# 机器学习-数据降维

2021年4月6日 11:20

参考：
数据降维 https://www.jianshu.com/p/f9811d1feee6

前言：正所谓每一个结果的出现都是一系列的原因导致的，当构建机器学习模型时候，有时候数据特征异常复杂，这就需要经常用到数据降维技术，下面主要介绍一些降维的主要原理

## 为什么要降维?

在实际的机器学习项目中，特征选择/降维是必须进行的，因为在数据中存在以下几个 方面的问题：

- 数据的多重共线性：特征属性之间存在着相互关联关系。多重共线性会导致解的空间不稳定， 从而导致模型的泛化能力弱；
- 高纬空间样本具有稀疏性，导致模型比较难找到数据特征；
- 过多的变量会妨碍模型查找规律；
- 仅仅考虑单个变量对于目标属性的影响可能忽略变量之间的潜在关系。
  通过特征选择/降维的目的是：
- 减少特征属性的个数
- 确保特征属性之间是相互独立的

当然有时候也存在特征矩阵过大， 导致计算量比较大，训练时间长的问题
常用的降维方法有：

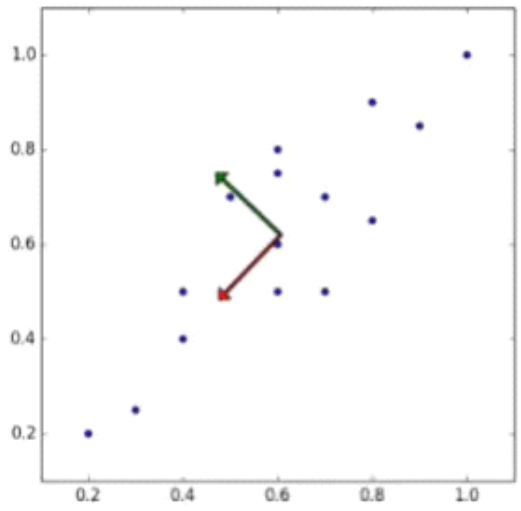
- PCA
- LDA
- 主题模型进行降维

## PCA原理

主成分析(PCA)：将高纬的特征向量合并称为低纬度的特征属性，是一种无监督 的降维方法。
算法目标是通过某种线性投影，将高维的数据映射到低维的空间中表 示，并且期望在所投影的维度上数据的方差最大（最大方差理论），以此使用较 少的数据维度，同时保留较多的原数据点的特性。

### 主成分选择

假设原来的特征数据是n维数据，首先选着方差最大方向为第一维数据。第二个坐标轴选择和第一个坐标轴垂直或者正交 的方向；第三个坐标轴选择和第一个、第二个坐标轴都垂直或者正交的方向；该 过程一直重复，直到新坐标系的维度和达到给定的值。 而这些方向所表示的数据特征就被称为“主成分”。




**数学原理**

目标函数：投影的维度上数据的方差最大

投影矩阵w，样本点xi在新空间中的超平面上的投影是：WTxi（假设X是已经中心化（z-score）过的数据矩阵）

若所有样本点的投影能 够尽可能的分开，则表示投影之后的点在各个维度上的方差应该最大化，那么投影样本点的各个维度方差和可以表示为：

$$\frac{1}{n}\sum_{i} W^T x_i x_i^T W$$

从而我们可以得到 PCA的最优目标函数是：

$$\max_{W} tr\left(W^T XX^T W\right)$$

$$s.t. W^T W = I$$

进行拉格朗日求解：

$$L = W^T XX^T W + \lambda\left(I - W^T W\right)$$

求偏导：

$$\frac{\partial L}{\partial W} = 2XX^T W - 2\lambda W$$

$$\xrightarrow{令\frac{\partial L}{\partial w}=0} \quad XX^T W = \lambda W$$

$$W^T XX^T W = W^T \lambda W = \lambda$$

可以发现如果，此时将XXT看成一个整体A，那么求解W的过程恰好就是求解矩 阵A的特征向量的过程，所以我们可以认为PCA的计算其实就是对进行去中心化 后的数据的协方差矩阵求解特征值和特征向量。

一般情况下特征值的求解都比较复杂，这里可以用SVD分解来求：

而且此时恰好XXT是对角矩阵，所以我们可以将其进行特征分解：

$$XX^T = W\Lambda W^T$$

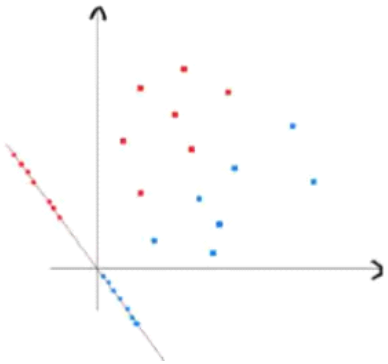
另外对矩阵X进行SVD矩阵分解，那么可以得到下列式子：

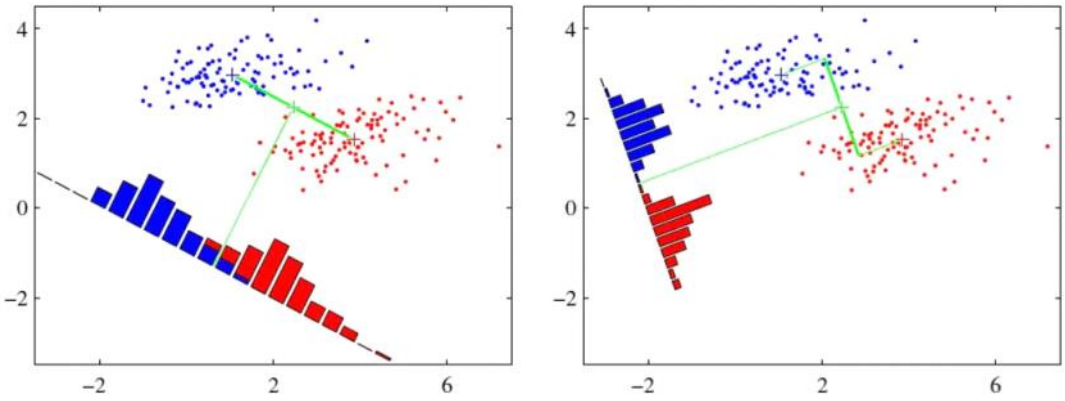
$$X = D\Sigma V^T \Rightarrow XX^T = D\Sigma V^T \left(D\Sigma V^T\right)^T = D\Sigma^2 D^T \Rightarrow W = D$$
$$\Downarrow$$
$$X' = W^T X = D^T D\Sigma V^T = \Sigma V^T$$

## LDA

线性判断分析(LDA)：LDA是一种基于分类模型进行特征属性合并的操作，是一 种有监督的降维方法。

LDA的原理是，将带上标签的数据（点），通过投影的方法，投影到维度更低的 空间中，使得投影后的点，会形成按类别区分，一簇一簇的情况，相同类别的点，将会在投影后的空间中更接近。用一句话概括就是： “投影后类内方差最小，类间方差最大”






**数学原理**（投影后类内方差最小，类 间方差最大）

假定转换为w，那么线性转换函数为：

$$x' = w^T x$$

并且转换后的数据是一维的

考虑二元分类的情况，认为转换后的值大于某个阈值，属于某个类别，小于等于 某个阈值，属于另外一个类别，使用类别样本的中心点来表示类别信息，那么这 个时候其实就相当于让这两个中心的距离最远：

$$\mu_j = \frac{1}{N_j}\sum_{x\in X_j} x \qquad \mu_j{}' = \frac{1}{N_j}\sum_{x\in X_j} x' = \frac{1}{N_j}\sum_{x\in X_j} w^T x = w^T \mu_j$$

$$J = \left|\mu_1{}' - \mu_2{}'\right| = w^T\left|\mu_1 - \mu_2\right|$$

同时又要求划分之后同个类别中的样本数据尽可能的接近，也就是同类别的投影 点的协方差要尽可能的小。

$$\Sigma_j = \sum_{x \in X_j} (x - \mu_j)(x - \mu_j)^T \qquad \Sigma_j{}' = \sum_{x \in X_j} \left(x' - \mu_j{}'\right)\left(x' - \mu_j{}'\right)^t = w^T \Sigma_j w$$

$$\Sigma = \Sigma_1 + \Sigma_2$$

结合着两者，那么我们最终的目标函数就是：

$$\max_w J(w) = \frac{\left\| w^T \mu_1 - w^T \mu_2 \right\|_2^2}{w^T \Sigma_1 w + w^T \Sigma_2 w} = \frac{w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w}{w^T (\Sigma_1 + \Sigma_2) w}$$

对目标函数进行转换（A、B为方阵，A为正定矩阵）：

$$\max_w J(w) = \frac{\left\| w^T \mu_1 - w^T \mu_2 \right\|_2^2}{w^T \Sigma_1 w + w^T \Sigma_2 w} = \frac{w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w}{w^T (\Sigma_1 + \Sigma_2) w}$$

$$\xrightarrow{令 A=\Sigma_0+\Sigma_1, B=(\mu_1-\mu_2)(\mu_1-\mu_2)^T} \max_w J(w) = \frac{w^T B w}{w^T A w}$$

$$\xrightarrow{令 w'=A^{\frac{-1}{2}} w} \max_w J(w) = \frac{w'^T A^{\frac{-1}{2}} B A^{\frac{-1}{2}} w'}{w'^T w'}$$

$$\xrightarrow{因 w'^T w'=1} \max_w J(w) = w'^T A^{\frac{-1}{2}} B A^{\frac{-1}{2}} w'$$

该式子和PCA降维中的优化函数一模一样，所以直接对中间的矩阵进行矩阵分解即可。

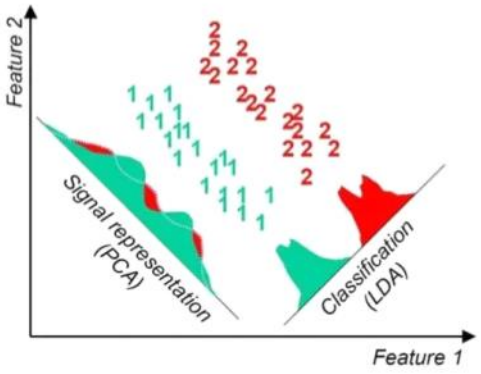
## 比较：

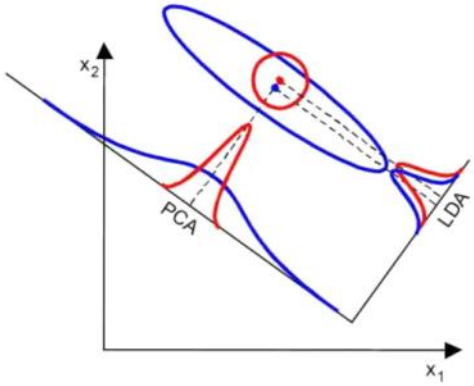
**相同点：**

- 两者均可以对数据完成降维操作
- 两者在降维时候均使用矩阵分解的思想
- 两者都假设数据符合高斯分布

不同点：

- LDA是监督降维算法，PCA是无监督降维算法
- LDA降维最多降到类别数目k-1的维数，而PCA没有限制
- LDA除了降维外，还可以应用于分类
- LDA选择的是分类性能最好的投影，而PCA选择样本点投影具有最大方差的方向


Feature 2
Signal representation
(PCA)
Classification
(LDA)
Feature 1


x2
PCA
LDA
x1