Московский государственный университет имени М. В. Ломоносова

Байсовские методы машинного обучения

Задание 1. Байесовские рассуждения
Вариант 1

Выполнил:
студент 4 курса 417 группы
Бабичев Дмитрий Олегович

Москва, 2017

Содержание

1 Описание моделей

Вероятностные модели посещаемости курса

Есть некоторый курс, которые могут посещать студенты двух категорий: те, у которых этот курс является профильным и все остальные. Общее количество студентов первой категории будем обозначать a, второй - b. Не все студенты посещают курс. Есть вероятность, что студент решит его прогулять. Обозначим за p_1 вероятность посещения курса студентом первой категории, за p_2 вероятность посещения курса студентом второй категории. Тогда общее количество пришедших на занятие студентов c будет зависить как и от a, так и от b. Обозначается это так: $c|a, b$. Далее мы рассмотрим две модели, которые отличаются именно характером зависимости c от a и b. Пусть также на занятиях ведется журнал посещаемости, и пусть студенты будут иметь возможность отметить отсутствующего товарища с вероятностью p_3. Общее количество записавшихся обозначим за d. Тогда случайная величина $d|c$ представляет собой сумму c и случайной величины, распределенной по биномиальному закону $\text{Bin}(c, p_3)$. И напоследок зададим априорные вероятности для a и b. Пусть они будут распределены равномерно на $[a_{min}, a_{max}]$ и $[b_{min}, b_{max}]$ соответственно.

1.1 Модель 1

Пусть $c|a, b \sim \text{Bin}(a, p_1) + \text{Bin}(b, p_2)$, тогда итоговая модель будет выглядеть следующим образом:

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

1.2 Модель 2

Допустим некоторое упрощение предыдущей модели, а именно заменим в распределении $c|a,b$ биномиальное распределение Пуассоновским распределением: $\text{Bin}(n,p) \to \text{Poiss}(np)$

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

2 Задания варианта 1

Рассматриваются приведенные выше модели 1.1 и 1.2 с параметрами $a_{min} = 75$, $a_{max} = 90,\ b_{min} = 500,\ b_{max} = 600,\ p_1 = 0.1,\ p_2 = 0.01,\ p_3 = 0.3$.

2.1 Пункт 1

Вывести формулы для всех необходимых далее распределений аналитически.

Все формулы в этом отчете верны для вычислений скаляров. В реализации значения высичляются по-другому. Также я подумал, что не буду выводить формулы математического ожидания и дисперсии для нетабличных случайных величин, потому что они легко считаются и по формулам общего вида:

$$\mathbb{E}\xi = \sum_{\xi=\xi_{min}}^{\xi_{max}} \xi p(\xi), \mathbb{E}\xi^2 = \sum_{\xi=\xi_{min}}^{\xi_{max}} \xi^2 p(\xi), \mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2$$

Выведем формулы вероятностей следующих величин: $p(a),\ p(b),\ p(c),\ p(d),\ p(c|a,b),\ p(c|a),\ p(c|b),\ p(c|d),\ p(d|c),\ p(c|a,b,d)$, используя знания некоторых распределений и правила произведения и суммирования и маргинализацию.

$\mathbf{p(a)},\ \mathbf{p(b)}$:

Дискретное равномерное распределение. Вероятность попасть в точку на отрезке равна единице, деленной на количество целых точек на отрезке

$p(a) = \frac{1}{a_{max} - a_{min} + 1},\ p(b) = \frac{1}{b_{max} - b_{min} + 1}$.

$\mathbf{p(c|a,b)}$ $(c = \xi_1 + \xi_2,\ \ \xi_1 \sim \mathrm{Bin}(a, p_1),\ \ \xi_2 \sim \mathrm{Bin}(b, p_2))$:
Условимся, что $\binom{n}{k} = 0,$ если $k > n$, тогда

$$p(c = k|a,b) = \sum_{i=0}^{k} p(\xi_1 = i)p(\xi_2 = k - i) = \\ = \sum_{i=0}^{k} \binom{a}{i} p_1^i (1-p_1)^{a-i} \binom{b}{k-i} p_2^{k-i} (1-p_2)^{b+i-k}$$

То, что величины ξ_1 и ξ_2 имеют биномиальные распределения позволяют прийти к формулам выше прямыми рассуждениями. А в общем, для нахождения плотности суммы двух случайных величин (не обязательно дискретных) используется свертка. Плотность суммы равна свертке плотностей слагаемых, а прямое преобразование Фурье плотности суммы равно произведению прямых преобразований Фурье плотностей слагаемых.

$\mathbf{p(c|a,b)}$ $(c|a,b \sim \mathrm{Pois}(ap_1 + bp_2))$:
Простое табличное распределение:

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

$\mathbf{p(c|a)},\ \mathbf{p(c|b)}$:

$$p(c|a) \overset{SR}{=} \sum_{b=b_{min}}^{b_{max}} p(c|a,b)p(b) = \frac{\sum\limits_{b=b_{min}}^{b_{max}} p(c|a,b)}{b_{max} - b_{min} + 1} = \\ = \begin{cases} \dfrac{\sum\limits_{b=b_{min}}^{b_{max}} \sum\limits_{i=0}^{c} \binom{a}{i} p_1^i (1-p_1)^{a-i} \binom{b}{c-i} p_2^{c-i} (1-p_2)^{b+i-c}}{b_{max} - b_{min} + 1}, & c|a,b \sim \mathrm{Bin}(a,p_1) + \mathrm{Bin}(b,p_2) \\ \dfrac{\sum\limits_{b=b_{min}}^{b_{max}} e^{-(ap_1+bp_2)}(ap_1 + bp_2)^c}{(b_{max} - b_{min} + 1)c!}, & c|a,b \sim \mathrm{Pois}(ap_1 + bp_2) \end{cases}$$

Аналогично

$$p(c|b) = \begin{cases} \dfrac{\sum\limits_{a=a_{min}}^{a_{max}} \sum\limits_{i=0}^{c} \binom{a}{i} p_1^i (1-p_1)^{a-i} \binom{b}{c-i} p_2^{c-i} (1-p_2)^{b+i-c}}{a_{max} - a_{min} + 1}, & c|a,b \sim \mathrm{Bin}(a,p_1) + \mathrm{Bin}(b,p_2) \\ \dfrac{\sum\limits_{a=a_{min}}^{a_{max}} e^{-(ap_1+bp_2)}(ap_1 + bp_2)^c}{(a_{max} - a_{min} + 1)c!}, & c|a,b \sim \mathrm{Pois}(ap_1 + bp_2) \end{cases}.$$

$\mathbf{p(c)}$:

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

$$= \frac{\sum\limits_{b=b_{min}}^{b_{max}} \sum\limits_{a=a_{min}}^{a_{max}} \sum\limits_{i=0}^{c} \binom{a}{i} p_1^i (1-p_1)^{a-i} \binom{b}{c-i} p_2^{c-i} (1-p_2)^{b+i-c}}{(b_{max} - b_{min} + 1)(a_{max} - a_{min} + 1)}$$

$\mathbf{p(d|c)}$:

Вспомним, что в обеих моделях $d|c \sim c + \mathrm{Bin}(c, p_3)$. Учитывая нашу договоренность о биномиальных коэффициентах, получаем следующуй формулу:

$$p(d|c) = \binom{c}{d-c} p_3^{d-c} (1-p_3)^{2c-d}.$$

$\mathbf{p(d)}$:

$$p(d) \overset{SR}{=} \sum_{c=0}^{a_{max}+b_{max}} p(d|c)p(c) = \sum_{c=0}^{a_{max}+b_{max}} \binom{c}{d-c} p_3^{d-c} (1-p_3)^{2c-d} p(c) =$$

$$= \sum_{c=0}^{a_{max}+b_{max}} \binom{c}{d-c} p_3^{d-c} (1-p_3)^{2c-d} \cdot$$

$$\cdot \frac{\sum\limits_{b=b_{min}}^{b_{max}} \sum\limits_{a=a_{min}}^{a_{max}} \sum\limits_{i=0}^{c} \binom{a}{i} p_1^i (1-p_1)^{a-i} \binom{b}{c-i} p_2^{c-i} (1-p_2)^{b+i-c}}{(b_{max} - b_{min} + 1)(a_{max} - a_{min} + 1)}$$

$\mathbf{p(c|d)}$:

$$p(c|d) \overset{Bayes' \ formula}{=} \frac{p(d|c)p(c)}{p(d)} \overset{SR}{=} \frac{p(d|c)p(c)}{\sum\limits_{k=0}^{a_{max}+b_{max}} p(d|c=k)p(c=k)}$$

$\mathbf{p(c|a,b,d)}$:

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

$$= \frac{p(d|c)p(c|a,b)p(a)p(b)}{\sum\limits_{k=0}^{a_{max}+b_{max}} p(d|c=k)p(c=k|a,b)p(a)p(b)} =$$

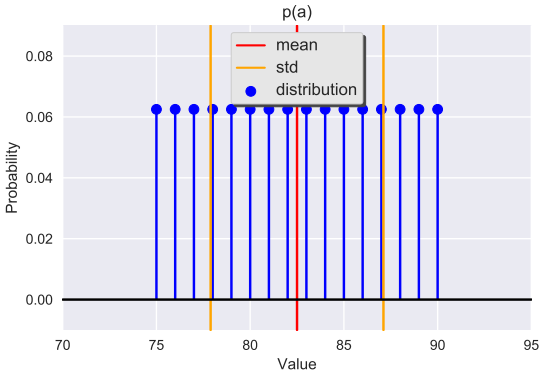
$$= \frac{p(d|c)p(c|a,b)}{\sum\limits_{k=0}^{a_{max}+b_{max}} p(d|c=k)p(c=k|a,b)}$$

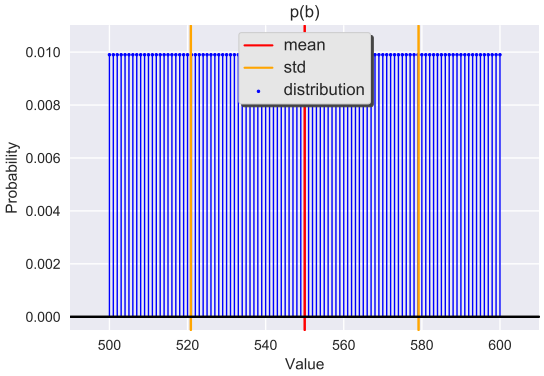
2.2 Пункт 2

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

$\mathbb{E}a = 82.5;$ $\mathbb{E}b = 550;$

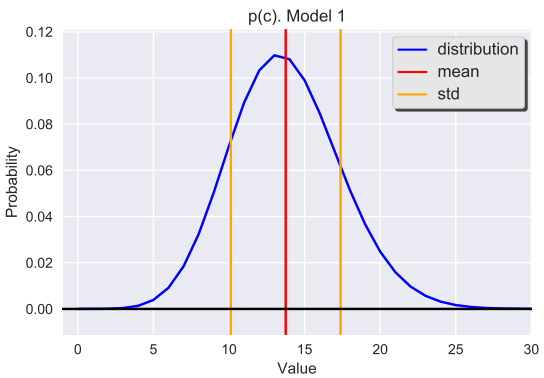
$\mathbb{D}a = 21.25;$ $\mathbb{D}b = 850;$

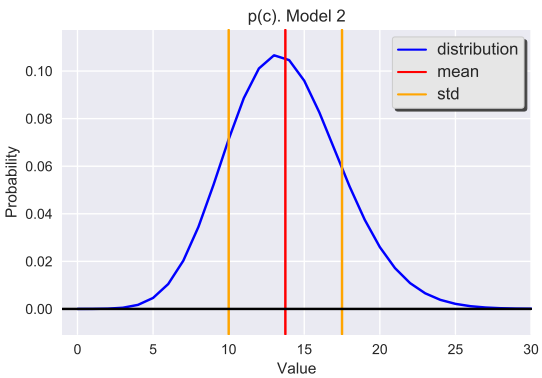




$\mathbb{E}c_1 = 13.75;$ $\mathbb{E}c_2 = 13.75;$

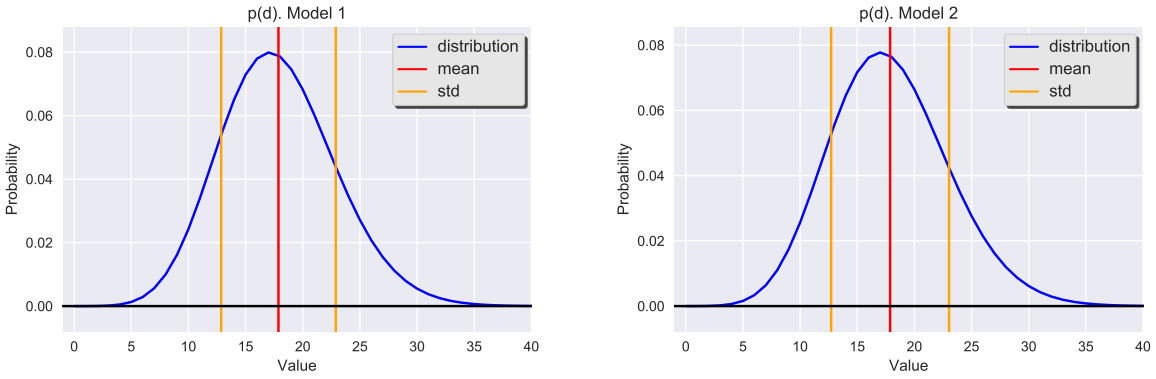
$\mathbb{D}c_1 = 13.1675;$ $\mathbb{D}c_2 = 14.0475;$





$\mathbb{E}d_1 = 17.875;$ $\mathbb{E}d_2 = 17.875;$

$\mathbb{D}d_1 = 25.140575;$ $\mathbb{D}d_2 = 26.627775;$



2.3 Пункт 3

Пронаблюдать, как происходит уточнение прогноза для величины c по мере прихода новой косвенной информации.

Предлагается внимательно изучить распределение и статистики следующих случайных величин: $p(c),\ p(c|b),\ p(c|d),\ p(c|a,b),\ p(c|a,b,d)$.

Для начала построим график всех распределений и отметим на нем математические ожидания и среднеквадратические отклонения от него:

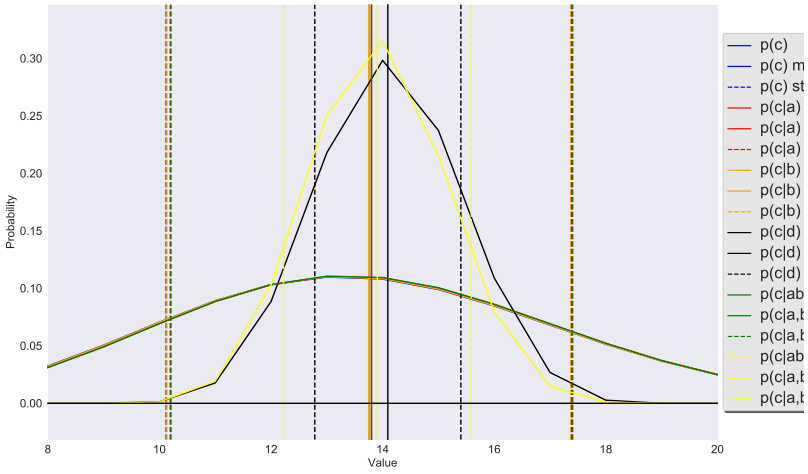


Рис. 1: Модель 1

Во-первых, заметим, что графики всех распределений очень похожи на гауссианы. Но увы, на этих графиках 4 распределения слились. Но уже по ним видно, что у распределений, для которых определено d ($p(c|d),\ p(c|a,b,d)$ дисперсия значительно

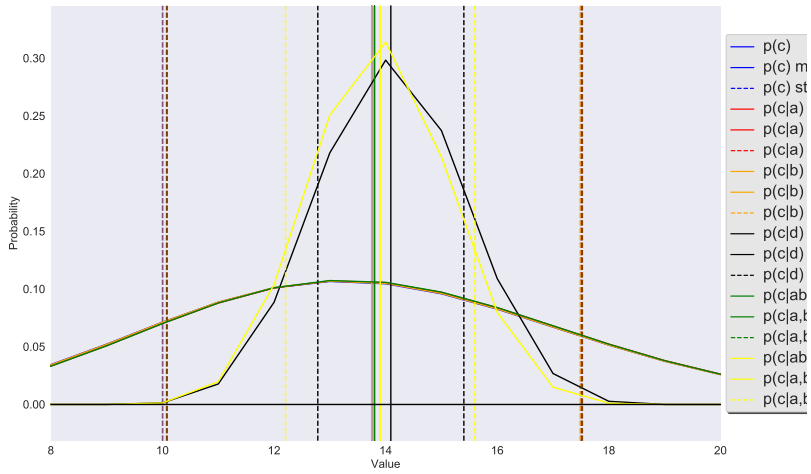


Рис. 2: Модель 2

ниже. Можно предположить, что d несет в себе значительно больше информации, чем a, b и их комбинации. Еще одно интеренсое наблюдение (надеюсь, что это не ошибка кода): дисперсия $p(c|d)$ меньше, чем дисперсия $p(c|a, b, d)$. Все эти догадки справедливы для обеих моделей. Более четко это можно увидеть на следующих графиках:

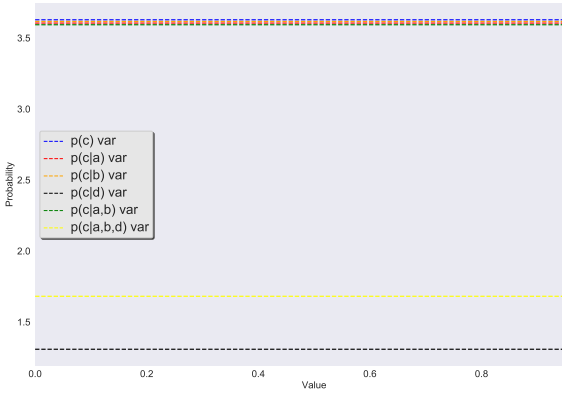


Рис. 3: Дисперсии. Модель 1

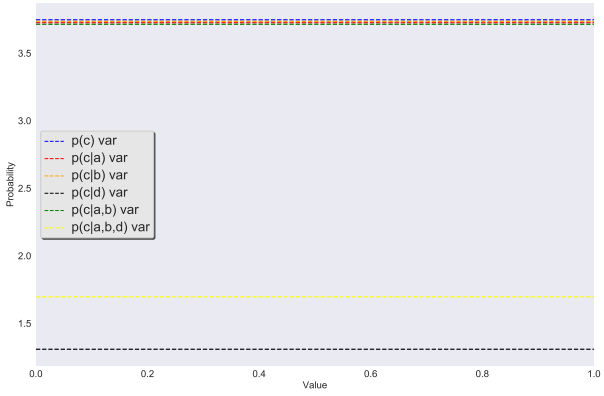
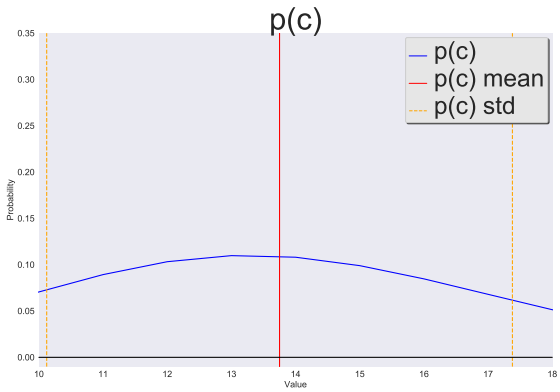


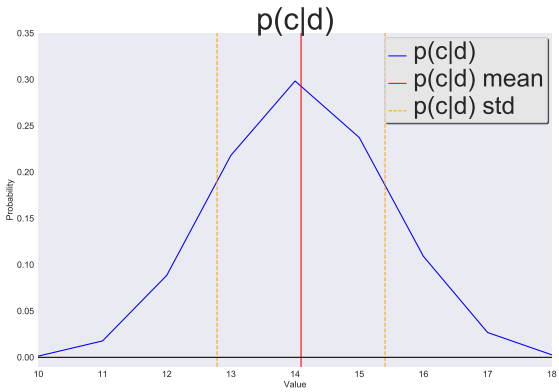
Рис. 4: Дисперсии. Модель 2

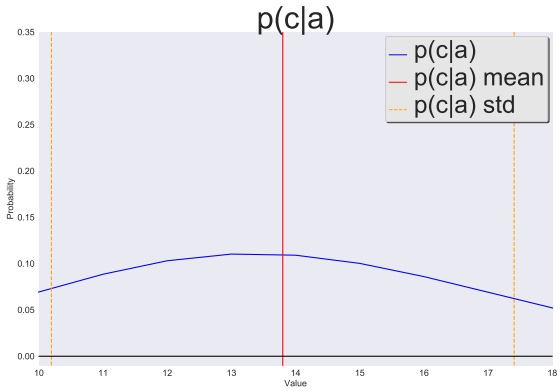
Посмотрим на графики распределений по отдельности в одинаковом масштабе.

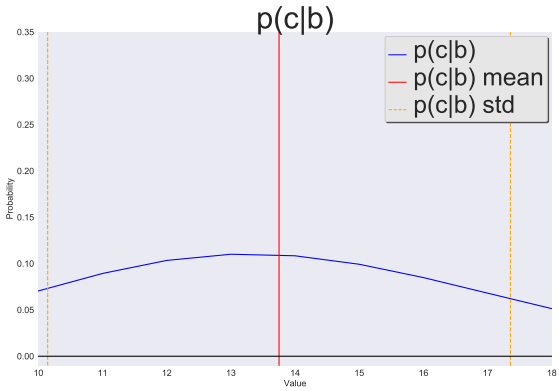
Модель 1

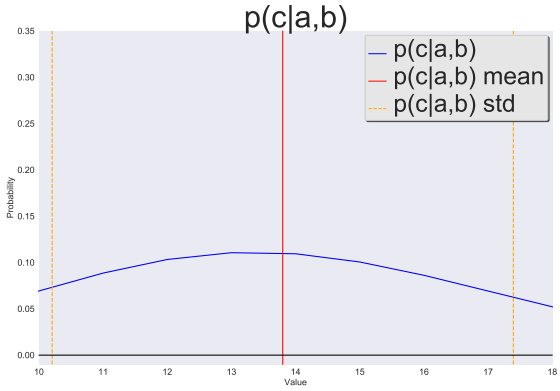
На этих графиках видно практически отсутствие отличий между распределениями

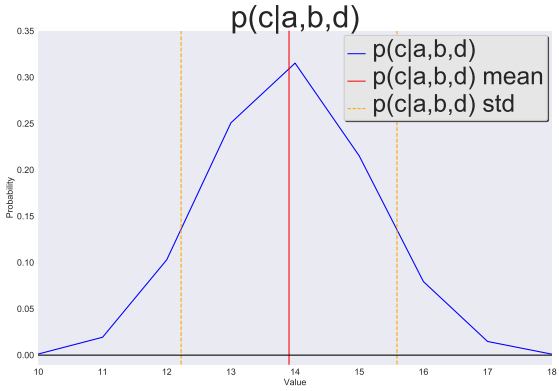








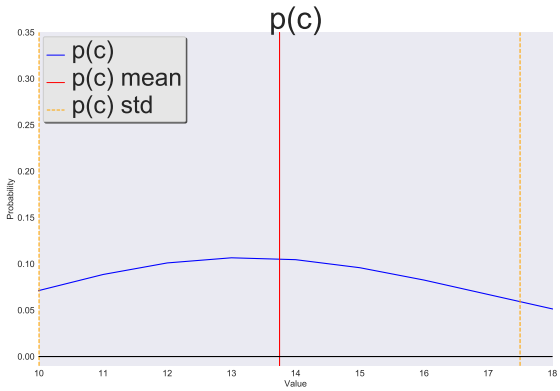


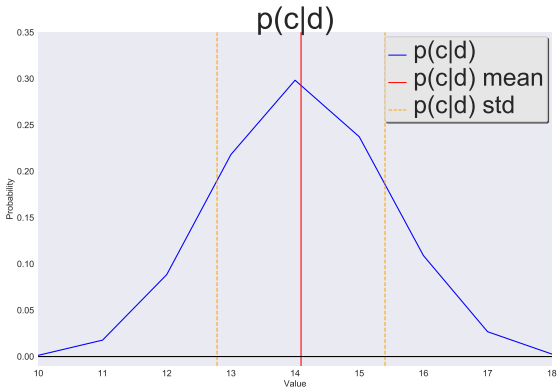


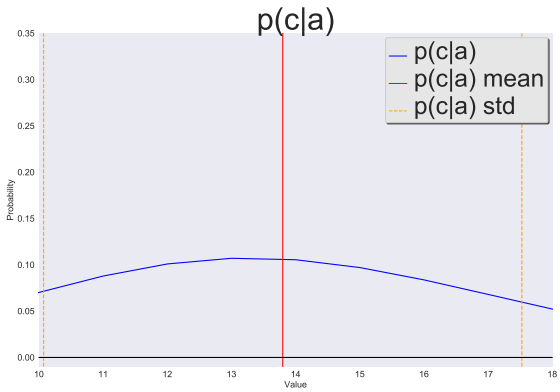
$p(c),\ p(c|a),\ p(c|b),\ p(c|a,b)$. А что во ворой модели?

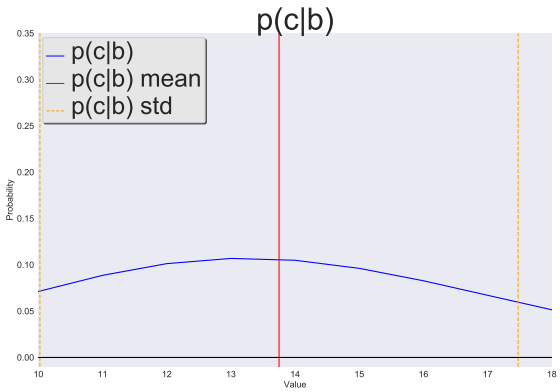
Модель 2

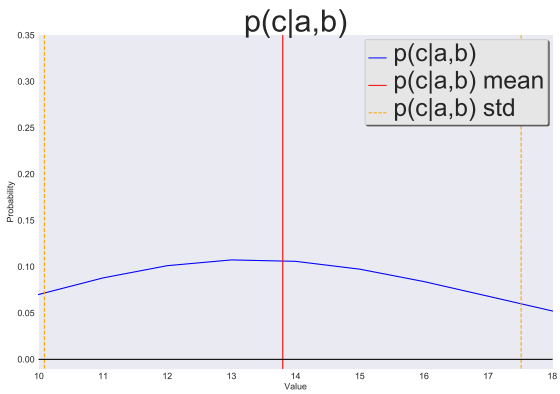
Аналогичная картина. Можно посмотреть на табличку.

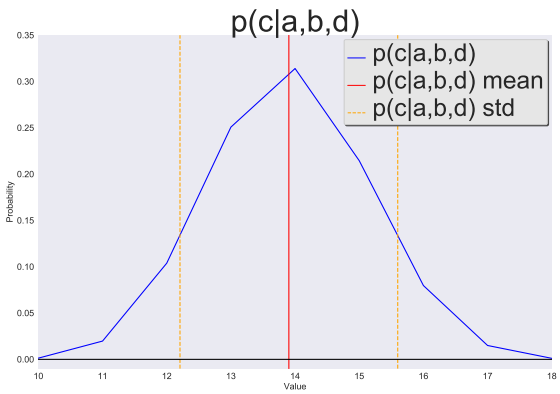












	$p(c)$	$p(c\|a)$	$p(c\|b)$	$p(c\|d)$	$p(c\|a,b)$	$p(c\|a,b,d)$
Модель 1						
$\mathbb{E}c$	13.75	13.8	13.75	14.09	13.8	13.91
$\mathbb{D}c$	13.17	13	13.08	1.71	12.92	2.83
Модель 2						
$\mathbb{E}c$	13.75	13.8	13.75	14.09	13.8	13.9
$\mathbb{D}c$	14.05	13.89	13.96	1.71	13.8	2.88

Получается, что a несет информации больше, чем b, вместе они информативнее, чем по отдельности, d вносит самый существенный вклад, а по непонятным мне причинам одновременно они вносят вклад меньше, чем отдельно d. Изучим этот эффект подробнее в пункте 4.

2.4 Пункт 4

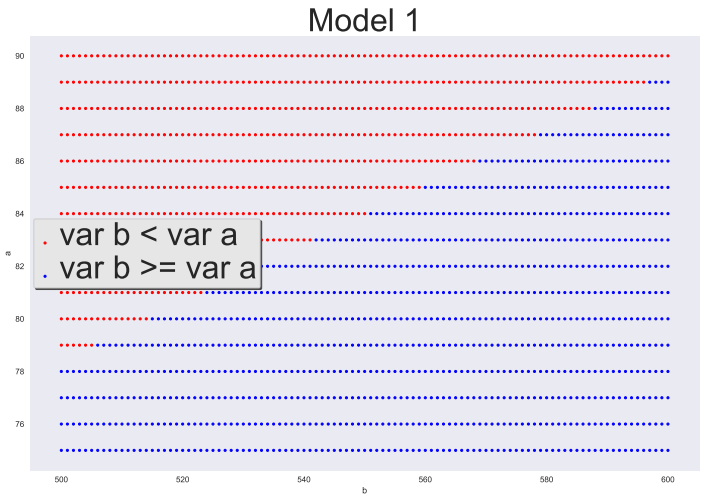
Определить, какая из величин $a,\ b,\ d$ вносит наибольший вклад в уточнение прогноза для величины c (в смысле дисперсии распределения).
При выполнении этого пункта возникло ряд проблем с точностью вычислений вероятности $p(c|d)$ и, соответственно, с дисперсиями. Наверно стоило попробовать увеличить разрядность до 128 бит, но я решил не делать этого, отчего мои результаты не совпали с ожидаемыми. В частности:

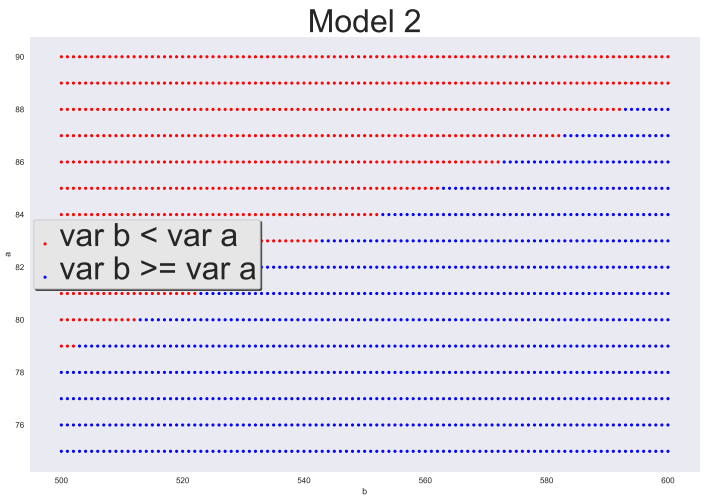
	$p(c\|a)$	$p(c\|b)$	$p(c\|d)$
Модель 1			
$\max \mathbb{D}c$	13.63	13.57	114.90
$\min \mathbb{D}c$	12.28	12.58	0.0
Модель 2			
$\max \mathbb{D}c$	14.58	14.4625	17.00
$\min \mathbb{D}c$	13.08	13.46	0.0

Ожидалось, что для всех d $\mathbb{D}[c|d]$ будет меньше $\min\limits_a \mathbb{D}[c|a]$ и $\min\limits_b \mathbb{D}[c|b]$, на деле же это условие выволняется при $d < 80$ и $d > 922$ в первой модели, $d < 400$ и $d > 520$

во второй модели.

Что касается множеств $\{(a, b) : \mathbb{D}[c|b] < \mathbb{D}[c|a]\}$ и $\{(a, b) : \mathbb{D}[c|b] \geq \mathbb{D}[c|a]\}$, то они оказались линейно разделимыми в обеих моделях. Наверно это можно доказать аналитически, но я решил применить к этим множествам метод опорных векторов, который выдававал 100% точность на этих данных, откуда следует, что существует прямая, идеально разделяющая эти два множества.





2.5 Пункт 5

Провести временные замеры по оценке всех необходимых распределений.

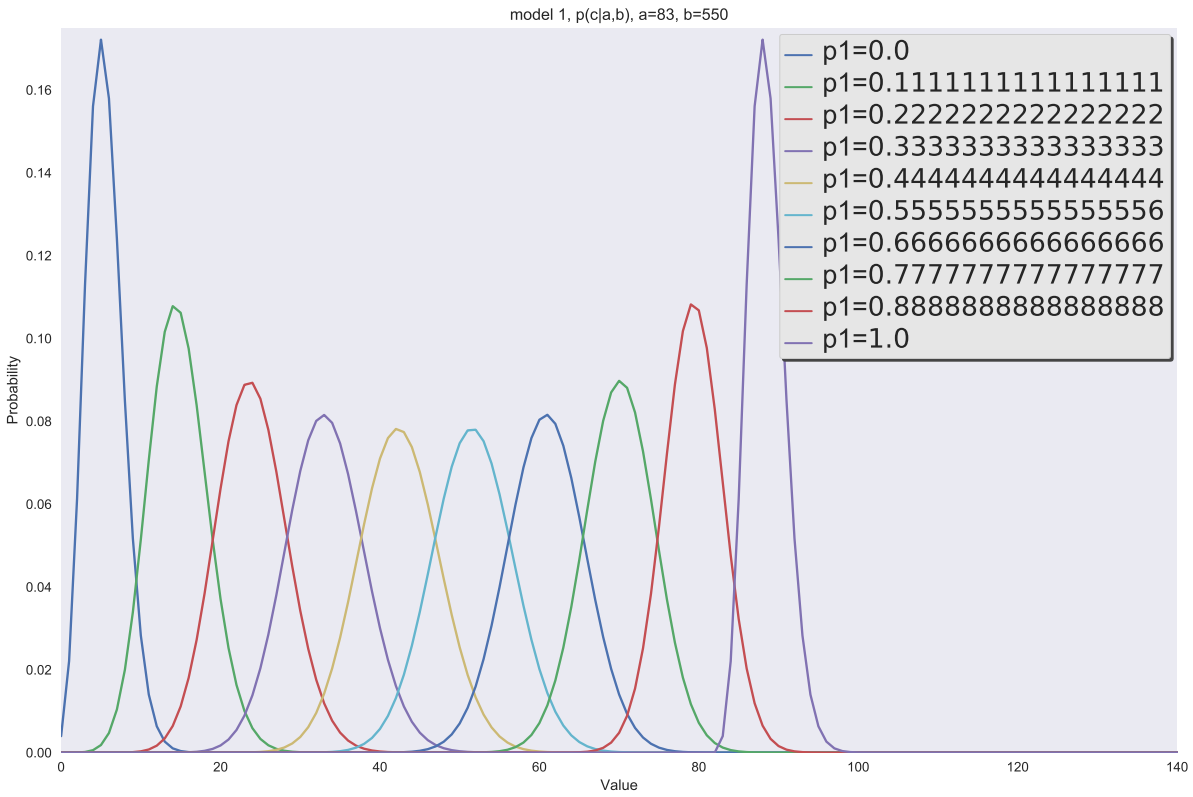
сек	$p(c)$	$p(c\|a)$	$p(c\|b)$	$p(c\|d)$	$p(c\|a,b)$	$p(c\|a,b,d)$	$p(d)$
Модель 1							
$median$	0.3124	0.3183	0.3162	0.4654	0.3268	0.4760	0.5507
min	0.3066	0.3056	0.3072	0.4579	0.3086	0.4417	0.5259
max	0.3599	0.6108	0.5954	0.5377	0.3776	0.6488	0.6817
Модель 2							
$median$	0.3017	0.3062	0.3067	0.4522	0.3104	0.4293	0.5413
min	0.2964	0.2972	0.2994	0.4427	0.2958	0.4544	0.5135
max	0.3683	0.4785	0.3905	0.5183	0.4080	0.6302	0.6990

Все вычисления укладываются в одну секунду. Самой емкой задачей является вычисление $p(c|a,b,d)$, где необходимо уметь эффективно работать с многомерными матрицами. Я использовал суммирование Эйнштейна, которое позволяет очень удобно манипулировать матрицами. Получилось эффективно по времени, но на моем устройстве были проблемы с памятью. Схожеть результатов для $p(c)$, $p(c|a)$, $p(c|b)$, $p(c|a,b)$ объясняется просто: все они опираются на вычисление $p(c|a,b)$ с последующей маргинализацией.

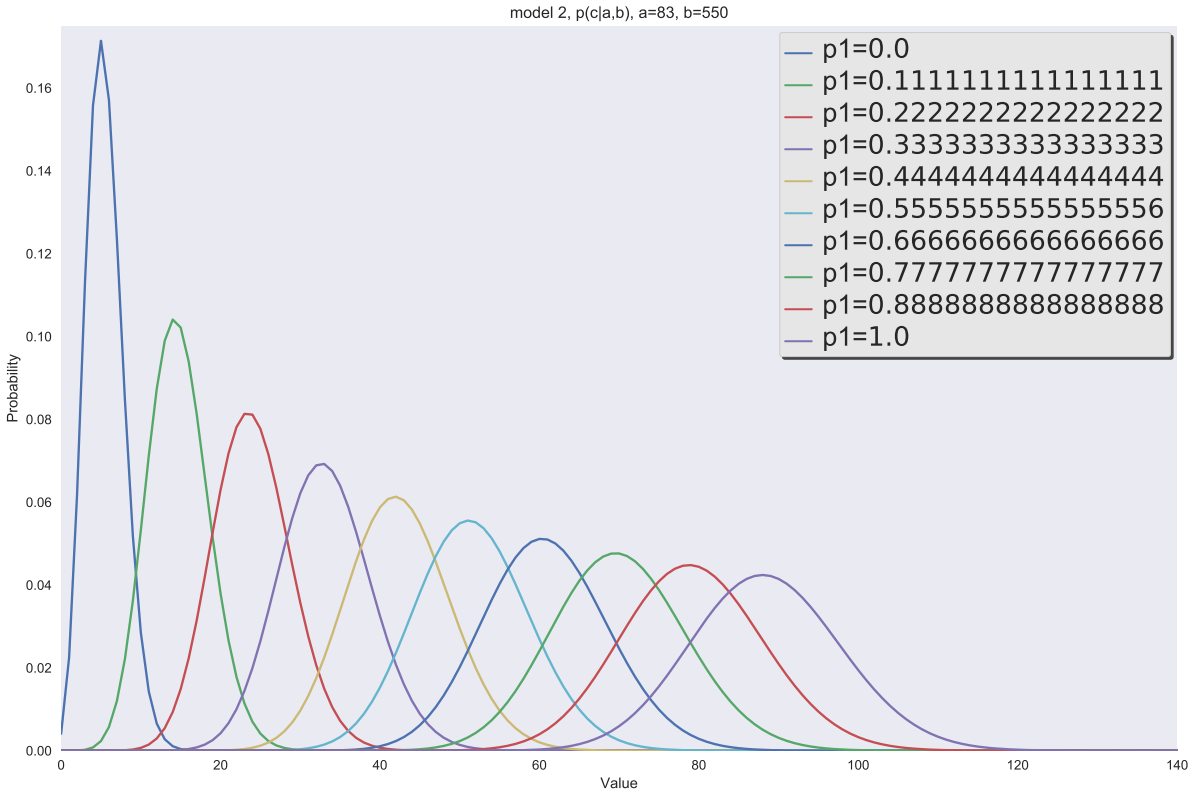
2.6 Пункт 6

Сравнить результаты для двух моделей. Показать где максимально проявляется разница между ними (привести конкретный пример, не обязательно из экспериментов выше). Объяснить причины подобного результата.

Как уже было сказано в задании, модель 2 похожа на модель 1 только при большом количестве наблюдений и небольшой вероятности успеха. Поэтому для того, чтобы найти существенные различия, достаточно менять эти параметры. Для иллюстрации я менял вероятность посещения лекции студентом профильного факультета p_1. Пронаблюдаем поведение распределений обеих моделей: Сразу же видны все раз-



личия. При росте вероятности успеха в биномильном распределении график плотности по-идее должен смещаться влево и становиться острее. Мы же видим сначала



сглаживание, затем уже рост. Оказывается так на поведение флияет свертка двух биномиальных случайных величин. Распределение пуассона же при росте λ Также должно смещаться вправо, но не заостряться, а наоборот, угасать, "размазываться". Что мы и наблюдаем. Вторая модель значительно проще, потому что параметры распределения b входят лишь как слагаемое $\lambda = ap_1 + bp_2$.