PCA, LDA, ICA的分析

学院	年级	班级	姓名	学号
控制科学与工程	2017级	人工智能与机器人	陈逸群	201700181055

目录

1 PCA回顾

1.1 基本原理

主成分分析是一种降维方法，其目的是由少数不相关的变量代替相关变量，它通常包括两步：

- 规范化：使得数据的均值为0，方差为1；
- 正交变换：将线性相关的变量表示的数据变成线性无关的新变量所表示的数据；

PCA算法所采取的正交变换矩阵是所有可能的正交变换中能够使原始数据的信息保存最大的，此处通过原始数据的各个维度的方差的和来表示信息量。

对数据进行规范化能够使数据分布在原点附近，减少由于各个维度的量纲带来的问题。规范化之后的各个维度的方差之和（信息量）即为数据的各个维度值得平方之和。

主成分分析用于对数据降维，在不知道原始数据各个维度的重要性的条件下，可以先将原始数据进行线性变换，使得变换后的新的数据在各个维度上都能够确定其重要程度，一个维度的重要程度可以用方差贡献率衡量，方差贡献率越大表示其保留的信息越多，其重要性也越大，因此可以从中选取所需的几个维度，达到降维的目的。

1.2 相关定理及概念的分析与理解

记某一个m维样本 $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$ 从某一个概率分布中采样得到，该概率分布的均值为 $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}$ 按照定义，其协方差矩阵为$\Sigma = cov(x,x) = E[(x-\mu)(x-\mu)^T]$，选定一个对所有样本的变换矩阵 $A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mm} \end{bmatrix}$ 则对于当前选定的样本x而言，可以得到一个新的数据表示 $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mm} \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$ 简记为$y = A^T x$，那么对于新的数据样本而言，其均值为$\mu_y = \alpha\mu$，其协方差矩阵为

$$\Sigma_y = E[(y-\mu_y)(y-\mu_y)^T] \tag{1}$$
$$= E[(A^Tx - A^T\mu)(A^Tx - A^T\mu)^T] \tag{2}$$
$$= A^T E[(x-\mu)(x-\mu)^T]A = A^T\Sigma A \tag{3}$$

进行主成分分析时，需要满足以下条件：

- 线性变换是正交变换，即A是正交矩阵，满足$\alpha_i^T\alpha_j = \begin{cases} 1, i=j \\ 0, i\neq j \end{cases}$；
- y_i, y_j互不相关，即协方差矩阵为对角阵；
- y_1是所有线性变换中方差最大的，y_i是除了$y_1, y_2, \ldots, y_m$以外所有线性变换中方差最大的，这时候分别称之为第一主成分，第二主成分，…，第m主成分；

下面定理给出了主成分分析的方法：

【定理】 若x是m维随机变量，Σ是其协方差矩阵且特征值满足 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$，这些特征值对应的单位 特征向量分别是列向量$\alpha_1, \alpha_2, \ldots, \alpha_m$，则第k主成分为 $y_k = \alpha_k^T x$，对应方差为$var(y_k) = \alpha_k^T \Sigma \alpha = \lambda_k$

该定理可由拉格朗日乘数法以及数学归纳法得证。由该定理可知，$A = [\alpha_1, \alpha_2, \ldots, \alpha_m], \Sigma_y = \Lambda$，$\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ 根据该定理同时可以得到推论：

【推论】 y是x的总体主成分的充要条件是：

- A是正交矩阵；
- 样本的协方差矩阵Σ对角化之后为对角阵$\Lambda = \Sigma_y$；

由以上定理可知，不同主成分之间相关系数为0，信息保留与Σ的特征值有关，由于已将特征值按照大小顺序进行排列，那么选取前q个特征值对应的特征向量构成变换矩阵，可以得到降维后的数据，含有q个主成分，变换后总体保留的信息量由累计协方差贡献率表示。定义第k主成分方差贡献率为： $\eta_k = \frac{\lambda_k}{\sum_{i=1}^m \lambda_i}$ 那么累计协方差贡献率为： $\sum_{i=1}^k \eta_i = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^m \lambda_i}$ 通常选择k使得累计贡献率能够达到80%以上。

以上分析未对原始数据进行规范化，而实际中不同变量往往具有不同量纲，直接求主成分有时候会产生不合理的结果，为了消除这一影响，通常要先对各个随机变量进行规范化，使之均值为0，方差为1.

对于原始数据分布不可知的情况，均值和协方差不可预测，这时候可以使用样本均值和样本方差来代替原始数据分布的均值和方差。

以上所讨论的主成分分析的方法，实际上是对（规范化之后的）样本数据的协方差矩阵进行分析以选择更加重要的样本成分（维度）。实际上求主成分的过程也就是样本协方差矩阵对角化的过程，与一般对角化稍有区别的是：主成分分析时要求将协方差矩阵的各个特征值按照由大到小排列以方便确定更加重要的主成分，同时对角化时所需要的线性变换矩阵为正交矩阵，每一个列向量均为单位向量对角化之后为了降低维度，通常选取最大的q个主成分对应的单位特征向量作为降维时的线性变换矩阵，线性变换之后样本冗余的维度就被减少了。

除了对样本协方差矩阵进行主成分分析以外，还可以采用奇异值分解的办法求出主成分，这时候采用截断奇异值分解以减少冗余的维度。对于$m \times n$为的样本矩阵X，其每一行元素的均值均为0（规范化后），确定主成分的个数，构造新的$m \times n$矩阵$X' = \frac{X^T}{\sqrt{n-1}}$，对构造的矩阵进行奇异值分解，有 $X' = U\Sigma V^T$，则降维之后的样本矩阵为：$Y = V^T X$

2 LDA简介

2.1 基本原理

线性判别分析是一种经典的线性学习方法，最早被应用于二分类问题上。线性判别分析的思想非常朴素：给定训练集，将训练样例投影到一条直线上，使得同样的样例的投影点尽可能接近、异类样例的投影点尽可能远离。在对新样本进行分类时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样例的类别。

由于多分类LDA将样本投影到N-1维空间，N-1通常远小于数据原有的属性数，而且投影的过程中使用了类别信息，因此可以通过投影来减小样本点的维数，此时LDA就变成了一种经典的有监督降维方法。

2.2 算法

给定数据集$D = \{(x_i, y_i)\}_{i=1}^m, y_i \in \{0, 1\}$，令$X_i, \mu_i, \Sigma_i$分别表示样例$i \in \{0, 1\}$类样例的集合、均值向量、协方差矩阵。若将数据投影到直线w上，则两类样本的中心在直线上的投影分别为$w^T\mu_0, w^T\mu_1$；若

将所有样本点都投影到直线上，则两类样本的协方差分别为 $w^T\Sigma_0 w, w^T\Sigma_1 w$。由于直线是一维空间，因此$w^T\mu_0, w^T\mu_1$和 $w^T\Sigma_0 w, w^T\Sigma_1 w$均为实数。

若想要同类样例的投影点尽可能接近，意味着它们的相似度尽可能高，因此同类样例投影点的协方差应该尽可能地小。若想要异类样例地投影点尽可能原理，意味着它们的类中心应该尽可能有区分度，即两类的类中心的距离尽可能大。同时考虑二者，得到优化的目标：

$$J = \frac{||w^T\mu_0 - w^T\mu_1||_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w} \tag{4}$$

$$= \frac{w^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w}{w^T(\Sigma_0 + \Sigma_1)w} \tag{5}$$

定义“类内散度矩阵”

$$S_w = \Sigma_0 + \Sigma_1 \tag{6}$$

$$= \sum\nolimits_{x\in X_0}(x-\mu_0)(x-\mu_0)^T + \sum\nolimits_{x\in X_1}(x-\mu_1)(x-\mu_1)^T \tag{7}$$

“类间散度矩阵”

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T \tag{8}$$

则上式可以改写为：$J = \frac{w^T S_b w}{w^T S_w w}$

为了确定w，可以利用拉格朗日乘数法，求解最优化问题：

$$\begin{cases} min_w \quad -w^T S_b w \\ s.t. \quad w^T S_w w = 1 \end{cases}$$

可以解得$w = S_w^{-1}(\mu_0 - \mu_1)$ 考虑到数值解的稳定性，在实际中通常对类内散度矩阵进行奇异值分解，然后求解类内散度矩阵的逆阵。 LDA可以从贝叶斯决策理论的角度来解释，并可以证明，当两类数据同先验，满足高斯分布且协方差相等时，LDA可达到最优分类。

定义全局散度矩阵 $S_t = S_b + S_w = \sum_{i=1}^m (x_i - \mu)(x_i - \mu)^T$ 其中μ是所有实例的均值向量，将类内散度矩阵定义为每个类别的散度矩阵之和，即

$$S_w = \sum\nolimits_{i=1}^N S_{w_i} = \sum\nolimits_{i=1}^N \sum\nolimits_{x\in X_i}(x-\mu_i)(x-\mu_i^T) \tag{9}$$

则可以得到类间散度矩阵为

$$S_b = S_t - S_w = \sum\nolimits_{i=1}^N m_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{10}$$

常用的一种方法就是采用优化目标 $max_W \quad \frac{tr(W^T S_b W)}{tr(W^T S_w W)}$ 通过如下广义特征值问题求解：$S_b W = \lambda S_w W$，若将W视为一个投影矩阵，则多分类LDA将样本投影到N-1维空间，N-1通常远小于数据原有的属性数，而且投影的过程中使用了类别信息，因此可以通过投影来减小样本点的维数，从而达到降维的目的。

3 ICA简介

3.1 基本原理

独立成分分析又称为“盲源分离”，被认为来自于“鸡尾酒会问题”：在一场嘈杂的鸡尾酒会中，有很多噪声，如何分辨出某个人说话的内容？

独立成分分析假设观察到的随机信号服从模型$x = As$，其中s维未知源信号，其分量相互独立，A是一个混合矩阵，它把源信号进行线性组合之后作为我们观测到的观测数据。独立成分分析的目的是通过且仅通过观察x来估计混合矩阵A以及源信号s。

3.2 算法

这里采用了极大似然算法。我们假设s，x是d维向量，即$s \in \Re^d, x \in \Re^d$，s各个维度相互独立，每个维度都视为从相互独立的分布中采样得到的源信号。假设我们在不同的时刻进行采样，得到n个样本$x^{(1)}, x^{(2)}, \ldots, x^{(n)}$，对应的未知的源信号为$s^{(1)}, s^{(2)}, \ldots, s^{(n)}$ 且第i个时刻的样本满足$x^{(i)} = As^{(i)}$，其中，A为混合矩阵，将A的逆矩阵称为去混和矩阵，即$W = A^{-1}$ 并记W的第i行为w_i^T，即

$$W = \begin{bmatrix} \text{˘˘} w_1^T \text{˘˘} \\ \vdots \\ \text{˘˘} w_2^T \text{˘˘} \end{bmatrix} \tag{11}$$

那么有$s_J^{(i)} = w_j^T x^{(i)}$

我们对于源信号没有任何先验知识的时候，有些情况下去混和矩阵W是没有办法估计的：

源信号位置的不确定性 我们假设有一个矩阵P由仅对单位矩阵施加初等行（列）变换得到，这时候实际上是无法区分算法得到的是W还是PW，也就是说，我们没有办法确定出s各个源信号的相对位置。好在大部分实际问题都不会涉及此。

A的放缩 当对混合矩阵A进行放缩时，对s进行缩放可以得到一样的观测数据。实际上，对A的某一列进行缩放，对相应的s的源信号进行放缩，仍然能够得到一样的结果，也就是说，我们没有办法确定实际的源信号的真实值，只能得到经过某一未知缩放因子后得到的源信号，亦即只能确定源信号之间的大小的相对关系。

源信号应该是非高斯的 假设s是高斯的，亦即$s \sim \mathcal{N}(0, I)$，则易知

$$E(x) = E(As) = AE(s) = 0$$

$$Cov(x) = E(xx^T) = E(Ass^T A^T) = AE(ss^T)A^T = ACov(s)A^T = AA^T$$

亦即$x \sim \mathcal{N}(0, AA^T)$ 即R为任意正交矩阵，令观测数据$x' = A's, A' = AR$，则同样的可以得到$x' \sim \mathcal{N}(0, AA^T)$，也就是说，我们无法得知混合矩阵到底是A还是A'。大部分情况下源信号都是非高斯的，因此大部分问题仍可求解。

假设每一个源信号s_j都有自己的概率分布$p_s(s_j)$，则将各个信号的联合分布假设为各个边缘分布的乘积，以满足源信号相互独立的假设，即$p(s) = \prod_{j=1}^{d} p_s(s_j), p(x) = \prod_{j=1}^{d} p_s(w_j^T) \cdot |W|$

现在确定累积分布函数以求得概率密度函数。由于概率密度函数必须是非高斯的，因此由很多其他的选择，这里选择sigmoid 函数，即$g(s) = \frac{1}{1+e^{-s}}, p_s(s) = g'(s)$。联合概率的对数似然函数为

$$l(W) = \sum\nolimits_{i=1}^{n} (\sum\nolimits_{j=1}^{d} log(g'(w_j^T x^{(i)})) + log(|W|))$$

这里将整个数据集的所有样本考虑进来，并且假设数据集中的每一个样本都相互独立（不是指一个样本的各个维度相互独立），显然在有些问题中并不能满足这样的假设，例如说话、演讲等。但可以证明，只要数据量够多，就可以无视其带来的影响。根据对数似然函数，可以求得关于模型参数W的梯度为

$$\nabla_W |W| = |W|(W^{-1})^T$$

，则W的更新规则为

$$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(2w_1^T x^{(i)}) \\ \vdots \\ 1 - 2g(2w_d^T x^{(i)}) \end{bmatrix} \right)$$

其中α为指定的学习率，算法将持续更新直至收敛。

另外还有FastICA等算法，由于篇幅所限不作介绍。

4 三者异同分析

PCA是一种无监督的降维方法，通常用于减少数据冗余的维度，筛选出数据的主要结构信息，采用的线性变换为正交变换，并保证投影在各个维度上的方差尽可能大，而且相互正交，它不需要太在意数据本身的类别信息，没有权重初始化，不易收敛到不同的值，有更强的鲁棒性，更不需要担心数据过拟合问题。

LDA是一种有监督的降维方法，也可以用于数据的分类，它假设数据分布为高斯分布，其基本思想为减小数据的类内距离并增大类间距离，并不保证投影所得到的新坐标系是正交的，不同的维度之间可能还存在着关联。

ICA并不认为最有用的信息体现在类间距离或者数据方差中，而是构成样本的独立成分，它假设数据分布为非高斯分布，在高斯分布的数据上效果反而不好。实际上ICA要求的源信号维度与观测信号维度一致，并不能起到降维的效果，但得到的各自独立的源信号，某种意义上会更有区分度，而且ICA一般也不单独使用，而是与PCA或者白化处理结合使用。

5 参考资料

[1] 机器学习，周志华；
[2] CS229，斯坦福大学公开课，吴恩达等；
[3] 统计学习方法，李航；