Лекция 1
Задачи Data Mining

Николай Анохин

1 октября 2014 г.

План занятия

Задача кластеризации

Смесь нормальных распределений и EM

K-means и его модификации

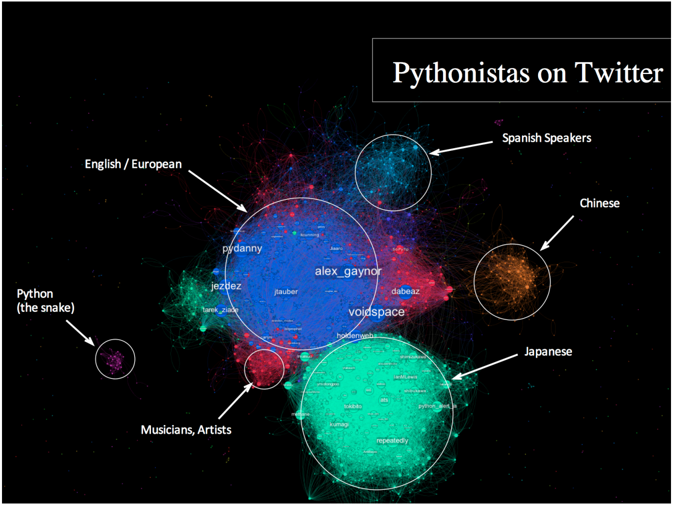
Задача кластеризации

В задачах кластеризации целевая переменная не задана. Цель – отыскать “скрытую структуру” данных.

Зачем вообще рассматривать задачи без целевой переменной?

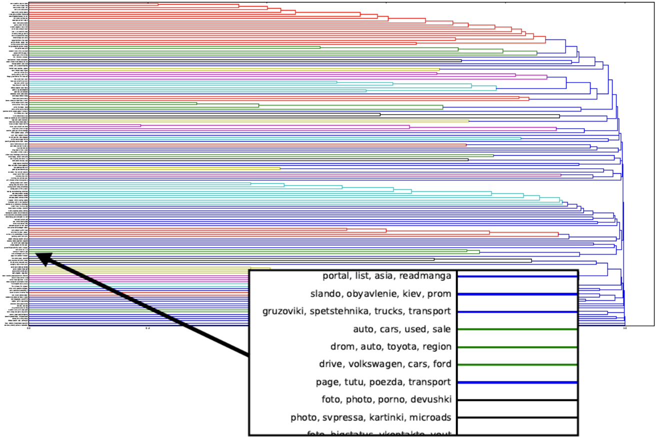
1. разметка данных – дорогое удовольствие
2. можно сначала поделить, а потом разметить
3. возможность отслеживать эволюционные изменения
4. построение признаков
5. exploratory data analysis

Программисты python в Twitter



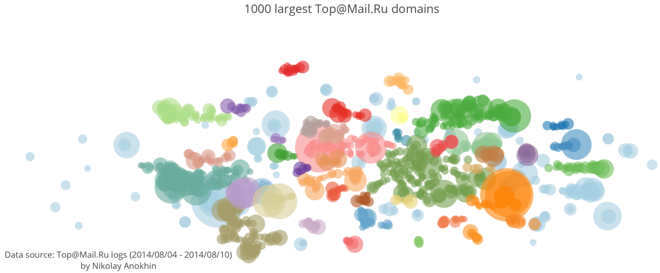
Графо-теоретические методы (источник)

Похожие тематики



Иерархическая кластеризия

Топ 1000 самых посещаемых доменов рунета



T-SNE + DBSCAN

Постановка задачи

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 $H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathbb{N}\}$, 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.

- Как определить похожесть объектов?
- Как оценить качество модели?
- Как выбрать K?

Многомерное нормальное распределение

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

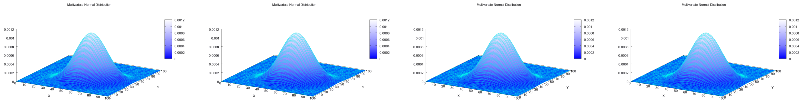
Параметры

D-мерный вектор средних

$$\mu = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$D \times D$-мерная матрица ковариации

$$\mathbf{\Sigma} = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]$$



(a) $D = 2$ (b) (c) $\mathbf{\Sigma} = \mathrm{diag}(\sigma_i)$ (d) $\mathbf{\Sigma} = \sigma \mathbf{I}$

Смесь нормальных распределений

"Скрытая" K-мерная переменная $\mathbf{z}$ – принадлежность объекта к одному из кластеров

$$p(z_k = 1) = \pi_k, \quad z_k \in \{0, 1\}, \quad \sum_k z_k = 1 \quad \rightarrow \quad p(\mathbf{z}) = \prod_k \pi_k^{z_k}$$

Распределение $\mathbf{x}$ для каждого из K кластеров

$$p(\mathbf{x}|\mathbf{z_k}) = \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \quad \rightarrow \quad p(\mathbf{x}|\mathbf{z}) = \prod_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)^{z_k}$$

Смесь нормальных распределений

$$p(\mathbf{x}) = \sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

Апостериорная вероятность принадлежности к k кластеру (априорная равна π_k)

$$\gamma(z_k) = p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} =$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \mathbf{\Sigma}_j)}$$

Maximum Likelihood (!)

ML принцип

Сформулировать принцип тут

Функция правдоподобия

$$\log(\mathbf{X}|\pi, \mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \log \sum_{k} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \to \max_{\pi, \mu, \mathbf{\Sigma}}$$

Сложности

- схлопывание компонент
- переименование кластеров
- невозможно оптимизировать аналитически

Дифференцируем функцию правдоподобия

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}), \quad \mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)^T (\mathbf{x}_n - \mu_k)$$

$$\pi_k = \frac{N_k}{N}$$

Expectation Maximization (!)

E Expectation: при фиксированных μ_k, Σ_k, π_k

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}$$

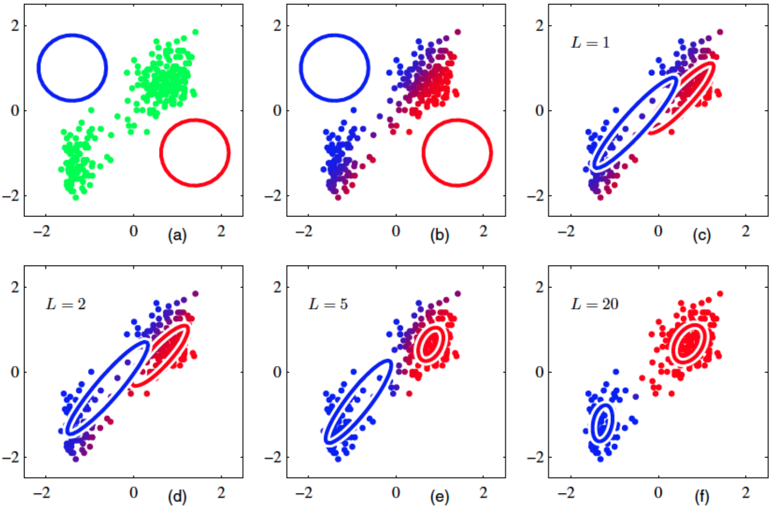
M Maximization: при фиксированных $\gamma(z_{nk})$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}), \quad \mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

S Остановиться при достижении сходимости


L = 1
L = 2
L = 5
L = 20
(a)
(b)
(c)
(d)
(e)
(f)

EM-алгоритм

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

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

Улучшение: ввести априорное распределение $p(\theta)$

K-means

Пусть $\Sigma_k = \epsilon I$, тогда

$$p(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi\epsilon}} \exp(-\frac{1}{2\epsilon}\|\mathbf{x} - \mu_\mathbf{k}\|^2)$$

Рассмотрим стремление $\epsilon \to 0$

$$\gamma(z_{nk}) = \frac{\pi_k \exp(-\frac{1}{2\epsilon}\|\mathbf{x}_n - \mu_\mathbf{k}\|^2)}{\sum_j \pi_j \exp(-\frac{1}{2\epsilon}\|\mathbf{x}_n - \mu_\mathbf{j}\|^2)} \to r_{nk} = \begin{cases} 1, \text{ для } k = \arg\min_j \|\mathbf{x}_n - \mu_\mathbf{j}\|^2 \\ 0, \text{ иначе} \end{cases}$$

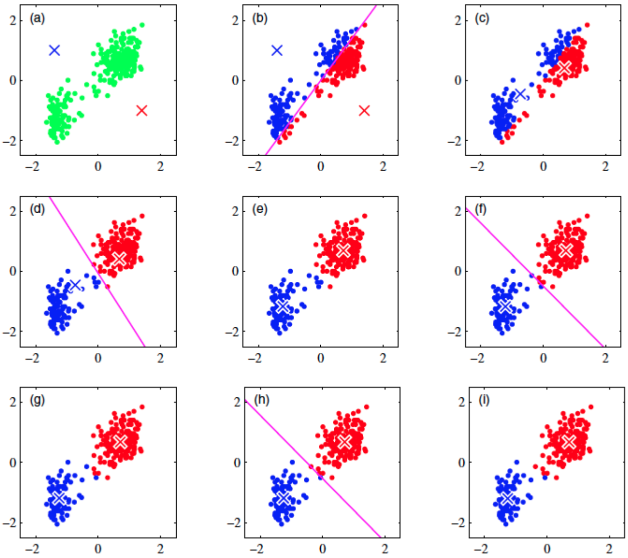
Функция правдоподобия

$$E_\mathbf{Z}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)] \to -\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \mu_k\|^2 + const$$

Вектор средних

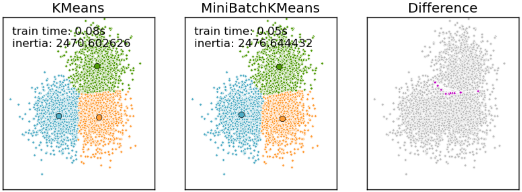
$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x_n}}{\sum_n r_{nk}}$$

K-means


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Модификации k-means

- На каждом шаге работаем с b случайно выбранными объектами из каждого кластера (mini-batch k-means)



- Критерий качества (k-medoids)

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} d(\mathbf{x}_n, \mu_k)$$

d – функция расстояния, μ_k – один из объектов в кластере

Альтернативные функции расстояния

Def

Функция $d(\mathbf{x}, \mathbf{y}) : \mathbf{X} \times \mathbf{X} \to R$ является функцией расстояния, определенной на пространстве $\mathbf{X}$ тогда и только тогда, когда $\forall \mathbf{x} \in \mathbf{X}$, $\forall \mathbf{y} \in \mathbf{X}$, $\forall \mathbf{z} \in \mathbf{X}$ выполнено:

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$
2. $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$
3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
4. $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})$

Расстояния 1

- Минковского

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

- Евклидово $r = 2$

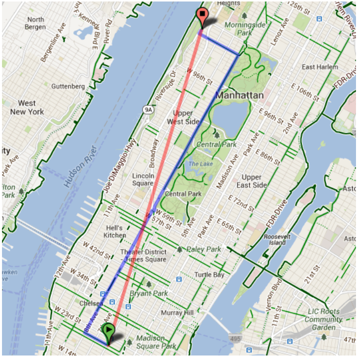
$$d_E(\mathbf{x}, \mathbf{y}) = d_2(\mathbf{x}, \mathbf{y})$$

- Манхэттэн $r = 1$

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

- $r = \infty$

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



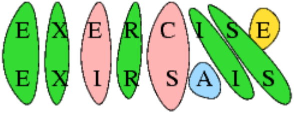
Проблема

Функции расстояния чувствительны к преобразоаниям данных

Решение

- Преобразовать обучающую выборку так, чтобы признаки имели нулевое среднее и единичную дисперсию – инвариантность к растяжению и сдвигу (stanartize)
- Преобразовать обучающую выборку так, чтобы оси совпадали с главными компонентами матрицы ковариации – инвариантность относительно поворотов (PCA)

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* $\mathbf{y}$*.*

Проклятие размерности

Задача

Даны два случайных вектора $\mathbf{x}$ и $\mathbf{y}$ в пространстве размерности D. Как зависит математическое ожидание косинус-расстояния между $\mathbf{x}$ и $\mathbf{y}$ от размерности D?

$$d_c(\mathbf{x}, \mathbf{y}) = \arccos \frac{\sum_{j=1}^{D} x_j y_j}{\sum_{j=1}^{D} x_j^2 \sum_{j=1}^{D} y_j^2}$$

Наблюдения:

- числитель стремится к нулю
- знаменатель положительный

Вывод: $d_c(\mathbf{x}, \mathbf{y}) \to \frac{\pi}{2}$.

Альтернативные критерии качества

Критерий

$$
\begin{aligned}
J &= \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2 = \\
&= \frac{1}{2} \sum_{k=1}^{K} n_k \left[\frac{1}{n_k^2} \sum_{\mathbf{x}_i \in C_k} \sum_{\mathbf{x}_j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right] = \\
&= \frac{1}{2} \sum_{k=1}^{K} n_k \left[\frac{1}{n_k^2} \sum_{\mathbf{x}_i \in C_k} \sum_{\mathbf{x}_j \in C_k} s(\mathbf{x}_i, \mathbf{x}_j) \right] = \frac{1}{2} \sum_{k=1}^{K} n_k \bar{s}_k
\end{aligned}
$$

Примеры $\bar{s}_i$

$$
\underline{s}_k = \min_{\mathbf{x}_i, \mathbf{x}_j} s(\mathbf{x}_i, \mathbf{x}_j); \quad \bar{s}_k = \max_{\mathbf{x}_i, \mathbf{x}_j} s(\mathbf{x}_i, \mathbf{x}_j)
$$

Кластеризация

Идея

Выбрать критерий качества кластеризации J и расстояние между объектами $d(\mathbf{x}_i, \mathbf{x}_j)$ и вычислить разбиение выборки на кластеры, которое которое соответствует оптимальному значению выбранного критерия.

Вопросы