



ТЕХНОСФЕРА

Лекция 4 Визуализация результатов кластеризации

Николай Анохин

20 марта 2015 г.

Краткое содержание предыдущих лекций

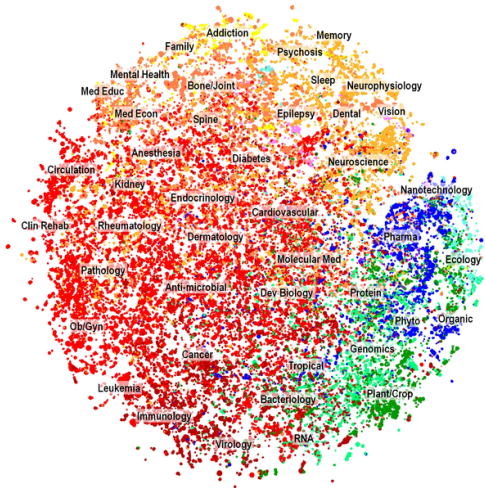
Дано. N обучающих D -мерных объектов $\mathbf{x}_i \in \mathcal{X}$, образующих тренировочный набор данных (training data set) X .

Найти. Модель $h^*(\mathbf{x})$ из семейства параметрических функций $H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \rightarrow \mathbb{N}\}$, ставящую в соответствие произвольному $\mathbf{x} \in \mathcal{X}$ один из K кластеров так, чтобы объекты внутри одного кластера были похожи, а объекты из разных кластеров различались.

Краткое содержание предыдущей лекции

Рассмотрели классические алгоритмы кластеризации

1. Смесь гауссовских распределений и k-means
2. Hierarchical Clustering
3. DBSCAN



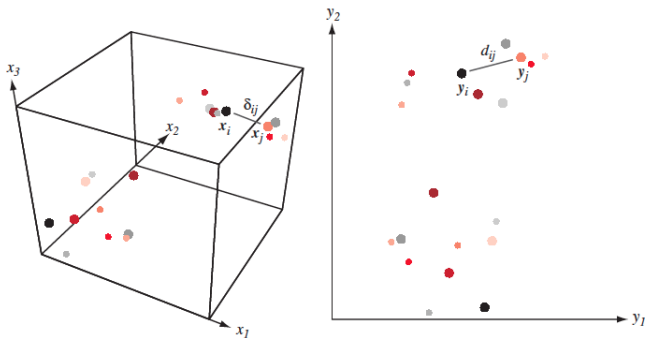
Multidimensional Scaling

Идея метода

Перейти в пространство меньшей размерности так, чтобы расстояния между объектами в новом пространстве были подобны расстояниям в исходном пространстве.

Обозначения

- ▶ $\mathbf{x}_i \in \mathcal{X} \subset R^D$ – объекты в исходном многомерном пространстве
- ▶ δ_{ij} – расстояние между \mathbf{x}_i и \mathbf{x}_j
- ▶ $\mathbf{x}_i \in \mathcal{Y} \subset R^E$ – объекты в целевом пространстве ($E = 2$ или $E = 3$)
- ▶ d_{ij} – расстояние между \mathbf{y}_i и \mathbf{y}_j



Критерии

Выбираем конфигурацию \mathbf{y}_i , соответствующую минимуму критерия

$$J_{ee} = \frac{\sum_{i < j} (d_{ij} - \delta_{ij})^2}{\sum_{i < j} \delta_{ij}^2}$$

$$J_{ff} = \sum_{i < j} \frac{(d_{ij} - \delta_{ij})^2}{\delta_{ij}^2}$$

$$J_{ef} = \frac{1}{\sum_{i < j} \delta_{ij}} \sum_{i < j} \frac{(d_{ij} - \delta_{ij})^2}{\delta_{ij}}$$

Градиентный спуск

Требуется найти минимум функции $f(\mathbf{a})$, при этом

1. мы умеем вычислять градиент функции $\nabla f(\mathbf{a})$
2. задана начальная точка \mathbf{a}_0
3. выбрана функция learning rate $\eta(k)$

```
1 function gd(grad, a0, epsilon):  
2     initialise eta(k)  
3     k = 0  
4     a = a0  
5     do:  
6         k = k + 1  
7         a = a - eta(k) grad(a)  
8     until |eta(k) grad(a)| < epsilon  
9     return a
```

(демо)

Градиенты критериев

$$\nabla_{\mathbf{y}_k} J_{ee} = \frac{2}{\sum_{i < j} \delta_{ij}^2} \sum_{j \neq k} (d_{kj} - \delta_{kj}) \frac{\mathbf{y}_k - \mathbf{y}_j}{d_{kj}}$$

$$\nabla_{\mathbf{y}_k} J_{ff} = 2 \sum_{j \neq k} \frac{d_{kj} - \delta_{kj}}{\delta_{kj}^2} \frac{\mathbf{y}_k - \mathbf{y}_j}{d_{kj}}$$

$$\nabla_{\mathbf{y}_k} J_{ef} = \frac{2}{\sum_{i < j} \delta_{ij}} \sum_{j \neq k} \frac{d_{kj} - \delta_{kj}}{\delta_{kj}} \frac{\mathbf{y}_k - \mathbf{y}_j}{d_{kj}}$$

Результаты применения

T-SNE

Идея метода

Та же, что в MDS, но определяется необычная (вероятностная) схожесть между объектами в исходном и целевом пространствах, а также критерий оптимизации.

Схожесть между объектами x_i и $x_j \sim$ вероятность того, что x_i “выберет” x_j из остальных соседей, будучи центром некоторого нормального распределения.

Схожесть между объектами

В исходном пространстве

$$p(j|i) = \frac{\exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_k - \mathbf{x}_i\|^2 / 2\sigma_i^2)}$$

В целевом пространстве

$$q(j|i) = \frac{\exp(-\|\mathbf{y}_j - \mathbf{y}_i\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_k - \mathbf{y}_i\|^2)}$$

Критерий оптимизации

Дивергенция Кульбака-Лейблера

Насколько распределение P отличается от распределения Q ?

$$KL(P\|Q) = \sum_i P(z) \log \frac{P(z)}{Q(z)}$$

Критерий

$$J_{SNE} = \sum_i KL(P_i\|Q_i) = \sum_i \sum_j p(j|i) \log \frac{p(j|i)}{q(j|i)} \rightarrow \min_{\mathbf{y}_1, \dots, \mathbf{y}_n}$$

Градиент

$$\nabla_{\mathbf{y}_i} J_{SNE} = 2 \sum_j (p(j|i) - q(j|i) + p(i|j) - q(i|j)) (\mathbf{y}_i - \mathbf{y}_j)$$

Параметры алгоритма

Идея

В областях высокой плотности выбрать σ_i маленьким, а в областях низкой плотности – большим.

$$Perp(P_i) = 2^{H(P_i)}, \quad H(P_i) = - \sum_j p(j|i) \log p(j|i)$$

На практике выбираем фиксированное perplexity в интервале (5, 50).

T-SNE

Недостатки SNE

- ▶ Трудно оптимизировать критерий
- ▶ “Crowding problem”

Отличия T-SNE от SNE

- ▶ Использует симметризованный критерий с более простым градиентом
- ▶ В целевом пространстве схожесть основана на t-распределении, а не на распределении Гаусса

Вопросы

