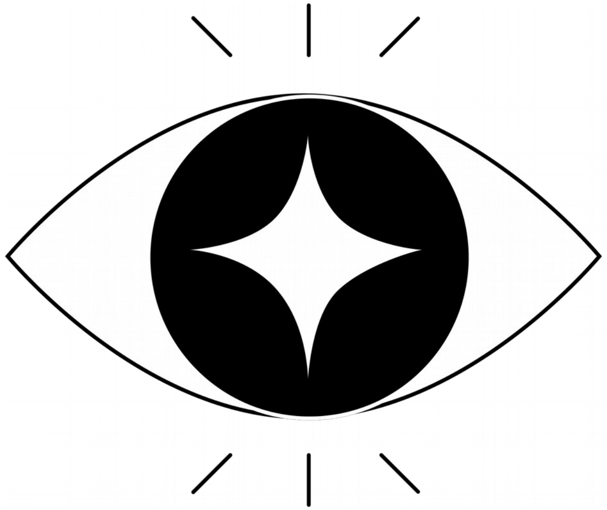
Тематическое моделирование ARTM

На этом уроке

- ARTM теория;
- Практика BigARTM.



Операционная задача

Функционал правдоподобия:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi,\Theta} \qquad (1)$$

Органические стохастичности на матрицы:

$$\left.\begin{array}{ll} \sum_{w \in W} \phi_{wt} = 1, & \phi_{wt} \geq 0 \\ \sum_{t \in T} \theta_{td} = 1, & \theta_{td} \geq 0 \end{array}\right\} \qquad (2)$$

Можно решить с помощью ЕМ-алгоритма:

Регуляризация модели PLSA

PLSA позволяет получить из коллекции D при заданном числе тем $|T|$ матрицы параметров Φ и Θ, но задача матричного разложения $F \approx \Phi \times \Theta$ имеет бесконечное множество решений!

Пусть $S \in \mathbb{R}^{|T| \times |T|}$ — произвольная невырожденная квадратная матрица. Тогда верно:

$$F \approx \Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

Задача PLSA является некорректно поставленной. Попробуем наложить на неё больше ограничений — это позволить получить Φ и Θ, которые будут обладать дополнительными полезными свойствами.

Additive Regularization of Topic Models

Стандартный приём доопределения некорректно поставленной задачи — регуляризация.

ARTM (аддитивная регуляризация тематических моделей) расширяет задачу PLSA слагаемыми-регуляризаторами в функционале правдоподобия:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta} \qquad (3)$$

$$R(\Phi, \Theta) = \sum_{i} \tau_i R_i \, (\Phi, \Theta)$$

Здесь R_i — выраженное в виде функционала ограничение, τ_i — весовой гиперпараметр.

EM-алгоритм для ARTM

Теорема: пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка (Φ, Θ) локального экстремума задачи (3) с ограничениями (2) удовлетворяет системе уравнений со вспомогательными переменными, если из решения исключить нулевые столбцы Φ и Θ:

E-шаг

$$p_{tdw} = \operatorname*{norm}_{t \in T}(\phi_{wt}\theta_{td})$$

M-шаг

$$\phi_{wt} = \operatorname*{norm}_{w \in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}$$

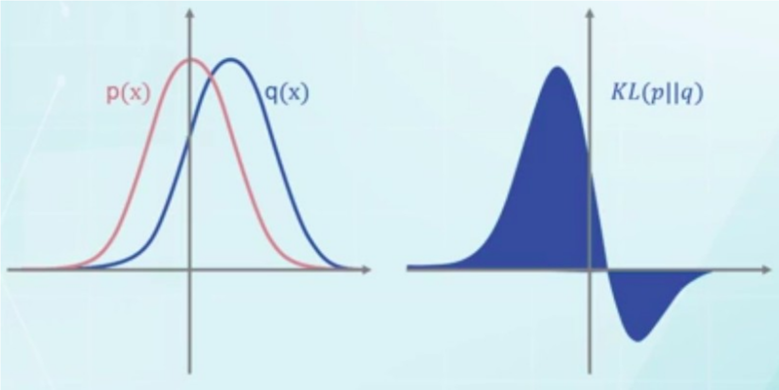
$$\theta_{td} = \operatorname*{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}$$

где $\operatorname*{norm}_{i \in I}(x_i) = \dfrac{\max\{x_i, 0\}}{\sum_{j \in I}\max\{x_j, 0\}}$

KL-дивергенция

KL-дивергенция определяется для пары непрерывных распределений *p* и *q* без нулевых элементов на одинаковом носителе:

$$KL(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$



KL-дивергенция

KL-дивергенция определяется для пары непрерывных распределений *p* и *q* без нулевых элементов на одинаковом носителе:

$$KL(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

и показывает степень вложенности одного распределения в другое. Это несимметричная мера различия распределений.

Для дискретных распределений *p* и *q* на одном носителе мощности *k* определение аналогичное:

$$KL(p||q) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}$$

Пример регуляризатора: сглаживание

Базовый вариант регуляризации — сглаживание параметров заданными вероятностными распределениями. Потребуем близости по KL-дивергенции столбцов ϕ_t и Θ_d к распределениям
$\beta = (\beta_w)_{w \in W}$ и $\alpha = (\alpha_t)_{t \in T}$:

$$\sum_{t \in T} KL(\beta || \phi_t) \to \min_{\Phi} \qquad \sum_{d \in D} KL(\alpha || \theta_d) \to \min_{\Theta}$$

Сложим две суммы с коэффициентами τ_1 и τ_2 и удалим константные слагаемые.

$$R(\Phi, \Theta) = \tau_1 \sum_{t \in T} \sum_{w \in W} \beta_w \log \phi_{wt} + \tau_2 \sum_{d \in D} \sum_{t \in T} \alpha_t \log \theta_{td} \to \max_{\Phi, \Theta}$$

Пример регуляризатора: сглаживание

Базовый вариант регуляризации — сглаживание параметров заданными вероятностными распределениями.

$$R(\Phi, \Theta) = \tau_1 \sum_{t \in T} \sum_{w \in W} \beta_w \log \phi_{wt} + \tau_2 \sum_{d \in D} \sum_{t \in T} \alpha_t \log \theta_{td} \to \max_{\Phi, \Theta}$$

Применим общие формулы ЕМ-алгоритмы из теоремы. Получим итоговые выражения для М-шага:

$$\phi_{wt} = \underset{w \in W}{\mathrm{norm}}(n_{wt} + \tau_1 \beta_w)$$

$$\theta_{td} = \underset{t \in T}{\mathrm{norm}}(n_{td} + \tau_2 \alpha_t)$$

Пример регуляризатора: разреживание

Противоположная стратегия регуляризации — разреживание. В ряде случаев оно приводит к получению более интерпретируемых тем и полезно с точки зрения оптимизации ресурсов при обучении.

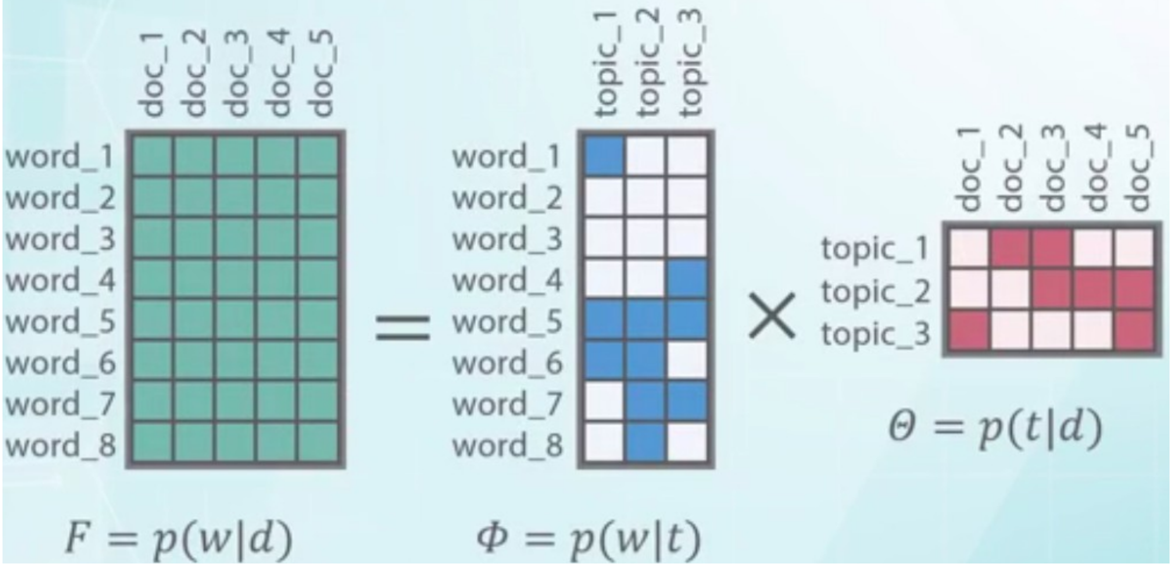
Потребуем, чтобы столбцы ϕ_t и Θ_d были далеки от заданных распределений KL-дивергенции.

Итоговый регуляризатор и формулы M-шага получатся такими же, как и при сглаживании, но с противоположным знаком.

Разреживание и сглаживание можно успешно комбинировать в разных частях одной модели.

Пример регуляризатора: разреживание

На практике разрежённые модели полезны, однако ещё лучше, если модель разрежена так, чтобы темы получались как можно более различными.



Пример регуляризатора: декорреляция

Формализуем требование различности тем с помощью такого регуляризатора:

$$R(\Phi, \Theta) = -\frac{\tau_3}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max_{\Phi, \Theta}$$

Применим теорему и получим формулы М-шага:

$$\phi_{wt} = \underset{w \in W}{\mathrm{norm}} \left(n_{wt} - \tau_3 \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right)$$

Получился разреживающий регуляризатор. Напрямую он работает только с матрицей Ф.

Основные выводы

- Задача тематического моделирования имеет бесконечно много решений;
- Регуляция позволяет наложить дополнительные требования на модель;
- Подход аддитивной регуляризации ARTM позволяет обучать модели с любым набором дополнительных требований;
- Простейшие регуляризаторы позволяют сглаживать или разреживать модель;
- Регуляризатор декорреляции даёт возможность обучать модели с различными темами.