**ВМК МГУ**

# Задание 1. Байесовские рассуждения

**Курс: Байесовские методы в машинном обучении, 2016**

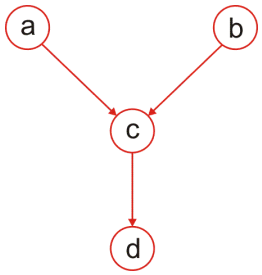
Начало выполнения задания: 3 сентября
Срок сдачи: **17 сентября (суббота), 23:59.**
Среда для выполнения задания – PYTHON 2.

## Содержание



## 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:



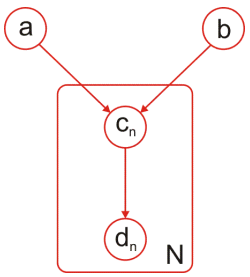

$$
\begin{aligned}
&p(a,b,c,d) = p(d|c)p(c|a,b)p(a)p(b),\\
&d|c \sim c + \text{Bin}(c,p_3),\\
&c|a,b \sim \text{Bin}(a,p_1) + \text{Bin}(b,p_2),\\
&a \sim \text{Unif}[a_{min},a_{max}],\\
&b \sim \text{Unif}[b_{min},b_{max}].
\end{aligned}
\tag{1}
$$

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 $\lambda_1$ и $\lambda_2$ 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:

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

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




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

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

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

Задание состоит из трёх вариантов.

## Вариант 1

Рассматривается модель 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 на компьютере следующие исследования:

1. Найти математические ожидания и дисперсии априорных распределений для всех параметров $a$, $b$, $c$, $d$.

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

3. Определить, какая из величин $a$, $b$, $d$ вносит наибольший вклад в уточнение прогноза для величины $c$ (в смысле дисперсии распределения). Для этого убедиться в том, что $\mathbb{D}[c|d] < \mathbb{D}[c|b]$ и $\mathbb{D}[c|d] < \mathbb{D}[c|a]$ для любых допустимых значений $a$, $b$, $d$. Найти множество точек $(a,b)$ таких, что $\mathbb{D}[c|b] < \mathbb{D}[c|a]$. Являются ли множества $\{\,(a,b) \mid \mathbb{D}[c|b] < \mathbb{D}[c|a]\,\}$ и $\{\,(a,b) \mid \mathbb{D}[c|b] \geq \mathbb{D}[c|a]\,\}$ линейно разделимыми?

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

5. Провести исследования из пп. 1–4 для точной модели 1 и сравнить результаты с аналогичными для модели 2. Привести пример оценки параметра, для которого проявляется разница между моделью 1 и 2. Объяснить причины подобного результата.

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

При оценке выполнения задания будет учитываться эффективность программного кода.

## Вариант 2

Рассматривается модель 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 на компьютере следующие исследования:

1. Найти математические ожидания и дисперсии априорных распределений для всех параметров $a$, $b$, $c$, $d$.
2. 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.
3. 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?
4. Провести временные замеры по оценке всех необходимых распределений $p(c)$, $p(c|a)$, $p(c|b)$, $p(b|a)$, $p(b|d)$, $p(b|a,d)$, $p(d)$.
5. Провести исследования из пп. 1–4 для точной модели 1 и сравнить результаты с аналогичными для модели 2. Привести пример оценки параметра, для которого проявляется разница между моделью 1 и 2. Объяснить причины подобного результата.

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

При оценке выполнения задания будет учитываться эффективность программного кода.

## Вариант 3

Рассматривается модель 4 с параметрами $a_{min} = 75$, $a_{max} = 90$, $b_{min} = 500$, $b_{max} = 600$, $p_1 = 0.1$, $p_2 = 0.01$, $p_3 = 0.3$, $N = 50$. Провести на компьютере следующие исследования:

1. Найти математические ожидания и дисперсии априорных распределений для всех параметров $a$, $b$, $c_n$, $d_n$.
2. Реализовать генератор выборки $d_1, \ldots, d_N$ из модели при заданных значениях параметров $a, b$.
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 $p(b)$, $p(b|d_1), \ldots, p(b|d_1, \ldots, d_N)$, где выборка $d_1, \ldots, d_N$ 1) сгенерирована из модели при параметрах $a, b$, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого и 2) $d_1 = \cdots = d_N$, где $d_n$ равно мат.ожиданию своего априорного распределения, округленного до ближайшего целого. Провести аналогичный эксперимент, если дополнительно известно значение $a$. Сравнить результаты двух экспериментов.
4. Провести временные замеры по оценке всех необходимых распределений $p(c_n)$, $p(d_n)$, $p(b|d_1, \ldots, d_n)$, $p(b|a, d_1, \ldots, d_n)$.
5. Провести исследования из пп. 1–4 для точной модели 3 и сравнить результаты с аналогичными для модели 4.

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

При оценке выполнения задания будет учитываться эффективность программного кода.

## Оформление задания

Выполненное задание следует отправить письмом по адресу *bayesml@gmail.com* с заголовком письма

«[БММО16] Практика 1, Фамилия Имя, Номер варианта».

Убедительная просьба присылать выполненное задание только один раз с окончательным вариантом. Также убедительная просьба строго придерживаться заданных ниже прототипов реализуемых функций (для проверки задания используются, в том числе, автоматические процедуры, которые являются чувствительными к неверным прототипам).

Номер варианта вычисляется как $(s \ mod \ 3)+1$, где $s$ — сумма кодов букв своей фамилии в кодировке UTF-8. В питоне это выглядит так:

$$sum([ord(x) \ for \ x \ in \ u'\text{Фамилия}']) \ \% \ 3 + 1$$

Присланный вариант задания должен содержать в себе:

- Текстовый файл в формате PDF с указанием ФИО и номера варианта, содержащий описание всех проведённых исследований (вывод необходимых формул, графики, анализ и выводы). Файл должен называться surname.pdf, где surname – фамилия студента.
- Python модуль со всеми требуемыми функциями в соответствии с прототипами, приведенными ниже. Модуль должен называться surname.py, где surname – фамилия студента.

## Требования к реализации

Все исходные коды должны располагаться в одном модуле surname.py. Распределения должны быть реализованы в виде **отдельных функций**. Прототип функции для оценки распределения $p(c|a,d)$ показан в таблице 1. Прототипы функций для других распределений выглядят аналогично. Если в распределении переменных до или после | несколько, то в названии функции они идут в алфавитном порядке. Функция для оценки распределения $p(b|a,d_1,\ldots,d_N)$ для модели 3 имеет название $pb\_ad$, а входной параметр $d$ является массивом длины $N$.

Таблица 1: Прототип функции для оценки распределения $p(c|a,d)$ для модели 1 и 2

| **p, c = pc_ad(a, d, params, model)** |
|---|
| ВХОД<br>$a$ – значение параметра $a$;<br>$d$ – значение параметра $d$;<br>*params* – набор параметров вероятностной модели, словарь с ключами 'amin', 'amax', 'bmin', 'bmax', 'p1', 'p2', 'p3';<br>*model* – номер модели; |
| ВЫХОД<br>$p$ – распределение вероятности, numpy array длины len(c);<br>$c$ – носитель распределения, numpy array. |

Таблица 2: Прототип функции для генерации выборки из распределения $p(d_1,\ldots,d_N|a,b)$ для модели 3 и 4

| **d = generate(N, a, b, params, model)** |
|---|
| ВХОД<br>$N$ – количество лекций;<br>$a$ – значение параметра $a$;<br>$b$ – значение параметра $b$;<br>*params* – набор параметров вероятностной модели, словарь с ключами 'amin', 'amax', 'bmin', 'bmax', 'p1', 'p2', 'p3';<br>*model* – номер модели; |
| ВЫХОД<br>$d$ – значения $d_1,\ldots,d_N$, numpy array длины $N$. |