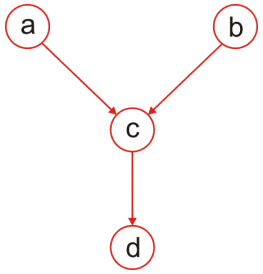
Задание 1. Байесовские рассуждения

Курс: Байесовские методы в машинном обучении 2020
Студент: Спирин Егор Сергеевич, HSE SPb MLDA

Вероятностные модели посещаемости курса

Рассмотрим модель посещаемости студентами ВУЗа одной лекции по курсу. Пусть аудитория данного курса состоит из студентов профильного факультета, а также студентов других факультетов. Обозначим через a количество студентов, поступивших на профильный факультет, а через b – количество студентов других факультетов. Пусть студенты профильного факультета посещают лекцию с некоторой вероятностью p_1, а студенты остальных факультетов – с вероятностью p_2. Обозначим через c количество студентов, посетивших данную лекцию. Тогда случайная величина $c|a,b$ есть сумма двух случайных величин, распределённых по биномиальному закону $\text{Bin}(a,p_1)$ и $\text{Bin}(b,p_2)$ соответственно. Пусть далее на лекции по курсу ведётся запись студентов. При этом каждый студент записывается сам, а также, быть может, записывает своего товарища, которого на лекции на самом деле нет. Пусть студент записывает своего товарища с некоторой вероятностью p_3. Обозначим через d общее количество записавшихся на данной лекции. Тогда случайная величина $d|c$ представляет собой сумму c и случайной величины, распределённой по биномиальному закону $\text{Bin}(c,p_3)$. Для завершения задания вероятностной модели осталось определить априорные вероятности для a и для b. Пусть обе эти величины распределены равномерно в своих интервалах $[a_{min},a_{max}]$ и $[b_{min},b_{max}]$ (дискретное равномерное распределение). Таким образом, мы определили следующую вероятностную модель:

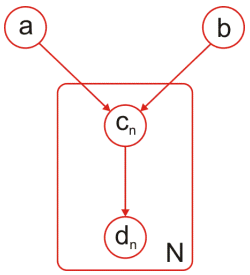


$$
\begin{aligned}
&p(a,b,c,d) = p(d|c)p(c|a,b)p(a)p(b),\\
&d|c \sim c + \text{Bin}(c,p_3),\\
&c|a,b \sim \text{Bin}(a,p_1) + \text{Bin}(b,p_2),\\
&a \sim \text{Unif}[a_{min},a_{max}],\\
&b \sim \text{Unif}[b_{min},b_{max}].
\end{aligned}
\tag{1}
$$

Рассмотрим несколько упрощённую версию модели 1. Известно, что биномиальное распределение $\text{Bin}(n,p)$ при большом количестве испытаний и маленькой вероятности успеха может быть с высокой точностью приближено пуассоновским распределением $\text{Poiss}(\lambda)$ с $\lambda = np$. Известно также, что сумма двух пуассоновских распределений с параметрами λ_1 и λ_2 есть пуассоновское распределение с параметром $\lambda_1 + \lambda_2$ (для биномиальных распределений это неверно). Таким образом, мы можем сформулировать вероятностную модель, которая является приближённой версией модели 1:

$$
\begin{aligned}
&p(a,b,c,d) = p(d|c)p(c|a,b)p(a)p(b),\\
&d|c \sim c + \text{Bin}(c,p_3),\\
&c|a,b \sim \text{Poiss}(ap_1 + bp_2),\\
&a \sim \text{Unif}[a_{min},a_{max}],\\
&b \sim \text{Unif}[b_{min},b_{max}].
\end{aligned}
\tag{2}
$$

Рассмотрим теперь модель посещений нескольких лекций курса. Будем считать, что посещения отдельных лекций являются независимыми. Тогда:



$$
\begin{aligned}
&p(a,b,c_1,\ldots,c_N,d_1,\ldots,d_N) = p(a)p(b)\prod_{n=1}^{N} p(d_n|c_n)p(c_n|a,b),\\
&d_n|c_n \sim c_n + \text{Bin}(c_n,p_3),\\
&c_n|a,b \sim \text{Bin}(a,p_1) + \text{Bin}(b,p_2),\\
&a \sim \text{Unif}[a_{min},a_{max}],\\
&b \sim \text{Unif}[b_{min},b_{max}].
\end{aligned}
\tag{3}
$$

По аналогии с моделью 2 можно сформулировать упрощённую модель для модели 3:

$$
\begin{aligned}
&p(a, b, c_1, \ldots, c_N, d_1, \ldots, d_N) = p(a)p(b)\prod_{n=1}^{N} p(d_n|c_n)p(c_n|a, b),\\
&d_n|c_n \sim c_n + \text{Bin}(c_n, p_3),\\
&c_n|a, b \sim \text{Poiss}(ap_1 + bp_2),\\
&a \sim \text{Unif}[a_{min}, a_{max}],\\
&b \sim \text{Unif}[b_{min}, b_{max}].
\end{aligned} \tag{4}
$$

Номер варианта

$$
f(\text{Spirin Egor}) = \sum\big(5, 7, 15, 18, 19, 16, 9, 18, 9, 14\big) \mod 3 + 1 = 130 \mod 3 + 1 = 2
$$

Вариант 2

Рассматриваются модели 1 и 2 с параметрами $a_{min} = 75$, $a_{max} = 90$, $b_{min} = 500$, $b_{max} = 600$, $p_1 = 0.1$, $p_2 = 0.01$, $p_3 = 0.3$. Провести следующие исследования для обеих моделей:

1. Вывести формулы для всех необходимых далее распределений аналитически.

$$P(a = k) = \frac{1}{a_{max} - a_{min} + 1}$$

$$P(b = k) = \frac{1}{b_{max} - b_{min} + 1}$$

$$P_{\text{Bin}}(c = k|a, b) = \sum_{i=0}^{k} P(\text{Bin}(a, p_1) = i) \cdot P(\text{Bin}(b, p_2) = k - i) =$$

$$= \sum_{i=0}^{k} \binom{i}{a} p_1^i (1 - p_1)^{a-i} \cdot \binom{k-i}{b} p_2^{k-i} (1 - p_2)^{b-k+i}$$

$$P_{\text{Poiss}}(c = k|a, b) = P(\text{Poiss}(ap_1 + bp_2) = k) = \frac{(ap_1 + bp_2)^k}{k!} e^{-ap_1 - bp_2}$$

$$P_{\text{Bin}}(c = k) = \sum_{a,b} P_{\text{Bin}}(c = k|a, b) \cdot P(a, b) = \sum_{a,b} P_{\text{Bin}}(c = k|a, b) \cdot P(a) \cdot P(b)$$

$$P_{\text{Poiss}}(c = k) = \sum_{a,b} P_{\text{Poiss}}(c = k|a, b) \cdot P(a, b) = \sum_{a,b} P_{\text{Poiss}}(c = k|a, b) \cdot P(a) \cdot P(b)$$

$$P(a|b) = P(a)$$

$$P(b|a) = P(b)$$

$$P(c = k|a) = \sum_{a,b} P(c = k|a, b) \cdot P(b|a) = \sum_{a,b} P(c = k|a, b) \cdot P(b)$$

$$P(c = k|b) = \sum_{a,b} P(c = k|a, b) \cdot P(a|b) = \sum_{a,b} P(c = k|a, b) \cdot P(a)$$

$$P(d = k) = \sum_{c} P(d = k|c) \cdot P(c) = \sum_{i=0}^{c_{max}} P(\text{Bin}(i, p_3) = k - i) \cdot P(c = i)$$

$$P(b|d) = \frac{P(d|b)P(b)}{P(d)} = \frac{\sum_c \big[P(d|c) \cdot P(c|b)\big] P(b)}{P(d)}$$

$$P(b|a, d) = \frac{\sum_c P(a, b, c, d)}{\sum_{b,c} P(a, b, c, d)} = \frac{\sum_c P(d|c)P(c|a, b)P(a)P(b)}{\sum_c P(d|c)P(c|a)P(a)} = \frac{\sum_c P(d|c)P(c|a, b)P(b)}{\sum_c P(d|c)P(c|a)}$$

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

$$\mathbb{E}(a) = \sum_a (a \cdot P(a)) = \frac{a_{min} + a_{max}}{2}$$

$$\mathbb{D}(a) = \mathbb{E}(a^2) - (\mathbb{E}(a))^2 = \frac{(a_{max} - a_{min} + 1)^2 - 1}{12}$$

$$\mathbb{E}(b) = \sum_b (b \cdot P(b)) = \frac{b_{min} + b_{max}}{2}$$

$$\mathbb{D}(b) = \mathbb{E}(b^2) - (\mathbb{E}(b))^2 = \frac{(b_{max} - b_{min} + 1)^2 - 1}{12}$$

$$\mathbb{E}(c) = \sum_c (c \cdot P(c)) = \sum_{k=0}^{a_{max}+b_{max}} k \cdot P(c = k)$$

$$\mathbb{D}(c) = \sum_{k=0}^{a_{max}+b_{max}} [k^2 \cdot P(c = k)] - (\mathbb{E}\, c)^2$$

$$\mathbb{E}(d) = \sum_d (d \cdot P(d)) = \sum_{k=0}^{2\cdot(a_{max}+b_{max})} k \cdot P(d = k)$$

$$\mathbb{D}(d) = \sum_{k=0}^{2\cdot(a_{max}+b_{max})} [k^2 \cdot P(d = k)] - (\mathbb{E}\, d)^2$$

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.

   Рисунок 2 показывает необходимую зависимость. Графики $p(b)$ и $p(b|a)$ совпадают так как события являются независимыми. Для распределений $p(b|d)$ и $p(b|a, d)$ можно заметить, что добавление косвенной информации повышает вероятность значения b, близкого к истинному мат. ожиданию. При этом использование второй модели ухудшает приближение. Аналогичные выводы можно сделать и из таблицы 1 с мат.ожиданиями и дисперсиями этих распределений.

	Model 1		Model 2	
	$\mathbb{E}$	$\mathbb{D}$	$\mathbb{E}$	$\mathbb{D}$
$p(b)$	550.0	850.0	550.0	850.0
$p(b\|a = \mathbb{E}(a))$	550.0	850.0	550.0	850.0
$p(b\|d = \mathbb{E}(d))$	550.07	848.04	550.1	848.13
$p(b\|a = \mathbb{E}(a), d = \mathbb{E}(d))$	550.04	848.03	550.06	848.12

Таблица 1: Мат. ожидание и дисперсия рассматриваемых распределений для первой и второй моделей.

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!

   Рисунок 1 показывает как меняется важность параметров a и b для оценки величины c при изменении p_1 и p_2. Из рисунка кажется, что множества линейно-разделимы, покажем это формально.

   Из свойств условной дисперсии:

$$\mathbb{D}[c|a] = \mathbb{E}\left(\mathbb{D}[c|a, b]\right) + \mathbb{D}\left(\mathbb{E}[c|a, b]\right)$$

   Распишем по отдельности слагаемые, воспользуемся независимостью a и b:

$$\mathbb{D}[c|a, b] = \mathbb{D}[\mathrm{Bin}(a, p_1) + \mathrm{Bin}(b, p_2)] = ap_1(1 - p_1) + bp_2(1 - p_2)$$

$$\mathbb{E}_b\left(\mathbb{D}[c|a, b]\right) = ap_1(1 - p_1) + \mathbb{E}[b]p_2(1 - p_2)$$

$$\mathbb{E}[c|a, b] = \mathbb{E}[\mathrm{Bin}(a, p_1) + \mathrm{Bin}(b, p_2)] = ap_1 + bp_2$$

$$\mathbb{D}_b\left(\mathbb{E}[c|a, b]\right) = \mathbb{D}[b]p_2^2$$

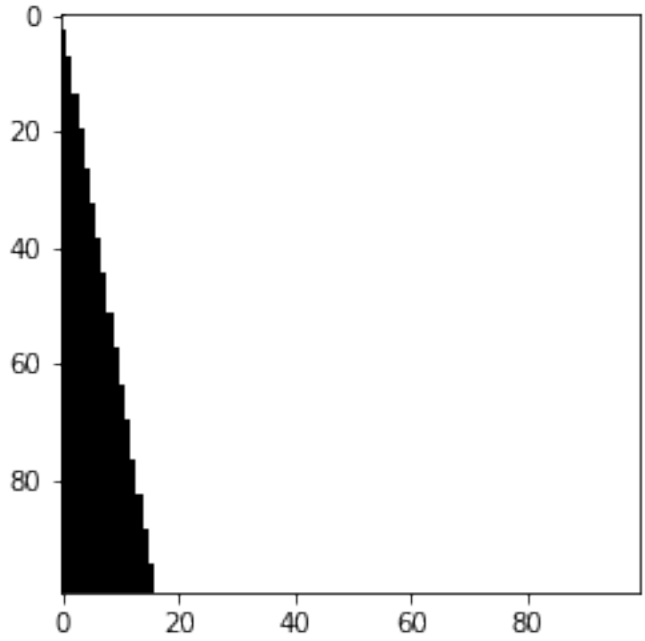


Рис. 1: Сравнение $\mathbb{D}[c|b]$ и $\mathbb{D}[c|a]$ при разных p_1 и p_2. Чёрная часть соответствует $\mathbb{D}[c|b] < \mathbb{D}[c|a]$.

Таким образом (C_1 и C_2 константы):

$$\mathbb{D}[c|a] = ap_1(1-p_1) + \mathbb{E}[b]p_2(1-p_2) + \mathbb{D}[b]p_2^2 = ap_1(1-p_1) + C_1 + C_2$$

Аналогично можно получить для $\mathbb{D}[c|b]$:

$$\mathbb{D}[c|b] = bp_2(1-p_2) + \mathbb{E}[a]p_1(1-p_1) + \mathbb{D}[a]p_1^2 = bp_2(1-p_2) + C_3 + C_4$$

Если $\mathbb{D}[c|a]$ и $\mathbb{D}[c|b]$ линейно-разделимы, то $\mathbb{D}[c|a] - \mathbb{D}[c|b] = 0$ является уравнением прямой.

$\mathbb{D}[c|a] - \mathbb{D}[c|b] = 0$
$ap_1(1-p_1) + C_1 + C_2 - bp_2(1-p_2) - C_3 - C_4$
$ap_1(1-p_1) + C = bp_2(1-p_2)$
$\frac{p_1(1-p_1)}{p_2(1-p_2)}a + \frac{C}{p_2(1-p_2)} = b$
$ka + c = b$

Таким образом, получили уравнение прямой. Значит множества действительно являются линейно-разделимыми.

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

	Model 1 mean time	Model 2 mean time
$p(c)$	430.03 ± 9.22 ms	28.22 ± 1.47 ms
$p(c\|a)$	454.81 ± 12.47 ms	26.22 ± 1.79 ms
$p(c\|b)$	440.81 ± 14.91 ms	27.69 ± 1.27 ms
$p(b\|a)$	11.73 ± 5.8 θs	15.0 ± 7.93 θs
$p(b\|d)$	520.08 ± 20.67 ms	113.98 ± 2.96 ms
$p(b\|a,d)$	2872.17 ± 157.69 ms	1892.33 ± 35.39 ms
$p(b)$	488.68 ± 14.86 ms	72.67 ± 1.64 ms

Таблица 2: Замеры времени вычислений распределений для первой и второй модели.

Как видно из таблицы 2, вычисление распределений для второй модели осуществляется гораздо быстрее. Скорее всего дело в имплементации, для вычисления многих распределений надо считать $p(c|a,b)$, однако именно для модели с пуассоновским распределением получилось полностью его векторизовать.

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.

   Модели практически не отличаются, по крайней мере для заданных параметров. Можно увидеть, что использование распределения Пуассона позволяет проводить вычисления гораздо быстрее. Возникающие при этом смещение, как видно из мат.ожидания, пренебрижительно мало.

Взять в качестве диапазона допустимых значений для величины c интервал $[0, a_{max} + b_{max}]$, а для величины d – интервал $[0, 2(a_{max} + b_{max})]$.

Исследование должно быть выполнено на компьютере, однако за дополнительные аналитические выкладки в пунктах 2-4 будут ставиться дополнительные баллы. При оценке выполнения задания будет учитываться эффективность программного кода - любая из функций должна работать быстрее секунды на скалярных входах (для этого код должен реализовываться векторно). По всем пунктам задания должен быть проведен анализ результатов и сделаны выводы.

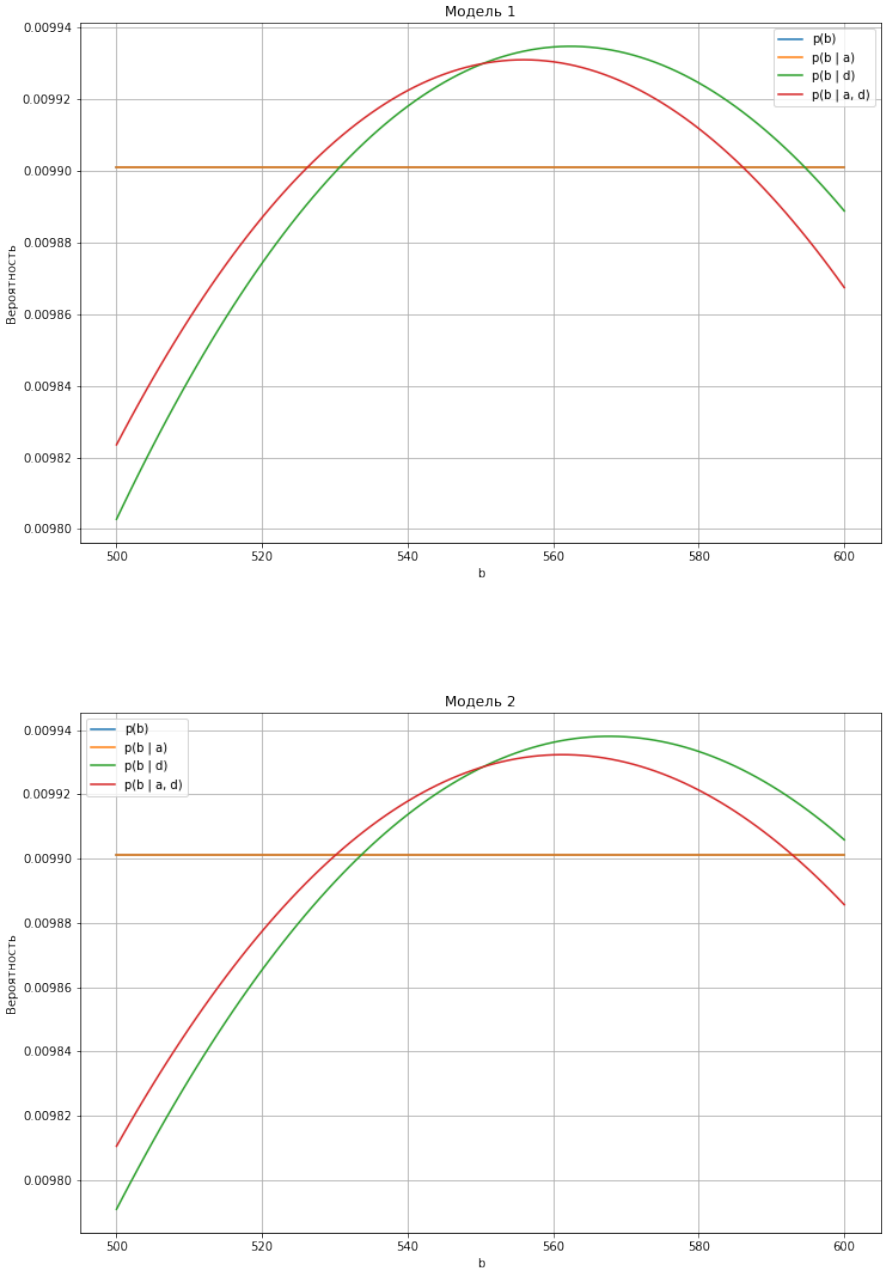


Рис. 2: Оценка влияния косвенной информации на величину b при использование первой и второй моделей.