

非監督式學習參考手冊

Afshine AMIDI 和 Shervine AMIDI

December 15, 2019

翻譯: kevingo. 審閱: imironhead, 徐承漢.

非監督式學習介紹

□ **動機** – 非監督式學習的目的是要找出未標籤資料 $\{x^{(1)}, \dots, x^{(m)}\}$ 之間的隱藏模式

□ **Jensen's 不等式** – 令 f 為一個凸函數、 X 為一個隨機變數，我們可以得到底下這個不等式：

$$E[f(X)] \geq f(E[X])$$

最大期望值

□ **潛在變數(Latent variables)** – 潛在變數指的是隱藏/沒有觀察到的變數，這會讓問題的估計變得困難，我們通常使用 z 來代表它。底下是潛在變數的常見設定

設定	潛在變數 z	$x z$	評論
k 元高斯模型	$\text{Multinomial}(\phi)$	$\mathcal{N}(\mu_j, \Sigma_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
因素分析	$\mathcal{N}(0, I)$	$\mathcal{N}(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

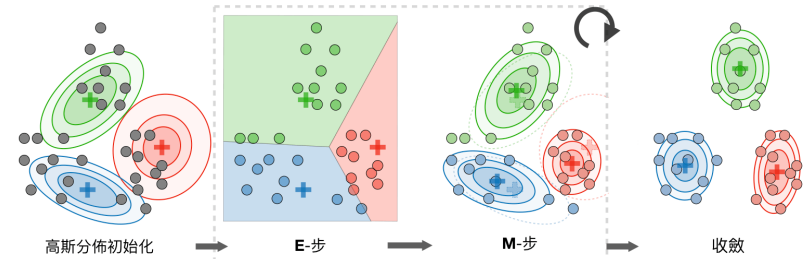
□ **演算法** – 最大期望演算法(EM Algorithm) 透過重複建構一個概似函數的下界(E-step) 和最佳化下界(M-step) 來進行最大概似估計給出參數 θ 的高效率估計方法：

- E-step: 評估後驗機率 $Q_i(z^{(i)})$ ，其中每個資料點 $x^{(i)}$ 來自於一個特定的群集 $z^{(i)}$ ，如下：

$$Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}; \theta)$$

- M-step: 使用後驗機率 $Q_i(z^{(i)})$ 作為資料點 $x^{(i)}$ 在群集中特定的權重，用來分別重新估計每個群集，如下：

$$\theta_i = \operatorname{argmax}_{\theta} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \left(\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) dz^{(i)}$$



k -平均算法分群法

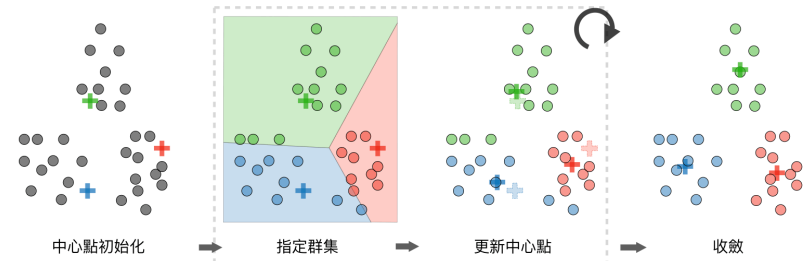
我們使用 $c^{(i)}$ 表示資料 i 屬於某群，而 μ_j 則是群 j 的中心

□ **演算法** – 在隨機初始化群集中心點 $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ 後， k -平均算法演算法重複以下步驟直到收斂：

$$c^{(i)} = \operatorname{argmin}_j \|x^{(i)} - \mu_j\|^2$$

和

$$\mu_j = \frac{\sum_{i=1}^m 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m 1_{\{c^{(i)}=j\}}}$$



□ **畸變函數** – 為了確認演算法是否收斂，我們定義以下的畸變函數：

$$J(c, \mu) = \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

階層式分群法

□ **演算法** – 階層式分群法是透過一種階層架構的方式，將資料建立為一種連續層狀結構的形式。

□ **類型** – 底下是幾種不同類型的階層式分群法，差別在於要最佳化的目標函式的不同，請參考底下：

Ward 鏈結距離	平均鏈結距離	完整鏈結距離
最小化群內距離	最小化各群彼此的平均距離	最小化各群彼此的最大距離

分群衡量指標

在非監督式學習中，通常很難去評估一個模型的好壞，因為我們沒有擁有像在監督式學習任務中正確答案的標籤

□ **輪廓係數(Silhouette coefficient)** – 我們指定 a 為一個樣本點和相同群集中其他資料點的平均距離、 b 為一個樣本點和下一個最接近群集其他資料點的平均距離，輪廓係數 s 對於此一樣本點的定義為：

$$s = \frac{b - a}{\max(a, b)}$$

□ **Calinski-Harabaz 指標** – 定義 k 是群集的數量， B_k 和 W_k 分別是群內和群集之間的離差矩陣(dispersion matrices)：

$$B_k = \sum_{j=1}^k n_{c(i)} (\mu_{c(i)} - \mu)(\mu_{c(i)} - \mu)^T, \quad W_k = \sum_{i=1}^m (x^{(i)} - \mu_{c(i)})(x^{(i)} - \mu_{c(i)})^T$$

Calinski-Harabaz 指標 $s(k)$ 指出分群模型的好壞，此指標的值越高，代表分群模型的表現越好。定義如下：

$$s(k) = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{N - k}{k - 1}$$

主成份分析

這是一個維度縮減的技巧，在於找到投影資料的最大方差

□ **特徵值、特徵向量** – 給定一個矩陣 $A \in \mathbb{R}^{n \times n}$ ，我們說 λ 是 A 的特徵值，當存在一個特徵向量 $z \in \mathbb{R}^n \setminus \{0\}$ ，使得：

$$Az = \lambda z$$

□ **譜定理** – 令 $A \in \mathbb{R}^{n \times n}$ ，如果 A 是對稱的，則 A 可以透過正交矩陣 $U \in \mathbb{R}^{n \times n}$ 對角化。當 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ ，我們得到：

$$\exists \Lambda \text{ 對角線, } A = U\Lambda U^T$$

注意：與特徵值所關聯的特徵向量就是 A 矩陣的主特徵向量

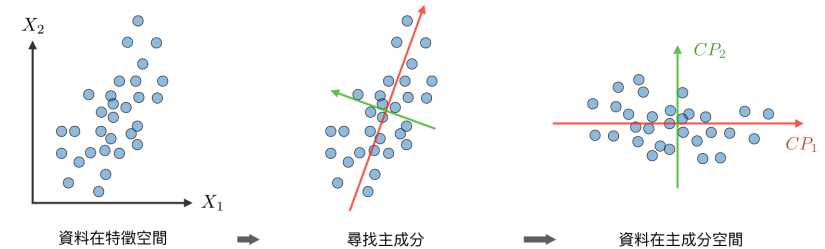
□ **演算法** – 主成份分析(PCA) 是一種維度縮減的技巧，它會透過尋找資料最大變異的方式，將資料投影在 k 維空間上：

- 第一步: 正規化資料，讓資料平均為0，變異數為1

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \quad \text{哪裡} \quad \mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)} \quad \text{和} \quad \sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- 第二步: 計算 $\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \in \mathbb{R}^{n \times n}$ ，即對稱實際特徵值

- 第三步: 計算 $u_1, \dots, u_k \in \mathbb{R}^n$ ， k 個正交主特徵向量的總和 Σ ，即是 k 個最大特徵值的正交特徵向量
- 第四部: 將資料投影到 $\text{span}_{\mathbb{R}}(u_1, \dots, u_k)$



獨立成分分析

這是用來尋找潛在生成來源的技巧

□ **假設** – 我們假設資料 x 是從 n 維的來源向量 $s = (s_1, \dots, s_n)$ 產生， s_i 為獨立變數，透過一個混合與非奇異矩陣 A 產生如下：

$$x = As$$

目的在於找到一個unmixing 矩陣 $W = A^{-1}$

□ **Bell 和 Sejnowski 獨立成份分析演算法** – 此演算法透過以下步驟來找到unmixing 矩陣：

- 紀錄 $x = As = W^{-1}s$ 的機率如下：

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

- 在給定訓練資料 $\{x^{(i)}, i \in [1, m]\}$ 的情況下，其對數概似估計函數與定義 g 為sigmoid 函數如下：

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

因此，梯度隨機下降學習規則對每個訓練樣本 $x^{(i)}$ 來說，我們透過以下方法來更新 W ：

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