Задание 1. Байесовские рассуждения

Курс: Байесовские методы в машинном обучении
Студент: Никулин Александр

Вероятностная модель

$$
\begin{aligned}
&p(a,b,c,d) = p(d|c)p(c|a,b)p(a)p(b),\\
&d|c \sim c + \mathrm{Bin}(c, p_3),\\
&c|a,b \sim \mathrm{Bin}(a, p_1) + \mathrm{Bin}(b, p_2) \text{ or } \mathrm{Poiss}(ap_1 + bp_2)\\
&a \sim \mathrm{Unif}[a_{min}, a_{max}],\\
&b \sim \mathrm{Unif}[b_{min}, b_{max}].
\end{aligned}
$$

Вариант 2

Параметры модели: $a_{min} = 75$, $a_{max} = 90$, $b_{min} = 500$, $b_{max} = 600$, $p_1 = 0.1$, $p_2 = 0.01$, $p_3 = 0.3$. Взять в качестве диапазона допустимых значений для величины c интервал $[0, a_{max} + b_{max}]$, а для величины d – интервал $[0, 2(a_{max} + b_{max})]$.

1. Вывести формулы для всех необходимых далее распределений аналитически.

$$
\begin{aligned}
&p(a) = \frac{1}{a_{max} - a_{min} + 1}\\
&p(b) = \frac{1}{b_{max} - b_{min} + 1}\\
&p_{model1}(c = k) = \sum_{a,b} p(c = k|a,b)p(a)p(b) = \sum_{a,b}\sum_{i=0}^{k} p(\mathrm{Bin}(a, p_1) = i)p(\mathrm{Bin}(b, p_2) = k - i)p(a)p(b)\\
&p_{model2}(c = k) = \sum_{a,b} p(\mathrm{Poiss}(ap_1 + bp_2) = k)p(a)p(b)\\
&p(b|a) = p(b), p(a|b) = p(a)\\
&p(c = k|a) = \sum_{b} p(c = k|a,b)p(b|a) = \sum_{b} p(c = k|a,b)p(b)\\
&p(c = k|b) = \sum_{a} p(c = k|a,b)p(a|b) = \sum_{a} p(c = k|a,b)p(a)\\
&p(d|b) = \frac{p(b,d)}{p(b)} = \frac{\sum_{ac} p(a,b,c,d)}{p(b)} = \frac{\sum_c p(d|c)p(c|b)p(b)}{p(b)} = \sum_{c} p(d|c)p(c|b)\\
&p(b|d) = \frac{p(d|b)p(b)}{p(d)} = \frac{[\sum_c p(d|c)p(c|b)]p(b)}{p(d)}\\
&p(b|a,d) = \frac{\sum_c p(a,b,c,d)}{\sum_{b,c} p(a,b,c,d)} = \frac{\sum_c p(d|c)p(c|a,b)p(a)p(b)}{[\sum_c p(d|c)p(c|a)]p(a)} = \frac{\sum_c p(d|c)p(c|a,b)p(b)}{\sum_c p(d|c)p(c|a)}\\
&p(d = k) = \sum_{c}^{c_{max}} p(d = k|c)p(c) = \sum_{c}^{c_{max}} p(\mathrm{Bin}(c, p_3) = k - c)p(c)
\end{aligned}
$$

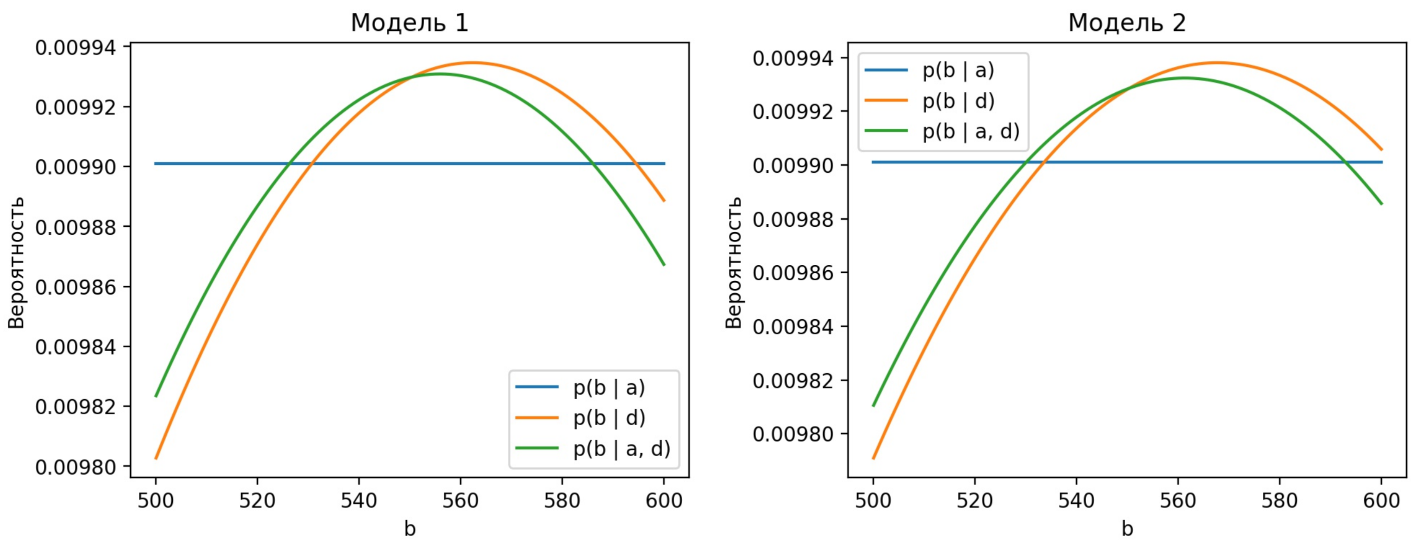
2. Найти математические ожидания и дисперсии априорных распределений $p(a)$, $p(b)$, $p(c)$, $p(d)$.

   В коде математические ожидания и дисперсии для всех распределений посчитаны по определениям:

$$
\mathbb{E}[x] = \sum_{x \in \mathbb{X}} xp(x)
$$
$$
\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2
$$

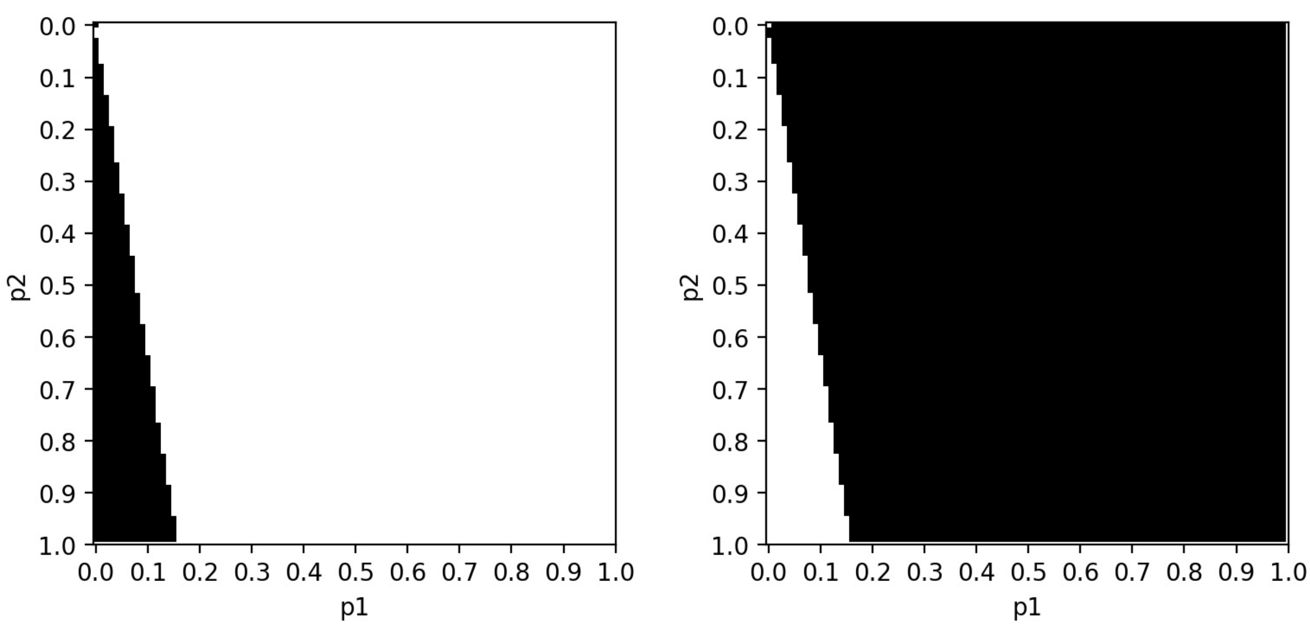
3. Пронаблюдать, как происходит уточнение прогноза для величины b по мере прихода новой косвенной информации. Для этого построить графики и найти мат.ожидание и дисперсию для распределений $p(b)$, $p(b|a)$, $p(b|d)$, $p(b|a,d)$ при параметрах a, d, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого.

   На графике не отображено распределение $p(b)$ т.к. оно идентично $p(b|a)$.



Распределение	Модель 1	Модель 2
$p(b\|a = \mathbb{E}[a])$	550 ± 850	550 ± 850
$p(b\|d = \mathbb{E}[d])$	550.07 ± 848.04	550.09 ± 848.13
$p(b\|a = \mathbb{E}[a], d = \mathbb{E}[d])$	550.04 ± 848.03	550.06 ± 848.12

4. Определить, при каких соотношениях параметров p_1, p_2 изменяется относительная важность параметров a, b для оценки величины c. Для этого найти множество точек $\{ (p_1, p_2) \mid \mathbb{D}[c|b] < \mathbb{D}[c|a] \}$ при a, b, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого. Являются ли множества $\{ (p_1, p_2) \mid \mathbb{D}[c|b] < \mathbb{D}[c|a] \}$ и $\{ (p_1, p_2) \mid \mathbb{D}[c|b] \geq \mathbb{D}[c|a] \}$ линейно разделимыми? Ответ должен быть обоснован!



   Считать $\mathrm{Var}[p(c|a)]$ и $\mathrm{Var}[p(c|b)]$ по честному сложно. Можно воспользоваться формулой полной дисперсии (law of total variance):

$$\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathbb{E}[Y|X]]$$

Пусть $Y = c|b, X = a|b$. Тогда:

$$\mathrm{Var}[Y|X] = \mathrm{Var}[c|a,b] = \mathrm{Var}[\mathrm{Bin}(a,p_1) + \mathrm{Bin}(b,p_2)] = ap_1(1-p_1) + bp_2(1-p_2)$$
$$\mathbb{E}[Y|X] = \mathbb{E}[c|a,b] = ap_1 + bp_2$$
$$\mathbb{E}[\mathrm{Var}[Y|X]] = \mathbb{E}[a]p_1(1-p_1) + bp_2(1-p_2)$$
$$\mathrm{Var}[\mathbb{E}[Y|X]] = p_1^2\mathrm{Var}[a]$$

Получаем нужное выражение для дисперсии. Для условия на a аналогично, в силу симметричности формулы и независимости a и b.
$\mathrm{Var}[c|b = \mathbb{E}[b]] = \mathbb{E}[a]p_1(1-p_1) + \mathbb{E}[b]p_2(1-p_2) + p_1^2\mathrm{Var}[a]$
$\mathrm{Var}[c|a = \mathbb{E}[a]] = \mathbb{E}[a]p_1(1-p_1) + \mathbb{E}[b]p_2(1-p_2) + p_2^2\mathrm{Var}[b]$
Рассмотрим разность, все константы после сокращения обозначим за c_1, c_2 и $c = \frac{c_2}{c_1}$ т.к дисперсии неотрицательны. Приравняв к нулю получим линейную функцию: $p_1 = \pm\sqrt{c}p_2$

5. Провести временные замеры по оценке всех необходимых распределений $p(c)$, $p(c|a)$, $p(c|b)$, $p(b|a)$, $p(b|d)$, $p(b|a,d)$, $p(d)$.

   Замеры времени с помощью **timeit**, в качестве значений для условий вся область определения c, a, b, d.

Распределение	Модель 1	Модель 2
$p(c)$	531 ms ± 10.9 ms	**31 ms ± 459 µs**
$p(d)$	690 ms ± 33.3 ms	**149 ms ± 4.59 ms**
$p(c\|a)$	531 ms ± 3.9 ms	**31.9 ms ± 581**
$p(c\|b)$	529 ms ± 6.98 ms	**31.8 ms ± 1.96 ms**
$p(b\|a)$	**3.62 µs ± 149 ns**	3.72 µs ± 307 ns
$p(b\|d)$	812 ms ± 30.9 ms	**281 ms ± 4.06 ms**
$p(b\|a,d)$	3.81 s ± 37.6 ms	**3.44 s ± 85.7 ms**

   Вычисления для второй модели происходят значительно быстрее. Вероятно это связано скорее с имлементацией, т.к. для вычисления $p(c|a,b)$ для второй модели можно обойтись одной функцией из **scipy**. Для первой же модели приходится совершать больше действий, т.к. не все из них можно полностью векторизовать (например сумму по k при $p(c = k)$). Точнее, можно, но это выйдет сильно дороже по памяти (за счет тензоров размерности ≥ 3). Дольше всего вычисялется распределение $p(b|a,d)$ как раз из-за тензоров большой размерности.

6. Используя результаты всех предыдущих пунктов, сравнить две модели. Показать где максимально проявляется разница между ними (привести конкретный пример, не обязательно из экспериментов выше). Объяснить причины подобного результата.

   При заданных параметрах модели практически не отличаются. Вторая модель оставляет больше плотности для хвостов распределения, что отражается на большей дисперсии. Однако, порядок отклонения для этой задачи (подсчет кол-во студентов) слишком мал, чтобы дать сильное смещение. Об этом говорит почти идентичное среднее (и чем более информации, тем меньше разница). С точки зрения имплементации также выгоднее использовать вторую модель, т.к. подсчеты для нее занимают меньше времени.