ICA

2015年7月27日 星期一　　下午10:32

一, 简介

ICA(independent component analyse)独立成分分析,是一种信号/数据分解的方法. 最早应用于 盲源信号分离（Blind Source Separation，BBS）,即 从一组混合的观测信号中分离出独立信号.

PCA是一种数据降维的方法,但是只对符合高斯分布的比较有效,ICA是针对非高斯分布的一种分解

二,详述

1,定义

是一种利用统计原理进行计算的方法。它是一个线性变换。这个变换把数据或信号分离成统计独立的非高斯的信号源的线性组合。独立成分分析是 盲信号分离 （Blind source separation）的一种特例。

2,一般定义

观察的数据或者信号用随机向量 $x = (x_1, \ldots, x_m)$ 表示，独立成分量可以定义为向量 $s = (s_1, \ldots, s_n)$ 。独立成分分析的目的是通过线性变换把观察的数据 x , 转换成独立成分向量 $s = Wx$, 而独立成分分量满足互相统计独立的特性。统计独立的量化通常通过某指定函数 $F(s_1, \ldots, s_n)$ 来衡量。也就是“鸡尾酒会问题”(cocktail party problem)

3,算法过程 (Ngnote和博客)

1) 我们确定下来已知和未知

$$x = As$$

S是每个演讲者发出的信号 x是每个麦克风接收到的信号 A是mixing matrix 这里s和x都是考虑时间的一个大的矩阵

$$x^{(i)} = As^{(i)}$$

$$s^{(i)} = Wx^{(i)}$$

这是i时刻的情况 xi和si都是一个向量 表示i时刻的各个演讲者/麦克风的情况

$x_j^{(i)}$
是i时刻j个麦克风
$s_j^{(i)}$
是i时刻j演讲者　每个麦克风和每个演讲者之间只有一个系数 现在看来是线性的
我们要求的就是S.

2) ICA的不确定性

但是我们现在只知道x,A s 都不知道.

1,对于两个都不知道的变量,是无法唯一的求得s和w的,同时扩大n倍就会得到另一个结果.

2,将A的列向量顺序换了就会导致s也跟着变幻,对于是哪个演讲者发出的声音也看不出来了

这种情况称为原信号的不确定性.

由于高斯分布是对称的,所以ICA不适用于高斯分布

3)密度函数和线性变换

我们要求的是 , 随机变量s的概率密度函数
$p_s(s)$
.一个naive的想法是,根据已知的概率密度函数Px(x). 直接替换
$p_x(x) = p_s(Wx)$
, 但是这个是不对的,应该按照累计概率分布来计算.

]公式应该是

$p_x(x) = p_s(Wx)|W|$

推导方法

$F_x(x) = P(X \le x) = P(AS \le x) = P(S \le Wx) = F_s(Wx)$

$p_x(x) = F_x'(x) = F_s'(Wx) = p_s(Wx)|W|$

更一般地，如果s是向量，A可逆的方阵，那么上式子仍然成立。

4)ICA算法

首先,根据每个演讲者的概率分布计算出所有演讲者的概 率分布.(这个公式代表一个假设前提：每个人发出的声音信号各自独立)

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

其次,继续化简

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

然后,为Ps(s) 选一个先验的概率分布,不能用高斯分布,我们发现可以用累积概率分布为sigmoid的概率分布,就是对sigmoid求个导,得到概率分布函数

$$p_s(s) = g'(s) = \frac{e^s}{(1+e^s)^2}$$

接着,根据训练样本 $\{\mathbf{x}^{(i)}(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}); i = 1, \ldots, m\}$,对样本做对数似然概率分布,

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

最后,求导之后的公式是这样的

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

三,其他

1,发展

ICA方法最早是由法国的J.Herault和C.Jutten于八十年代中期提出来的，现在常称他们的方法为H-J算法，可以说是最经典的ICA算法之一。目前比较流行的ICA算法又Infomax算法（信息最大化）、FastICA算法（定点算法，Fixed-point、快速ICA算法），方法分类的依据主要是求取分离矩阵W的方法不同。

线性独立成分分析可以分为无噪声模型和有噪声模型，其中无噪声模型可看作有噪声模型的特例。非线性独立成分分析的情况应该单独处理。

2,注意

是基于非高斯分布的

四,参考资料

http://www.cnblogs.com/jerrylead/archive/2011/04/19/2021071.html 和Ng公开课很像的博客,入门不错

http://blog.sina.com.cn/s/blog_73402e3c0101gqy0.html 一些补充的博客

http://zh.wikipedia.org/wiki/%E7%8B%AC%E7%AB%8B%E6%88%90%E5%88%86%E5%88%86%E6%9E%90 维基