1. 请简述主成分分析（PCA）或线性回归（或对应的核版本）有哪些方面的应用（选择本学期一种方法即可）？简述你所叙述方法的优缺点，并 Google 学术搜索你所叙述方法有哪些好的改进版本?

一 为什么需要 PCA

我们在学习过程中，为了计算方便，所用的数据的属性维数一般很低，但现实应用中属性维数经常成千上万，此时要满足密采样条件，所需要的样本数据将十分庞大。此外，许多学习方法都涉及距离计算，而高维空间会给距离计算带来很大的麻烦，例如当维数很高时甚至连计算内积都十分困难。

事实上，在高维情形下出现的数据样本稀疏、距离计算困难等问题。是所有机器学习方法共同面临的严重障碍，被称为“维度灾难”（curse of dimensionality）。

缓解维度灾难的一个重要途径就是降维（dimension reduction），即通过某种数学变换将原始高维属性空间转化为一个低维子空间，在子空间中样本密度大幅提高，距离计算也变得更加容易。那么为什么可以降维呢？这是因为人们观测或收集到的数据样本虽然是高维的，但与学习任务密切相关的也许只是某个低维分布，即我们真正关心的，也许只是众多维度中的几个维度下的分布情况。而主成分分析（Principal Component Analysis，简称 PCA）正是最常用的一种降维方法。实际应用中，我们可以利用 PCA 来实现人脸识别、数据可视化、数据压缩和预处理等等。

二 PCA 的方法及算法

首先引入单位正交阵的概念，单位正交阵可以解释为坐标轴。比如对于单位正交阵 $E^T=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$ 和一个三维向量 $X=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}$，$E^TX$ 可表示为 X 在平面直角坐标系下的坐标表示。

那么类似地，对于任意一个单位正交阵 W^T 而言,W^TX 表示 X 在该坐标系下的坐标表示。

一般的，对于 n 个 D 纬的样本点 $X_{D\times n}=[x_1,x_2,\cdots,x_n]$，和一个 D×d 纬的单位正交阵 $W_{D\times d}=[w_1,w_2,\cdots,w_d]$，其中 $\|w_i\|^2=1$。令 $y_i=W^Tx_i$，则 y_i 是 x_i 在新坐标系下的坐标表示，$Y_{d\times n}=[y_1,y_2,\cdots,y_n]$，这样就实现了将 D×n 纬的原始数据降到了 d×n 纬。但是单位正交阵有无数个，我们应该选择哪个呢?即我们应该把数据投影到哪个坐标轴上，这就引出了 PCA 的核心思想---降维后数据尽可能分散，即选择一个尽可能保留最多原始信息的方向。（直观来看，数据重叠越严重，我们丢失的原始信息就越多）

那么我们如何来找这个降维后令数据最分散的方向呢？很容易想到，如果一组点中任意两点之间的距离和最大，那么这些点应该是最分散的。即要令 $\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2$ 最大。

而 $\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2=\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(2x_i\cdot x_i^T-2x_i\cdot x_j^T)$

$$=\frac{1}{n^2}\sum_{i=1}^{n}x_i\cdot x_i^T\cdot\sum_{j=1}^{n}-\frac{1}{n^2}\sum_{i=1}^{n}x_i\sum_{j=1}^{n}x_j$$

$$=\frac{1}{n}\sum_{i=1}^{n}x_i\cdot x_i^{T}-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}x_j^{T}\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}x_i\cdot x_i^{T}-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}x_j^{T}\right)-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}x_j^{T}\right)$$

$$+\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}x_j^{T}\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}x_i\cdot x_i^{T}-\frac{1}{n}\sum_{i=1}^{n}x_i\cdot\bar{x}^{T}-\frac{1}{n}\sum_{j=1}^{n}\bar{x}\cdot x_j^{T}+\bar{x}\cdot\bar{x}^{T}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(x_i\cdot x_i^{T}-x_i\cdot\bar{x}^{T}-\bar{x}\cdot x_i^{T}+\bar{x}\cdot\bar{x}^{T}\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)^{T}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^{2}$$

即任意两点之间的距离和最小等价于点到样本中心的距离和最小，有了这个结论后，我们开始推导 PCA。因为 $y_i=W^Tx_i$，在新的坐标系下，我们要让样本尽可能分散，即要

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^{2}=\frac{1}{n}\sum_{i=1}^{n}\left(W^Tx_i-W^T\bar{x}\right)\left(W^Tx_i-W^T\bar{x}\right)^{T}$$

$$=W^T\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(x_i-\bar{x}\right)^{T}\right]W$$

最大，令 $A=[a_1,a_2,\cdots a_n]$，其中 $a_i=\left(x_i-\bar{x}\right)$，$\sum_{i=1}^{n}a_i\cdot\overline{a_i}=A\cdot\overline{A}$，上式转化为 $\frac{1}{n}W^TAA^TW$

注意到 W 是单位正交阵，所以问题等价于在 $W^TW=I$ 的条件下，求 W^TAA^TW 的最大值。

引入拉格朗日函数 $f(w,\lambda)=W^TAA^TW+\lambda(I-W^TW)$，分别对 W 和 λ 求偏导数，令其为零，可得 $AA^TW=\lambda W$,可以看出 λ 和 W 分别是 X 的协方差矩阵 AA^T 的特征值和特征向量。

现在只须对 AA^T (X 的协方差矩阵)进行特征值分解，取前 d 个最大的特征值对应的特征向量，构成的矩阵即为所求的投影矩阵 W。这就是主成分分析（PCA）的解。PCA 降维的过程中虽然舍弃了 $D-d$ 个特征向量，但是舍弃这些样本信息往往是必要的。一方面舍弃这些信息后能使样本的采样密度增大，这正是降维的重要动机。另一方面，当数据受到噪声影响时，最小的特征值对应的特征向量往往与噪声有关，将它们舍弃往往能在一定程度上起到去噪的

作用。

三 PCA 在数据可视化方面的应用

实际应用过程中，使用的数据往往是高纬度的，如果想在二维平面上直观地观测样本的属性情况，就需要用到 PCA 进行降维。现在我们使用以下链接中的葡萄酒数据集(http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data),共有 178 个样本，分成三类，每个样本有 13 个特征（13 纬）。我们调用 python 中 sklearn 包中的函数来实现这个过程。主要代码及注释如下。

首先对数据进行中心化，使它们均值为零

```
import pandas as pd
from sklearn.preprocessing import StandardScaler

df = pd.read_csv('http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data', header=No
ne)
x, y = df.iloc[:, 1:].values, df.iloc[:, 0].values
sc = StandardScaler()
x = sc.fit_transform(x)
```

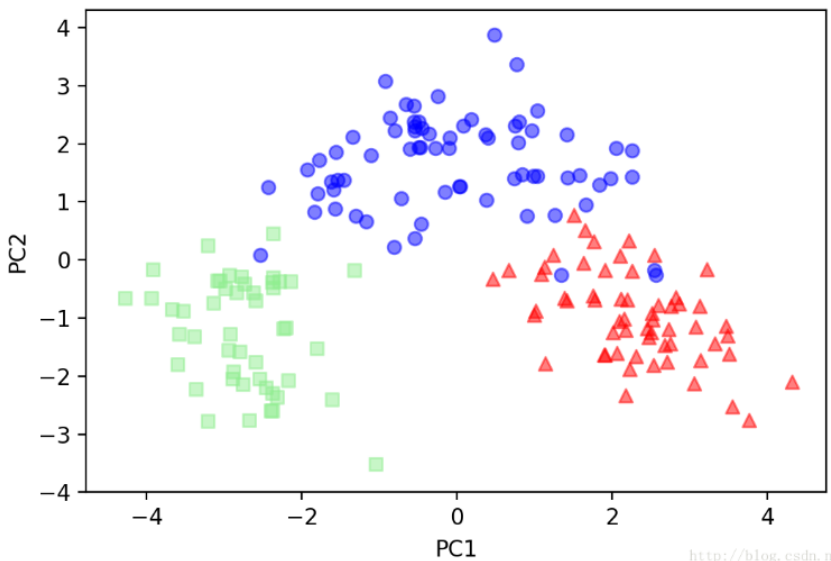
然后我们利用 PCA 将样本的十三个特征降到两个特征（可以在二维平面上观察）

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
x_pca = pca.fit_transform(x)
```

接下来我们用不同的颜色及形状来标记不同类别的样本

```
import matplotlib.pyplot as plt

plt.scatter(x_pca[y==1, 0], x_pca[y==1, 1], color='red', marker='^', alpha=0.5)
plt.scatter(x_pca[y==2, 0], x_pca[y==2, 1], color='blue', marker='o', alpha=0.5)
plt.scatter(x_pca[y==3, 0], x_pca[y==3, 1], color='lightgreen', marker='s', alpha=0.5)
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.show()
```

可以得到以下图像



可以看出，降维后的数据清晰得分成了三类，并且通过图形的方式清晰有效地传达了样本的信息。

四 PCA 的优缺点

优点：PCA 可以对高纬度的线性相关的数据进行降维，并且尽可能保持较好的原始信息。PCA 是一种无参数技术，不需要复杂的调参过程，便于通用实现。

缺点：当数据不是线性可分时，直接使用 PCA 进行线性降维，会丢失数据的低维结构。另外 PCA 假设数据各主特征是分布在正交方向上的，如果在非正交方向上存在几个数据方差较大的方向，PCA 的效果就大打折扣了。其次在不需要调参的同时，也无法进行个性化的优化。

五 PCA 的改进

这里我们讨论数据非线性可分以及数据维度很高而样本数很少时的优化方案。

首先当数据不是线性可分的时候，我们可以通过引入核函数，将数据投影到更高维的空间，使其线性可分（幸运的是，如果原始数据属性数有限，一定存在一个高维的特征空间使样本线性可分），然后再通过 PCA 进行降维。

设z_i是样本点x_i在高维特征空间(m纬)中的像，$z_i=\phi(x_i)$，ϕ就是从低维空间到高维空间的映射，设降维时的投影矩阵为W

根据$\left(\sum_{i=1}^{m} z_i {z_i}^T\right)W=\lambda W$ ①

易得$W=\frac{1}{\lambda}\left(\sum_{i=1}^{m} z_i {z_i}^T\right)W=\sum_{i=1}^{m} z_i \frac{{z_i}^T W}{\lambda}=\sum_{i=1}^{m} z_i \alpha_i$

又因为$z_i=\phi(x_i)$，所以$W=\sum_{i=1}^{m}\phi(x_i)\cdot\alpha_i$ ②

但是我们并不知道这个映射的具体形式，所以无法将ϕ显式表达出来，$\phi(x_i)$也无法求出。也就是我们不知道映射到高维空间后数据的协方差矩阵，自然无法进行特征值分解。所以我们引入核函数

$k(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\phi(x_i)^T\cdot\phi(x_j)$ ③

将②③代入①式可得

$$\left(\sum_{i=1}^{m}\phi(x_i)\phi(x_i)^T\right)\sum_{i=1}^{m}\phi(x_i)\alpha_i=\lambda\sum_{i=1}^{m}\phi(x_i)\alpha_i$$

在等式两端同时乘$\sum_{i=1}^{m}\phi(x_i)^T$

则有$K\cdot K^T\alpha=\lambda K\alpha$，继而可得$K\alpha=\lambda\alpha$

其中核矩阵 $K=\begin{bmatrix}\phi(x_1)^T\\ \cdots\\ \phi(x_n)^T\end{bmatrix}\left[\phi(x_1),\cdots,\phi(x_n)\right]$，此时$\lambda$和$\alpha$分别是核矩阵$K$的特征值和特征向量，而核矩阵$K$是已知的，取它的前 d 个（假设数据降到 d 纬）最大的特征值对应的特征向量即可求出α。由于$\phi(x)$未知，我们无法求出降维投影矩阵W，但是降维后的坐标 $Q_i=W^T\cdot\phi(x)=\sum_{i=m}^{m}\alpha_i\phi(x_i)^T\phi(x)=\sum_{i=1}^{m}\alpha_i k(x_i,x)$ 是可求的，而这也是我们的最终目的。

在实际应用中，我们很难确定合适的核函数使得训练样本在特征空间（上文所提到的高维空间）中线性可分，即便恰好找到了某个核函数使训练集在特征空间中线性可分，也很难确定这是不是由于过拟合所造成的。对此，我们可以引入软间隔和正则项来缓解这个问题。

当数据维度很高而样本数很少时（D>>N），数据的协方差矩阵XX^T（D×D）很难计算，这时我们注意到矩阵X^TX（N×N）要简单的多。

设$X^TXW=\lambda W$，λ和W分别为矩阵X^TX的特征值和特征向量构成的矩阵，可以通过特征值分解求出W，我们在等式左右两端同时$\times X$，可以得到$XX^TXW=\lambda XW$，可以发现λ和XW分别为矩阵X^TX的特征值和特征向量构成的矩阵，为了保证特征向量为标准正交基，取$\frac{1}{\sqrt{\lambda}}XW$为数据协方差矩阵$XX^T$的特征向量矩阵。这就巧妙地解决了样本维度过高的问题。