Статистические критерии адекватности вероятностных тематических моделей коллекции текстовых документов

Целых Влада

Московский физико-технический институт

Научный руководитель:
ст.н.с. ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

24 июня 2013 г.

Вероятностные тематические модели

Вероятностная тематическая модель описывает
каждую тему t — дискретным распределением $p(w \mid t)$,
каждый документ d — дискретным распределением $p(t \mid d)$.

Обозначения:
D — коллекция документов,
W — словарь (множество слов),
T — конечное множество тем.

Основное предположение тематического моделирования

- гипотеза условной независимости $p(w \mid d, t) = p(w \mid t)$.

Вероятностная модель порождения документа d

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d).$$

Постановка задачи

Большинство алгоритмов тематического моделирования оценивают вероятности $p(t \,|\, d, w)$, по которым вычисляются счетчики:

$$n_{dwt} = n_{dw} p(t \,|\, d, w), \qquad n_{dt} = \sum_{w \in W} n_{dwt},$$

$$n_{wt} = \sum_{d \in D} n_{dwt}, \qquad n_t = \sum_{d \in D} \sum_{w \in W} n_{dwt},$$

затем вычисляются частотные оценки условных вероятностей:

$$\hat{p}(w \,|\, t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(w \,|\, d, t) = \frac{n_{dwt}}{n_{dt}}.$$

Цель работы

Разработать вычислительно эффективный статистический тест для проверки гипотезы условной независимости:

$$p(w \,|\, d, t) = p(w \,|\, t).$$

Критерий согласия хи-квадрат Пирсона

Пусть $\{x_1, \ldots, x_n\}$ — выборка из n независимых наблюдений случайной величины X, принимающей значения из конечного множества Ω. Эмпирическое распределение:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} [x_i = x], \quad x \in \Omega.$$

Гипотеза H_0: X имеет дискретное распределение $p(x), x \in \Omega$.
Статистика хи-квадрат:

$$X^2 = n \sum_{x \in \Omega} \frac{(\hat{p}(x) - p(x))^2}{p(x)}.$$

При $n \geqslant 50$ и $np(x) \geqslant 5 \; \forall x \in \Omega$ применима асимптотика:

$$X^2 \sim \chi^2(|\Omega| - 1).$$

Проблема разреженности распределения

Распределение $p(x)$ *разрежено*, если вероятности $p(x)$ малы для многих $x \in \Omega$ или $|\Omega| \gg n$. В таких случаях условие $np(x) \geqslant 5$ может не выполняться даже на больших выборках.

Пример (Закон Ципфа — описывает распределение частот слов в языке):

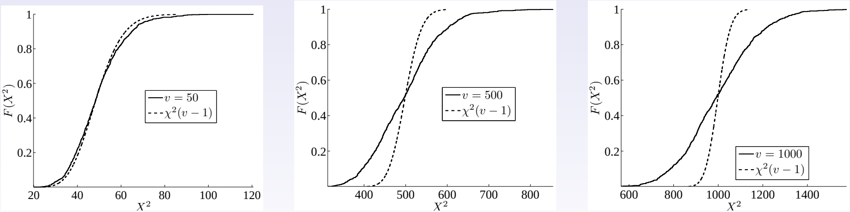
$$p(x) = Ax^{-s}, \quad x \in \Omega = \{1, \dots, |\Omega|\},$$

где $A = \left(\sum_{x=1}^{|\Omega|} x^{-s}\right)^{-1}$ — нормировочный множитель,
s — параметр (обычно $s \approx 1$).
Чем больше s и $|\Omega|$, тем более разрежено распределение $p(x)$.

Неприменимость асимптотики

Сгенерировано $N = 1000$ выборок длины $n = 100$ из распределения Ципфа при $s = 1$ и $v = |\Omega| \in \{50,\ 500,\ 1000\}$.

Эмпирические и асимптотические функции распределения статистики хи-квадрат:



Построение теста на основе сэмплирования

Алгоритм вычисления $(1-\alpha)$-квантиля эмпирического распределения статистики хи-квадрат:

1. для всех $j := 1, \ldots, N$:
2. сгенерировать выборку длины n из распределения $p(x)$;
3. вычислить значение статистики X_j^2;
4. по полученным значениям статистики $X_1^2, \ldots, X_N^2$ построить эмпирическую функцию распределения $\hat{F}_n(X^2)$ и вычислить её $(1-\alpha)$-квантиль $\hat{F}_{n,1-\alpha}$.

Число N рекомендуется брать не менее 1000 при типичном значении $\alpha = 0.05$.

Применение теста на основе сэмплирования

Для выборки длины n с эмпирическим распределением $\hat{p}(x)$:

$$X^2 = n \sum_{x \in \Omega} \frac{(\hat{p}(x) - p(x))^2}{p(x)}.$$

Если $X^2 > \hat{F}_{n,1-\alpha}$, то гипотеза $H_0 : \hat{p}(x) \sim p(x)$ отвергается.

Применение теста для тематической модели:

$$X^2_{dt} = n_{dt} \sum_{w:\, n_{wt} > 0} \frac{\big(\hat{p}(w \,|\, d, t) - \hat{p}(w \,|\, t)\big)^2}{\hat{p}(w \,|\, t)}.$$

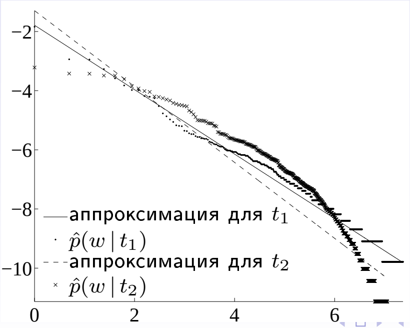
Недостаток: требуется около 1000 сэмплирований выборок длины n_{dt} из распределения $\hat{p}(w \,|\, t)$, что занимает много времени.

Параметрический тест для закона Ципфа

Если $p(x)$ — распределение Ципфа с параметром s, то при $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$, $\alpha = 0.05$ построена 5-параметрическая аппроксимация:

$$\tilde{F}_{1-\alpha}(s, v, n) = Av(1 + Bn^{-c})(1 + GH^s).$$

Недостаток: $p(w|t)$ плохо подчиняется закону Ципфа:



Точный тест Фишера

Гипотеза H_0: $p(w \,|\, d,t) = p(w \,|\, t)$ для фиксированных (d,w,t).
Альтернатива H_1: w встречается в d слишком часто.
Таблица сопряженности X:

	w	$W \backslash w$	$\sum$
d	n_{dwt}	$n_{dt} - n_{dwt}$	n_{dt}
$D \backslash d$	$n_{wt} - n_{dwt}$	$n_t - n_{dt} - n_{wt} + n_{dwt}$	$n_t - n_{dt}$
$\sum$	n_{wt}	$n_t - n_{wt}$	n_t

Вероятность реализации n_{dwt} *при фиксированных* n_{dt}, n_{wt}, n_t:

$$P(X) = \frac{C_{n_{dt}}^{n_{dwt}} C_{n_t - n_{dt}}^{n_{wt} - n_{dwt}}}{C_{n_t}^{n_{wt}}}.$$

Уровень значимости: $\mathrm{Pvalue}(d,w,t) = \sum_Z P(Z)$,
где Z — таблица с теми же маргинальными суммами, что и X,
и не меньшим элементом первой строки первого столбца.

Множественное использование теста Фишера

Гипотеза H_0: $p(w \,|\, d, t) = p(w \,|\, t)$ для всех (d, w) в теме t.

Алгоритм проверки гипотезы:

1. для каждой пары (d, w) в теме t провести точный тест Фишера и вычислить $\mathrm{Pvalue}(d, w, t)$;
2. с помощью биномиального теста проверить гипотезу: вероятность того, что $\mathrm{Pvalue}(d, w, t) < \alpha$ равна α.

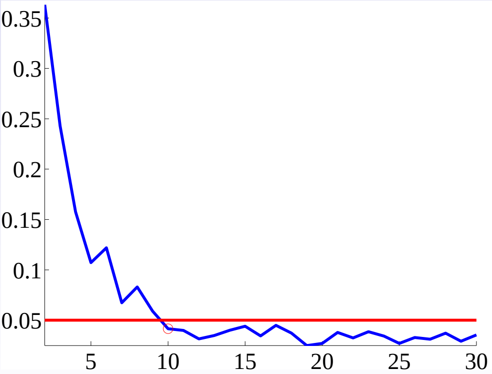
Определение оптимального числа тем

Проверка гипотезы условной независимости используется для определения оптимального числа тем в коллекции.

Если после достижения сходимости алгоритма LDA гипотеза условной независимости принимается, то задано число тем, больше либо равное оптимальному.

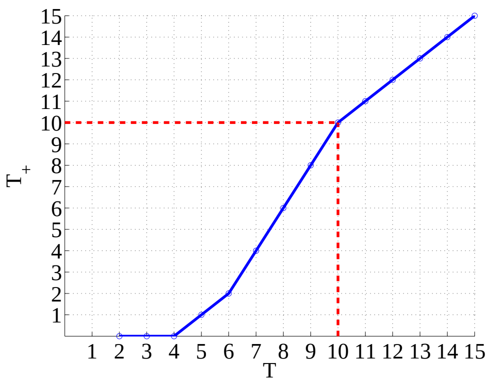
Тест хи-квадрат на основе сэмплирования

Модельная коллекция: $D = 500$, $W = 200$, $n_d = 120$, $T = 10$.
Зависимость доли слов, для которых гипотеза отвергается на уровне значимости 0.05, от числа тем, задаваемого в алгоритме LDA:



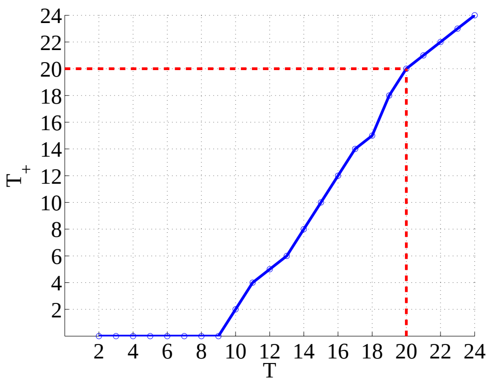
Множественный тест Фишера

Модельная коллекция: $D = 500$, $W = 200$, $n_d = 120$, $T = 10$.
Зависимость числа тем T_+, для которых гипотеза принимается, от общего числа тем T, задаваемого в алгоритме LDA:



Множественный тест Фишера

Модельная коллекция: $D = 900$, $W = 300$, $n_d = 120$, $T = 20$.
Зависимость числа тем T_+, для которых гипотеза принимается, от общего числа тем T, задаваемого в алгоритме LDA:



Выводы

- разработан критерий согласия на основе сэмплирования для разреженных дискретных распределений.
- разработан критерий проверки независимости на основе множественного использования точного теста Фишера.
- рассмотрено применение предложенных тестов для проверки адекватности вероятностных тематических моделей.