Задачи оценивания значимости выравнивания при помощи скрытых марковских моделей

Власенко Даниил Владимирович
Научный руководитель: к.ф.-м.н. Коробейников А.И.

Санкт-Петербургский государственный университет
Кафедра "Статистического моделирования"

Санкт-Петербург
Декабрь 2021

Выравнивание последовательностей

Определение

Выравнивание последовательностей — размещение двух или более последовательностей друг под другом таким образом, чтобы было легче увидеть их схожие участки.

A	C	E	A	A	F	A	E
C	E	A	F	D	C	E	

A	C	E	A	A	F	A	—	E
—	C	E	A	—	F	D	C	E

Определение

Значимость выравнивания — действительное число s, отражающее сходство последовательностей.

Ложноположительная вероятность

- достаточно ли высокая значимость, чтобы считать последовательность не шумом, или шум мог добиться такой значимости.
- достаточно ли низкая значимость, чтобы считать последовательность шумом, или не шум мог получить такую значимость.

Определение

Ложноположительная вероятность значимости s — это вероятность того, что шум получит значимость равную или выше s.

Модели

Определение

Пусть X_n и Y_n дискретные стохастические процессы, $n \geq 1$. Пара (X_n, Y_n) называется скрытой марковской моделью, если

- *X_n — марковский процесс, поведение которого напрямую не наблюдается ("скрытый");*
- $\mathrm{P}(Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n) = \mathrm{P}(Y_n | X_n = x_n)$ *для любого $n \geq 1$, где $x_1, \ldots, x_n$ — значения, принимаемые процессом X_n (**состояния модели**), y_n — значение, принимаемое процессом Y_n (**наблюдаемый символ модели**).*

Модели

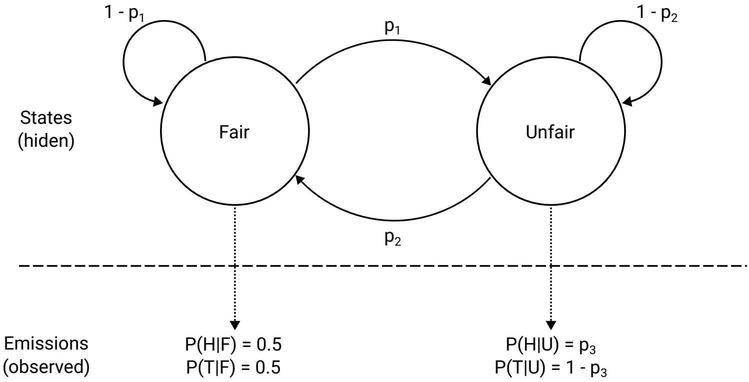

Рис. 1: Простая скрытая марковская модель.

Модели

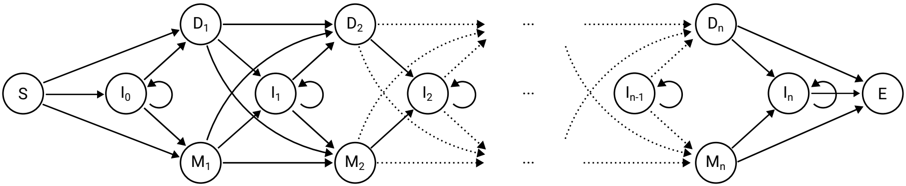

Рис. 2: Профильная скрытая марковская модель.

Модели

Определение

Вероятность последовательности D может интерпретироваться и считаться по-разному — алгоритмом Витерби или Форвард алгоритмом.

$$s_{max}(D) = \max_{\pi \in \pi_D}(s(\pi));$$

$$s_{fw}(D) = \sum_{\pi \in \pi_D} s(\pi);$$

$$Z(D, T) = \sum_{\pi \in \pi_D} s(\pi)^{\frac{1}{T}}.$$

Модели

Мы предполагаем наличие простой фоновой модели B для последовательностей длины L такой, что все L символьных позиций независимы и одинаково распределены в соответствии с некоторым распределением $\mathrm{P}(d|B)$, где d отражает возможный наблюдаемый символ:

$$\mathrm{P}(D|B) = \prod_{i=1}^{L} \mathrm{P}(d_i|B),$$

где d_i — это i-ый наблюдаемый символ последовательности D.

Постановка математической задачи

Определение

Ложноположительная вероятность значимости s_0 для строк длины L:

$$fpr(s_0) = \sum_{D \in D_L} \mathrm{P}(D|B)\Theta(s(D) \geq s_0), \tag{1}$$

где $\mathrm{P}(D|B)$ — *условная вероятность последовательности D, описываемая фоновой моделью, $s(D)$ — вероятность последовательности D, считаемая профильной СММ, и*

$$\Theta(s(D) \geq s_0) = \begin{cases} 1, & s(D) \geq s_0 \\ 0, & s(D) < s_0 \end{cases}.$$

Алгоритм

Вычисление $fpr(s_0)$ через формулу (1) обычно неосуществимо, значение $fpr(s_0)$ может быть оценено через выборку по значимости.

Пусть $\mathrm{P}(D|T)$ — это условная вероятность строки D относительно некоторой модели строк длины L параметризованной значением T. Тогда можно переписать $fpr(s_0)$:

$$fpr(s_0) = \sum_{D \in D_L} \mathrm{P}(D|T) f(D, s_0),$$

где

$$f(D, s_0) = \frac{\mathrm{P}(D|B)\Theta(s(D) \geq s_0)}{\mathrm{P}(D|T)}. \tag{2}$$

Алгоритм

Определим модель, используемую для выборки по важности параметризованную T:

$$\mathrm{P}(D|T) = \frac{\mathrm{P}(D|B)Z(D,T)}{Z(T)},$$

где

$$Z(T) = \sum_{D \in D_L} \mathrm{P}(D|B)Z(D,T).$$

Подставив определение $\mathrm{P}(D,T)$ в уравнение (2) получим

$$f(D,s_0) = \frac{Z(T)\Theta(s(D) \geq s_0)}{Z(D,T)}.$$

Результаты

Вычислим оценку $\widehat{fpr}(s_0)$ для строк длины $L = 100$, состоящих из 5 символов, и доверительные интервалы уровня $\gamma = 0.99$:

Таблица 1: Результаты

s_0	T	$\widehat{fpr}(s_0)$	$[c_1(\gamma); c_2(\gamma)]$
10^{-85}	7	0.0000000183	[0.0; 0.00066349]
10^{-90}	7	0.003175	[0.001884; 0.004779]
10^{-100}	7	0.615709	[0.597540 0.622677]