主成分分析（PCA）方法步骤以及代码详解

前言

> 上一节我们了解到在构建神经网络模型，除了掌握如何搭建神经网络架构，了解参数具体含义，规避风险等方法。第一步是要对采用数据集的详细了解，无需接触任何神经网络代码，而是从彻底检查数据开始。这一步是非常关键的一步，往往我们在数据处理的某一个步骤会一定程度上的影响实验结果。本节将讲述常见的数据降维方法**PCA**，减少数据集的变量数量，同时保留尽可能多的信息。

1. 什么是主成分分析?

PCA（Principal Component Analysis） 是一种常见的数据分析方式，常用于高维数据的降维，可用于提取数据的主要特征分量。PCA通常用于降低大型数据集的维数，方法是数据集中的指标数量变少，并且保留原数据集中指标的大部分信息。总而言之：减少数据指标数量，保留尽可能多的信息。

1.1 PCA适用范围

- 在已标注与未标注的数据上都有降维技术
- 主要关注未标注数据上的降维技术，将技术同样也可以应用于已标注的数据。

1.2 优缺点

PCA优点在于数据降维，便于提取数据的主要特征，使得数据更容易使用，减少计算开销，去除噪音等等。缺点在于不一定需要，有可能损失有用信息，只针对训练集保留主要信息，可能造成过拟合。适用于结构化数据.**PCA**不仅能将数据压缩，也使得降维之后的数据特征相互独立。

2. PCA的方法步骤

PCA作为一个传统的机器学习算法，可以通过基础的线代知识推导（协方差矩阵计算，计算特征向量，特征值，正交...）。主要涉及的数学方法不在本节过多描述，有兴趣的读者可以参考花书中的线性代数部分，做推导。**PCA**的步骤主要分为五步；

2.1 标准化连续初始变量的范围（非结构化转成结构化）

此步骤的目的是标准化结构化指标的范围，因为**PCA**对于初始变量的方差非常敏感，如果初始变量的范围之间存在较大差异，则会造成很大变差，使用标准化可以将数据转换为可比较的尺度。最常用的方法主要有一下两种

- 线性函数归一化。将原始数据进行线性变换，使结果映射到[0,1]的范围，实现对原始数据的等比缩放。归一化公式如下：

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- 零均值归一化。它会将原始数据映射到均值为0，标准差为1的分布上。具体来说，假设原始特征的均值**μ**，标准差为**σ**，那么归一化公式定义为：

$$z = \frac{x - \mu}{\sigma}.$$

此方法仅限于结构化数据，对于类别型特征主要是指男，女，血型等只在有限选项内取值的特征。类别型特征原始输入通常是字符串形式，可以使用序号编码，独热编码，二进制编码等进行预处理。

- 序号编码：序号编码通常用于处理类别间具有大小关系的数据。例如成绩，可以分为 低、中、高三档，并且存在“高>中>低”的排序关系。序号编码会按照大小关系对 类别型特征赋予一个数值ID，例如高表示为3、中表示为2、低表示为1，转换后依 然保留了大小关系。

- 独热编码：独热编码通常用于处理类别间不具有大小关系的特征。例如血型，一共有4个 取值（A型血、B型血、AB型血、O型血），独热编码会把血型变成一个4维稀疏 向量，A型血表示为（1, 0, 0, 0），B型血表示为（0, 1, 0, 0），AB型表示为（0, 0, 1, 0），O型血表示为（0, 0, 0, 1）。（对于类别值较多的情况注意使用稀疏向量来节省空间，以及配合特征选择来降低维度）
- 二进制编码：二进制编码主要分为两步，先用序号编码给每个类别赋予一个类别ID，然后 将类别ID对应的二进制编码作为结果。以A、B、AB、O血型为例，表1.1是二进制 编码的过程。A型血的ID为1，二进制表示为001；B型血的ID为2，二进制表示为 010；以此类推可以得到AB型血和O型血的二进制表示。可以看出，二进制编码本 质上是利用二进制对ID进行哈希映射，最终得到0/1特征向量，且维数少于独热编 码，节省了存储空间。
- **Categorical Embedder**：通过神经网络编码分类变量，有兴趣的朋友可以参考这篇文章（这个以后可能会单独列出一章讲述，不能占篇幅过大...）

对于文本类型的非结构化数据，主要使用的是词袋模型**（Bag of Words）**，**TF-IDF**，主题模型**（Topic Model）**，词嵌入模型**（Word Embedding）**，这个也不做过多叙述了简单叙述一下即可，对于专攻**NLP**的朋友就是关公面前耍大刀了...

2.2 计算协方差矩阵以识别相关性

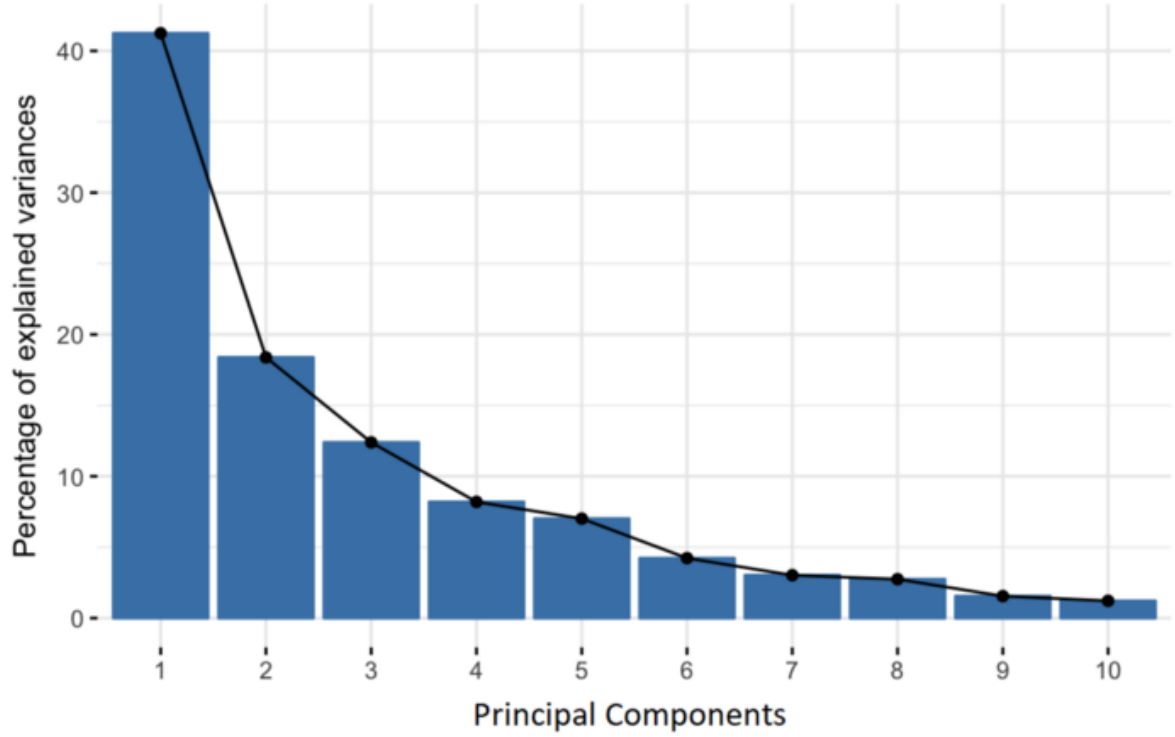
此步骤的目的是观察数据标签彼此是否存在相关性，观察指标间是否包含冗余信息。使用协方差矩阵是一个**p * p**对称矩阵（其中p是维数），它具有与所有可能的初始变量对相关联的协方差作为条目。假设三个变量**x，y，z**三维数据集，协方差矩阵是**3 * 3**矩阵如下图所示：

$$\begin{bmatrix} \mathrm{Cov}(x,x) & \mathrm{Cov}(x,y) & \mathrm{Cov}(x,z) \\ \mathrm{Cov}(y,x) & \mathrm{Cov}(y,y) & \mathrm{Cov}(y,z) \\ \mathrm{Cov}(z,x) & \mathrm{Cov}(z,y) & \mathrm{Cov}(z,z) \end{bmatrix}$$

自身协方差是自身的方差，**（Cov(a,b)=Cov(b,a)）**是可以交换的，意味着上三角部分和下三角部分相等。如果协方差为正，则两个变量正相关，如果协方差为负，则两个变量呈负相关。

2.3 计算协方差矩阵的特征向量和特征值以识别主成分

通过计算协方差矩阵的特征向量和特征值来确定数据的主成分。首先解释一下主成分定义：主成分是由初始变量的线性组合或混合构成的新变量。新变量是互不相关的，并且初始变量中的大部分信息被挤压或压缩到第一成分中。通俗来讲，十维数据给十个主成分，**PCA**试图将最大可能信息放在第一个组件中，然后第二组件中放置最大的剩余信息，以此类推，直到出现下图所示内容。



通过这种方式在主成分中组织信息，可以在不丢失太多信息的情况下降低维数生成新的指标。此时的新指标互不相关且无可解释性。它们是初始变量的线性组合。主成分表示解释最大方差量的数据的方向。方差与信息的关系是，一条携带的方差越大，沿线的数据点的离散度越高，沿线的离散度越大，它所包含的信息越多。计算协方差矩阵的特征值其实就是计算最大的方差，计算其对应的特征向量就是最佳投影方向，计算协方差矩阵特征值需要将其对角化，为了满足变化后的指标间协方差为0且指标方差尽可能大，因此要求解最大化问题，可表示为：

$$\begin{cases} \max \left\{\omega^{\mathrm{T}} \Sigma \omega\right\} \\ \text{s.t.} \quad \omega^{\mathrm{T}} \omega = 1 \end{cases}$$

此时引用拉格朗日乘子法将问题转化为最优化问题，并对对ω求导令其等于0，便可以推出Σ ω=λ ω，此时：

$$D(\boldsymbol{x}) = \boldsymbol{\omega}^{\mathrm{T}} \Sigma \boldsymbol{\omega} = \lambda \boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{\omega} = \lambda$$

将计算好的特征值的顺序对特征向量进行排序，从高到低，可以按照重要程度顺序获得主成分。

2.4 创建特征向量来决定保留那些主成分

计算特征向量并按照特征值降序对他们进行排序，使我们可以按照重要性顺序找到主成分。在这一步骤我们选择保留所有特征值还是丢弃那些重要程度较低的特征值。并与剩余的特征值形成一个成为特征向量的向量矩阵。特征向量只是一个矩阵，列为我们决定保留的特征向量。此步骤根据我们的需求来决定。通常是取特征值前**d**对应的特征向量量**ω1 ,ω2 ,...,ωd**，通过以下映射将**n**维样本映射到**d**维；

$$\boldsymbol{x}_i' = \begin{bmatrix} \boldsymbol{\omega}_1^{\mathrm{T}} \boldsymbol{x}_i \\ \boldsymbol{\omega}_2^{\mathrm{T}} \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{\omega}_d{}^{\mathrm{T}} \boldsymbol{x}_i \end{bmatrix}$$

新的**xi '**的第**d**维就是**xi**在第**d**个主成分**ωd**方向上的投影，通过选取最大的**d**个特 征值对应的特征向量，我们将方差较小的特征（噪声）抛弃，使得每个**n**维列向量 **xi**被映射为**d**维列向量**xi '**，定义降维后的信息占比为；

$$\eta = \sqrt{\frac{\sum_{i=1}^{d} {\lambda_i}^2}{\sum_{i=1}^{n} {\lambda_i}^2}}$$

2.5 沿主成分轴重铸数据

在这一步骤，使用协方差矩阵的特征向量形成的特征值，将数据从原始轴重新定向到主成分表示的轴。可以将原始数据集的转置乘以特征向量的转置来完成。

2.6 总结

除了使用目标函数求解最大方差外，可以考虑其他思路分析，例如最小回归误差得到新的目标函数。实际上对应原理和求解方法与该方法是等价的。**PCA**是一种线性降维方法，具有一定的局限性，可以考虑通过核映射对**PCA**机械能扩展得到核主成分分析（**KPCA**），可以通过流形映射降维方法，比如等距映射，局部线性嵌入，拉普拉斯特征映射等，对一些**PCA**效果不好的复杂数据集进行非线性降维操作。

下面将会举例使用**PCA**降维处理;

参考文献

1. 《百面机器学习》
2. aces recognition example using eigenfaces and SVMs
3. A Step-by-Step Explanation of Principal Component Analysis (PCA)
4. 利用**PCA**来简化数据