Vũ Quang Minh
18110150

Bài tập Xử lý đa chiều

Câu a

Câu b

IV t-Distribution Stochastic Neighbor Embedding (t-SNE)

t-SNE là một công cụ giảm chiều không tuyến tính, thích hợp cho việc trực quan hoá dữ liệu đa chiều. Nó được dùng trong xử lí ảnh, NLP, giọng nói.

Đầu tiên xét ma trận $N \times N$ $\mathbf{P}$ trong không gian nhiều chiều có các phần tử

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

với xác xuất để x_j là neighbor của x_i

$$p_{j|i} = \frac{e^{-\|x_i - x_j\|^2/(2\sigma^2)}}{\sum_{k \neq i} e^{-\|x_i - x_k\|^2/(2\sigma^2)}}$$

Ma trận $\mathbf{Q}$ $N \times N$ trong không gian thấp chiều hơn ban đầu với các phần tử có phân phối t-Student

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_i - y_l\|^2)^{-1}}$$

Hàm đánh giá sẽ là

$$C = \sum_i \mathrm{KL}(\mathbf{P}_i|\mathbf{Q}_i) = \sum_{i=1}^{N}\sum_{i=1}^{N} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

Gradient của hàm đánh giá này là:

$$\begin{aligned}
\frac{\delta C}{\delta y_i} &= 4 \sum_{j=1,j}^{N} (p_{ij} - q_{ij})(1 + \|y_i - y_j\|^2)^{-1}(y_i - y_j) \\
&= 4 \sum_{j=1,j}^{N} (p_{ij} - q_{ij}) q_{ij} Z (y_i - y_j) \\
&= 4 \left[\sum_{j \neq i}^{N} p_{ij} q_{ij} Z (y_i - y_j) - \sum_{j \neq i}^{N} q_{ij}^2 Z (y_i - y_j) \right] \\
&= 4(F_{\text{attraction}} + F_{\text{repulsion}})
\end{aligned}$$

với $Z = \sum_{l,s=1,l\neq s}^{N} (1 + \|y_l - y_s\|^2)^{-1}$

Thuật toán t-SNE:
Input: Tập dữ liệu $X = x_1, \ldots, x_n \in \mathbb{R}^d$, perplexity k, exaggeration parameter α, kích cỡ bước nhảy $h > 0$, số lượng $T \in \mathbb{N}$
Tính $\{p_{ij} | i, j \in [n], i \neq j\}$
Khởi tạo $y_1^0, y_2^0, \ldots, y_n^0$ i.i.d từ phân phối uniform trên $[-0.01, 0.01]^2$
for $t = 0$ to $T - 1$ do

$$Z^{(t)} \leftarrow \sum_{i,j \in [n], i \neq j} \left(1 + \left\| y_i^{(t)} - y_j^{(t)} \right\|^2\right)^{-1}$$

$$q_{ij}^{(t)} \leftarrow \frac{\left(1 + \left\| y_i^{(t)} - y_j^{(t)} \right\|^2\right)^{-1}}{Z^{(t)}}, \forall i, j \in [n], i \neq j$$

$$y_i^{(t)} \leftarrow y_i^{(t)} + h \sum_{j \in [n]/\{i\}} \left(\alpha p_{ij} - q_{ij}^t\right) q_{ij}^t Z^t \left(y_i^{(t)} - y_j^{(t)}\right), \forall i \in [n]$$

Output: dữ liệu có số chiều thấp hơn ban đầu $Y^{(T)} = \left\{y_1^{(T)}, y_2^{(T)}, \ldots, y_n^{(T)}\right\} \in \mathbb{R}^2$

Các giả định trong thuật toán t-SNE:
* t-SNE sẽ không hoạt động tốt trong vấn đề đa chiều tổng quát, khi mà lớn hơn 2D hay 3D nhưng trung bình khoảng cách giữa các điểm cần được giữ nguyên giống như cấu trúc tổng quát
* Curse of Dimensionality
* $O(n^2)$ computational complexity
* Perplexit number, số lượng lặp, giá trị của tham số exaggeration phải được chọn thủ công.

Sự so sánh giữa PCA và t-SNE được cho trong bảng 1.