机器学习 第二次课程作业

UNIkeEN

2024 年 5 月 12 日

问题解答

1. Give at least two algorithms that could take data set $X = \{x_1, \ldots, x_N\}, x_t \in \mathbb{R}^{n\times 1}, \forall t$, as input, and output the first principal component w. Specify the computational details of the algorithms, and discuss the advantages or limitations of the algorithms.

解 方法一：SVD（奇异值分解）

Algorithm 1 SVD 法计算第一主成分

输入: 数据集 $X = \{x_1, \ldots, x_N\}, x_t \in \mathbb{R}^{n\times 1}, \forall t$

输出: 第一主成分 w

1: 计算数据集 X 的均值向量 $\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$

2: 对每个观测值 x_j 进行中心化：$x_j \leftarrow x_j - \bar{x}$

3: 对中心化后的数据矩阵 X 执行奇异值分解：$X = U\Sigma V^T$

4: U 的列向量是 X 的左奇异向量，Σ 的对角元素是奇异值

5: 选择 Σ 中最大奇异值对应的右奇异向量 v_1 作为第一主成分

6: 设置 $w = v_1$

7: **return** w

优点

- 可以处理非方阵的矩阵（另一方法，特征分解只适用于方阵）
- 不需要计算协方差矩阵，降低计算量
- SVD 可以考虑两个不同方向

缺点

- SVD 的压缩结果缺乏可解释性

解 方法二：特征分解

Algorithm 2 特征分解法计算第一主成分

输入: 数据集 $X = \{x_1, \ldots, x_N\}, x_t \in \mathbb{R}^{n\times 1}, \forall t$
输出: 第一主成分 w

1: 计算数据集 X 的均值向量 $\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$
2: 对每个 x_j 执行中心化操作: $x_j \leftarrow x_j - \bar{x}$
3: 计算中心化后的数据矩阵 X 的协方差矩阵: $C = \frac{1}{n}XX^T$
4: 利用特征分解计算协方差矩阵 C 的特征值 λ_i 和特征向量 α_i
5: 选择最大特征值 λ_m 对应的特征向量 α_m
6: 计算第一主成分: $w = \alpha_m X$
7: **return** w

优点

- 最简单的 PCA，易于理解并实现、计算结果精确
- 通过方差测量，不受样本标签约束

缺点

- 当数据维度非常高时，计算和存储协方差矩阵代价非常昂贵
- 解释主成分含义时有一些歧义，不如原始样本完整
- 对异常值敏感

2. Calculate the Bayesian posterior $q(y|x)$ of the Factor Analysis model $x = Ay + \mu + e$, with $q(x|y) = G(x|Ay + \mu, \Sigma_e)$, $q(y) = G(y|0, \Sigma_y)$, where $G(z|\mu, \Sigma)$ denotes Gaussian distribution density with mean μ and covariance matrix Σ.

解

根据贝叶斯定理，后验概率可以表示为:

$$q(y|x) = \frac{q(x|y)q(y)}{q(x)} \tag{1}$$

$$= \frac{G(x|Ay + \mu, \Sigma_e) \cdot G(y|0, \Sigma_y)}{q(x)} \tag{2}$$

对数似然可以写为:

$$-\frac{1}{2}\left[(x - Ay - \mu)^T \Sigma_e^{-1}(x - Ay - \mu) + y^T \Sigma_y^{-1} y\right] \tag{3}$$

其中，常数项与 x 和 y 无关。为了简化，设 $\mu_{y|x}$ 为条件均值，$\Sigma_{y|x}$ 为条件协方差，我们有:

$$\Sigma_{y|x}^{-1}\mu_{y|x} = A^T \Sigma_e^{-1}(x - \mu) \tag{4}$$

$$\Sigma_{y|x}^{-1} = A^T \Sigma_e^{-1} A + \Sigma_y^{-1} \tag{5}$$

求逆得到：

$$\Sigma_{y|x}^{-1} = (A^T \Sigma_e^{-1} A + \Sigma_y^{-1}) \tag{6}$$

$$\mu_{y|x} = \Sigma_{y|x} A^T \Sigma_e^{-1} (x - \mu) \tag{7}$$

因此，后验概率 $q(y|x)$ 可表示为：

$$q(y|x) = G(y|\Sigma_{y|x} A^T \Sigma_e^{-1}(x-\mu), (A^T \Sigma_e^{-1} A + \Sigma_y^{-1})^{-1}) \tag{8}$$

3. Explain why maximizing non-Gaussianity could be used as a principle for ICA estimation.

解 假设有两个声源信号 s 遵从多元高斯分布，即 $S \sim N(0, I)$，其中 I 是 2x2 的单位矩阵，表示信号是独立同分布的。这里的 $x = As$ 表示观测信号 x 是源信号 s 的线性混合，其中 A 是混合矩阵。

根据多元高斯分布的性质，x 也遵从多元高斯分布：

$$x \sim N(0, AA^T)$$

现在考虑一个正交矩阵 R，满足 $RR^T = R^T R = I$。定义一个新的混合矩阵 $A' = AR$。使用新的混合矩阵，我们得到新的观测信号：

$$x' = A's = ARs$$

由于 R 是正交矩阵，它不改变 s 的分布（即 Rs 仍然是 $N(0, I)$ 分布），因此 x' 也是高斯分布：

$$x' \sim N(0, ARA^T)$$

但由于 R 是正交的，我们有 $ARA^T = AA^T$。这说明观测信号 x 和 x' 有相同的分布，即：

$$x \sim N(0, AA^T) = x' \sim N(0, ARA^T)$$

由于高斯分布源信号的线性组合（通过任意正交变换）仍然是高斯分布，我们无法仅通过观测信号确定唯一的混合矩阵 A。如果源信号是非高斯的，它们的线性组合不太可能仍然是高斯分布，这样通过观测信号就可以更容易地恢复出原始的非高斯源信号。

因此，最大化非高斯性成为一种有效的策略，通过这种方式，ICA 能够找到一个适当的解混矩阵 W，使得 $y = Wx$ 的分量尽可能地独立（即尽可能地非高斯）。这样，每个组分的非高斯性度量将被用来辅助找到正确的 W，从而有效分离独立源信号。

4. Consider the following Factor Analysis (FA) model,

$$x = Ay + \mu + e, \tag{1}$$
$$q(x|y) = G(x|Ay + \mu, \sigma^2 I), \tag{2}$$
$$q(y) = G(y|0, I), \tag{3}$$

where the observed variable $x \in \mathbb{R}^n$, the latent variable $y \in \mathbb{R}^m$, and $G(z|\mu, \Sigma)$ denotes Gaussian distribution density with mean μ and covariance matrix Σ. Write a report on experimental comparisons on model selection performance by BIC, AIC on selecting the number of latent factors, i.e., $\dim(y) = m$.

Specifically, you need to randomly generate datasets based on FA, by varying some setting values, e.g., sample size N, dimensionality n and m, noise level σ^2, and so on. For example, set $N = 100$, $n = 10$, $m = 3$, $\sigma^2 = 0.1$, $\mu = 0$, and assign values for $A \in \mathbb{R}^{n \times m}$. The generation process is as follows:

1. Randomly sample a y_t from Gaussian density $G(y|0, I)$, with $\dim(y) = m = 3$;
2. Randomly sample a noise vector e_t from Gaussian density $G(e|0, \sigma^2 I)$, with $\sigma^2 = 0.1$, $e_t \in \mathbb{R}^n$;
3. Get $x_t = Ay_t + \mu + e_t$.

Collect all the x_t as the dataset $X = \{x_t\}_{t=1}^N$.

The two-stage model selection process for BIC, AIC is as follows:

Stage 1: Run EM algorithm on each dataset X for $m = 1, \ldots, M$, and calculate the log-likelihood value $\ln p(X|\hat{\Theta}_m)$, where $\hat{\Theta}_m$ is the maximum likelihood estimate for parameters;

Stage 2: Select the optimal m^* by

$$m^* = \arg \max_{m=1,\ldots,M} J(m), \tag{4}$$
$$J_{\text{AIC}}(m) = \ln p(X|\hat{\Theta}_m) - d_m, \tag{5}$$
$$J_{\text{BIC}}(m) = \ln p(X|\hat{\Theta}_m) - \frac{\ln N}{2} d_m, \tag{6}$$

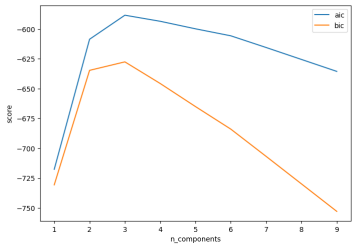
where d_m denotes the number of free parameters of FA model with m latent factors. You may set $M = 5$, if you generate the dataset X based on $n = 10$, $m = 3$.

解 在本次实验中，我们对因子分析（FA）模型进行了测试。实验中的相关参数如下：数据集的样本量 $N = 100$，维度 $n = 10$，方差 $\sigma^2 = 0.1$，均值 $\mu = 0$，潜在因子数 m 分别设置为 3、4、5、6。

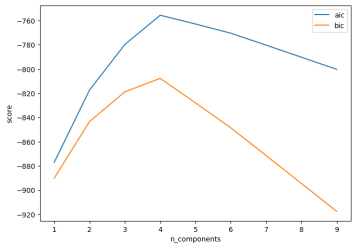
第一步是生成数据集，我们借助 `np.random.multivariate_normal` 中的方法，使用如下公式生成数据集：

$$y_l \sim G(y|0, I) \tag{9}$$
$$e_l \sim G(e|0, \sigma^2 I), \quad e_l \in \mathbb{R}^n \tag{10}$$
$$X_l = Ay_l + \mu + e_l \tag{11}$$

接下来，分别应用 Akaike 信息判据（AIC）、Bayes 信息判据（BIC）。如图 1 所示，当样本量较小且向量维度较低时，对数似然值明显在设定的 m 值处有一个转折点：即随着 AIC、BIC 的 `n_components` 参数上升，两条曲线均先上升后下降，在设定的 m 值处达到峰值。BIC 曲线的下降速度比 AIC 更快，这是由于 BIC 考虑了样本量。



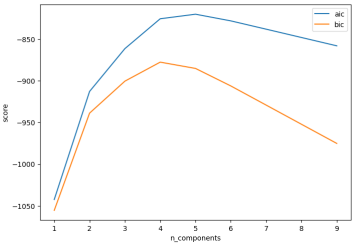
(a) $n = 10, m = 3$



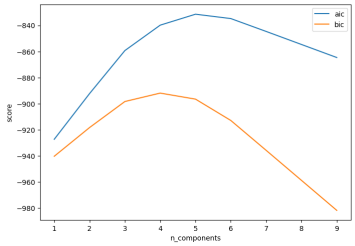
(b) $n = 10, m = 4$

图 1: Correct Result

增大因子数 m 时，出现一些错误情况，如图 2 所示，$m = 5$ 时，AIC 结果正确，但 BIC 似乎对高维组成部分过度惩罚，因此选择了 4 个组成部分。



(a) $n = 10, m = 5$

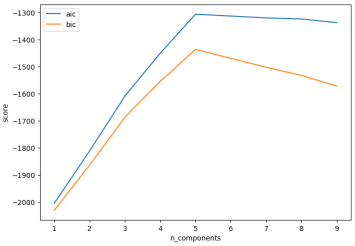


(b) $n = 10, m = 6$

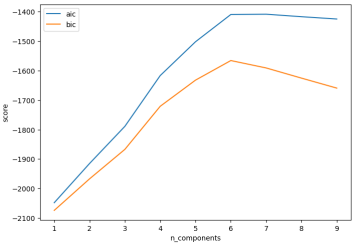
图 2: Wrong Result

.

这可能是因为原始样本因子的维度很低，将 n 修改为 20，这个问题得到了解决，如图 3 所示。此时，AIC 和 BIC 都取得了良好的结果。



(a) $n = 20, m = 5$



(b) $n = 20, m = 6$

图 3: Fixed Result of m=5,6