Отчёт по практическому заданию по БММО «Байесовские рассуждения»

Вариант 1

Аят Оспанов
617 гр., ММП, ВМК МГУ, Москва
20 сентября 2017 г.

Содержание

1 Математические ожидания и дисперсии априорных распределений

1.1 $p(a)$, $p(b)$

По условию $a \sim \text{Unif}[a_{min}, a_{max}]$, $b \sim \text{Unif}[b_{min}, b_{max}]$. Тогда матожидания и дисперсии считаются по определению:

$$\mathbb{E}a = \frac{a_{min} + a_{max}}{2} \tag{1}$$

$$\mathbb{E}b = \frac{b_{min} + b_{max}}{2} \tag{2}$$

$$\mathbb{D}a = \frac{(a_{max} - a_{min} + 1)^2 + 1}{12} \tag{3}$$

$$\mathbb{D}b = \frac{(b_{max} - b_{min} + 1)^2 + 1}{12} \tag{4}$$

1.2 $p(c)$, $p(d)$

Воспользуемся следующими свойствами условных матожидания и дисперсии:

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

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

По условию, для модели 1, $c|a, b \sim \text{Bin}(a, p_1) + \text{Bin}(b, p_2)$; $d|c \sim c + \text{Bin}(c, p_3)$. Тогда:

$$\mathbb{E}c = \mathbb{E}_{a,b}\mathbb{E}_c[c|a, b] = \mathbb{E}_{a,b}[ap_1 + bp_2] = p_1\mathbb{E}a + p_2\mathbb{E}b \tag{5}$$

$$\mathbb{E}d = \mathbb{E}_c\mathbb{E}_d[d|c] = \mathbb{E}_c[c + cp_3] = \mathbb{E}[c] + \mathbb{E}[cp_3] = (1 + p_3)\mathbb{E}c \tag{6}$$

$$\begin{aligned}\mathbb{D}c = \mathbb{E}_{a,b}\mathbb{D}_c[c|a, b] + \mathbb{D}_{a,b}\mathbb{E}_c[c|a, b] = \mathbb{E}_{a,b}[ap_1(1 - p_1) + bp_2(1 - p_2)] + \mathbb{D}_{a,b}[ap_1 + bp_2] = \\ = p_1(1 - p_1)\mathbb{E}a + p_2(1 - p_2)\mathbb{E}b + p_1^2\mathbb{D}a + p_2^2\mathbb{D}b\end{aligned} \tag{7}$$

$$\begin{aligned}\mathbb{D}d = \mathbb{E}_c\mathbb{D}_d[d|c] + \mathbb{D}_c\mathbb{E}_d[d|c] = \mathbb{E}_c[0 + cp_3(1 - p_3)] + \mathbb{D}_c[c + cp_3] = \\ = p_3(1 - p_3)\mathbb{E}c + (1 + p_3)^2\mathbb{D}c\end{aligned} \tag{8}$$

Для модели 2 $c|a, b \sim \text{Poiss}(ap_1 + bp_2)$. При этих условиях меняется только дисперсия:

$$\begin{aligned}\mathbb{D}c = \mathbb{E}_{a,b}\mathbb{D}_c[c|a, b] + \mathbb{D}_{a,b}\mathbb{E}_c[c|a, b] = \mathbb{E}_{a,b}[ap_1 + bp_2] + \mathbb{D}_{a,b}[ap_1 + bp_2] = \\ = p_1\mathbb{E}a + p_2\mathbb{E}b + p_1^2\mathbb{D}a + p_2^2\mathbb{D}b\end{aligned} \tag{9}$$

1.3 Численные значения

	Модель 1	Модель 2
$\mathbb{E}a$	82.5	82.5
$\mathbb{E}b$	550.0	550.0
$\mathbb{E}c$	13.75	13.75
$\mathbb{E}d$	17.875	17.875
$\mathbb{D}a$	21.25	21.25
$\mathbb{D}b$	850.0	850.0
$\mathbb{D}c$	13.1675	14.0475
$\mathbb{D}d$	25.1405	26.6277

2 Уточнение прогноза для величины c по мере прихода новой косвенной информации

2.1 Распределение $p(c)$

$$p(c) = \sum_{a=a_{min}}^{a_{max}} \sum_{b=b_{min}}^{b_{max}} p(a,b,c) = \sum_{a=a_{min}}^{a_{max}} \sum_{b=b_{min}}^{b_{max}} p(c|a,b)p(a,b) = \sum_{a=a_{min}}^{a_{max}} \sum_{b=b_{min}}^{b_{max}} p(c|a,b)p(a)p(b) =$$

$$= p(a)p(b) \sum_{a=a_{min}}^{a_{max}} \sum_{b=b_{min}}^{b_{max}} p(c|a,b) = \{\text{формула свертки + смена порядка суммирования}\} =$$

$$= p(a)p(b) \sum_{k=0}^{a_{max}+b_{max}} \sum_{a=a_{min}}^{a_{max}} p_A(k) \sum_{b=b_{min}}^{b_{max}} p_B(a_{max} + b_{max} - k). \tag{10}$$

Где:

$$A \sim \text{Bin}(a, p_1), B \sim \text{Bin}(b, p_2) \text{ (для 1 модели)}$$

$$A \sim \text{Poiss}(ap_1), B \sim \text{Poiss}(bp_2) \text{ (для 2 модели)}$$

2.2 Распределение $p(c|a)$

$$p(c|a) = \sum_{b=b_{min}}^{b_{max}} p(c|a,b)p(b) = p(b) \sum_{b=b_{min}}^{b_{max}} p(c|a,b) =$$

$$\{\text{аналогично с } p(c)\} = p(b) \sum_{k=0}^{a_{max}+b_{max}} p_A(k) \sum_{b=b_{min}}^{b_{max}} p_B(a_{max} + b_{max} - k). \tag{11}$$

2.3 Распределение $p(c|b)$

Аналогично с $p(c|a)$

$$p(c|b) = p(b) \sum_{k=0}^{a_{max}+b_{max}} p_B(a_{max} + b_{max} - k) \sum_{a=a_{min}}^{a_{max}} p_A(k). \tag{12}$$

2.4 Распределение $p(c|d)$

$$p(c|d) = \frac{p(d|c)p(c)}{p(d)} \propto p(d|c)p(c) \tag{13}$$

2.5 Распределение $p(c|a, b, d)$

$$p(c|a, b, d) = \frac{p(a, b, c, d)}{p(a, b, d)} = \frac{p(d|c)p(c|a, b)p(a)p(b)}{p(a, b, d)} \propto p(d|c)p(c|a, b) \tag{14}$$

2.6 Наблюдение

Из Рис. 1 видно, что добавление информации о количестве студентов (a, b) не уточняет прогноз для величины c. Но добавление информации о записавшихся на лекцию (d) существенно уточняет прогноз. Также видна похожесть графиков $p(c|d)$ и $p(c|a, b, d)$, что подтверждает, что a и b не влияют на прогноз.

Из Таблиц 1 и 2 можно понять насколько идет уточнение. В случае добавления a и b дисперсия уменьшается по сравнению с $p(c)$, но меняется незначительно. А при добавлении d дисперсия значительно снижается, что показывает хорошее уточнение прогноза.

Таблица 1: Модель 1

	p(c)	**p(c|a)**	**p(c|b)**	**p(c|d)**	**p(c|a,b)**	**p(c|a,b,d)**
Матожидание	13.7500	13.8	13.7500	13.895971	13.800	13.902756
Дисперсия	13.1675	13.0	13.0825	1.533582	12.915	1.530140

Таблица 2: Модель 2

	p(c)	**p(c|a)**	**p(c|b)**	**p(c|d)**	**p(c|a,b)**	**p(c|a,b,d)**
Матожидание	13.7500	13.8	13.7500	13.893834	13.8	13.900175
Дисперсия	14.0475	13.885	13.9625	1.543943	13.8	1.540884

Также по таблицам видно, что 2 модель показывает примерно те же матожидания, что и 1 модель, но дисперсии чуть больше, т.к. 2 модель является приближением 1 модели, что приводит к потере точности.

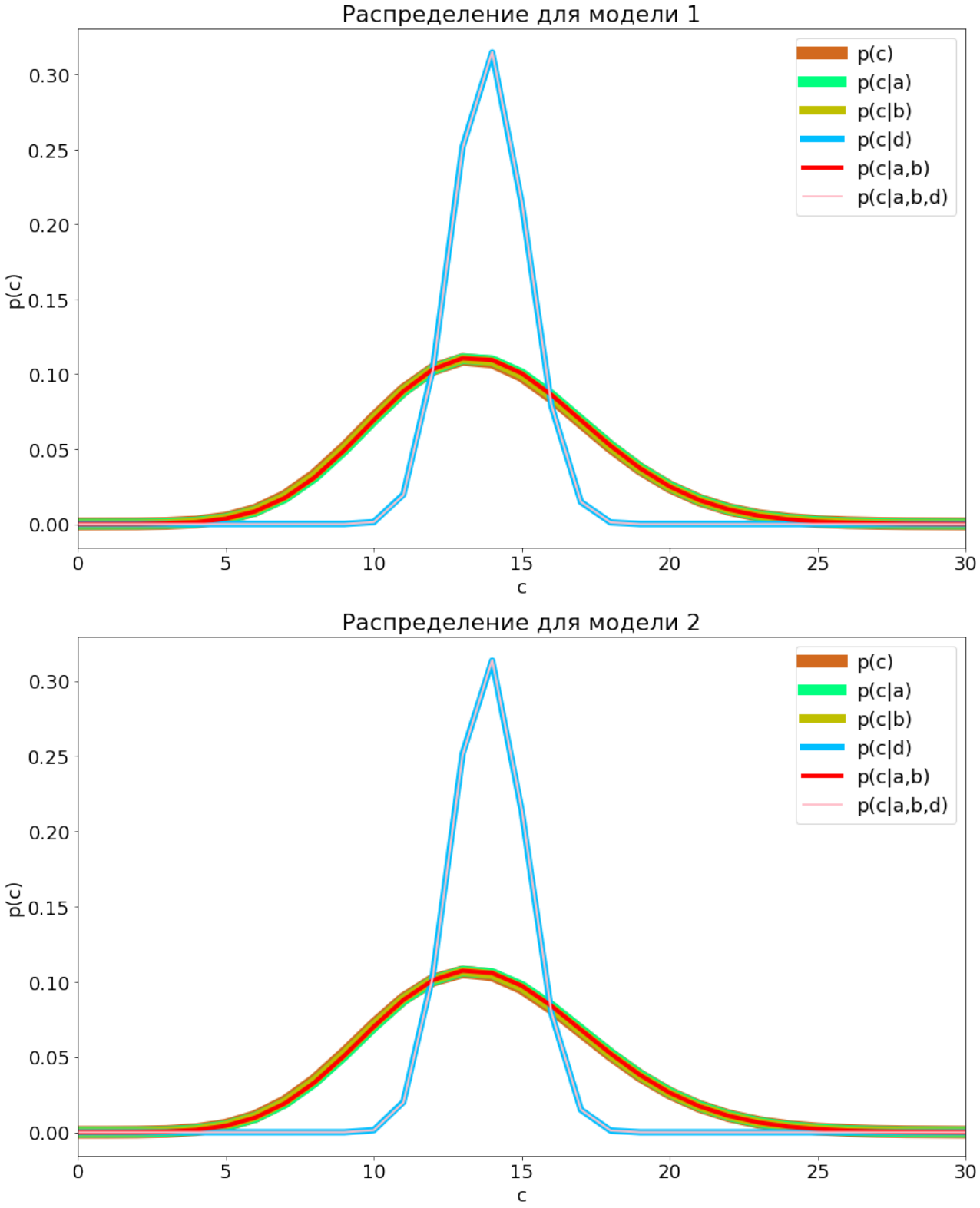

Рис. 1: Распределения вероятностей

3 Наибольший вклад в уточнение прогноза для величины c

Программным путем было выяснено, что для допустимых значений a, b, d верны выражения $\mathbb{D}[c|d] < \mathbb{D}[c|a], \mathbb{D}[c|d] < \mathbb{D}[c|b]$:

Модель	$\max \mathbb{D}[c\|d]$	$\min \mathbb{D}[c\|a]$	$\min \mathbb{D}[c\|b]$
1	10.2986909051	12.28	12.5875
2	12.8941557054	13.085	13.4625

Далее вычислим множество точек (a, b) таких, что $\mathbb{D}[c|b] < \mathbb{D}[c|a]$:

$$\mathbb{D}_c[c|a] = \mathbb{E}_b\mathbb{D}_c[c|a,b] + \mathbb{D}_b\mathbb{E}_c[c|a,b] = \mathbb{E}_b[ap_1(1-p_1) + bp_2(1-p_2)] + \mathbb{D}_b[ap_1 + bp_2] = \\ = ap_1(1-p_1) + p_2(1-p_2)\mathbb{E}b + p_2^2\mathbb{D}b. \tag{15}$$

Аналогично:

$$\mathbb{D}_c[c|b] = p_1(1-p_1)\mathbb{E}a + bp_2(1-p_2) + p_1^2\mathbb{D}a. \tag{16}$$

Далее решим $\mathbb{D}[c|b] < \mathbb{D}[c|a]$:

$$p_1(1-p_1)\mathbb{E}a + bp_2(1-p_2) + p_1^2\mathbb{D}a < ap_1(1-p_1) + p_2(1-p_2)\mathbb{E}b + p_2^2\mathbb{D}b$$
$$bp_2(1-p_2) - ap_1(1-p_1) < p_2(1-p_2)\mathbb{E}b + p_2^2\mathbb{D}b - p_1(1-p_1)\mathbb{E}a - p_1^2\mathbb{D}a$$

$$\text{Пусть } A = p_1(1-p_1); B = p_2(1-p_2);$$
$$C = p_2(1-p_2)\mathbb{E}b + p_2^2\mathbb{D}b - p_1(1-p_1)\mathbb{E}a - p_1^2\mathbb{D}a$$

Тогда видно, что неравенство линейное относительно (a, b):

$$Bb - Aa < C \tag{17}$$

Следовательно для $\mathbb{D}[c|b] \geq \mathbb{D}[c|a]$: $Bb - Aa \geq C$. Это и означает линейную разделимость множеств $\{(a,b)|\mathbb{D}[c|b] < \mathbb{D}[c|a]\}$ и $\{(a,b)|\mathbb{D}[c|b] \geq \mathbb{D}[c|a]\}$ прямой $Bb - Aa = C$.

Для модели 2 линейность сохраняется и меняются лишь константы.

4 Временные замеры

Таблица 3: Время вычислений распределений, в сек

Модель	**p(c)**	**p(c|a)**	**p(c|b)**	**p(c|d)**	**p(c|a,b)**	**p(c|a,b,d)**	**p(d)**
1	0.012417	0.011621	0.000789	0.011432	0.000493	0.000585	0.072148
2	0.005038	0.004840	0.000393	0.005094	0.000146	0.000253	0.068229

5 Сравнение результатов для двух моделей

При апроксимации биномиального распределения пуассоновским распределением, мы отметили, что можем с высокой точностью приблизить при большом количестве испытаний и маленькой вероятности успеха. Следовательно, максимальная разница проявляется при высоких вероятностях успеха. Например, возьмем $p_1 = p_2 = 0.99$ ($a = 100, b = 200$). Тогда:

$$\mathbb{D}c = 2.0746983826247742 \text{ для модели 1}$$
$$\mathbb{D}c = 296.99999999816646 \text{ для модели 1}$$

Но, как показывает таблица из пункта 4, такая аппроксимация дает прирост в среднем на 2 раза. Если выполнены условия аппроксимации и важно время выполнения, то можно использовать модель 2, но в целом лучше использовать модель 1.