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.

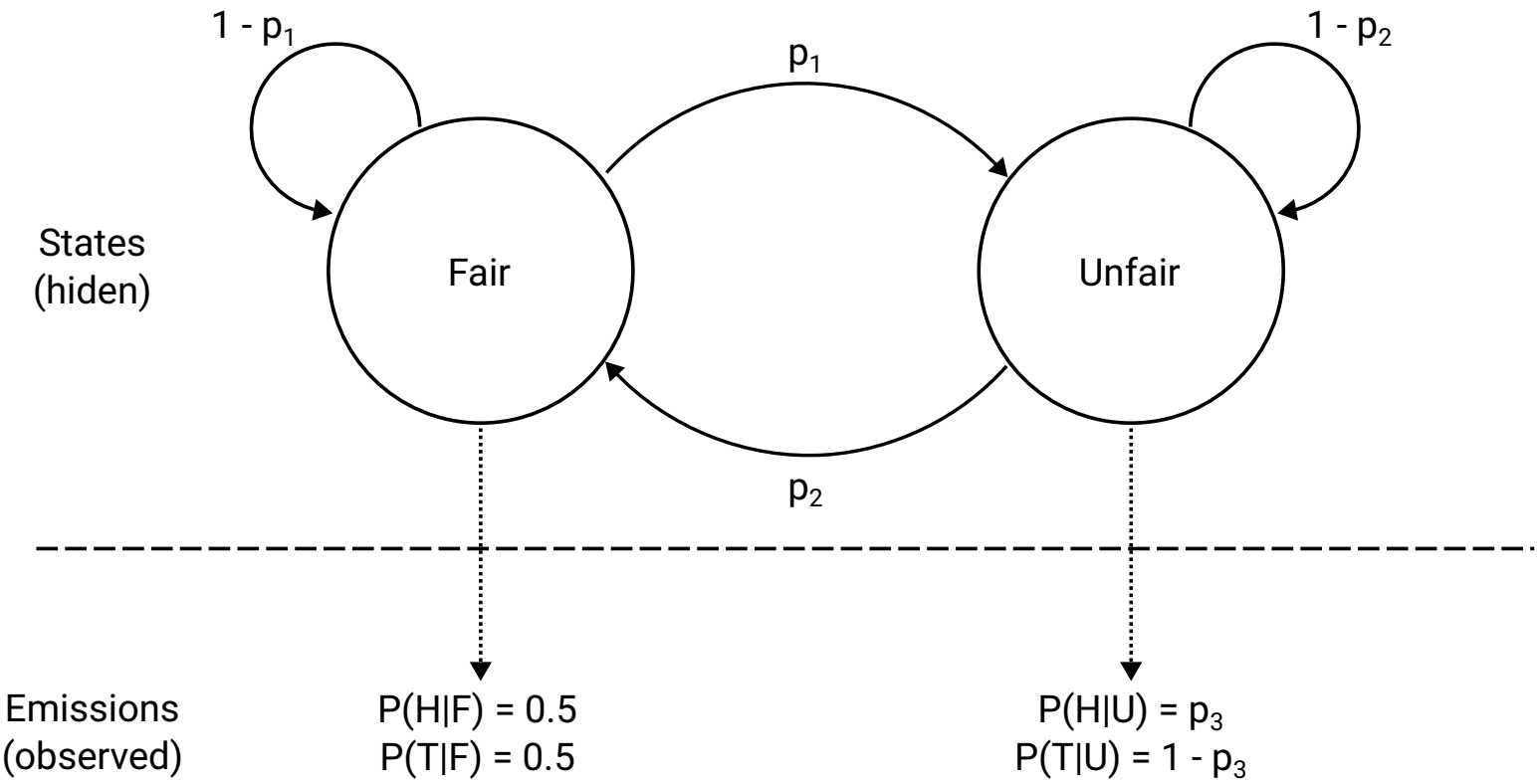
Модели

Определение

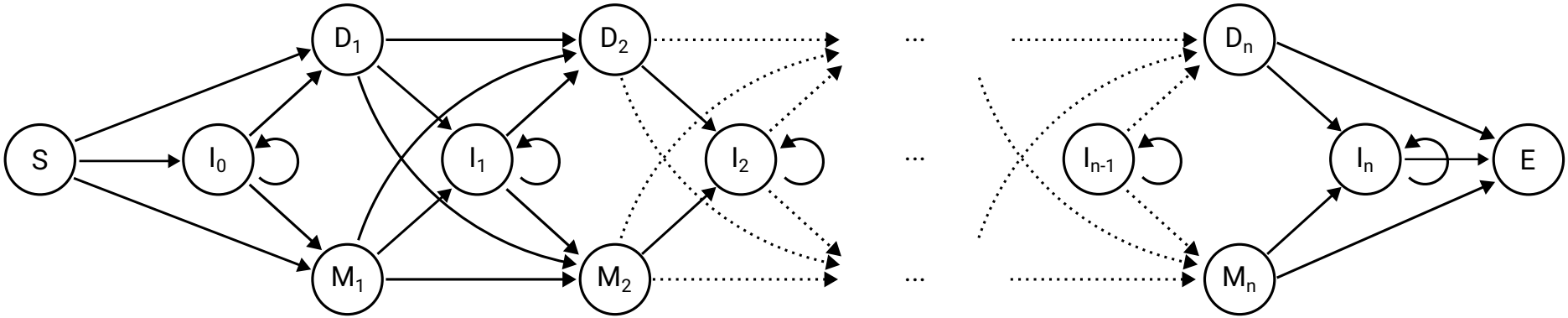
Пусть X_n и Y_n дискретные стохастические процессы, $n \geq 1$. Пара (X_n, Y_n) называется скрытой марковской моделью, если

- *X_n — марковский процесс, поведение которого напрямую не наблюдается ("скрытый");*
- *$P(Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n) = P(Y_n | X_n = x_n)$ 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


1 - p_1
p_1
1 - p_2
States
(hiden)
Fair
Unfair
p_2
Emissions
(observed)
P(H|F) = 0.5
P(T|F) = 0.5
P(H|U) = p_3
P(T|U) = 1 - p_3

Модели



Модели

Определение

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

$$s_{max}(D) = \max_{\pi \in \pi_D}(s(\pi)), \tag{1}$$

$$s_{fw}(D) = \sum_{\pi \in \pi_D} s(\pi). \tag{2}$$

$$Z(D, T) = \sum_{\pi \in \pi_D} s(\pi)^{\frac{1}{T}}. \tag{3}$$

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 $Pr(d|B)$, где d отражает возможный наблюдаемый символ:

$$Pr(D|B) = \prod_{i=1}^{L} Pr(d_i|B), \tag{4}$$

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

Постановка математической проблемы

Ложноположительная вероятность значимости s_0 для строк длины L:

$$fpr(s_0) = \sum_{D \in D_L} Pr(D|B)\Theta(s(D) \geq s_0), \tag{5}$$

где $Pr(D|B)$ — условная вероятность последовательности D, описываемая фоновой моделью, $s(D)$ — вероятность последовательности D, считаемая профильной СММ, и

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