# ВИКОРИСТАННЯ ВАРІАЦІЙНОГО ВИВЕДЕННЯ ДЛЯ ЛАТЕНТНОГО РОЗМІЩЕННЯ ДІРІХЛЕ В ЗАДАЧІ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Деркач О. С.
Науковий керівник – к.т.н., доц. Гибкіна Н. В.
Харківський національний університет радіоелектроніки
61166, Харків, просп. Науки, 14, каф. прикладної математики,
тел. (057) 702-14-36, e-mail: oleksii.derkach@nure.ua

Topic modeling is one of the modern directions of the statistic processing of natural language, which has been actively developing since the late 1990s. The topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods for empirical Bayes parameter estimation.

У зв'язку з розвитком масової та ділової комунікації, розповсюдженням соціальних мереж та інших інтернет-ресурсів все більш актуальними стають задачі вилучення інформації з текстів для її подальшого аналізу. Для розв'язання подібних задач перспективними є статистичні методи обробки текстів, до яких відносяться генеративні ймовірнісні моделі.

Метою ймовірнісного моделювання є визначення тематики документів та пов'язаних з ними об'єктів.

Перед будуванням подібних моделей текст природньої мови зазвичай підлягає серії перетворень: лематизації, стемінгу, видаленню стоп-слів та рідкісних слів. Ці перетворення мають на меті спрощення подальшої обробки тексту за рахунок зменшення обсягу вихідних даних.

Будемо вважати, що слово – одиниця дискретних даних, яка визначається як елемент зі словникового запасу. Слова будемо представляти як одиничні базисні вектори, які мають один компонент, що дорівнює одиниці, та інші компоненти, що дорівнюють нулю. Документом будемо вважати послідовність з $N$ слів, і позначатимемо її як $W = (w_1,...,w_n)$, де $w_n$ – $n$-те слово послідовності. Корпус – це колекція $M$ документів, що позначається як $D = \{W_1,...,W_M\}$. Також позначатимемо множину тем як $Z = (z_1,...,z_n)$.

Латентний розподіл Дірехле (LDA) – це генеративна ймовірнісна модель корпусу. Основна ідея полягає в тому, що документи представлені як випадкові суміші прихованих тем, де кожна тема характеризується розподілом слів.

Враховуючи параметри $\alpha$ і $\beta$, спільний розподіл тематичної суміші $\theta$, набір з $N$ тем та набір з $N$ слів, де $N$ – випадкова величина, можемо отримати ймовірність корпусу [1]:

$$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)\,p(w_{dn}|z_{dn},\beta)\right)d\theta_d\,.$$

Ключовою проблемою, яку треба вирішити, щоб використовувати LDA, є обчислення апостеріорного розподілу прихованих змінних для документа [1]:

$$p(\theta,Z|W,\alpha,\beta)=\frac{p(\theta,Z,W|\alpha,\beta)}{p(W|\alpha,\beta)}\,.$$

Цей вираз неможливо обчислити у явному вигляді, оскільки $p(z_{dn}|\theta_d)$ невідомі. Варіаційне виведення є методом наближення $p(\theta,Z|W,\alpha,\beta)$, сутність якого полягає у представленні шуканого розподілу у вигляді добутку [1]

$$q(\theta,Z|\gamma,\phi)=q(\theta|\gamma)\prod_{n=1}^{N}q(z_n|\phi_n),$$

де параметр Діріхле $\gamma$ і поліноміальні параметри $(\phi_1,...,\phi_N)$ є вільними варіаційними параметрами. Мінімізуючи дивергенцію Кульбака-Лейбнера між варіаційним та дійсним апостеріорним розподілами [1]

$$(\gamma^*,\phi^*)=\arg\min_{(\gamma,\phi)} D(q(\theta,Z|\gamma,\phi)\,\|\,p(\theta,Z|W,\alpha,\beta))$$

отримуємо шукане наближення $q(\theta,Z|\gamma^*,\phi^*)$. Цю задачу мінімізації можна розв'язати ітеративним обчисленням наступної пари рівнянь [1]:

$$\phi_{ni}\propto\beta_{iw_n}\exp\{E_q[\log(\theta_i)|\gamma]\},$$

$$\gamma_i=\alpha_i+\sum_{n=1}^{N}\phi_{ni}\,.$$

Математичне сподівання обчислюємо за формулою:

$$E_q[\log(\theta_i)|\gamma]=\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right),$$

де $\Psi$ – перша похідна логарифма гамма-функції, яка може бути апроксимована рядом Тейлора.

**Список використаних джерел:**

1. David M. Blei, Andrew Y. Ng., Michael I. Jordan (2003) Latent Dirichlet Allocation. The Journal of Machine Learning Research, volume 3. PP.993-1022.