Московский государственный университет им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Задание № 1.
Вероятностные модели посещаемости курса

Автор: Арбузова Дарья
Группа: 417

Москва, 2014

1 Цели задания

В рамках данного задания мы познакомимся с байесовскими сетями...

2 Описание моделей

Рассмотрим модель посещаемости студентами одного курса лекции. Пусть аудитория данного курса состоит из студентов профильной кафедры, а также студентов других кафедр. Обозначим через a количество студентов, распределившихся на профильную кафедру, а через b — количество студентов других кафедр на курсе. Пусть студенты профильной кафедры посещают курс с некоторой вероятностью p_1, а студенты остальных кафедр — с вероятностью p_2. Обозначим через c количество студентов на данной лекции. Тогда случайная величина $c|a, b$ есть сумма двух случайных величин, распределенных по биномиальному закону $B(a, p_1)$ и $B(b, p_2)$ соответственно. Пусть далее на лекции по курсу ведется 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 с и случайной величины, распределенной по биномиальному закону $B(c, p_3)$. 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:

Рассмотрим несколько упрощенную версию модели 1. Известно, что биномиальное распределение $B(n, p)$ при большом количестве испытаний и маленькой вероятности успеха может быть с высокой точностью приближено пуассоновским распределением $Poiss(\lambda)$ 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 аналогичное неверно). Таким образом, мы можем сформулировать вероятностную модель, которая является приближенной версией модели 1:

3 Модель 1

3.1 Вывод формул для расчёта распределений

Проведём вывод необходимых распределений для рассматриваемой модели, пользуясь фактами из теории вероятностей:

$a \sim R[a_{min}; a_{max}]$

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

$$\mathbb{E}[a] = \frac{a_{min} + a_{max}}{2}$$

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

$b \sim R[b_{min}; b_{max}]$ аналогично:

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

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

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

$p(b|a)$

$$p(b|a) = \frac{p(a,b)}{p(a)} = \frac{\sum\limits_{c=0}^{a+b}\sum\limits_{d=c}^{2c} p(a,b,c,d)}{p(a)} = p(b) \cdot \sum_{c=0}^{a+b} p(c|a,b) \cdot \sum_{d=c}^{2c} p(d|c) = p(b)$$

Величины a и b независимы в рамках данной модели.

$c|a,b \sim B(a,p_1) + B(b,p_2)$

Пусть $c = x + y$, где $x \sim B(a,p_1), y \sim B(b,p_2)$, тогда

$$p(c|a,b) = \sum_{k=0}^{c} p(x=k; a, p_1) \cdot p(y = c-k; b, p_2) =$$

$$= \sum_{k=0}^{c} C_a^k p_1^k (1-p_1)^{a-k} \cdot C_b^{c-k} p_2^{c-k} (1-p_2)^{b+k-c}$$

$d|c \sim c + B(c, p_3)$

$$p(d|c) = C_c^{d-c} p_3^{d-c} (1-p_3)^{2c-d}$$

$c|a$

$$p(c|a) = \frac{p(a,c)}{p(a)} = \frac{\sum\limits_{b=b_{min}}^{b_{max}} \sum\limits_{d=0}^{2(a+b)} p(a,b,c,d)}{p(a)} =$$

$$= p(b) \cdot \sum_{b=b_{min}}^{b_{max}} p(c|a,b) \cdot \sum_{d=0}^{2(a+b)} p(d|c) = p(b) \cdot \sum_{b=b_{min}}^{b_{max}} p(c|a,b)$$

$c|b$ аналогично:

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

---

c

$$p(c) = \sum_{a=a_{min}}^{a_{max}} \sum_{b=b_{min}}^{b_{max}} \sum_{d=0}^{2(a+b)} p(a,b,c,d) =$$

$$= p(a) \cdot p(b) \cdot \sum_{a=a_{min}}^{a_{max}} \sum_{b=b_{min}}^{b_{max}} p(c|a,b) \cdot \sum_{d=0}^{2(a+b)} p(d|c) = p(a) \cdot p(b) \cdot \sum_{a=a_{min}}^{a_{max}} \sum_{b=b_{min}}^{b_{max}} p(c|a,b)$$

---

d

$$p(d) = \sum_{c=0}^{a_{max}+b_{max}} p(d|c) p(c)$$

3.2 Априорные распределения

Требуется рассчитать математические ожидания и дисперсии априорных распределений a, b, c и d.

Пусть для некоторый случайной величины x известно её распределение, тогда

$$\mathbb{E}[x] = \sum_{x=x_{min}}^{x_{max}} xp(x)$$

$$\mathbb{D}[x] = \sum_{x=x_{min}}^{x_{max}} x^2p(x) - (\mathbb{E}[x])^2$$

В пункте 3.1 было показано, как получить априорные распределения, имея $p(c|a, b)$ и $p(d|c)$.

Результаты приведены в таблице 1:

Величина	$\mathbb{E}$	$\mathbb{D}$
a	22.5	21.25
b	300	850
c	26.25	27.3125
d	39.375	68.0156

Таблица 1: Априорные распределения

3.3 Прогноз величины b

Требуется пронаблюдать, как происходит уточнение прогноза для величины b с приходом новой информации.

Рассмотрим распределения $b, b|a, b|a, d$. Как было показано выше в пункте 3.1, b и $b|a$ распределены одинаково, поэтому остаётся сравнить только b с $b|a, d$.

На рисунке ? показано распределение соответствующих величин, а также указаны их матожидания и дисперсии.

3.4 Влияние параметров p_1 и p_2

3.5 Временные замеры

4 Модель 2

4.1 Априорные распределения

Аналогично пункту для модели 1.

Результаты приведены в таблице 2:

Величина	$\mathbb{E}$	$\mathbb{D}$
a	22.5	21.25
b	300	850
c	26.25	33.6875
d	39.375	82.3594

Таблица 2: Априорные распределения

4.2 Прогноз величины b

4.3 Влияние параметров p_1 и p_2

4.4 Временные замеры

5 Сравнение моделей 1 и 2

Модель 2 получена из первой путём предельного перехода... Значит, первая описывает всё лучше и правильней при малых значениях.

6 Выводы

Печаль-беда

7 Список литературы

- http://www.machinelearning.ru/wiki/index.php?title=
- Лекции и семинары по графическим моделям