1 Свойства статистики Хмелёва

Напомним определение статистики Хмелёва. Пусть даны два текста T_1 и T_2 (под текстом понимаем упорядоченный набор символов). Рассмотрим некоторый алфавит $\mathcal{A}$ и при всех $i, j \in \mathcal{A}$ определим

$$\nu_{i,j} = \#\{(i,j) \in T_1\}, \qquad p^*_{i,j} = \frac{\nu_{i,j}}{\sum_{j\in\mathcal{A}} \nu_{i,j}}$$

– частоты и оценки вероятностей перехода от символа к символу в первом тексте, и

$$\mu_{i,j} = \#\{(i,j) \in T_2\}, \qquad q^*_{i,j} = \frac{\mu_{i,j}}{\sum_{j\in\mathcal{A}} \mu_{i,j}}$$

– частоты и оценки вероятностей перехода от символа к символу во втором тексте. Отметим, что

$$\sum_{j\in\mathcal{A}} \nu_{i,j} = \nu_i \qquad \text{и} \qquad \sum_{j\in\mathcal{A}} \mu_{i,j} = \mu_i,$$

при всех $i \in \mathcal{A}$, где ν_i и μ_i – частоты символа i в текстах T_1 и T_2 соответственно.

Тогда статистика Хмелёва $H(T_1||T_2)$ определяется следующим образом:

$$H(T_1||T_2) = \sum_{i,j\in\mathcal{A}} \nu_{i,j} \log\left(\frac{p^*_{i,j}}{q^*_{i,j}}\right).$$

В рамках летней школы показали, что эта статистика неотрицательна:

$$\begin{aligned} H(T_1||T_2) &= \sum_{i,j\in\mathcal{A}} \nu_{i,j} \log\left(\frac{p^*_{i,j}}{q^*_{i,j}}\right) \\ &= \sum_{i,j\in\mathcal{A}} \nu_i p^*_{i,j} \log\left(\frac{p^*_{i,j}}{q^*_{i,j}}\right) \\ &= \sum_{i\in\mathcal{A}} \nu_i \sum_{j\in\mathcal{A}} p^*_{i,j} \log\left(\frac{p^*_{i,j}}{q^*_{i,j}}\right) \\ &= \sum_{i\in\mathcal{A}} \nu_i d_{KL}(P^*(i,\cdot)||Q^*(i,\cdot)) \\ &\geq 0, \end{aligned} \tag{1}$$

поскольку d_{KL} – дивергенция Кульбака – Лейблера – неотрицательна. Здесь через $P^*(i,\cdot)$ обозначена оценка распределения перехода из символа i.

Основная проблема состоит в том, что дивергенция $d_{KL}(P^*(i,\cdot)||Q^*(i,\cdot)))$ определена в случае, когда $P^*(i,\cdot)$ абсолютно непрерывно относительно $Q^*(i,\cdot)$, что, вообще говоря, в нашем случае не всегда выполнено, особенно для текстов маленькой длины.

Чтобы обойти эту проблему мы решили учитывать в $H(T_1||T_2)$ только те слагаемые, у которых и $p^*_{i,j} \neq 0$ и $q^*_{i,j} \neq 0$. **Из-за этого и возникает отрицательность.**

Приведём простой пример. Пусть распределения случайных величин ξ и η заданы следующими таблицами

ξ	1	2	3	4
P	1/4	1/4	1/4	1/4

и

η	-1	2	3	5
P	1/6	1/3	1/3	1/6

.

Используя наш метод подсчёта, мы учитывем лишь те атомы, которые принадлежат общему носителю (2 и 3) и получаем результат:

$$\begin{aligned} d_{KL}(P_\xi||P_\eta) &= 1/4\log(3/4) + 1/4\log(3/4) \\ &= 1/2\log(3/4) < 0. \end{aligned}$$

Возможные решения:

1. Считать дивергенцию Кульбака – Лейблера между распределениями $\widetilde{\xi} = \xi \cdot \mathbf{I}(\xi \in CS)$ и $\widetilde{\eta} = \eta \cdot \mathbf{I}(\eta \in CS)$, где CS – общий носитель ($\{2,3\}$ в нашем примере). В этом случае мы получаем следующие таблицы

$\widetilde{\xi}$	0	2	3
P	1/2	1/4	1/4

и

$\widetilde{\eta}$	0	2	3
P	1/3	1/3	1/3

и следующее расстояние Кульбака – Лейблера:

$$d_{KL}(P_{\widetilde{\xi}}||P_{\widetilde{\eta}}) = 1/2\log(3/2) + 1/4\log(3/4) + 1/4\log(3/4) = 1/2\log(3/2) + 1/2\log(3/4) > 0.$$

2. Считать дивергенцию Кульбака – Лейблера между **условными распределениями** $\widehat{\xi} = (\xi|\xi \in CS)$ и $\widehat{\eta} = (\eta|\eta \in CS)$. При таком способе подсчёта в нашем примере $\widehat{\xi}$ и $\widehat{\eta}$ одинаково распределены, а значит и
$$d_{KL}(P_{\widehat{\xi}}||P_{\widehat{\eta}}) = 0.$$

3. **Байесовская постановка.** Будем предполагать, что вектор истинных вероятностей перехода $(p_{i,1}, \ldots, p_{i,n})$ случаен и имеет распределение Дирихле с параметрами $\alpha_{i,1}, \ldots, \alpha_{i,n}$, где n – длина алфавита $\mathcal{A}$. Тогда апостериорная функция вероятности имеет вид
$$\pi_i(\boldsymbol{t}|T_1) \propto t_1^{\alpha_{i,1}-1+\nu_{i,1}} \ldots t_n^{\alpha_{i,n}-1+\nu_{i,n}}.$$

   Беря в качестве оценок апостериорное среднее, получаем
$$p^*_{i,j} = \frac{\alpha_{i,j} + \nu_{i,j}}{\alpha_i + \nu_i}$$
   и
$$q^*_{i,j} = \frac{\alpha_{i,j} + \mu_{i,j}}{\alpha_i + \mu_i}.$$

   Байесовскую статистику Хмелёва, видимо, уместно в таком случае определить так:
$$H(T_1||T_2) = \sum_{i,j \in \mathcal{A}} (\alpha_{i,j} + \nu_{i,j}) \log\left(\frac{p^*_{i,j}}{q^*_{i,j}}\right).$$

   Такое определение делает её неотрицательной (см. (1)). Байесовский подход, по-видимому, решает две проблемы:

   - Проблемы с абсолютной непрерывнотью, описанные выше, которые возникают из-за недостатка информации.
   - Делает «важными» слагаемые, соответствующие наиболее «важным» переходам во всём языке (за счёт добавления коэффициента $\alpha_{i,j}$ к $\nu_{i,j}$ перед логарифмом).

   Кажется, что у подхода есть и минусы. Прежде всего, он как будто будет работать на довольно длинных текстах: иначе коэффициенты $\nu_{i,j}$ могут быть значительно меньше, чем $\alpha_{i,j}$, которые оцениваются по большому корпусу текстов.