Практическое задание по теме «Байесовские рассуждения»

Солоткий Михаил, 417 группа ВМК МГУ

25 сентября 2018 г.

1 Вывод формул в модели посещаемости

1.1 Модель (1)

Рассмотрим вероятностную модель

$$\mathsf{P}(a,b,c,d) = \mathsf{P}(d|c)\,\mathsf{P}(c|a,b)\,\mathsf{P}(a)\,\mathsf{P}(b),$$

$$d|c \sim c + \mathrm{Bin}(c,p_3),$$

$$c|a,b \sim \mathrm{Bin}(a,p_1) + \mathrm{Bin}(b,p_2),$$

$$a \sim \mathrm{Unif}[a_{min}, a_{max}],$$

$$b \sim \mathrm{Unif}[b_{min}, b_{max}].$$

В варианте №2 задания предлагается вычислить следующие распределения: $\mathsf{P}(a)$, $\mathsf{P}(b)$, $\mathsf{P}(c)$, $\mathsf{P}(d)$, $\mathsf{P}(b|a)$, $\mathsf{P}(c|a)$, $\mathsf{P}(c|b)$, $\mathsf{P}(b|d)$, $\mathsf{P}(b|a,d)$.

- $\mathsf{P}(a) = \dfrac{1}{a_{max} - a_{min} + 1}$
- $\mathsf{P}(b) = \dfrac{1}{b_{max} - b_{min} + 1}$
- $\mathsf{P}(c) = \sum\limits_{a,b} \mathsf{P}(c|a,b) \cdot \mathsf{P}(a,b) = \sum\limits_{a,b} \mathsf{P}(c|a,b) \cdot \mathsf{P}(a) \cdot \mathsf{P}(b)$

  $\mathsf{P}(c=k|a,b) = \sum\limits_{i=0}^{k} \mathsf{P}(\ \mathrm{Bin}(a,p_1) = i) \cdot \mathsf{P}(\mathrm{Bin}(b,p_2) = k-i)$

  $\mathsf{P}(c=k) = \sum\limits_{a=a_{min}}^{a_{max}} \sum\limits_{b=b_{min}}^{b_{max}} \sum\limits_{i=0}^{k} \mathsf{P}(\mathrm{Bin}(a,p_1) = i) \cdot \mathsf{P}(\mathrm{Bin}(b,p_1) = k-i) \cdot \mathsf{P}(a) \cdot \mathsf{P}(b)$

  $\mathsf{P}(c=k) = \sum\limits_{a=a_{min}}^{a_{max}} \sum\limits_{b=b_{min}}^{b_{max}} \Big[\sum\limits_{i=0}^{k} \binom{a}{i} \cdot p_1^i \cdot (1-p_1)^{a-i} \cdot \binom{b}{k-i} \cdot p_2^i \cdot (1-p_2)^{b-k+i} \Big] \mathsf{P}(b)\,\mathsf{P}(a)$
- $\mathsf{P}(d) = \sum\limits_{c} \cdot \mathsf{P}(d|c) \cdot \mathsf{P}(c)$

  $\mathsf{P}(d=k|c) = \mathsf{P}(\mathrm{Bin}(c,p_3) = k-c)$

  $\mathsf{P}(d=k) = \sum\limits_{c=0}^{a_{max}+b_{max}} \mathsf{P}(\mathrm{Bin}(c,p_3) = k-c) \cdot \mathsf{P}(c)$
- $\mathsf{P}(b|a) = \mathsf{P}(b)$
- $\mathsf{P}(c|a) = \sum\limits_{b} \mathsf{P}(c|a,b) \cdot \mathsf{P}(b|a) = \sum\limits_{b} \mathsf{P}(c|a,b) \cdot \mathsf{P}(b)$

  $\mathsf{P}(c=k|a) = \sum\limits_{b=b_{min}}^{b_{max}} \Big[\sum\limits_{i=0}^{k} \mathsf{P}(\ \mathrm{Bin}(a,p_1) = i) \cdot \mathsf{P}(\mathrm{Bin}(b,p_2) = k-i) \Big] \cdot \mathsf{P}(b)$
- $\mathsf{P}(c|b) = \sum\limits_{a} \mathsf{P}(c|a,b) \cdot \mathsf{P}(a|b) = \sum\limits_{a} \mathsf{P}(c|a,b) \cdot \mathsf{P}(a)$

  $\mathsf{P}(c=k|b) = \sum\limits_{a=a_{min}}^{a_{max}} \Big[\sum\limits_{i=0}^{k} \mathsf{P}(\ \mathrm{Bin}(a,p_1) = i) \cdot \mathsf{P}(\mathrm{Bin}(b,p_2) = k-i) \Big] \cdot \mathsf{P}(a)$
- $\mathsf{P}(b|d) = \dfrac{\mathsf{P}(d|b) \cdot \mathsf{P}(b)}{\mathsf{P}(d)} = \dfrac{\Big[\sum\limits_{c} \mathsf{P}(d|b,c) \cdot \mathsf{P}(c|b) \Big] \cdot \mathsf{P}(b)}{\mathsf{P}(d)} = \dfrac{\Big[\sum\limits_{c} \mathsf{P}(d|c) \cdot \mathsf{P}(c|b) \Big] \cdot \mathsf{P}(b)}{\mathsf{P}(d)}$

- $\mathsf{P}(b|a,d) = \dfrac{\sum\limits_c \mathsf{P}(d|c)\cdot\mathsf{P}(c|a,b)\cdot\mathsf{P}(a)\cdot\mathsf{P}(b)}{\sum\limits_{b,c} \mathsf{P}(d|c)\cdot\mathsf{P}(c|a,b)\cdot\mathsf{P}(a)\cdot\mathsf{P}(b)}$

  $\sum\limits_{b,c} \mathsf{P}(d|c)\cdot\mathsf{P}(c|a,b)\cdot\mathsf{P}(a)\cdot\mathsf{P}(b) = \sum\limits_c \mathsf{P}(d|c)\Big[\sum\limits_b \mathsf{P}(c|a,b)\cdot\mathsf{P}(a)\cdot\mathsf{P}(b)\Big] = \sum\limits_c \mathsf{P}(d|c)\Big[\sum\limits_b \mathsf{P}(a,b,c)\Big] = \sum\limits_c \mathsf{P}(d|c)\,\mathsf{P}(a,c) =$
  $= \sum\limits_c \mathsf{P}(d|c)\cdot\mathsf{P}(c|a)\cdot\mathsf{P}(a)$

  $\mathsf{P}(b|a,d) = \dfrac{\sum\limits_c \mathsf{P}(d|c)\cdot\mathsf{P}(c|a,b)\cdot\mathsf{P}(a)\cdot\mathsf{P}(b)}{\sum\limits_c \mathsf{P}(d|c)\cdot\mathsf{P}(c|a)\cdot\mathsf{P}(a)} = \dfrac{\sum\limits_c \mathsf{P}(d|c)\cdot\mathsf{P}(c|a,b)\cdot\mathsf{P}(b)}{\sum\limits_c \mathsf{P}(d|c)\cdot\mathsf{P}(c|a)}$

1.2 Модель (2)

Теперь изменим немного модель: заменим распределение $c|a,b$ на $\text{Poiss}(ap_1 + bp_2)$. Тогда поменяются формулы для $\mathsf{P}(c)$, $\mathsf{P}(c|a)$, $\mathsf{P}(c|b)$.

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

  $\mathsf{P}(c=k) = \sum\limits_{a=a_{min}}^{a_{max}} \sum\limits_{b=b_{min}}^{b_{max}} \dfrac{e^{-ap_1+bp_2}\cdot(ap_1+bp_2)^k}{k!}\cdot\mathsf{P}(b)\cdot\mathsf{P}(a)$

- $\mathsf{P}(c|a) = \sum\limits_b \mathsf{P}(c|a,b)\cdot\mathsf{P}(b|a) = \sum\limits_b \mathsf{P}(c|a,b)\cdot\mathsf{P}(b)$

  $\mathsf{P}(c=k|a) = \sum\limits_{b=b_{min}}^{b_{max}} \mathsf{P}(\text{Poiss}(ap_1+bp_2)=k)\cdot\mathsf{P}(b)$

- $\mathsf{P}(c|b) = \sum\limits_a \mathsf{P}(c|a,b)\cdot\mathsf{P}(a|b) = \sum\limits_a \mathsf{P}(c|a,b)\cdot\mathsf{P}(a)$

  $\mathsf{P}(c=k|b) = \sum\limits_{a=a_{min}}^{a_{max}} \mathsf{P}(\text{Poiss}(ap_1+bp_2)=k)\cdot\mathsf{P}(a)$

2 Вывод математических ожиданий и дисперсий случайных величин

- $\mathsf{E}(a) = \sum\limits_a \Big[a\cdot\mathsf{P}(a)\Big] = \sum\limits_{a=a_{min}}^{a_{max}} a\cdot\dfrac{1}{a_{max}-a_{min}+1} = \dfrac{1}{a_{max}-a_{min}+1}\cdot\sum\limits_{a=a_{min}}^{a_{max}} a =$
  $= \dfrac{1}{a_{max}-a_{min}+1}\cdot\left(\sum\limits_{a=1}^{a_{max}} a - \sum\limits_{a=1}^{a_{min}-1} a\right) = \dfrac{1}{a_{max}-a_{min}+1}\cdot\left(\dfrac{a_{max}\cdot(a_{max}+1)}{2} - \dfrac{(a_{min}-1)\cdot a_{min}}{2}\right) =$
  $= \dfrac{1}{a_{max}-a_{min}+1}\cdot\left(\dfrac{a_{max}^2}{2}+\dfrac{a_{max}}{2}-\dfrac{a_{min}^2}{2}+\dfrac{a_{min}}{2}\right) = \dfrac{1}{a_{max}-a_{min}+1}\cdot$
  $\cdot\left(\dfrac{(a_{max}-a_{min})\cdot(a_{max}+a_{min})}{2}+\dfrac{a_{max}+a_{min}}{2}\right) = \dfrac{a_{max}+a_{min}}{2\cdot(a_{max}-a_{min}+1)}\cdot(a_{max}-a_{min}+1) =$
  $\dfrac{a_{max}+a_{min}}{2}$

  $\mathsf{D}(a) = \mathsf{E}(a^2) - \mathsf{E}(a)^2 = \dfrac{1}{a_{max}-a_{min}+1}\cdot\sum\limits_{a=a_{min}}^{a_{max}} a^2 - \dfrac{1}{(a_{max}-a_{min}+1)^2}\cdot\left(\sum\limits_{a=a_{min}}^{a_{max}} a\right)^2 =$
  $= \dfrac{(a_{max}-a_{min}+1)^2-1}{12}$

- $\mathsf{E}(b) = \dfrac{b_{max}+b_{min}}{2}$

  $\mathsf{D}(b) = \dfrac{(b_{max}-b_{min}+1)^2-1}{12}$

- $c_{max} = a_{max} + b_{max}$

  $\mathsf{E}(c) = \sum\limits_{c=0}^{c_{max}} c\cdot\mathsf{P}(c)$

  $\mathsf{D}(c) = \mathsf{E}(c^2) - \mathsf{E}(c)^2 = \sum\limits_{c=0}^{c_{max}} \Big[c^2\cdot\mathsf{P}(c)\Big] - \Big(\sum\limits_{c=0}^{c_{max}} c\cdot\mathsf{P}(c)\Big)^2$

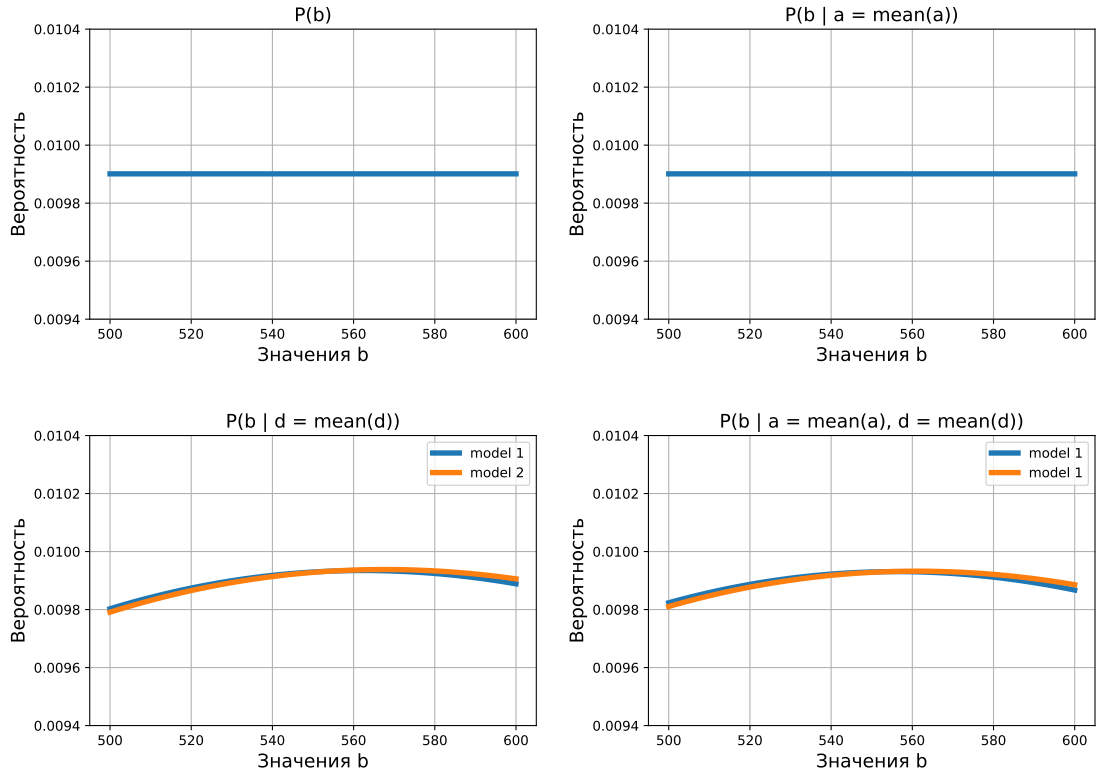
- $d_{max} = 2 \cdot c_{max}$

  $\mathsf{E}(d) = \sum\limits_{d=0}^{d_{max}} d \cdot \mathsf{P}(d) = \sum\limits_{d=0}^{d_{max}} \sum\limits_{c=0}^{c_{max}} d \cdot \mathsf{P}(\mathrm{Bin}(c, p_3) = d - c) \cdot \mathsf{P}(c)$

  $\mathsf{D}(d) = \mathsf{E}(d^2) - \mathsf{E}(d)^2 = \sum\limits_{d=0}^{d_{max}} \left[d^2 \cdot \mathsf{P}(d)\right] - \left(\sum\limits_{d=0}^{d_{max}} d \cdot \mathsf{P}(d)\right)^2$

3 Влияние косвенной информации на прогноз величиины b

Ниже приведены графики распределений величины b.

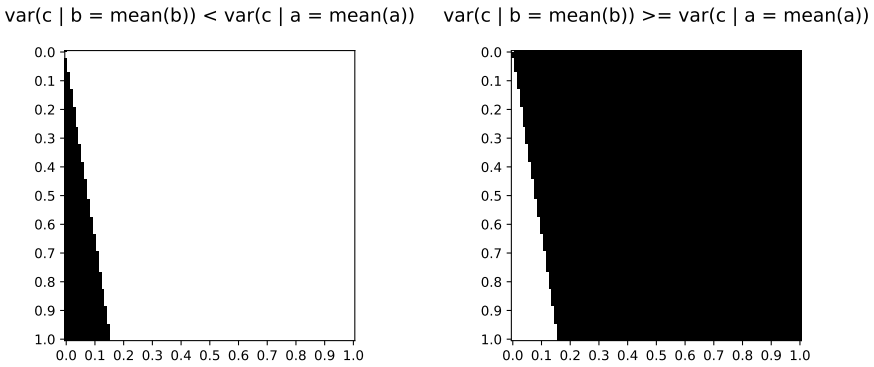


Распределения $\mathsf{P}(b)$ и $\mathsf{P}(b|a)$ не отличаются, так как величины a и b независимы. Распределения $\mathsf{P}(b \mid d = \mathsf{E}(d))$ и $\mathsf{P}(b \mid a = \mathsf{E}(a), d = \mathsf{E}(d))$ отличаются не сильно, как можно видеть на графиках. Различия в вероятностях при одинаковых значениях b между ними порядка 10^{-5}. Как видно, само по себе наличие косвенной информации (зависимой) чуть-чуть меняет вид распределения, то есть более ярко выражаются наиболее правдоподобные значения b. Уточнение происходит относительно небольшое, такой же вывод можно сделать по значениям дисперсии – она не сильно уменьшилась. Ещё можно заметить, что для каждой отдельной модели добавление величины a в качестве условия уменьшает дисперсию и меняет мат. ожидание.

Величина	Модель	Мат. ожидание	Дисперсия
b	1, 2	550	850
$b \mid d = \mathsf{E}(d)$	1	550.07	848.037
$b \mid d = \mathsf{E}(d)$	2	550.09	848.128
$b \mid a = \mathsf{E}(a), d = \mathsf{E}(d)$	1	550.03	**848.031**
$b \mid a = \mathsf{E}(a), d = \mathsf{E}(d)$	2	550.06	**848.123**

4 Влияние параметров вероятностей на относительную важность параметров a и b для оценки велчины c

Проведём вычисление значений $\mathsf{D}(c \mid b = \mathsf{E}(b))$ и $\mathsf{D}(c \mid a = \mathsf{E}(a))$ при различных значениях параметров p_1, p_2 вероятностей бинмиальных распределений в модели (1). На равномерной сетке квадрата $[0, 1] \times [0, 1]$ (по 100 значений для каждой размерности). Ниже приведены графики соотношений дисперсий, белым обозначена область, где условие выполняется. По графикам кажется, что множества линейно разделимы.



5 Доказательство линейной разделимости

Формула полной дисперсии:

$$\mathsf{D}(Y) = \mathsf{E}(\mathsf{D}(Y|X)) + \mathsf{D}(\mathsf{E}(Y|X))$$

Возьмём $Y = c|b$, $X = a|b$
$\mathsf{D}(Y|X) = \mathsf{D}(c|a,b) = \mathsf{D}(\mathrm{Bin}(a,p_1) + \mathrm{Bin}(b,p_2)) = a \cdot p_1 \cdot (1-p_1) + b \cdot p_2 \cdot (1-p_2)$
$\mathsf{E}(Y|X) = \mathsf{E}(c|a,b) = \mathsf{E}(\mathrm{Bin}(a,p_1) + \mathrm{Bin}(b,p_2)) = p \cdot p_1 + b \cdot p_2$
$\mathsf{E}(\mathsf{D}(Y|X)) = \mathsf{E}(a) \cdot p_1 \cdot (1-p_1) + b \cdot p_2 \cdot (1-p_2)$
$\mathsf{D}(\mathsf{E}(Y|X)) = \mathsf{D}(a) \cdot p_1$

$\mathsf{D}(c|b = \mathsf{E}(b)) = \mathsf{E}(a) \cdot p_1 \cdot (1-p_1) + \mathsf{E}(b) \cdot p_2 \cdot (1-p_2) + \mathsf{D}(a) \cdot p_1$
В силу симметрии:
$\mathsf{D}(c|a = \mathsf{E}(a)) = \mathsf{E}(a) \cdot p_1 \cdot (1-p_1) + \mathsf{E}(b) \cdot p_2 \cdot (1-p_2) + \mathsf{D}(b) \cdot p_2$

Пусть $c_1 = \mathsf{D}(a), c_2 = \mathsf{D}(b)$, заметим, что $c_1 > 0, c_2 > 0$.
$f(p_1,p_2) = D(c|b = \mathsf{E}(b)) - D(c|a = \mathsf{E}(a)) = c_1 \cdot p_1 - c_2 \cdot p_2$
Если приравнять к 0, получим уравнение прямой, которая разбивает плоскость на 2 полуплоскости, причём в одной $f > 0$, а в другой $f < 0$. Заметим, что на квадрате $[0,1] \times [0,1]$ тоже есть точки, где функция f разного знака.

6 Замеры времени

Ниже приведены замеры времени вычисления распределений. Для величин, стоящих в условиях вычислялись вероятности для всех значений от минимального до максимального, то есть, к примеру $0 \le d \le 2 \cdot (a_{max} + b_{max})$. Замеры проводились с помощью утилиты **timeit** ipython notebook, которая запускает код 10 раз и оценивает по ним стандартное отклонение и среднее затраченного времени.

Распределение	c	$c\|a$	$c\|b$	$b\|a$	$b\|d$	$b\|a,d$	d
Время вычисления 1 модели: t_1	99.2 ms	111 ms	108 ms	**2.48 μs**	233 ms	2.42 s	165 ms
Стандартное отклонение t_1	1.46 ms	2.74 ms	1.35 ms	114 ns	13.9 ms	198 ms	2.18 ms
Время вычисления 2 модели: t_2	**59.7 ms**	**58.6 ms**	**63.8 ms**	2.63 μs	**172 ms**	**2.19 s**	**115 ms**
Стандартное отклонение t_2	1.55 ms	345 μs	2.31 ms	72.2 ns	4.29 ms	41 ms	2.17 ms

Можно заметить, что распределние $b \mid a,d$ вычисляется сильно дольше, чем остальные, но это и понятно, ведь оно единственное выводилось по определению условной вероятности и разложению совместного распределения. В процессе вычисления появлялись трёхмерные тензоры, в которых надо поменять некоторые размерности перед умножением. Быстрее всего вычислялось $b \mid a$, так как величины a и b независимы, а b распределено равномерно, то есть достаточно было заполнить матрицу одинаковым значением. Почти все распределения второй модели вычисляются быстрее соттветствующих распределений первой модели. В одном случае получилось наоборот, но это не стат. значимый результат.

7 Сравнение моделей

Как видно из графиков распределений величины $b \mid a$ и величины $b \mid a,d$ для разных моделей разница в вероятностях незначительная, это значит, что можно не считать биномиальные распределения

и приближать их распределениями Пуассона (в случае конкретных использовавшихся в экспериментах параметрах вероятностей $p_1 = 0.1$ и $p_2 = 0.01$). Исходя из проведённых экспериментов нельзя сказать, что какая-то из моделей существенно лучше предсказывает какие-либо величины. Исходя из на практике использовать модель (2) предпочтительней (не факт, что тоже самое можно сказать в случае других значений параметров вероятностей посещения).