УДК 519.237.8

Д. С. Фоменко

ЕМ-АЛГОРИТМ ЯК МЕТОД ОПТИМІЗАЦІЇ У ЗАДАЧАХ КЛАСТЕРИЗАЦІЇ

Вінницький національний технічний університет;

Анотація

Розглянуто принципи роботи ЕМ-алгоритма, переваги та недоліки в задачах кластеризації з неповними змінними та наведено детальних математичний опис.

Ключові слова: ЕМ-алгоритм, кластеризація, ймовірнісні моделі, оптимізація, машинне навчання.

Abstract

EM algorithm, its advantages and disadvantages in clustering tasks with incomplete variables, with a detailed mathematical description.

Keywords: EM algorithm, clustering, probabilistic models, optimization, machine learning.

Вступ

Кластеризація даних є однією з фундаментальних задач машинного навчання без учителя, метою якої є групування об'єктів таким чином, щоб об'єкти в одній групі (кластері) були більш подібними один до одного, ніж до об'єктів з інших груп. ЕМ-алгоритм, запропонований Демпстером, Лердом і Рубіном у 1977 році[1], є ітеративним методом для знаходження оцінок максимальної правдоподібності параметрів ймовірнісних моделей, коли модель залежить від прихованих змінних. У задачах кластеризації ЕМ-алгоритм часто застосовується для моделювання суміші розподілів, де приналежність спостереження до конкретного кластера розглядається як прихована змінна. ЕМ-алгоритм складається з двох основних кроків: Е-крок (Expectation) та М-крок (Maximization), які циклічно повторюються до збіжності алгоритму.

ЕМ-АЛГОРИТМ

Нехай набір спостережуваних даних:

$$X = \{x_1, x_2, ..., x_n\}$$

Рисунок 1 – Набір спостережуваних даних

Маємо також набір прихованих змінних:

$$Z = \{z_1, z_2, ..., z_n\}$$

Рисунок 2 – Набір прихованих змінних

За параметри моделі, які потрібно оцінити, візьмемо Θ. Мета полягає в максимізації функції подібності:

$$L(\theta|X) = p(X|\theta)$$

Рисунок 3 – Функція подібності

1. Для обчислення Е-кроку потрібно обчислити очікуване значення логарифмічної функції правдоподібності з урахуванням поточної оцінки параметрів θ^t [2]:

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t}[\log p(X, Z|\theta)]$$

Рисунок 4 – Логарифмічна функція правдоподібності

Тут $E_{(x)}[y]$ позначає математичне сподівання y відносно змінної x. Формула визначає Q-функцію, яка представляє очікуване значення логарифмічної функції правдоподібності повної моделі, враховуючи приховані змінні Z.

2. На М-кроці знаходиться нова оцінка параметрів θ^{t+1}, що максимізує Q-функцію:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t)$$

Рисунок 5 – Функція оцінки параметрів

Формула демонструє, як оновлюються параметри моделі на кожній ітерації. Це стандартна задача максимізації, яка може бути розв'язана аналітично для багатьох моделей.

У контексті кластеризації ЕМ-алгоритм часто застосовується до моделі суміші Гаусових розподілів (Gaussian Mixture Model, GMM). Припустимо, що наші дані генеруються з k різних нормальних розподілів, кожен з яких представляє окремий кластер.

Ймовірність спостереження x_i визначається як:

$$p(x_i|\theta) = \sum_{j=1}^{k} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j) \quad (1)$$

де:

- π_j – вага j-го компонента суміші, причому

$$\sum_{j=1}^{k} \pi_j = 1$$

- μ_j і Σ_j – середнє значення та коваріаційна матриця j-го компонента,
- $N(x|\mu, \Sigma)$ – функція щільності багатовимірного нормального розподілу.

У випадку GMM, Е-крок включає обчислення відповідальностей (responsibilities) – ймовірностей приналежності кожного спостереження до кожного кластера:

$$\gamma_{ij} = \frac{\pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i|\mu_l, \Sigma_l)}$$

Рисунок 6 – Формула ймовірності належності

Формула, яку зображено на рисунку 6, визначає ймовірність $\gamma_{(ij)}$ того, що спостереження $x_{(i)}$ належить до j-го кластера, враховуючи поточні оцінки параметрів.

На М-кроці оновлюються параметри моделі:

$$\pi_j^{new} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}$$

Рисунок 7 – Формула оновлення ваги компонентів

$$\mu_j^{new} = \frac{\sum_{i=1}^{n} \gamma_{ij} \boldsymbol{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

Рисунок 8 – Формула оновлення середніх значень

$$\Sigma_j^{new} = \frac{\sum_{i=1}^{n} \gamma_{ij} (\boldsymbol{x}_i - \mu_j^{new})(\boldsymbol{x}_i - \mu_j^{new})^T}{\sum_{i=1}^{n} \gamma_{ij}}$$

Рисунок 9 – Формула оновлення коваріаційних матриць

На рисунках 7, 8 і 9 зображено формули, що демонструють, як оновлюються ваги компонентів, середні значення та коваріаційні матриці на кожній ітерації алгоритму.

З переваг можна виділити такі, як: математична обґрунтованість є ключовою перевагою алгоритму, що базується на строгих статистичних принципах максимізації правдоподібності, забезпечуючи надійну теоретичну базу для інтерпретації результатів, гнучкість моделювання дозволяє застосовувати ЕМ-алгоритм до різних ймовірнісних моделей, таких як суміші Гаусових розподілів і розподілів Бернуллі, адаптуючи модель під специфіку даних для підвищення точності кластеризації, м'яка кластеризація надає інформацію про ступінь приналежності спостереження до кожного кластера [3], що особливо корисно для даних з нечіткими границями між кластерами, ефективність для опрацювання неповних даних дозволяє обробляти відсутні значення як приховані змінні без необхідності попереднього введення.

З недоліків можна виділити такі, як: чутливість до початкової ініціалізації призводить до можливої збіжності до різних локальних максимумів, що впливає на стабільність результатів і вимагає методів вибору оптимальної ініціалізації, необхідність попереднього визначення кількості кластерів вимагає додаткових методів для оцінки оптимальної кількості, таких як AIC або BIC, обчислювальна складність для великих наборів даних з високою розмірністю виникає через необхідність обчислення коваріаційних матриць для кожного компонента, припущення про форму розподілу обмежує застосовність, оскільки, наприклад, GMM припускає нормальний розподіл даних у кластерах, що може бути невідповідним для нестандартних форм кластерів.

Висновки

ЕМ-алгоритм є потужним ймовірнісним методом для розв'язання задач кластеризації, який базується на принципі максимізації очікуваної правдоподібності. Його математична обґрунтованість та здатність до м'якої кластеризації роблять його важливим інструментом у машинному навчанні.

Незважаючи на певні обмеження, такі як чутливість до початкової ініціалізації та обчислювальна складність для великих наборів даних, ЕМ-алгоритм залишається популярним методом, особливо для задач з нечіткими границями між кластерами та при наявності неповних даних. Подальші дослідження спрямовані на подолання існуючих обмежень та розширення застосовності алгоритму в різних областях аналізу даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. ML. Expectation-Maximization Algorithm. [Електронний ресурс] — Режим доступу: https://www.geeksforgeeks.org/ml-expectation-maximization-algorithm/ (дата звернення: 22.03.2025). — Назва з екрана.
2. Maya R. Gupta. Theory and Use of the EM Algorithm / Maya R. Gupta, Yihua Chen // Foundations and Trends in Signal Processing.-2011.-Volume 4, Issue 3.-P.223-296.

3. Mirkin B. G. Mathematical classification and clustering / Mirkin B. G. - Dordrecht : Kluwer Academic Publishers, 1996. - 428 p.

Фоменко Денис Сергійович — студент групи 4ПІ-21б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: dfomenko571@gmail.com

Науковий керівник: Катєльніков Денис Іванович - к.т.н., доцент кафедри програмного забезпечення, Вінницький національний технічний університет, Вінниця, fuzzy2dik@gmail.com.

Fomenko Denys S. — *student of group 4PI-21b, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: dfomenko571@gmail.com*

Supervisor: Katielnikov Denys Ivanovich - Ph.D., Associate Professor of the Department of Software Engineering, Vinnytsia National Technical University, Vinnytsia, fuzzy2dik@gmail.com.