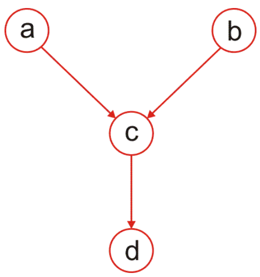
Задание 1. Байесовские рассуждения

Курс: Байесовские методы в машинном обучении 2020

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:

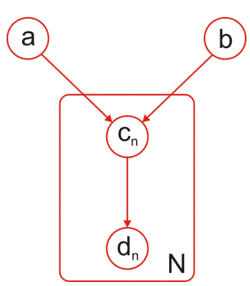


$$\begin{aligned}
&p(a, b, c, d) = p(d|c)p(c|a, b)p(a)p(b),\\
&d|c \sim c + \text{Bin}(c, p_3),\\
&c|a, b \sim \text{Bin}(a, p_1) + \text{Bin}(b, p_2),\\
&a \sim \text{Unif}[a_{min}, a_{max}],\\
&b \sim \text{Unif}[b_{min}, b_{max}].
\end{aligned} \tag{1}$$

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:



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

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

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

Задание состоит из трёх вариантов. Схема присвоения варианта приложена к заданию.

Вариант 1

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.
2. Найти математические ожидания и дисперсии априорных распределений $p(a)$, $p(b)$, $p(c)$, $p(d)$.
3. 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.
4. Определить, какая из величин a, b, d вносит наибольший вклад в уточнение прогноза для величины c (в смысле дисперсии распределения). Для этого проверить верно ли, 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? Ответ должен быть обоснован!
5. Провести временные замеры по оценке всех необходимых распределений $p(c)$, $p(c|a)$, $p(c|b)$, $p(c|d)$, $p(c|a, b)$, $p(c|a, b, d)$, $p(d)$.
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.

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.

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

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

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.

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.

Вариант 3

Рассматриваются модели 3 и 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 следующие исследования для обеих моделей:

1. Вывести формулы для всех необходимых далее распределений аналитически.

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

3. Реализовать генератор выборки $d_1, \ldots, d_N$ из модели при заданных значениях параметров a, b.

4. 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.

5. Провести временные замеры по оценке всех необходимых распределений $p(c_n)$, $p(d_n)$, $p(b|d_1, \ldots, d_n)$, $p(b|a, d_1, \ldots, d_n)$.

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.

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 пункте 2 будут ставиться дополнительные баллы. При оценке выполнения задания будет учитываться эффективность программного кода - любая из функций должна работать быстрее секунды на скалярных входах a, b и одномерных входных векторах c_n, d_n длины около 50 (для этого код должен реализовываться векторно). По всем пунктам задания должен быть проведен анализ результатов и сделаны выводы. За качественный анализ в пункте 4 также могут быть выставлены дополнительные баллы.

Оформление задания

На проверку в ejudge нужно отправить Python модуль со всеми требуемыми функциями в соответствии с прототипами, приведенными в отдельном файле. Модуль должен называться {name}_{surname}_v{variant}.py, например, petr_ivanov_v1.py. Модуль не должен содержать никакого main! То есть при импорте модуля никакие вычисления производиться не должны.

Перед отправкой кода в ejudge его нужно проверить с помощью выдаваемых открытых тестов. Если какой-то из них выдает предупреждение (кроме тестов по времени), то ваш код не соответствует прототипам и не может быть проверен. Предупреждения по времени говорят о том, что ваш код не достаточно эффективен, что может привести к понижению оценки.

На проверку в anytask нужно отправить:

- Тот же Python модуль, который был отправлен в ejudge.
- Отчет в формате PDF с указанием ФИО и номера варианта, содержащий описание всех проведённых исследований (вывод необходимых формул, графики, анализ и выводы). Отчет не должен содержать листинга кода и подобных вещей! Желательно для составления отчета использовать latex. Файл должен называться {name}_{surname}_v{variant}.pdf.

Будьте внимательны к формату названий файлов!