Independent Components Analysis

独立成分分析

算法目的

- 与PCA类似，也是找到一组新的基向量来表征样本数据，但两者目的不同
- 回到我们前面说过的“鸡尾酒问题”。在一个聚会场合中，有 n 个人同时说话，而屋子里的任意一个话筒录制的都是叠加在一起的这 n 个人的声音。但如果我们也有 n 个不同的话筒安装在屋子里，并且这些话筒与每个说话人的距离都各自不同，那么录下来的也就是不同 的组合形式的所有人声音的叠加。使用这样布置的 n 个话筒来录音，能不能区分开原始的 n 个说话者每个人的声音信号呢？
- 答案是能，用的就是ICA算法。所以ICA算法可以用来分离信号

鸡尾酒问题形式化描述

- 将上述问题形式化描述
- 假设有某些样本数据 $s \in \mathbb{R}^n$，由n个独立来源生成
- 假设 $x = As$，A是未知正方形矩阵，叫做混合矩阵(mixing matrix)
- 通过观察我们得到了训练集 $\{x^{(i)}; i = 1, \ldots, m\}$
- **我们的目的就是通过训练集恢复源数据** sources $s^{(i)}$，**其中** $(x^{(i)} = As^{(i)})$
- 对应到鸡尾酒问题，$s^{(i)}$是一个n维向量，$s_j^{(i)}$ 表示第j个演讲者在第i次录音时发出的声音，$x_j^{(i)}$ 表示第j个话筒在第i次录制到的声音。
- 记A的逆矩阵为 $W = A^{-1}$，称为还原矩阵(unmixing matrix)
- **这样我们的目的就变成了找出矩阵W，这样就可以通过话筒录音** $x^{(i)}$**恢复出源数据** $s^{(i)} = Wx^{(i)}$
- 为了方便起见，我们记 w_i^T 为W矩阵的第i行，所以有 $$W = \begin{bmatrix} — w_1^T — \\ \vdots \\ — w_n^T — \end{bmatrix}, \; w_i \in \mathbb{R}^n$$
- 这样，就可以通过 $s_j^{(i)} = w_j^T x^{(i)}$ 来恢复出第j个声源了

ICA ambiguities(ICA的模糊性)

- 因为在公式 $x = As$ 中，S,A都是未知的，所以在没有先验知识的情况下，不可能同时确定这两个参数。例如当W扩大两倍时，S缩小两倍，等式仍然成立。
- 还有，若将人的排列顺序打乱，换成另外一种排列，那么只需调整一下A的列向量顺序即可，所以也无法单独确定S
- 还有一种不确定性，那就是S分布不能是高斯分布
- 若S是高斯分布
  - 例如一个样本中，n=2，$s \sim \mathcal{N}(0, I)$
    - 这是一个标准正态分布，轮廓图是以原点为中心，密度是旋转对称的。
  - 接下来我们观测到 $x = As$
    - 那么EX=0并且协方差 $\mathrm{E}[xx^T] = \mathrm{E}[Ass^TA^T] = AA^T$
  - 设R为任意的一个正交矩阵，所以满足 $RR^T = R^TR = I$
  - 令 $A' = AR$
  - 如果混合矩阵使用A'，那么得到 $x' = A's$
  - x'的分布也是高斯分布，EX'=0并且协方差 $\mathrm{E}[x'(x')^T] = \mathrm{E}[A'ss^T(A')^T] = \mathrm{E}[ARss^T(AR)^T] = ARR^TA^T = AA^T$
  - 所以无论混合矩阵用的是哪一个，得到的分布都是 $\mathcal{N}(0, AA^T)$
  - 所以若S是高斯分布，则无法求出混合矩阵
  - 上面这些论证，是基于多元标准正态分布（multivariate standard normal distribution）是旋转对称（rotationally symmetric）的这个定理。

密度和线性变换

- 先把ICA放到一边，假设我们有某随机变量S，它的密度函数为 $p_s(s)$
- 为了简单起见，我们设 $s \in \mathbb{R}$，随机变量X通过 $x = As$ (here, $x \in \mathbb{R}, A \in \mathbb{R}$)被定义
- *那么x的密度函数p_x是多少呢？*
- 错误推理：设 $W = A^{-1}$，为了计算x的一个特值的概率，首先要计算对于 $s = Wx$，这一点上的 p_s，所以得出结论 $p_x(x) = p_s(Wx)$.
- 反例：$s \sim \text{Uniform}[0, 1]$，则 $p_s(s) = 1\{0 \le s \le 1\}$，若 $A = 2$, so that $x = 2s$.
  众所周知，$p_x(x) = (0.5)1\{0 \le x \le 2\}$，这并不等于 $p_s(Wx)$, where $W = 0.5 = A^{-1}$
- *正确结论是：* $p_x(x) = p_s(Wx) \cdot |W|$ where $W = A^{-1}$
- 推导方法：
  $$F_X(x) = P(X \le x) = \mathrm{P}(AS \le x) = \mathrm{P}(S \le Wx) = F_S(Wx)$$
  $$\mathrm{p}_x(\mathrm{x}) = F_X'(x) = F_S'(Wx) = \mathrm{p}_s(Wx)|\mathrm{W}|$$

ICA算法

- 我们假设每个声源的分布 si 都是通过密度函数 ps 给出，然后 联合分布 s 则为：$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$
- 利用上一节的推导 $p_x(x) = p_s(Wx) \cdot |W|$，可以得出 $$p(x) = \prod_{i=1}^{n} p_s(w_i^T x) \cdot |W|$$
- 剩下的就是需要确定每个独立的声源的密度函数ps了。
- 回顾知识点：给定一个随机变量z，它的累计分布函数(cdf)为 $F(z_0) = P(z \le z_0) = \int_{-\infty}^{z_0} p_z(z)dz$，密度函数为 $p_z(z) = F'(z)$
- 所以要确定声源的密度函数，首先要确定它的cdf，根据概率论中cdf的性质，这个函数是一个从0到1的单调函数。我们就选用常见的sigmoid函数作为cdf函数
- $g(s) = 1/(1 + e^{-s})$，$p_s(s) = g'(s)$
- W是方形矩阵，是模型中的参数，给定训练集合 $\{x^{(i)}; i = 1, \ldots, m\}$，对数似然函数为 $$\ell(W) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log g'(w_j^T x^{(i)}) + \log |W| \right)$$
- 借助行列式性质 $\nabla_W |W| = |W|(W^{-1})^T$ 对似然函数求导得到更新规则
- $$W := W + \alpha \left(\begin{bmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{bmatrix} x^{(i)T} + (W^T)^{-1} \right)$$
- 注：我们对ICA算法的推导过程是假设xi彼此独立，在实际中这个假设并不总是成立，但是只要有充足的数据，这个假设并不影响算法的性能。
- 附：行列式求导公式证明
  - 对行列式求导，设矩阵 A 是 n×n 的，我们知道行列式与代数余子式有关，
    $$|A| = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A_{\backslash i, \backslash j}| \quad (\text{for any } j \in 1, \ldots, n)$$
    $A_{\backslash i, \backslash j}$是去掉第 i 行第 j 列后的余子式，那么对$A_{k,l}$求导得
    $$\frac{\partial}{\partial A_{k\ell}} |A| = \frac{\partial}{\partial A_{k\ell}} \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A_{\backslash i, \backslash j}| = (-1)^{k+\ell} |A_{\backslash k, \backslash \ell}| = (\text{adj}(A))_{\ell k}.$$
    adj(A)跟我们线性代数中学的A*是一个意思，因此
    $$\nabla_A |A| = (\text{adj}(A))^T = |A| A^{-T}.$$
  - adj(A)就是伴随矩阵