

Введение в Data Science

Занятие 8. Expectation Maximization

Николай Анохин Михаил Фирулик

25 апреля 2014 г.

ТЕХНОСФЕРА @mail.ru

K-Means и EM

Задача модуля

Задача кластеризации

Дано

- ▶ обучающая выборка $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$
- ▶ количество кластеров K

Задача

Разбить обучающую выборку на K непересекающихся кластеров так, чтобы точки внутри одного кластера были близки, а точки из разных кластеров отдалены

Критерий качества

Пусть

- ▶ μ_k – “типичный” представитель кластера k (центроид)
- ▶ r_{nk} – переменная принадлежности \mathbf{x}_n к кластеру C_k

$$r_{nk} = \begin{cases} 1, & \text{если } \mathbf{x}_n \in C_k \\ 0, & \text{если } \mathbf{x}_n \notin C_k \end{cases}$$

Требуется выбрать μ_k и r_{nk} , которые **минимизируют**

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

Оптимизация критерия

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

Наблюдение: оптимизировать одновременно и по μ_k и по r_{nk} трудно

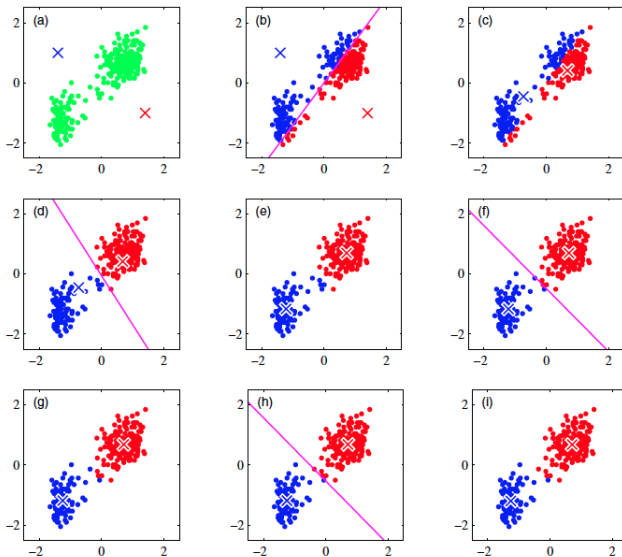
Идея: оптимизируем итеративно по очереди

- Е При фиксированных μ_k подбираем оптимальные r_{nk}
Члены с разными n друг от друга не зависят, откуда

$$r_{nk}^* = \begin{cases} 1, & \text{если } k = \arg \min_j \|\mathbf{x}_n - \mu_j\|^2 \\ 0, & \text{иначе} \end{cases}$$

- М При фиксированных r_{nk} подбираем оптимальные μ_k

$$\mu_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$



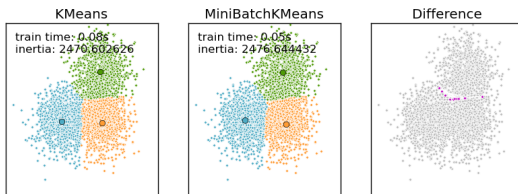
Алгоритм k-means

```
kmeans( $\mathbf{X}$ ,  $K$ ):  
    Случайно задать  $\mu_k$   
    while(not converged):  
        for  $n \in 1 \dots N$ :  
             $r_{nk}^* = \text{int}(k == \arg \min_j \|\mathbf{x}_n - \mu_j\|^2)$   
        for  $k \in 1 \dots K$ :  
             $\mu_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$   
    return  $\mu_k, r_{nk}$ 
```

Алгоритмическая сложность: $O(NK)$

Модификации

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



- ▶ Критерий качества (k-medoids)

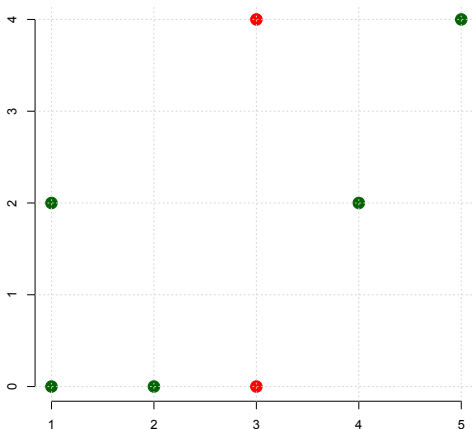
$$\tilde{J} = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \nu(\mathbf{x}_n, \mu_k)$$

μ_k — один из объектов в кластере

- ▶ Что если вхождение \mathbf{x} в кластер C_k описывается вероятностной функцией $p(\mathbf{x}|\theta_k)$?

Задача

Кластеризовать объекты алгоритмом k-means

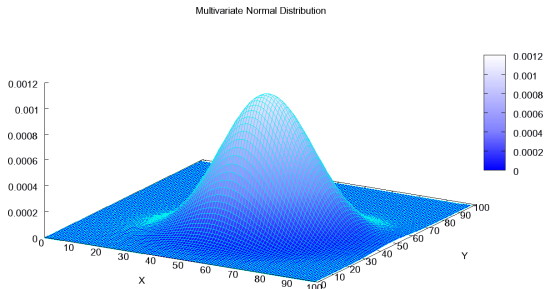


Многомерное гауссовское распределение

Плотность вероятности

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

μ – среднее, Σ – матрица ковариации



Смесь гауссовских распределений

Введем скрытую переменную $\mathbf{z} = (z_1, \dots, z_K)$ – бинарный случайный вектор размерности K , такой что

1. $z_k \in \{0, 1\}$

2. $\sum_k z_k = 1$

Тогда

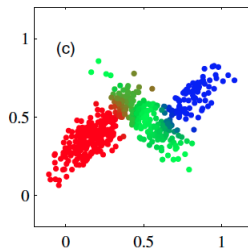
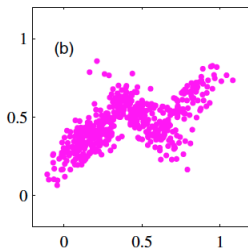
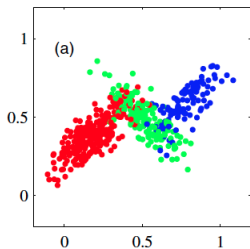
$$p(z_k = 1) = \pi_k, \quad p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

Гауссовское предположение

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k), \quad p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

“Ответственность” k -го компонента за расположение объекта \mathbf{x}

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



Принцип максимального правдоподобия

Log-likelihood

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^N \ln \left[\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right]$$

(Промежуточное) решение

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

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

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

ЕМ-алгоритм: гауссовский случай

E Expectation: при фиксированных μ_k, Σ_k, π_k

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

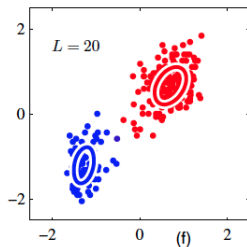
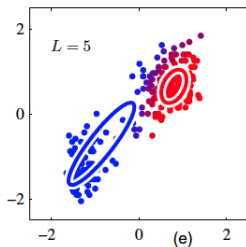
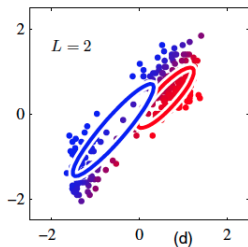
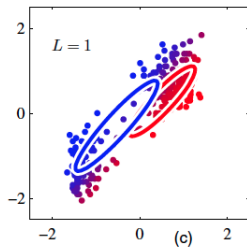
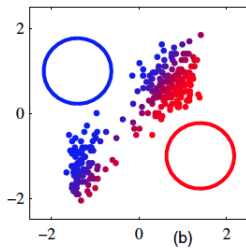
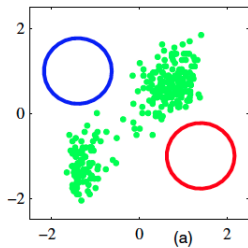
M Maximization: при фиксированных $\gamma(z_{nk})$

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

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

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

S Остановиться при достижении сходимости



ЕМ-алгоритм: общий случай

Задача

Пусть известно распределение $P(\mathbf{X}, \mathbf{Z}|\theta)$, где \mathbf{x} – наблюдаемые переменные, а \mathbf{z} – скрытые. Требуется найти θ , максимизирующее $P(\mathbf{X}|\theta)$.

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

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

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

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

EM и K-means

Пусть $\Sigma_k = \epsilon I$, тогда

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

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left(-\frac{1}{2\epsilon} \|\mathbf{x} - \mu_k\|^2\right)}{\sum_j \pi_j \exp\left(-\frac{1}{2\epsilon} \|\mathbf{x} - \mu_j\|^2\right)} \rightarrow r_{nk}$$

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

k-means: итог

- + Скорость
- + Простота
- Локальная сходимость
- Эллиптические кластеры

Домашнее задание 1



k-means и модификации

Реализовать алгоритм кластеризации k-means или одну из его модификаций (k-medoids, mini-batch) и протестировать на данных задачи модуля.

Ключевые даты

- ▶ До 2014/04/26 00.00 выбрать ответственного
- ▶ До 2014/05/03 00.00 предоставить решение (после – половина очков)

Задача модуля

Источник	Цель	Признаки
	Кластеризовать музыкальных исполнителей так, чтобы в кластерах находились исполнители одного жанра	Текстовое описание, совместное участие в фестивалях
	Кластеризовать фильмы, так чтобы в кластерах находились фильмы одного жанра	Текстовое описание, общий актерский состав

На выходе. Построение и отбор признаков, реализация 2-3 алгоритмов кластеризации, визуализация кластеров, мини-отчет

Выполнение. Группы по 2-3 человека

Баллы. 20 (индивидуально) + 10 (групповые – поровну)

На сегодня

1. Запускаем код из папки clustering
2. Делимся на группы
3. Выбираем, к какой задаче лежит душа
 - ▶ Для LastFM
Документация <http://www.lastfm.ru/api/intro>
Задача: для каждого исполнителя загрузить список фестивалей
 - ▶ Для Rotten Tomatoes
Документация <http://developer.rottentomatoes.com/docs>
Задача: для каждого фильма загрузить список жанров
4. Думаем о добром метриках расстояния