ТЕХНОСФЕРА

Лекция 4
Различные аспекты кластеризации

Стройкова Ксения

10 октября 2016 г.

Краткое содержание предыдущих лекций

Дано. Признаковые описания N объектов $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}$, образующие тренировочный набор данных X

Найти. Модель из семейства параметрических функций

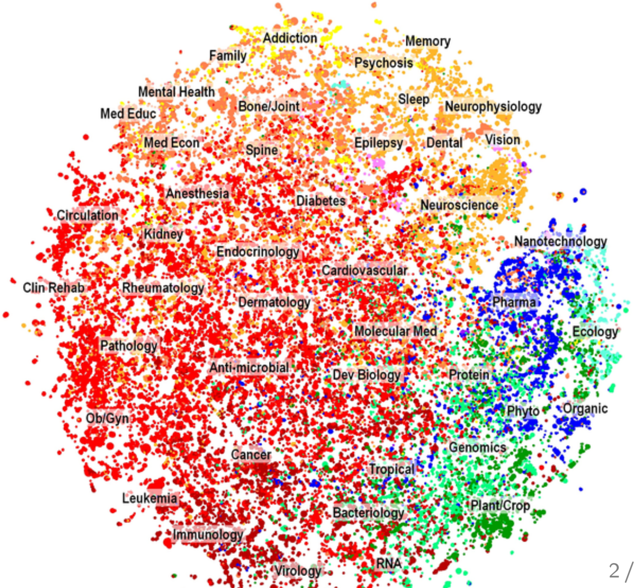
$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathcal{Y} \mid \mathcal{Y} = \{1, \ldots, K\}\},$$

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

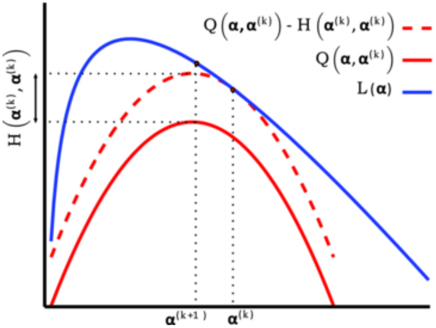
Краткое содержание предыдущих лекций

Рассмотрели классические алгоритмы кластеризации

1. Hierarchical Clustering
2. dbscan, OPTICS
3. Смесь гауссовских распределений и k-means



Байесовская кластеризация + EM



Expectation Maximization

Дано.
Известно распределение $P(\mathbf{X}, \mathbf{Z}|\theta)$, где $\mathbf{x}$ – наблюдаемые переменные, а $\mathbf{z}$ – скрытые.
Найти.
θ, максимизирующее $P(\mathbf{X}|\theta)$.

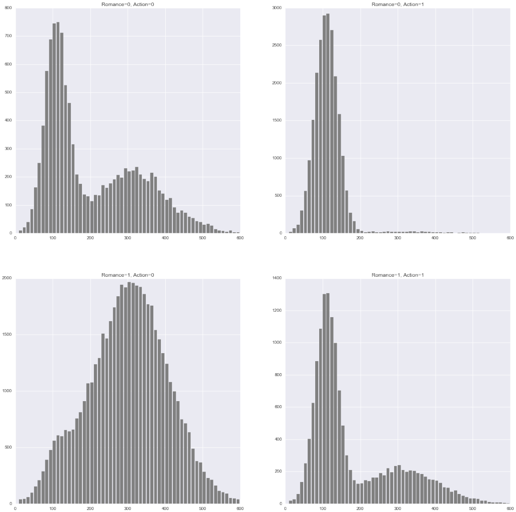
- E вычислить $P(\mathbf{Z}|\mathbf{X}, \theta^{old})$ при фиксированном θ^{old}
- M вычислить $\theta^{new} = \arg\max_\theta \mathcal{Q}(\theta, \theta^{old})$, где

$$\mathcal{Q}(\theta, \theta^{old}) = E_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta))$$

Кластеризация пользователей стримингового сервиса

	romance	action	avg. price
0	1	0	264.563366
1	1	1	100.852569
2	1	0	337.576899
3	0	1	105.545184
4	1	0	430.988385
5	1	0	284.593125
6	0	1	58.789076
7	0	1	116.824524
8	1	0	317.829967
9	1	1	146.660413

Для каждого пользователя известно, есть ли у него/нее интерес к романтике, экшену и средняя цена купленных фильмов



Romance=0, Action=0
Romance=0, Action=1
Romance=1, Action=0
Romance=1, Action=1

Предположения модели

Априорное распределение

$$p(C_k) = \pi_k$$

Распределение интересов

$$p(I_i|C_k) \sim Ber(P_{ki})$$

Распределение средней цены фильма

$$p(x|C_k) \sim \mathcal{N}(x|\mu_k, \sigma_k)$$

Итерации EM

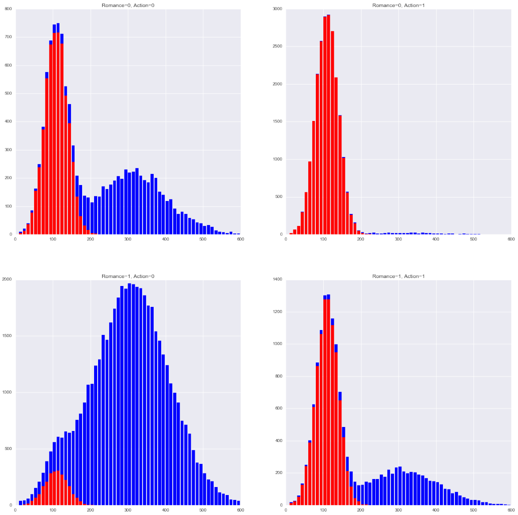
E

$$\gamma_{nk} = p(z_n = k | u_n, \pi, P, \mu, \sigma) = \frac{\pi_k \mathcal{N}(x_n | \mu_k, \sigma_k) \prod_{i=1}^{2} P_{ki}^{I_{ni}} (1 - P_{ki})^{1 - I_{ni}}}{\sum_j \pi_j \mathcal{N}(x_n | \mu_j, \sigma_j) \prod_{i=1}^{2} P_{ji}^{I_{ni}} (1 - P_{ji})^{1 - I_{ni}}}$$

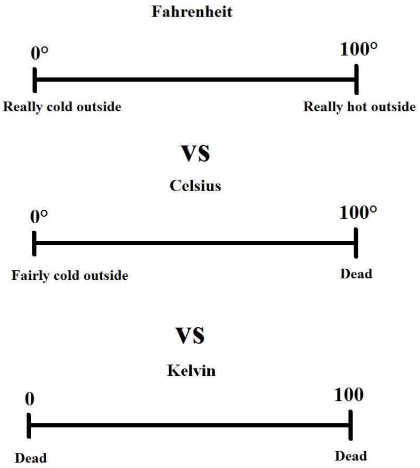
M

$$N_k = \sum_{n=1}^{N} \gamma_{nk}, \qquad \pi_k = \frac{N_k}{N}$$

$$P_{ki} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} I_{ni}, \quad \mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x_n, \quad \sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k)^2$$


Romance=0, Action=0
Romance=0, Action=1
Romance=1, Action=0
Romance=1, Action=1

Функции расстояния



Модификации алгоритма

Взять уже известную нам функцию потерь (инерцию) и “поиграть” с функцией расстояния.

$$\tilde{J}(\mu) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} d(\mathbf{x}_n, \mu_k), \quad r_{nk} = \begin{cases} 1, \text{ для } k = \arg\min_j d(\mathbf{x}_n, \mu_j) \\ 0, \text{ иначе} \end{cases}$$

Расстояния 1

- Минковского

$$d_r(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{N} |x_j - y_j|^r \right]^{\frac{1}{r}}$$

- Евклидово $r = 2$

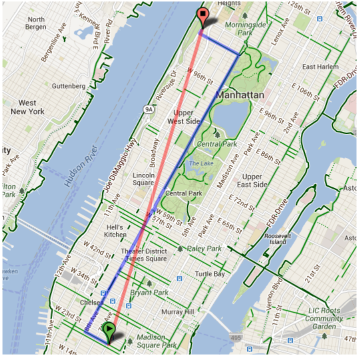
$$d_E(\mathbf{x}, \mathbf{y}) = d_2(\mathbf{x}, \mathbf{y})$$

- Манхэттэн $r = 1$

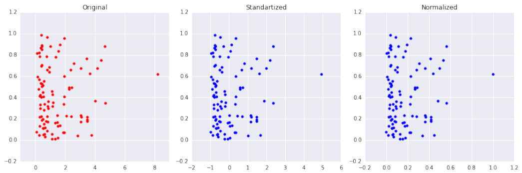
$$d_M(\mathbf{x}, \mathbf{y}) = d_1(\mathbf{x}, \mathbf{y})$$

- $r = \infty$

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_j |x_j - y_j|$$

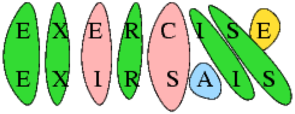


1. Функции расстояния чувствительны к "масштабу" данных
    - Преобразовать обучающую выборку так, чтобы признаки имели нулевое среднее и единичную дисперсию (standartization)
    - Преобразовать обучающую выборку так, чтобы значения признаков лежали на отрезке $[0, 1]$ (normalization)



2. Есть шанс улучшить качество, применив монотонное преобразование ($\log$, $\sqrt{}$)

Расстояния 2



- Жаккар

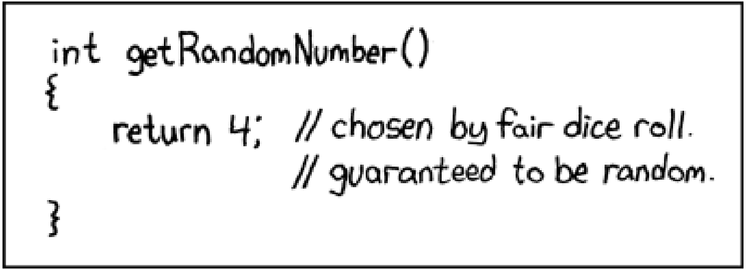
$$d_J(\mathbf{x}, \mathbf{y}) = 1 - \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}$$

- Косинус

$$d_c(\mathbf{x}, \mathbf{y}) = \arccos \frac{\mathbf{x}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

- Правки
  d_e – наименьшее количество удалений и вставок, приводящее $\mathbf{x}$ *к* $\mathbf{y}$.
- Хэмминг
  d_H – количество различных компонент в $\mathbf{x}$ *и* $\mathbf{y}$.

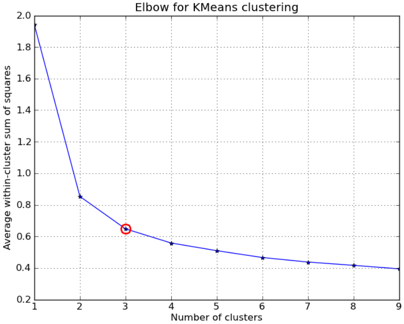
Выбор количества кластеров



Выбор наилучшего K

Идея. Выбрать критерий качества кластеризации и построить его значение для $K = 1, 2, \ldots$

- средняя сумма квадратов расстояния до центроида
- средний диаметр кластера



Критерий Silhouette

Пусть дана кластеризация в K кластеров, и объект i попал в C_k

- $a(i)$ – среднее расстояние от i объекта до объектов из C_k
- $b(i) = min_{j \neq k} b_j(i)$, где $b_j(i)$ – среднее расстояние от i объекта до объектов из C_j

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

Средний silhouette для всех точек из $\mathbf{X}$ является критерием качества кластеризации.

Качество кластеризации

Качество кластеризации[1]

Пусть дана обучающая выборка, для которой правильная кластеризация C известна. С помощью выбранного алгоритма получена кластеризация K. Проверить, насколько K совпадает с C.

- Adjusted Rand Index
  a – кол-во пар объектов, попавших в один кластер и в C, и в K
  b – кол-во пар объектов, попавших в разные кластеры и в C, и в K

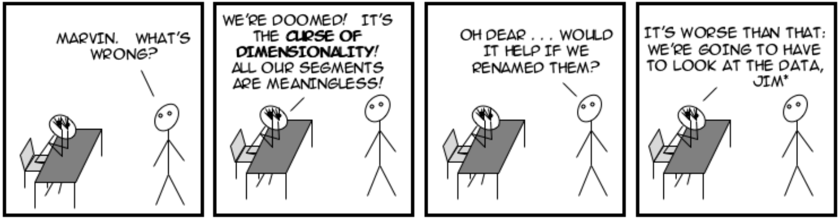
$$RI = \frac{a+b}{C_2^N}, \quad ARI = \frac{RI - E_{rdm}[RI]}{\max(RI) - E_{rdm}[RI]}$$

- Mutual Information

$$MI = \sum_{c \in C} \sum_{k \in K} p(c,k) \log \frac{p(c,k)}{p(k)p(c)}$$

[1] scikit-learn docs

Multidimensional Scaling



HTTP://SCIENTIFICMARKETER.COM

* WITH APOLOGIES TO MR SPOCK & STAR TREK.

Идея метода

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

t-Stochastic Neighbour Embedding (t-SNE)[2]

Схожесть между объектами в исходном пространстве

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

Схожесть между объектами в целевом пространстве

$$q(i,j) = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

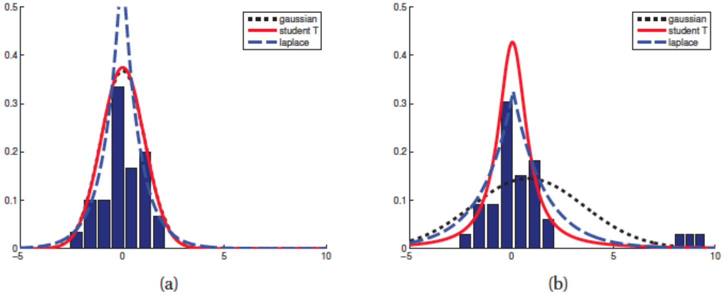
Критерий

$$J_{t-SNE} = KL(P\|Q) = \sum_i \sum_j p(i,j)\log\frac{p(i,j)}{q(i,j)} \to \min$$

[2] `http://lvdmaaten.github.io/tsne/`

t-распределение

$$\tau(\mu, \sigma^2, \nu) \propto \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}$$



Уильям Госсет 1908 (Student)

Дивергенция Кульбака-Лейблера[3]

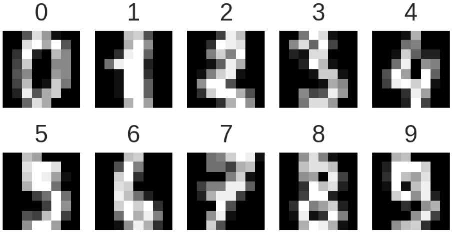
Насколько распределение P отличается от распределения Q?

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

[3] Visual Information Theory

Digits Dataset

около 1800 картинок 8x8 с рукописными цифрами



t-SNE

MNIST Dataset

70000 картинок 20x20 с рукописными цифрами

t-SNE

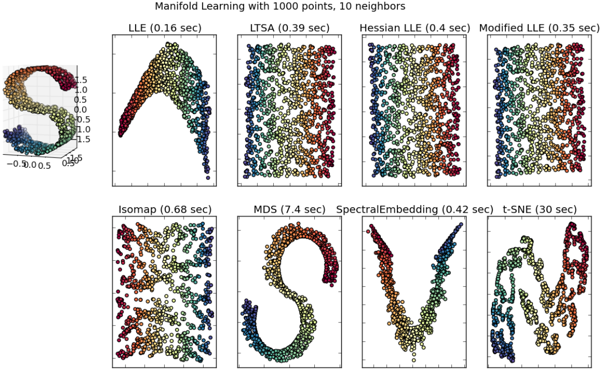
Еще примеры

CalTech

S&P 500

Words

Еще алгоритмы



Вопросы