Санкт-Петербургский государственный университет
Прикладная математика и информатика

Отчет о научно-исследовательской работе

Применение цепей Маркова в анализе данных

Выполнил:
Саттаров Никита Дмитриевич
группа 19.Б04-мм

Научный руководитель:
к. ф.-м. н., доцент
Н. Э. Голяндина

Санкт-Петербург
2021

Введение

В книге [1, Гл. 6] вводится следующее определение цепи Маркова:

Определение 1. *Конечная последовательность дискретных случайных величин* $\{X_n\}_{n\geq 0}$ *называется конечной цепью Маркова, если*

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n). \quad (1)$$

Таким образом, в простейшем случае условное распределение последующего состояния цепи Маркова зависит только от текущего состояния и не зависит от всех предыдущих состояний.

Область значений случайных величин $\{X_n\}$ *называется* ***пространством состояний*** *цепи, а номер* n *— номером шага.*

Определение 2. *Матрица* $P(n)$*, где*

$$P_{ij}(n) \equiv \mathbb{P}(X_{n+1} = j \mid X_n = i) \quad (2)$$

называется ***матрицей переходных вероятностей*** *на* n*-м шаге, а вектор* $\mathbf{p} = (p_1, p_2, \ldots)^\top$*, где*

$$p_i \equiv \mathbb{P}(X_0 = i) \quad (3)$$

— ***начальным распределением*** *цепи Маркова.*

В общем случае имеем конечное множество состояний системы S и конечное множество объектов O. Объекты могут перемещаться из одного состояния системы в другое. Наблюдением назовем перемещение некоторого объекта $o \in O$ из состояния $s_{start} \in S$ в состояние $s_{end} \in S$. Объекты перемещаются независимо друг от друга.

Имеем таблицу наблюдений:

$\forall i: \ s_{start_i}, s_{end_i} \in S, \ \forall k, j, m: \ o_k, o_j, o_m \in O$

Хотим извлечь из таблицы (1) информацию: вероятность попасть из s_i в s_j, среднее количество шагов, чтобы попасть из s_i в s_j и т.д. $\forall i, j: \ s_i, s_j \in S$

Можем это сделать с помощью преобразования матрицы переходных вероятностей (2).

Номер наблюдения	Объект	Начальное состояние	Конечное состояние
1	o_k	s_{start_1}	s_{end_1}
2	o_j	s_{start_2}	s_{end_2}
3	o_m	s_{start_3}	s_{end_3}
...	...	...	...

Таблица 1. Данные о наблюдениях в общем виде

В качестве таблицы наблюдений (1) берем таблицу данных с сайта *Kaggle*: `https://www.kaggle.com/pronto/cycle-share-dataset?select=trip.csv`. Она содержит наблюдения о поездках прокатных велосипедов от одной станции к другой за конечный промежуток времени. Станции будут являться состояниями цепи Маркова. А поездки — переходами от из одного состояния в другое. Всего станций: 58. Всего поездок: 286858.

Текущая задача заключается в проверки корректности цепи Маркова, построенной по данной таблице данных.

Для этого мы:

- Построим цепь Маркова, которая будет соответствовать вероятностям попасть из состояния s_i в состояние s_j для нашей таблицы данных
- Сгенерируем новые наблюдения на основе вероятностей из цепи Маркова
- Построим новую цепь Маркова по сгенерированным наблюдениям
- Сверим две получившиеся цепи

Глава 1

Построение цепи Маркова по таблице данных

1.1. Импорт таблиц и устранение ошибок в таблице данных

Импортируем две таблицы данных: *poezdki.csv*, в которой содержатся даные о поездках, и *station.csv*, в которой содержатся данне о станциях.

Сталкиваемся с ошибкой, что первые 50793 поездки записаны в таблицу данных дважды. Исправляем это, сделав срез таблицы.

Для текущей задачи необходимо использовать столбец *station_id* с идентификаторами станций из таблицы *station.csv* и столбцы $from\_station\_id$ и $to\_station\_id$ с идентификаторами станций, откуда была совершена поездка, и идентификаторами станций, куда была совершена поездка соответственно, из таблицы *poezdki.csv*.

На данный момент работы опустим столбец "Объект"из таблицы данных (1) и будем считать, что поездки совершались на одном велосипеде.

Далее берем необходимые нам столбцы из таблиц, а конкретно *station_id*, $from\_station\_id$ и $to\_station\_id$, и сортируем их в алфавитном порядке.

	0	1	2	3	4
station_id	BT-01	BT-03	BT-04	BT-05	CBD-03

Рис. 1.1. Идентификаторы первых 5 станций, отсортированных в алфавитном порядке

Здесь возникает несколько проблем с обработкой таблиц. Во-первых, не все станции, которые описаны в столбцах from_station_id и to_station_id (1.2), присутствуют в списке station_id (1.1). Во-вторых, были такие ячейки столбца to_station_id, которые содержали «мусор» или не содержали ничего вообще. Эти проблемы будем решать на стадии построения марковской цепи.

	from_station_id	to_station_id
0	8D OPS 02	8D OPS 02
1	8D OPS 02	8D OPS 02
2	BT-01	BT-01
3	BT-01	BT-01
4	BT-01	BT-01
...	...	...
236060	WF-04	WF-04
236061	WF-04	WF-04
236062	WF-04	WF-04
236063	WF-04	WF-04
236064	WF-04	WF-04

Рис. 1.2. Таблица поездок, отсортированная в алфавитном порядке станций

1.2. Алгоритм построения цепи Маркова по таблице данных

Обозначение 1. $\exists S$ — *множество кодов станций.*

$S = \{BT-01, BT-03, \ \dots \ , WF-04\}$

$N = |S|$ — *количество станций.*

Обозначение 2. $\exists d : S \to \mathbb{N}$ — *словарь кодов станций такой, что*

$$d(\text{"}BT-01\text{"}) = 0$$
$$d(\text{"}BT-03\text{"}) = 1$$
$$\dots$$
$$d(\text{"}WF-04\text{"}) = 57$$

Обозначение 3. *Под поездкой будем подразумевать упорядоченную пару* (s_{start}, s_{end}), *где* s_{start} — *код станции старта,* s_{end} — *код станции финиша,* $s_{start}, s_{end} \in S$.
Тогда множество $t_n = \{(s_{start_i}, s_{end_i}) \mid s_{start_i}, s_{end_i} \in S, \ i \in \{0, \ \dots, n\}\}$ — *таблица поездок размера* n.

Марковская цепь, основанная на таблице даннных, строится следующим образом: за переходные состояния принимаем станции, откуда выезжали велосипедисты. А сумма поездок из станции s_i в станцию s_j, деленная на общее количество поездок из станции

s_i, как раз дает нам вероятность попасть из станции s_i в станцию s_j для случайного велосипедиста, $s_i, s_j \in S \ \forall i, j \in \{0, \ \ldots, N\}$.

Чтобы построить Марковскую цепь, необходимо создать словарь d (2), который идентификаторам станций из таблицы будет сопоставлять число (индекс строки/столбца ячейки переходной матрицы).

С проблемой того, что в столбце $to\_station\_id$ может находиться непонятно что, справляемся следующим образом: пробегаем по списку всех поездок, если не получилось взять значение по ключу словаря d, значит поездка состоит из «мусора»и мы её обнуляем, в дальнейшем не используя для построения Марковской Цепи.

Обозначение 4. $\beth M : N \times N \to \mathbb{N}$ — *матрица, в каждой ячейке которой стоит количество поездок из* s_i *в* s_j *в таблице поездок* t_n.

$M_{ij} = \sum\limits_{\substack{n \geq k \geq 1, \\ d(s_{start_k}) = i, \\ d(s_{end_k}) = j}} 1$, *где* $(s_{start_k}, s_{end_k}) \in t_n$.

Обозначение 5. $\beth P_{ij} = \frac{M_{ij}}{\sum\limits_{j=0}^{N-1} M_{ij}}$ — *отнормированная по строкам матрица* M_{ij}.

Так как $\forall i \in \{0, 1, \ \ldots, N-1\} : \sum\limits_{j=0}^{N-1} (P_{ij}) = 1$, *то* P_{ij} — *построенная по таблице* t_n *матрица переходных вероятностей, в каждой ячейке* (i, j) *которой стоит вероятность попасть из станции* s_i *в станцию* s_j, $i, j \in \{0, 1, \ \ldots, N-1\}$.

В нашем случае $N = 58$, $n = 236044$.

Для построения матрицы переходных вероятностей P (2) пробегаем по таблице поездок и строим матрицу M (4), в каждой ячейке которой стоит количество поездок, совершенных из станции s_i в станцию s_j, где (i, j) — индексы ячейки матрицы. Далее отнормировав каждую строку матрицы M получаем матрицу P, в каждой ячейке которой стоит вероятность попасть из станции s_i в станцию s_j, где (i, j) — индексы ячейки матрицы. Таким образом мы получили матрицу переходных вероятностей (1.3) и построили Марковскую цепь (1).

	0	1	2	3	4
0	0.066759	0.019378	0.012138	0.042909	0.019059
1	0.045828	0.043240	0.012028	0.031060	0.028471
2	0.046609	0.027756	0.038230	0.015711	0.018853
3	0.074809	0.030886	0.015443	0.044124	0.033293
4	0.051306	0.033967	0.018290	0.028979	0.058195

Рис. 1.3.

Вероятности попасть из первых 5 станций в первые 5 станций в получившейся матрице переходных вероятностей

Глава 2

Генерация новых наблюдений и построение цепи Маркова по новым наблюдениям

2.1. Выбор начального состояния

Для начала нужно определиться, с какой станции мы будем начинать генерацию новых данных. Для этого построим вектор nss размерности N такой, что $nss_i = \sum_{j=0}^{N-1} M_{ij}$. То есть в каждой i-ой ячейке находится сумма всех исходящих поездок из станции s_i. Отнормировав вектор nss получим вектор $pss = \frac{nss}{||nss||_1}$, в каждой i-ой ячейке которого будет стоять вероятность начать поездку в станции s_i.

2.2. Генерация поездок в новой таблице данных

Будем генерировать 10.000.000 поездок. Для этого создадим новую матрицу P_{new}, изначально заполненную нулями, и новую таблицу $t'_{10.000.000}$ (2.1, a) со столбцами $station\_start$ и $station\_end$. k-ую поездку в таблице $t'_{10.000.000}$ будем обозначать упорядоченной парой (s_{start_k}, s_{end_k}), где $s_{start_k}, s_{end_k} \in S,\ k \in \{1, 2, \ \ldots, 10.000.000\}$.

Определение 3. *Функция $\xi : \Omega \to X$ называется случайной величиной, если для любого борелевского множества $B \in \mathfrak{B}(X)$ множество $\xi^{-1}(B)$ является событием, т.е. принадлежит σ-алгебре $\mathfrak{F}$.*

Обозначение 6. *$\beth\, \Omega = \{0, 1, \ \ldots, N-1\}$, случайная величина $\xi : \Omega \to S$ такая, что $\xi(\omega) = d^{-1}(\omega)\ \forall \omega \in \Omega$ (3), и*

ξ	$BT-01$	$BT-03$	$BT-04$	$\ldots$	$WF-04$
P	pss_0	pss_1	pss_2	$\ldots$	pss_{N-1}

Таблица 2.1. Распределение случайной величины ξ

Выберем случайную начальную станцию $s_{initial}$, с которой будем стартовать генерацию, по вероятности из вектора pss, то есть по распределению случайной величины ξ. Запишем её в таблицу данных. Таким образом $s_{start_1} = s_{initial} = \xi$.

Обозначение 7. $\exists\, \Omega = \{0, 1, \ldots, N-1\},\ \forall i \in \Omega$: *случайная величина* $\eta_i : \Omega \to S$ *такая, что* $\eta_i(\omega) = d^{-1}(\omega)\ \forall \omega \in \Omega$ (3), *и*

η	$BT-01$	$BT-03$	$BT-04$	$\ldots$	$WF-04$
P	P_{i_0}	P_{i_1}	P_{i_2}	$\ldots$	$P_{i_{N-1}}$

Таблица 2.2. Распределение случайной величины η_i

Далее выберем случайную следующую станцию s_{end_1} по вероятности из вектора P_i, то есть по распределению случайной величины η_i, где $i = d(s_{start_1})$, запишем её в таблицу. Первая поездка таблицы $t'_{10.000.000}$ равна (s_{start_1}, s_{end_1}). Следующая поездка будет начинаться из станции s_{end_1}, таким образом $s_{start_2} = s_{end_1}$. Продолжаем генерировать следующие поездки: $s_{end_i} = \eta_{d(s_{start_i})},\ s_{start_i} = s_{end_{i-1}},\ i = 1 \ldots 10.000.000$.

2.3. Построение новой цепи Маркова по сгенерированным поездкам

Параллельно с генерацией поездок строим матрицу M_{new}, в каждой ячейке которой стоит количество поездок из s_i в s_j в таблице поездок $t'_{10.000.000}$, по аналогии с матрицей M (4).

Определение 4. *В пространстве m-мерных векторов* R^m *введена фиксированная норма* $||x||$ *(например, одна из норм* $||x||_p$, $1 \le p \le \infty$, *в зависимости от того, какая ошибка нам нужна: линейная, среднеквадратичная и т.п.). В этом случае в качестве меры степени близости векторов* x *и* x^* *естественно использовать величину* $||x - x^*||$, *являющуюся аналогом расстояния между точками* x *и* x^*.
Введём абсолютную и относительную ошибки вектора x^* *с помощью формул:*

$$\Delta(x^*) = ||x - x^*||,\ \delta(x^*) = \frac{||x - x^*||}{||x||}.$$

Определение 5. *Нормой матрицы A называется величина* $||A|| = \max\limits_{x \neq 0} \frac{||Ax||}{||x||}$. *Абсолютная и относительная погрешности матрицы вводятся аналогично погрешностям вектора с помощью формул:*

$$\Delta\left(A^*\right) = ||A - A^*||, \ \delta\left(A^*\right) = \frac{||A - A^*||}{||A||}.$$

Через каждое количество поездок, кратное порядку 10, начиная с 10.000 и до 10.000.000, нормируем по строкам матрицу M_{new}, таким образом получаем матрицу P_{new} (2.1, *б*) по аналогии с получением матрицы переходных вероятностей P из матрицы M (5) и считаем абсолютную и относительную ошибки по формуле (5), где A — начальная матрица переходных вероятностей P, A^* — новая матрица P_{new}, построенная на сгенерированных данных.

	station_start	station_end
0	CBD-05	SLU-15
1	SLU-15	BT-01
2	BT-01	WF-01
3	WF-01	PS-05
4	PS-05	SLU-19

(*а*)

Первые 5 строк в таблице сгенерированных данных t'

	0	1	2	3	4
0	0.067347	0.019364	0.012303	0.043233	0.019254
1	0.046165	0.042317	0.011781	0.031573	0.028216
2	0.045980	0.027391	0.038047	0.015695	0.018689
3	0.075738	0.030146	0.015622	0.044520	0.033332
4	0.051179	0.034043	0.018512	0.028764	0.058755

(*б*)

Вероятности попасть из первых 5 станций в первые 5 станций в новой построенной по сгенерированным данным матрице переходных вероятностей P_{new}

Глава 3

Сравнение двух цепей Маркова

Программа выдала следующие результаты:

Количество поездок	Абсолютная погрешность	Относительная погрешность
10.000	0.895783	0.439285
100.000	0.277931	0.136295
1.000.000	0.092052	0.045583
10.000.000	0.026711	0.013099

Таблица 3.1. Абсолютные и относительные погрешности рассчетов

В учебном пособии по теории вероятностей [2, Гл. 8] вводится следующая теорема:

Теорема 1 (Центральная предельная теорема Ляпунова)
$\exists\ \xi_1, \xi_2, \ldots$ — *независимые и одинаково распределённые случайные величины с конечной и ненулевой дисперсией:* $0 < \mathbf{D}\xi_1 < +\infty$. $S_n = \xi_1 + \ldots + \xi_n$. *Тогда имеет место слабая сходимость*

$$\frac{S_n - n\mathbf{E}\xi_1}{\sqrt{n\mathbf{D}\xi_1}} \Rightarrow N_{0,1} \tag{3.1}$$

Откуда не трудно вывести соотношение:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbf{E}\xi_1\right| \leq k\frac{\sqrt{\mathbf{D}\xi_1}}{\sqrt{n}}\right) \Rightarrow 2\Phi(k) - 1 \tag{3.2}$$

Заметим, что для больших n*, при увеличении* n *отклонение* $\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbf{E}\xi_1\right|$ *уменьшается в* $\sqrt{n}$ *раз.*

Как мы видим (3.1), при увеличении числа поездок в 100 раз абсолютная и относительная погрешности уменьшаются примерно в 10 раз. Значит ошибки обратно пропорциональны корню из количества поездок. Ссылаясь на центральную предельную теорему (1) и принимая во внимание тот факт, что в нашем случае можем считать каждый элемент матрицы P_{new} преобразованием величин ξ и η_i, а каждый элемент матрицы P — математическим ожиданием этого преобразования, можем заключить, что изначально построенная матрица переходных вероятностей P построена корректно.

Глава 4

Следующая задача

Следующей задачей я ставлю рассчет матрицы первых достижений по новой матрице переходных вероятностей Марковской цепи, построенной на сгенерированных данных, и проверку корректности её значений с помощью алгоритма подсчёта времен первых достижений через таблицу сгенерированных данных.

Список литературы

1. Stirzaker DR, Grimmett GR. Probability and random processes. — Clarendon Press, 1992.
2. Чернова Н.И. Теория вероятностей. — Изд-во МЦНМО, 2007.