ТЕХНОСФЕРА

Лекция 2
K-means и EM-алгоритм

Николай Анохин

10 марта 2016 г.

План занятия

Смесь нормальных распределений и EM

K-means и его модификации

На предыдущей лекции

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

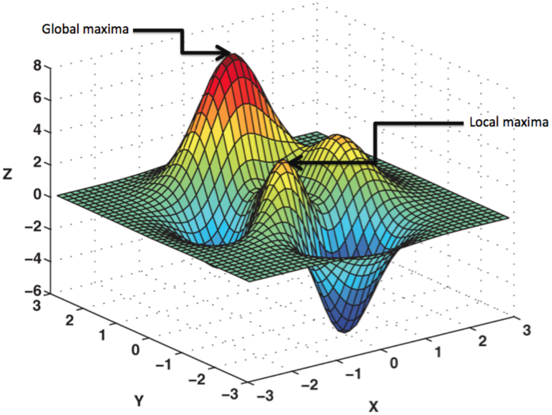
Найти. Модель из семейства параметрических функций

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

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

Алгоритмы. Hierarchical Clustering, dbscan, OPTICS

Смесь нормальных распределений и ЕМ



Гаусс, Карл Фридрих (1777-1855)

- Не открыл распределение Гаусса
- Открыл все остальное

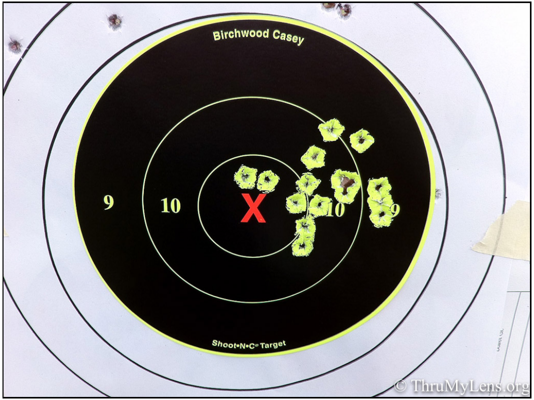
Начнем с простого (смоделируем один кластер)

Данные

Координаты точек попаданий по мишени из гауссовской пушки

Задача

Определить, куда смещен прицел



Многомерное нормальное распределение

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

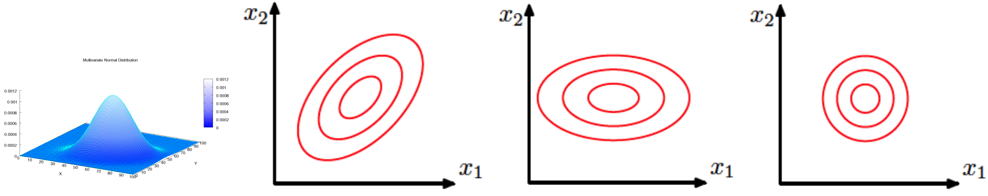
Параметры

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) $\mathbf{\Sigma} = \sigma I$

Формализуем задачу

Имеется набор данных

$$X = \{\mathbf{x}_n \in R^2\}$$

Предположение

$$p(\mathbf{x}_n) \sim \mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}), \quad \mu \in R^2, \;\; \boldsymbol{\Sigma} \in R^{2\times 2}$$

Требуется найти вектор средних μ и матрицу ковариации $\boldsymbol{\Sigma}$

Maximum likelihood (!)

Принцип максимального правдоподобия

Пусть дано семейство параметрических моделей $h(\mathbf{x}, \theta)$. Выбираем вектор параметров θ, максимизирующий функцию правдоподобия (likelihood) $p(\mathcal{D}|\theta)$, соответствующую рассматриваемому семейству моделей.

Правдоподобие

$$L(X|\mu, \mathbf{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x_n}|\mu, \mathbf{\Sigma}) \to \max_{\mu, \mathbf{\Sigma}}$$

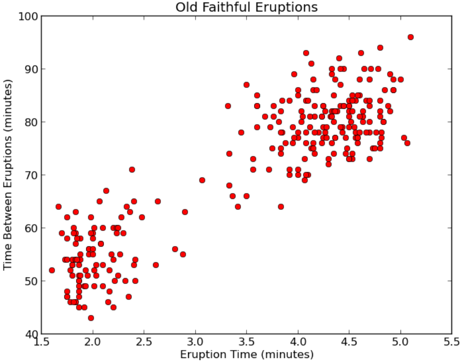
Решение

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x_n}, \quad \mathbf{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x_n} - \mu_{ML})(\mathbf{x_n} - \mu_{ML})^T$$

Old Faithful data set

D = *date of recordings in month (in August)*
X = *duration of the current eruption in minutes*
Y = *waiting time until the next eruption in minutes*

(a) Yellowstone Park



(b)

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

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

Смесь нормальных распределений

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

Апостериорная вероятность принадлежности к k кластеру (априорная равна π_k)

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

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

Maximum Likelihood

Функция правдоподобия

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

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

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

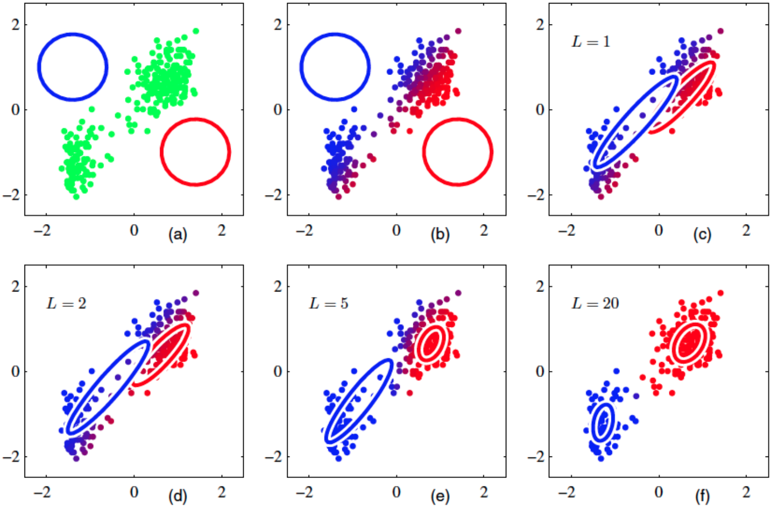
M Maximization: при фиксированных $\gamma(z_{nk})$

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

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

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

S Остановиться при достижении сходимости



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

EM-алгоритм[1]

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

E вычислить $P(\mathbf{Z}|\mathbf{X}, \theta^{old})$ при фиксированном θ^{old}

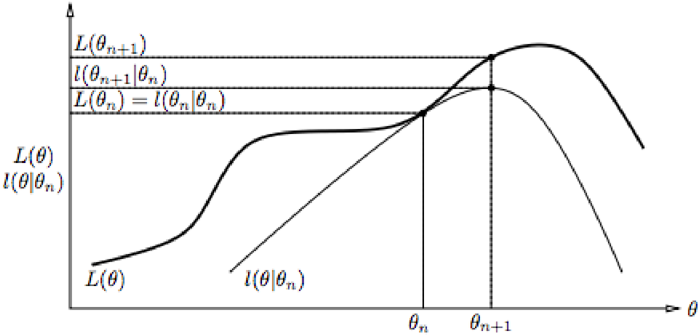
M вычислить $\theta^{new} = \arg\max_\theta \mathcal{Q}(\theta, \theta^{old})$, где

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

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

[1] Maximum Likelihood from Incomplete Data via the EM Algorithm

Сходимость[2]

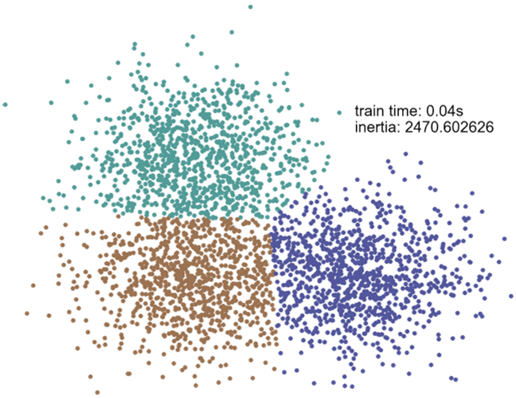


[2] The Expectation Maximization Algorithm: A short tutorial

Различные смеси

$p(\mathbf{x}_i\|\mathbf{z}_i)$	$p(\mathbf{z}_i)$	Name
MVN	Discrete	Mixture of Gaussians
Prod. Discrete	Discrete	Mixture of multinomials
Prod. Gaussian	Prod. Gaussian	Factor analysis/ probabilistic PCA
Prod. Gaussian	Prod. Laplace	Probabilistic ICA/ sparse coding
Prod. Discrete	Prod. Gaussian	Multinomial PCA
Prod. Discrete	Dirichlet	Latent Dirichlet allocation
Prod. Noisy-OR	Prod. Bernoulli	BN20/ QMR
Prod. Bernoulli	Prod. Bernoulli	Sigmoid belief net

K-means и его модификации



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

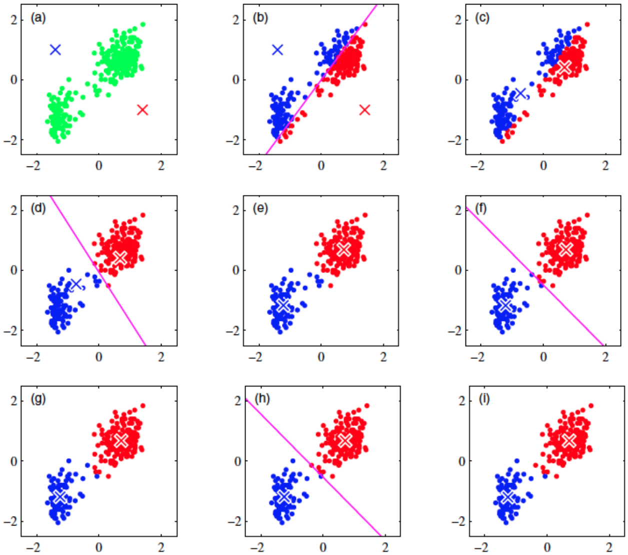
$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x_n}}{\sum_n r_{nk}}$$

K-means

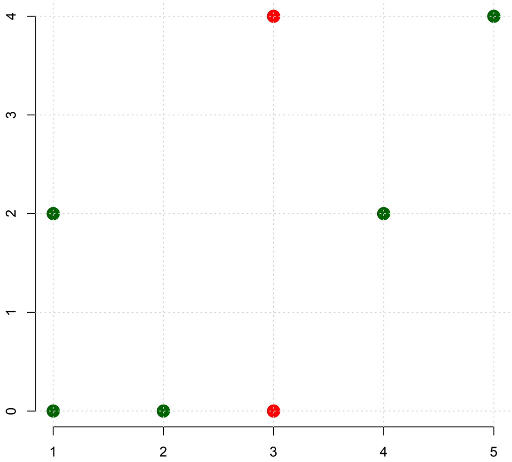
```
function kmeans(X, K):
    initialize N # number of objects
    initialize Mu = (mu_1 ... mu_K) # random centroids
    do:
        # E step
        for k in 1..K:
            for x in 1..N:
                compute r_nk # Cluster assignment
        # M step
        for k in 1..K:
            recompute mu_k # Update centroids
    until Mu converged
    J = loss(X, Mu)
    return Mu, J
```

Сложность $O(NK)$

Локальная оптимизация (!)

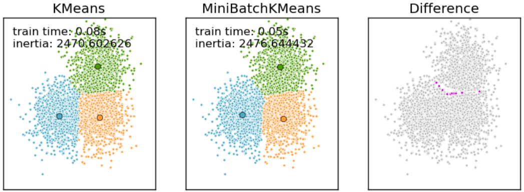

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
2
0
−2
−2
0
2

Задача



Модификации k-means (1)

- 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

Модификации k-means (2)

- k-means++ – "умная" инициализация[3]
- онлайн k-means

$$\mu_k^{new} = \mu_k^{old} + \eta_n(\mathbf{x}_n - \mu_k^{old})$$

[3] k-means++: The Advantages of Careful Seeding

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.

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 $\bar{s}_i$

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

Задача на дом

Рассмотреть смесь из D-мерных распределений Бернулли. В такой смеси $\mathbf{x}$ – D-мерный бинарный вектор, каждый компонент x_i которого подчиняется распределению бернулли с параметром μ_{ki} при заданном векторе μ_k:

$$p(\mathbf{x}|\mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1 - \mu_{ki})^{(1-x_i)}$$

Вероятность k-го вектора μ_k равна π_k. Выписать выражения для E и M шагов при исользовании EM алгоритма для нахождения неизвестных параметров μ_k и π_k.

Вопросы