斯坦福 ML 公开课笔记 14

上一篇笔记中，介绍了因子分析模型，因子分析模型使用 d 维子空间的隐含变量 z 来拟合训练数据，所以实际上因子分析模型是一种数据降维的方法，它基于一个概率模型，使用 EM 算法来估计参数。

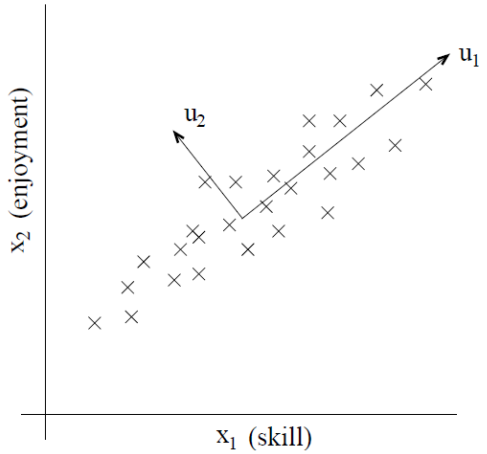
本篇主要介绍 PCA(Principal Components Analysis, 主成分分析)，也是一种降维方法，但是该方法比较直接，只需计算特征向量就可以进行降维了。本篇对应的视频是公开课的第 14 个视频，该视频的前半部分为因子分析模型的 EM 求解，已写入笔记 13，本篇只是后半部分的笔记，所以内容较少。

引入

PCA 解决的是什么问题呢？下面举一个例子来回答。

设想有一个数据集$\{x^{(i)}; i = 1, \ldots, m\}$，其中$x^{(i)} \in \mathbb{R}^n$。比如每个 x 代表一辆车，x 的属性可能是车的最高速度，每公里耗油量等。如果有这样两个属性，一个是以千米为单位的最大速度，一个是以英里为单位的最大速度。这两个速度很显然是线性成比例的，可能会因为数字取整的缘故有一些小小的扰动，但不影响比例。所以实际上，数据的信息量是 n-1 维的，多 1 维并不包含更多的信息。PCA 解决的就是将多余的属性去掉的问题。

视频中还举了一个不是很直观的例子，那就是直升飞机驾驶员的例子，每个驾驶员都有两个属性值，其一是驾驶员的技能评估，其二是驾驶员对驾驶的兴趣程度。由于丫驾驶 RC 直升飞机（啥是 RC 直升飞机？）难度比较大，所以一般只有对其有很大兴趣，才能较好的掌握这项技能。所以属性 1 和属性 2 是强相关的。实际上，根据已有的数据，可以将这两个属性使用坐标图进行展示。如下：



由图可以看到，u1 是展示出了数据的相关性，称之为“主方向”，u2 则反映了一些主方向之外的噪声，那么，如何计算得到 u1 这个方向呢？

预处理

运行 PCA 算法之前，数据一般都需要预处理。预处理步骤如下：

1) 令$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$

2) 使用$x^{(i)} - \mu$来替代$x^{(i)}$

3) 令$\sigma_j^2 = \frac{1}{m}\sum_i (x_j^{(i)})^2$

4) 使用$x_j^{(i)}/\sigma_j$来替代$x_j^{(i)}$

步骤 1-2 将数据的均值变为 0，当已经知道数据的均值为 0 的时候，可以省略这两步。步骤 3-4 将数据的每个维度的方差变为 1，从而使得每个维度都在同一个尺度下被衡量，不会造成某些维度因数值较大而有大的影响的情况。当预先知道数据处于同一尺度下时，可以忽略 3-4 步，比如图像处理中，已经预知了图像的每个像素都在 0-255 范围内，因而没有必要再进行归一化了。

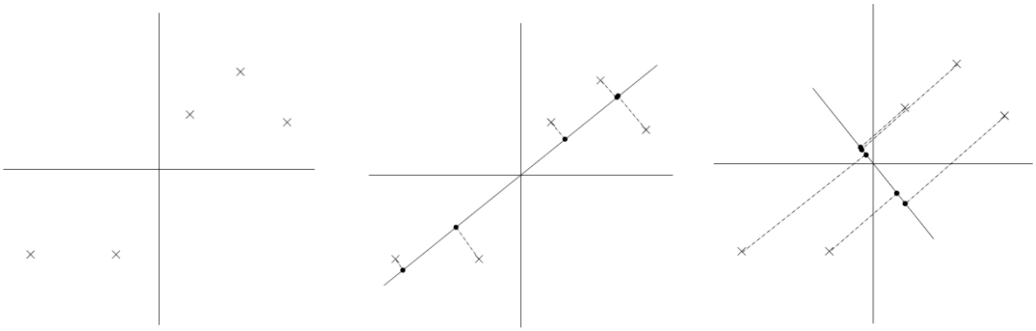
PCA 模型的定义

如何找到数据的主方向呢？在二维空间下可以这样理解，有一个单位向量 u，若从原点出发，这样定义 u 以后就相当于定义了一条直线。每个数据点在该直线上都有一个投影点，寻找主方向的的任务就是寻找一个 u 使得投影点的方差最大化。

那么问题就来了。问题 1，能不能不从原点出发？可以，但那样计算就复杂了，我们归一化时候将均值设为 0，就是为了在寻找方向的时候使向量从原点出发，方便计算。问题 2，多维空间下多个主方向时怎么办？那就是不止寻找一个单位向量咯，找到一个主方向后，将该主方向的方差影响去掉，然后再找主方向。如何去掉前一个主方向的方差影响呢？对于二维数据来说，是将所有数据点在垂直于该主方向的另一个方向上做投影，比如上图，要去掉主方向 u1 的方差影响，需要在 u2 方向上进行投影，多维空间上也可以类推。

以方差最大化来理解寻找主方向的依据是什么？直观上看，数据初始时会有一个方差，我们把这个方差当做数据包含的信息，我们找主方向的时候尽量使方差在子空间中最大化，从而能保留更多的信息。

再举一个例子来说明如何寻找主方向。比如下面左图中的五个点。



其中一个方向如中图所示，另一个方向如右图所示。显然中图的投影点的方差最大。

下面，给出寻找主方向的数学定义。

设$x^{(i)}$为数据集中的点，u 是要求解的单位向量，那么方差最大化可以形式化为最大化:

$$\frac{1}{m}\sum_{i=1}^{m}(x^{(i)^T}u)^2 = \frac{1}{m}\sum_{i=1}^{m}u^T x^{(i)}x^{(i)^T}u = u^T\left(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)^T}\right)u$$

注意到，对于归一化后的数据，其投影点的均值也为 0，因而在方差计算中直接平方。该公式有一个约束条件，即$||u||_2 = 1$。这个最大化问题的解就是矩阵$\Sigma = \frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)^T}$的特征向量。这是如何得到的呢？且看下式。

使用拉格朗日方程来求解该最大化问题，则：

$$\ell = u^T\left(\frac{1}{m}\sum_{i=1}^{m} x^{(i)}x^{(i)^T}\right)u - \lambda\left(||u||_2 - 1\right) = u^T\Sigma u - \lambda(u^Tu - 1)$$

对 u 求导。

$$\begin{aligned}\nabla_u\ell &= \nabla_u(u^T(\Sigma)u - \lambda(u^Tu - 1)) = \nabla_u u^T\Sigma u - \lambda\nabla_u u^Tu \\ &= \nabla_u tr(u^T\Sigma u) - \lambda\nabla_u tr(u^Tu) = (\nabla_{u^T} tr(u^T\Sigma u))^T - \lambda(\nabla_{u^T} tr(u^Tu))^T \\ &= (\Sigma u)^{T^T} - \lambda u^{T^T} = \Sigma u - \lambda u\end{aligned}$$

令导数为 0，可知 u 就是Σ的特征向量。上式的推导所用到的性质与上一篇笔记 13B 中那个 12 步的推导相似，在此不赘述了。

因为$\Sigma = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}x^{(i)^T}$是对称矩阵，因而可以得到相互正交的 n 个特征向量$\{u^1, u^2, \ldots, u^n\}$，那么，如何达到降维的效果呢？选取最大的 k 个特征值所对应的特征向量即可。降维后的数据可以用下式来表达：

$$y^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ \vdots \\ u_k^T x^{(i)} \end{bmatrix} \in \mathbb{R}^k$$

以上就是 PCA 完整的数学表达。

视频中，Ng 说，PCA 有 9 到 10 种解释方法，这种子空间方差最大法只是其中一种，另一种比较常见的理解方法就是最小化原始点到投影点的距离的平方和。

PCA 的应用

压缩与可视化：如果将数据由高维降至 2 维或 3 维，那么可以使用一些可视化工具进行查看。同时数据的量也减少了。

预处理与降噪：很多监督算法在处理数据前都对数据进行降维，降维不仅使数据处理更快，还去除了数据中的噪声。使得数据的稀疏性变低，减少了模型假设的复杂度，从而降低了过拟合的的概率。

具体的应用，比如图片处理中，对于一个 100*100 的图片，其原始特征长度为 10000，使用 PCA 将为后，大大减少了维度，形成了“特征脸”图片。而且还减小了噪声如光照等影响，使用 PCA 降维后的数据可以进行图片的相似度计算，在图片检索中和人脸检测中都能达到很好的效果。