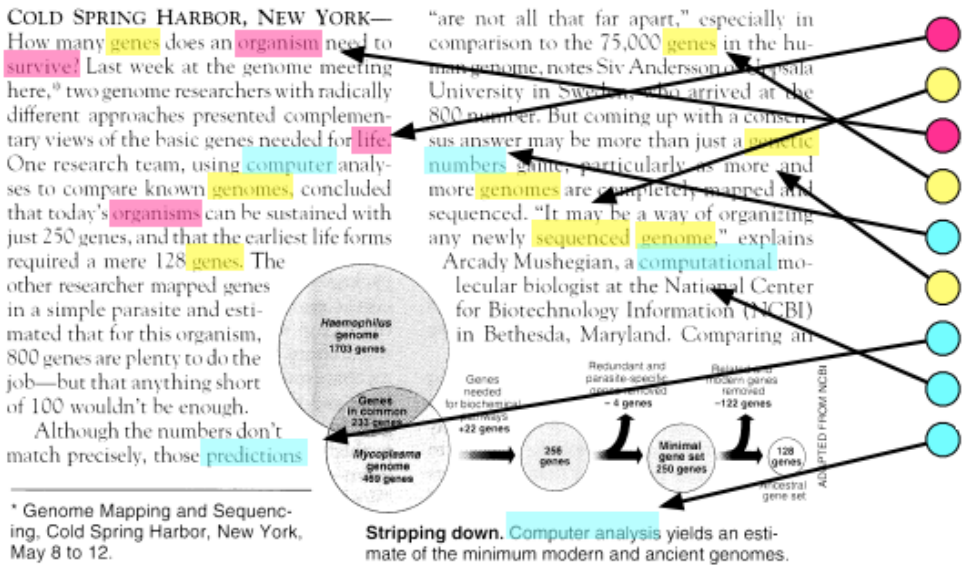
Тематическое моделирование

Виктор Китов

v.v.kitov@yandex.ru



COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Содержание

Тематическое моделирование

- Предположим, мы наблюдаем D документов=последовательность слов.
- $D = \#$[документов], $N = \#$[слов во всех документах], $W = \#$[уникальных слов языка].
- (d_n, w_n) - (индекс документа, индекс слова) на словопозиции $n = \overline{1, N}$.
- Текстовая коллекция: цепочка пар $\{(d_n, w_n)\}_{n=\overline{1,N}}$.
- Предположим:
  - каждое слово порождено какой-то ненаблюдаемой «темой» $z_n \in \{1, 2, ...Z\}$.
    - слово генерируется темой независимо от документа
  - объекты $\{(d_n, z_n, w_n)\}_{n=\overline{1,N}}$ независимы

Процесс порождения данных (модель pLSA)

- Для каждой словопозиции итеративный процесс генерации данных (d->z->w):
  - генерируется номер документа: $d \sim p(d)$
  - генерируется тема: $z \sim p(z|d)$
  - генерируется слово: $w \sim p(w|z)$
- Комментарии:
  - Каждый документ определяет распределение тем: $z \sim p(z|d)$
  - Каждая тема определяет распределение слов: $w \sim p(w|z)$
  - Только тема определяет распределение слов: $p(w|z,d) = p(w|z)$

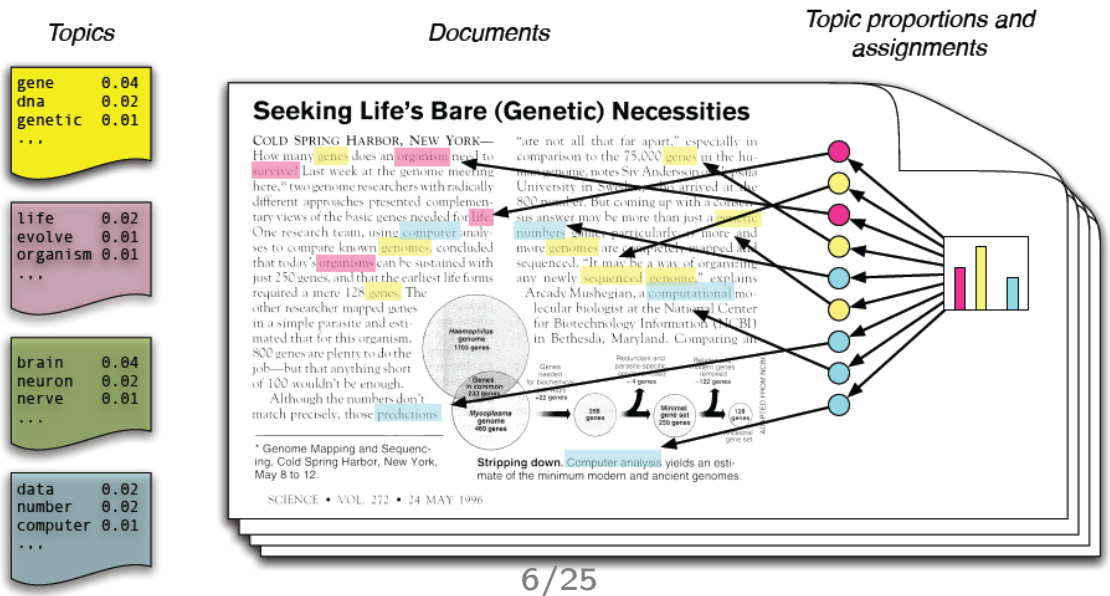
Применения

- Исходное представление документа
  $d \to [p(w=1|d), ...p(w=W|d)] \in \mathbb{R}^W$.
- Новое представление документа
  $d \to [p(z=1|d), ...p(z=Z|d)] \in \mathbb{R}^Z$
  - высокоуровневое семантическое представление: $Z << W$.
- Получаем темы,определяющие коллекцию документов:
  - тема=распределения на словах
    $z \to [p(w=1|z), ...p(w=W|z)] \in \mathbb{R}^W, \quad z = \overline{1,Z}$.
  - интерпретация темы: самые частотные в ней слова.
  - #тем необходимо задавать заранее.

Сегментация текста

Обозначим каждую словопозицию (d, w) самой вероятной темой:

$$(d,w) \to \arg\max_z p(z|d,w) = \arg\max_z \frac{p(z,d,w)}{p(d,w)} =$$
$$= \arg\max_z p(d)p(z|d)p(w|z)$$
$$= \arg\max_z p(z|d)p(w|z)$$



Применения

- Применения тематического моделирования:
  - снижение размерности
  - извлечение высокоуровневых семантических признаков
  - кластеризация документов
  - суммаризация документов
  - сегментация тем внутри документов
  - кластеризация слов (по темам, в которых они совстречаются)

Др. области

- Тематическое моделирование может быть применено к любым объектам=последовательностям сущностей.
  - ДНК=последовательности генов
  - видеозапись=последовательность событий
  - химические вещества=последовательности молекул
  - финансовые расчеты=последовательность транзакций

pLSA как матричное разложение

- Тематическая модель: $p(w|d) = \sum_z p(z|d)p(w|z)$
- В матричной форме:

$$X = BC, \quad X = \{p(w|d)\} \in \mathbb{R}^{DxW},$$
$$B = \{p(z|d)\} \in \mathbb{R}^{DxZ}, \; C = \{p(w|z)\} \in \mathbb{R}^{ZxW}$$

- B, C - неотрицательные матрицы распределений.
  - удовлетворяют $sum(A[d,:]) = 1, \; sum(C[z,:]) = 1 \; \forall d, z.$

pLSA как матричное разложение

- Тематическая модель: $p(w|d) = \sum_z p(z|d)p(w|z)$
- В матричной форме:

$$X = BC, \quad X = \{p(w|d)\} \in \mathbb{R}^{DxW},$$
$$B = \{p(z|d)\} \in \mathbb{R}^{DxZ}, \; C = \{p(w|z)\} \in \mathbb{R}^{ZxW}$$

- B, C - неотрицательные матрицы распределений.
  - удовлетворяют $sum(A[d,:]) = 1, \; sum(C[z,:]) = 1 \; \forall d, z.$
- Тематическая модель: $p(d, w) = \sum_z p(d)p(z|d)p(w|z)$
- В матричной форме:

$$X = ABC,$$
$$A = \text{diag}\{p(d)\} \in \mathbb{R}^{DxD}$$

Содержание

E-шаг для независимых (x_n, z_n)

- Рассмотрим частный случай независимых наблюдений $\{(x_n, z_n)\}_{n=1}^N$, x_n - наблюдаемые, z_n - латентные
  - пример: смесь Гауссиан, z_n-#компоненты, x_n-реализация.
- E-шаг становится:

$$q(Z) = p(Z|X,\theta) = p(z_1|x_1,\theta)...p(z_N|x_N,\theta) = q_1(z_1)...q_N(z_N)$$

$$q_n(z_n) = p(z_n|x_n,\theta)$$

M-шаг для независимых (x_n, z_n)

Для независимых объектов (x_n, z_n):

$$\sum_Z q(Z) \ln p(X, Z|\theta)\} = \sum_{z_1,...z_N} q_1(z_1)...q_N(z_N) \ln \prod_{n=1}^N p(x_n, z_n|\theta)$$

$$= \sum_{z_1,...z_N} q_1(z_1)...q_N(z_N) \ln p(x_n, z_n|\theta) =$$

$$= \sum_{n=1}^N q_n(z_n) \ln p(x_n, z_n|\theta) \prod_{k\neq n} \left(\underbrace{\sum_{z_k} q_k(z_k)}_{=1} \right)$$

$$= \sum_{n=1}^N q_n(z_n) \ln p(x_n, z_n|\theta) \to \max_\theta$$

Содержание

EM алгоритм для pLSA

- $\theta = \left(\{p(d)\}_{d=\overline{1,D}}; \{p(z|d)\}_{z=\overline{1,Z}}^{d=\overline{1,D}}; \{p(w|z)\}_{w=\overline{1,W}}^{z=\overline{1,Z}} \right)$
- E-шаг:

$$p(z|d,w) = \frac{p(z,d,w)}{p(d,w)} = \frac{p(d)p(z,w|d)}{p(d)p(w|d)} = \frac{p(z|d)p(w|z)}{p(w|d)}$$

- M-шаг:

$$\mathbb{E}_{q(z)} \ln p\left(\{d_n, z_n, w_n\}_{n=\overline{1,N}} \right) \to \max_{p(d),p(z|d),p(w|z)}$$

M-шаг

$$\mathbb{E}_{q(z)} \ln p\left(\{d_n, z_n, w_n\}_{n=\overline{1,N}}\right) = \sum_{n=1}^{N} \sum_{z_n=1}^{Z} p\left(z_n|d_n, w_n\right) \ln p\left(d_n, z_n, w_n\right)$$

$$= \sum_{d=1}^{D} \sum_{w=1}^{W} N_{dw} \sum_{z=1}^{Z} p\left(z|d, w\right) \ln p\left(d, z, w\right)$$

$$= \sum_{d=1}^{D} \sum_{w=1}^{W} N_{dw} \sum_{z=1}^{Z} p\left(z|d, w\right) \ln \left[p(d)p(z|d)p(w|z)\right]$$

$$\to \max_{p(d),p(z|d),p(w|z)}$$

$$\sum_{d} p(d) = 1, \quad \sum_{z} p(z|d) = 1, \quad \sum_{w} p(w|z) = 1 \quad \forall d, z, w.$$

M-шаг

$$\mathcal{L} = \sum_{d=1}^{D}\sum_{w=1}^{W} N_{dw} \sum_{z=1}^{Z} p\left(z|d,w\right) \ln p(d)$$

$$+\sum_{d=1}^{D}\sum_{w=1}^{W} N_{dw} \sum_{z=1}^{Z} p\left(z|d,w\right) \ln p(z|d)+$$

$$+\sum_{d=1}^{D}\sum_{w=1}^{W} N_{dw} \sum_{z=1}^{Z} p\left(z|d,w\right) \ln p(w|z)+$$

$$+\alpha\left(1-\sum_{d} p(d)\right)+\sum_{d}\beta_d\left(1-\sum_{z} p(z|d)\right)$$

$$+\sum_{z}\gamma_z\left(1-\sum_{w} p(w|z)\right) \to \text{extr}_{p(d),p(z|d),p(w|z)}$$

M-шаг: p(d)

$$\frac{\partial \mathcal{L}}{\partial p(d)} = \sum_w N_{dw} \sum_z p(z|d,w) \frac{1}{p(d)} - \alpha = 0$$

$$p(d) = \frac{1}{\alpha} \sum_w N_{dw} \sum_z p(z|d,w) = \frac{1}{\alpha} \sum_w N_{dw} = N_d$$

$$1 = \sum_d p(d) = \frac{1}{\alpha} \sum_d N_d = \frac{N}{\alpha} \implies \alpha = N$$

$$p(d) = \frac{N_d}{N} \quad \text{(constant)}$$

M-шаг: p(z|d)

$$\frac{\partial \mathcal{L}}{\partial p(z|d)} = \sum_w N_{dw} p(z|d,w) \frac{1}{p(z|d)} - \beta_d = 0$$

$$p(z|d) = \frac{1}{\beta_d} \sum_w N_{dw} p(z|d,w) = \frac{N_{dz}}{\beta_d}$$

$$1 = \sum_z p(z|d) = \frac{1}{\beta_d} \sum_w N_{dw} \sum_z p(z|d,w) = \frac{1}{\beta_d} \sum_w N_{dw} = \frac{N_d}{\beta_d}$$

$$\beta_d = N_d \implies p(z|d) = \frac{N_{dz}}{N_d}$$

M-шаг: p(w|z)

$$\frac{\partial \mathcal{L}}{\partial p(w|z)} = \sum_d N_{dw} p(z|d,w) \frac{1}{p(w|z)} - \gamma_z = 0$$

$$p(w|z) = \frac{1}{\gamma_z} \sum_d N_{dw} p(z|d,w) = \frac{N_{wz}}{\gamma_z}$$

$$1 = \sum_w p(w|z) = \frac{1}{\gamma_z} \sum_d \sum_w N_{dw} p(z|d,w)$$

$$\gamma_z = \sum_d \sum_w N_{dw} p(z|d,w) = N_z$$

$$p(w|z) = \frac{N_{wz}}{N_z}$$

EM алгоритм - реализация

- Инициализируем $N_{wd}, N_d, p(w|d) = \frac{N_{wd}}{N_d}$, инициализируем случайно $p(z|d)$ и $p(w|z)$, так что $\sum_z p(z|d) = 1, \sum_w p(w|z) = 1$.
- Повторять до сходимости:
  - $p(z|d,w) := \frac{p(z|d)p(w|z)}{p(w|d)} \quad \forall d,z,w.$
  - $N_{dz} := 0; \quad N_{wz} := 0; \quad N_z := 0 \quad \forall d,z,w.$
  - для $d = \overline{1,D}$:
    - для $w = \overline{1,W}$:
      $$N_{dwz} := N_{dw} p(z|d,w)$$
      $$N_{dz} := N_{dz} + N_{dwz},\ N_{wz} := N_{wz} + N_{dwz}$$
    - $N_z := N_z + N_{dz}$

$$p(z|d) := \frac{N_{dz}}{N_d}; \quad p(w|z) := \frac{N_{wz}}{N_z}$$

Содержание

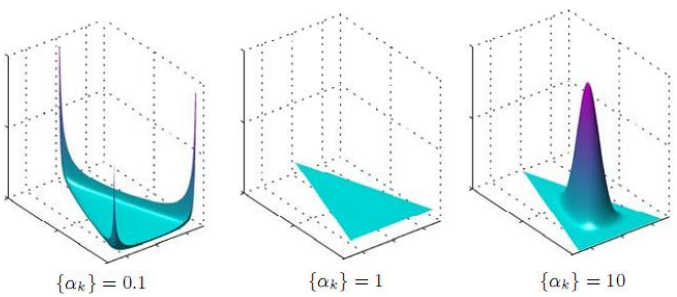
Метод Latent Dirichlet Allocation (LDA)

- Байесовское расширение pLSA
- Распределения $p(z|d)$ и $p(w|z)$ - сл. величины с априорными распределениями:

$$p(z|d) \sim Dir(\alpha), \quad p(w|z) \sim Dir(\beta)$$

- Распределение Дирихле $x \sim Dir(\alpha)$ определено для $x \in \mathbb{R}^K$: $x_n \in (0,1)$, $\sum_{n=1}^K x_n = 1$: $p(x) \propto x_1^{\alpha-1} x_2^{\alpha-1} ... x_K^{\alpha-1}$.

Dirichlet$(\alpha_1, \alpha_2, ...\alpha_K)$ для $\alpha_1 = ... = \alpha_K = \alpha$.



Переменные в LDA

Параметры:

- u-параметр априорного распределения тем $p(z|d)$
- v-параметр априорного распределения слов $p(w|z)$

Оцениваемые переменные:

- $\varphi_z = p(w|z)$, $w = \overline{1,W}$, $z = \overline{1,Z}$
- $\theta_d = p(z|d)$, $z = \overline{1,Z}$, $d = \overline{1,D}$

Латентные переменные:

- темы для каждой словопозиции:

$$z_i^d, \quad d = \overline{1,D},\ i = \overline{1,n_d}$$

Наблюдаемые величины:

- слова и документы на каждой словопозиции:

$$w_i^d, \quad d = \overline{1,D},\ i = \overline{1,n_d}$$

Процесс порождения данных в LDA[1]

- Инициализация:
  - сгенерировать $p(z|d) \sim Dir(u), \quad d = \overline{1, D}$
  - сгенерировать $p(w|z) \sim Dir(v), \quad z = \overline{1, Z}$
- Для каждой словопозиции:
  - сгенерировать номер документа: $d \sim p(d)$
  - сгенерировать тему: $z \sim p(z|d)$
  - сгенерировать слово: $w \sim p(w|z)$

[1] Выведите MAP-оценки для LDA при $\alpha \geq 1$, $\beta \geq 1$.

Расширения тематических моделей

- Автоматический выбор числа тем (например HDP)
  - но нужно задавать «склонность создавать новую тему»
- Иерархическая система тем
  - жадная послойная настройка
  - одновременная настройка всей иерархии
- Помимо слов в документе могут моделироваться др. сущности:
  - дискретные: заголовок, автор, ключевые слова, ссылки, читатели документа.
  - непрерывные: длина документа, время создания.
- Темы с требуемыми свойствами (регуляризация)
  - темы из слов общей лексики, научных терминов
  - регуляризатор разреженности тем (как LDA), непохожести тем и др.