独立成分分析[1]

简介

在PCA中我们学习了如何从原始数据中找到主成分，在这个算法的最后一步，我们需求解协方差矩阵Σ的top k特征向量。当数据的维度特别高时，求解Σ的特征向量可能变得异常困难，这时我们可以考虑PCA算法的另一种实现——奇异值分解（Sigular Value Decomposition），这里不详述。在本notes中，我们将学习另一种从数据中分解主元（basis）的方法——独立成分分析（Independent Components Analysis）。

1 问题引入

在经典的『鸡尾酒聚会问题』中，假设在一个房间里n个人在说话，同时放置在房间各个位置的n个仪器不断对房间的声音进行m组同步采样，也就是说每一组采样数据是n维的。我们该如何分离出m组采样数据中每个人说的话呢？这个问题用数学语言描述如下

假设采样数据$s \in \mathbb{R}^n$是由n个独立源信号混合产生，我们观察到的是

$$x = As$$

其中，A是一个未知的n维方阵，称为混合矩阵（mixing matrix）。重复观察并得到数据集$\left\{x^{(i)}; i = 1, \cdots, m\right\}$，我们的目标是找到源信号$s^{(i)}$。

具体到『鸡尾酒聚会问题』中，$s^{(i)}$是一个n维向量，$s_j^{(i)}$是说话者j在时刻i的发出的信号；$x^{(i)}$也是一个n为向量，$x_j^{(i)}$是仪器j在时刻i采样的数据。记分离矩阵（unmixing matrix）$W = A^{-1}$，显然我们的目标就是找到W，在给定采样数据$x^{(i)}$后，我们便可计算出源信号$s^{(i)} = Wx^{(i)}$。

令w_i^T表示W的第i行，那么W可表示为

$$W = \begin{bmatrix} -w_1^T- \\ \vdots \\ -w_n^T- \end{bmatrix}$$

此时$w_i \in \mathbb{R}^n$，源信号的第j维可由$s_j^{(i)} = w_j^T x^{(i)}$计算出。

2 ICA的不确定性

我们利用$s = Wx$从采样数据中恢复源信号，如果没有关于s或者W的先验知识，我们很难同时确定这两个参数。因为我们可以很容易地对s和W乘一个相等的因子α使等式仍然满足。此外，如果有人将s的维度打乱，我们也很容易通过改变W的行向量w_i的位置来使等式仍然满足。具体到『鸡尾酒聚会问题』中，我们可以理解为说话者的音量、符号和所在位置不能确定。

ICA的另一个不确定性表现为源信号s必须是非高斯分布。假设$n = 2$，且$s \sim \mathcal{N}(0, I)$，这里I表示2*2的单位矩阵。回顾多元高斯分布的知识可知s在平面上的投影是一个以原点为圆心的圆，是一个旋转对称的形状。由于$x = As$，所以x也服从高斯分布，且均值为0，协方差矩阵为$E[xx^T] = E[Ass^TA^T] = AA^TE[xx^T] = AA^T$。令$R$是某正交矩阵（$RR^T = I$），记$A' = AR, x' = A's$。此时$x'$服从高斯分布，且均值为0，协方差矩阵为$E[x'(x')^T] = E[A'ss^T(A')^T] = E[ARss^T(AR)^T] = ARR^TA = AA^T$。这说明对服从高斯分布的源信号，经过不同的混和矩阵做线性变换都可以得到相同的采样数据，在这种情况下无法确定W及s。

3 概率密度和线性变换

在讨论ICA算法之前，我们首先来回顾一下概率密度和线性变换的知识。

假定随机变量s概率密度为p_s，其中$s \in \mathbb{R}$是一个实数。令随机变量$x = As$，那么x的概率密度p_x该如何表示呢？

记$W = A^{-1}$是一个实数，那么$s = Wx$。x的累积分布函数（cumulative distribution function, cdf）

$$
\begin{aligned}
& F_X(t) = P\{X \le t\} = P\{AS \le t\} = P\{S \le Wt\} = F_S(Wt) \\
\Rightarrow & \int_{-\infty}^{t} p_x(x)dx = \int_{-\infty}^{Wt} p_s(x)dx \\
\Rightarrow & p_x(t) = p_s(Wt)\frac{d(Wt)}{dt} = p_s(Wt) \cdot |W|
\end{aligned}
$$

其中，$|W|$表示W的行列式，对于实数$|W| = W$。一般地，当$s \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}$且$A$可逆时上式仍成立。

我们也可以从线性变换角度得到p_x的公式。令$C_1 = [0,1]^n$表示n维超立体（hypercube），定义$C_2 = \{AS : s \in C_1\} \subseteq \mathbb{R}^n$表示$C_1$中的元素经线性变换$A$构成的集合。根据线性代数的结论，$C_2$的体积[2]等于$|A|$。现在假设$s \sim \text{Uniform}[0,1]^n$，其概率密度$p_s(s) = 1\{s \in C_1\}$。经过线性变换后，$x$在$C_2$中也均匀分布，其概率密度$p_x(x) = 1\{x \in C_2\}/\text{vol}(C_2) = 1\{x \in C_2\}/|A| = 1\{x \in C_2\}|W| = 1\{Wx \in C_1\}|W| = p_s(Wx)|W|$。

4 ICA算法

假设源信号的每一维s_i的概率密度都是p_s，那么源信号的联合概率可由下式给出

$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$

在上式中我们令联合概率等于所有边缘分布的乘积，这是因为我们假设所有的源都是独立的。此时采样数据x概率等于

$$p(x) = \prod_{i=1}^{n} p_s\left(w_i^T x\right) \cdot |W|$$

我们该如何给出s的累积分布函数F_S呢？首先s是非高斯分布，且累计分布函数要求单调从0递增到1，可以发现sigmoid函数$g(s) = 1/(1+e^{-s})$很适合，且此时$p_s(s) = g'(s) = e^s/(1+e^s)^2$。

现在我们模型中只剩下方阵W一个参数了，给定训练集$\left\{x^{(i)}; i = 1, \cdots, m\right\}$的对数估计[3]如下

$$l(W) = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} \log g'\left(w_j^T x^{(i)}\right) + \log|W|\right)$$

我们的目标是找到使$l(W)$最大化的参数W。由随机梯度上式算法得更新法则

$$W := W + \alpha\left(\begin{bmatrix} 1 - 2g\left(w_1^T x^{(i)}\right) \\ 1 - 2g\left(w_2^T x^{(i)}\right) \\ \vdots \\ 1 - 2g\left(w_n^T x^{(i)}\right) \end{bmatrix} x^{(i)^T} + (W^T)^{-1}\right)$$

为得到上式我们用到了两个结论，$\left(\log g'(s)\right)' = 1 - 2g(s)$及$\nabla_W |W| = |W|(W^{-1})^T$，具体过程请自己动手计算，并不是很难。

迭代多次得到收敛的W后，可利用$s^{(i)} = Wx^{(i)}$还原源信号。

1. Written by Jimmy on 2016/03/17.↩
2. 超立体的体积指的是超立体的密度在所有维度上的积分值，可以理解为向量、矩阵的范数，是一种衡量超立体大小的参数。↩
3. 在应用最大似然估计时，我们假定不同时刻观察到的数据$x^{(i)}$与$x^{(j)}$是相互独立的，然而对于语音信号或者其他具有时间依赖关系的信号（如温度），这个假设不成立。↩