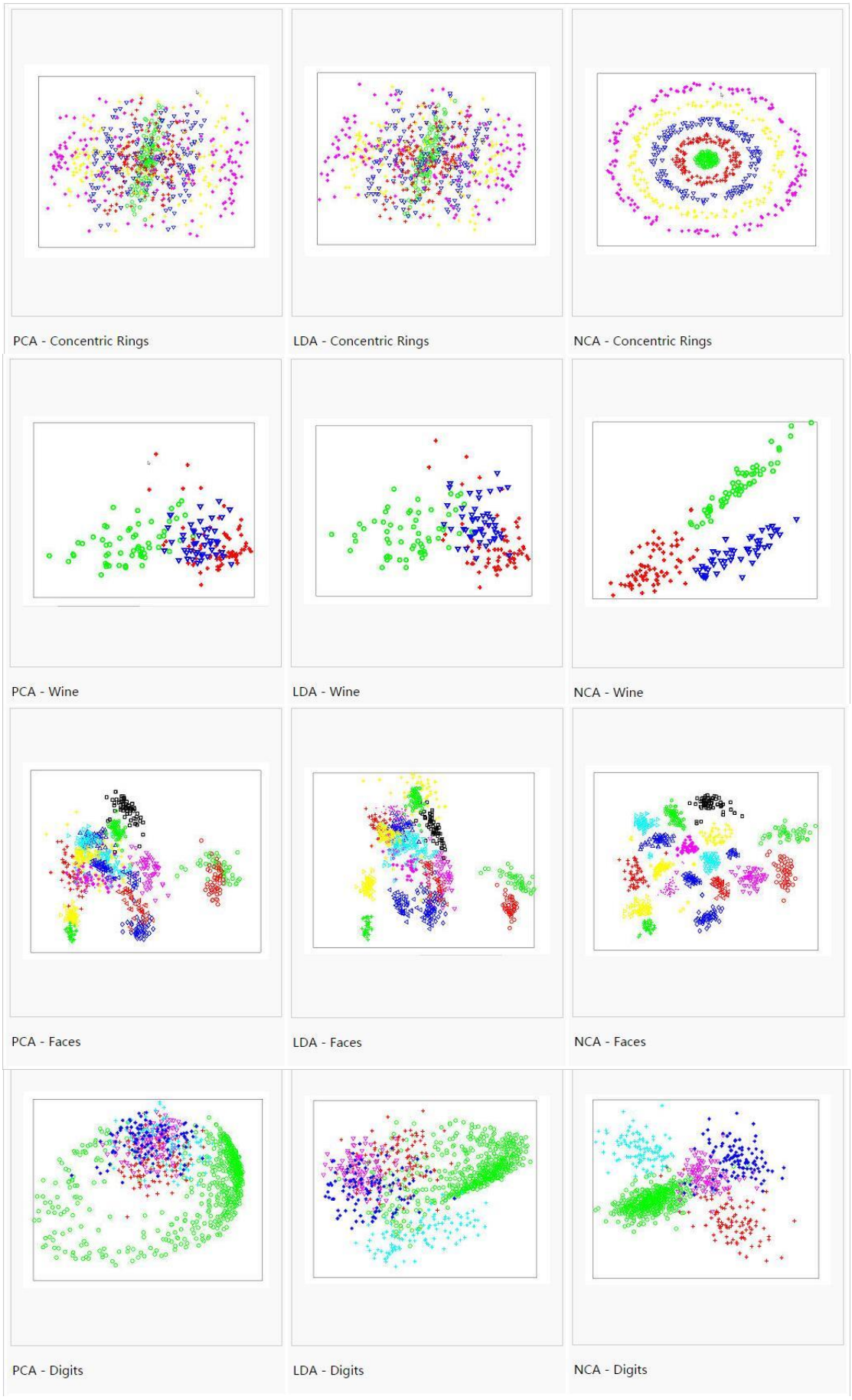
LDA 与 PCA 之间的区别就是 PCA 是一种无监督的映射方法而 LDA 是一种监督的映射方法。PCA 是无监督的，它所作的只是将整组数据整体映射到最方便表示这组数据的坐标轴上，映射时没有利用任何数据内部的分类信息，它用主要的特征代替其他相关的非主要的特征，所有特征之间的相关度越高越好。因此，虽然做了 PCA 后，整组数据在表示上更加方便(降低了维数并将信息损失降到最低)，但在分类上也许会变得更加困难，因为分类任务的特征可能是相互独立的；而 LDA 是有监督的，它使得类别内的点距离越近越好（集中），类别间的点越远越好，在增加了分类信息之后，输入映射到了另外一个坐标轴上，有了这样一个映射，两组数据之间的就变得更易区分了(在低维上就可以区分，减少了很大的运算量)。如果仅仅用于分类的话，LDA 降维的效果要比 PCA 好一些，PCA 更适合于解释样本在不同方向上的变化幅度大小。

NCA 与 PCA 之间的区别于 LDA 与 PCA 的类似，NCA 是一种监督学习方法而 PCA 是无监督的。

以下是从参考文献[4]中截取的三者比较的例子：

The Figures below show comparison of PCA, LDA, and NCA applied to "concentric ring", "wine", "faces", and "digits" data sets. The output show the reduced dimensionality of d= 2 3 from their original dimensionality which was D = 3 for concentric rings, D = 13 for wine, D = 560 for faces, and D = 256 for digits. Note that NCA produced better 2D projections of the data. Using these 2D projections for classification, NCA performed better consistently for both training and test data. When a two-dimensional projection is used, the classes are

consistently much better separated by the NCA transformation than by either PCA, or LDA.



5. 参考文献

[1]http://blog.csdn.net/ffeng271/article/details/7353834
[2]https://en.wikipedia.org/wiki/Neighbourhood_components_analysis
[3]http://blog.csdn.net/sunmenggmail/article/details/8071502
[4]http://www.wikicoursenote.com/wiki/Neighbourhood_Components_Analysis