主成分分析 (Principle Compnets Analysis, PCA)

胡琛

2017 年 7 月 1 日

目录

1 场景

在拿到一个数据集的时候，有时候会遇到一些尴尬，譬如数据中，某些特征是重复的，或者特征是无关紧要的，有时候则是数据点在高维情况下会遇到稀疏性问题，譬如：

1. 拿到一个关于汽车性能的数据集，其中对于汽车速度的表述，既有“公里/时”的表述，又有“英里/时”的特征，无疑，我们只需要其中一个特征即可。

2. 拿到数学系本科生期末成绩表和他们对于数学的感兴趣程度，平均在数学上每周花费的时间，很显然，他们对于数学的感兴趣程度和在数学上花费的时间是强相关的，而在数学上花费的时间又与他们的数学

期末成绩强相关，此时，将他们对数学感兴趣程度和在数学上花费时间合并起来是不是会更好。

3. 拿到某个样本，特征很多，但是样例却非常少，此时，直接用回归去做拟合很容易造成过拟合，譬如，北京的房价与房子的朝向、楼层、建造年代，是否学区房，是否二手，大小，位置等有关系，然而我们只有十个房子的样例，直接去做回归无疑是不合适的。

4. 譬如在建立的文档-词项矩阵中，出现了'study' 和'learn' 这样的两个词项，在传统向量空间中，这两个词项无疑是独立的，但是在语义上则是相似的，出现频率也近似。

5. 信号传输过程中，由于信道不理想，信道另一端收到的信号会有扰动，如何去过滤噪音：

   - 剔除和类标签无关的特征，譬如“学生的名字”与“学生的成绩”无关
   - 剔除和类标签有关但里面存在噪声或冗余的特征，在这种情况下，需要一种特征降维的方法来降低特征数，减少噪声和冗余，减少过拟合的可能。

2 PCA 的思想

将 n 维特征映射到 $k(k<n)$ 维特征上，这 k 维是全新的正交特征。这 k 维特征称为主元，是重新构造出来的 k 维特征，而不是简单地从 n 维特征中去除 $n-k$ 维特征得到。

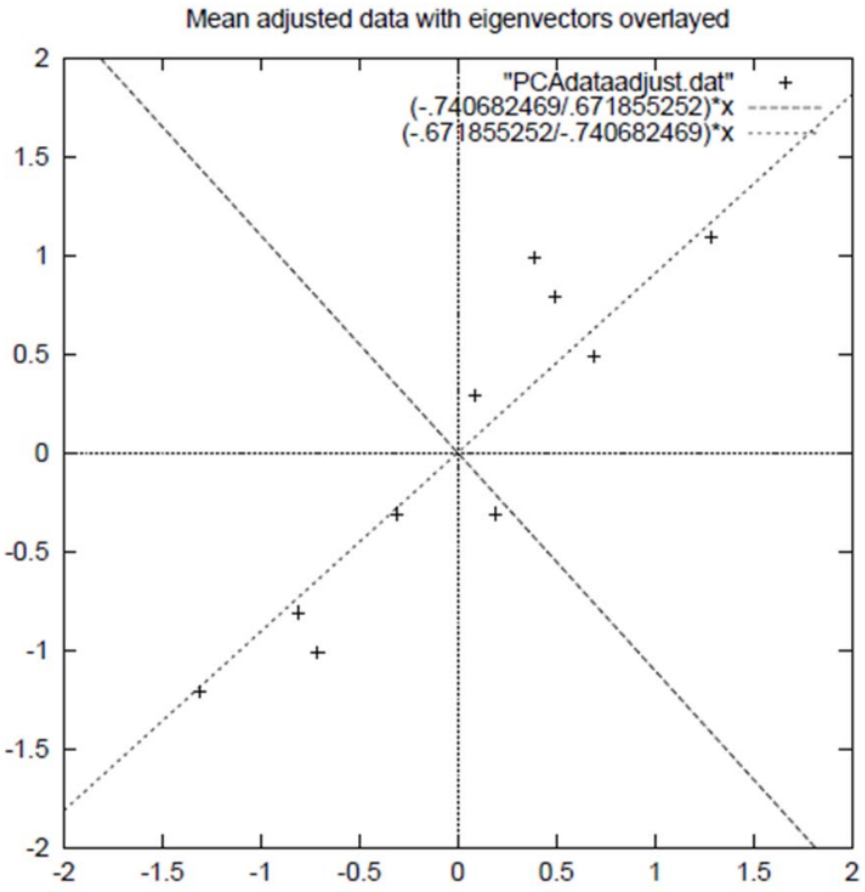
算法思想：最大方差理论，最小平方误差理论，坐标轴相关度理论

3 理论基础

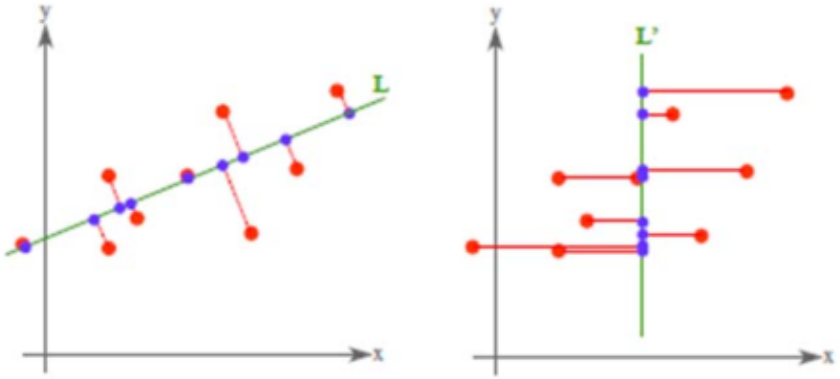
3.1 最大方差理论

在信号处理中，认为信号具有较大方差，噪声具有较小方差，信噪比就是信号与噪声的方差比，越大越好。

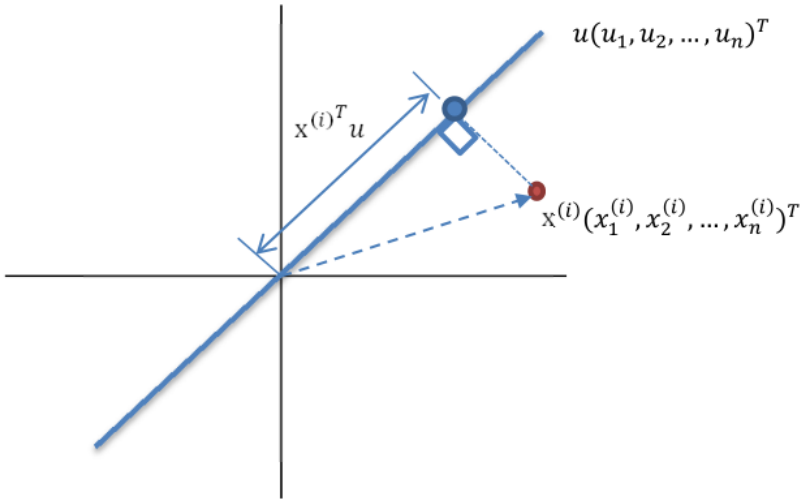
如下图所示，样本在横轴上投影方差相比在纵轴上投影的方差较大，那么，认为在纵轴上的投影是由噪声引起的。因此，我们认为，最好的 k 维特征，应该是将 n 维样本点转为 k 维样本点后，每一维上的投影方差都很大。



如前所述，在下图中，根据方差最大化理论，显然应该是选左边直线做投影更好。



3.1.1 投影



最佳的投影向量，应该使得投影后的样本点方差最大

在上图中，已知均值维 0，因此方差为：

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}u)^2 &= \frac{1}{m}\sum_{i=1}^{m}(u^T x^{(i)T} x^{(i)}u) \\ &= u^T(\frac{1}{m}\sum_{i=1}^{m} x^{(i)T}x^{(i)})u \end{aligned} \tag{1}$$

其中，m 表示数据集中数据点的个数，u 表示某个投影轴，我们的目的

是要找出足够的投影轴，譬如 $k(k < n)$ 个投影轴，并以此作为新的坐标轴来表示数据点，由于 $k < n$, 显然此时伴随着信息的丢失，因此我们需要取方差最大的组合，以尽可能地保留信息，丢弃噪声等冗余信息。

按之前的讨论，此时应该对 1 取极大，为此，对该公式进行一些处理。由于 u 是单位向量，满足 $u^T u = 1$, 因此，

$$\begin{aligned} u\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}u)^2 &= uu^T\frac{1}{m}(\sum_{i=1}^{m}{x^{(i)}}^T x^{(i)})u \\ &= (\frac{1}{m}\sum_{i=1}^{m}{x^{(i)}}^T x^{(i)})u \end{aligned} \tag{2}$$

如果令 $\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}u)^2 = \lambda$, $(\frac{1}{m}\sum_{i=1}^{m}{x^{(i)}}^T x^{(i)}) = \Sigma$, 于是，有

$$\lambda u = \Sigma u \tag{3}$$

λ 是 Σ 的本征值，u 是特征向量，最佳的投影向量，显然需要 λ 取最大的时候对应的 u. 因此，PCA 就是对 Σ, 也就是协方差矩阵，求本征值，并按从大到小排列，取比较大的 k 个本征值对应的本征向量组成新的 $n \times k$ 矩阵空间就是新的表示空间。