Итоги семинара 31.03.20

На семинаре мы обсудили введение в теорию цепей Маркова и их приложения к задачам страхования и ранжирования выдачи поискового запроса.

Цепь Маркова (ЦМ) – это последовательность случайных чисел $(X_n)_{n\geqslant 0} = (X_0, X_1, X_2, \dots)$ с конечным (или счетным) числом исходов, (то есть в каждый момент времени система может принимать одно из состояний $E = \{1, 2, \dots\}$), обладающая *марковским свойством*:

$$\mathbb{P}\left(X_{n+1} = j \mid X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i\right) = \mathbb{P}\left(X_{n+1} = j \mid X_n = i\right)$$

для любого n и любых $i, j, i_0, \dots, i_{n-1} \in E$.

Если множество E конечно, $E = \{1, 2, \dots, N\}$, то цепь называется ЦМ с *конечным числом состояний.*

ЦМ называется *однородной* (по времени), если вероятности перехода из i-го состояния в j-е не зависят от момента времени n, в который мы наблюдаем процесс. Обозначим эти вероятности через p_{ij}:

$$p_{ij} = \mathbb{P}\left(X_1 = j \mid X_0 = i\right) = \mathbb{P}\left(X_{n+1} = j \mid X_n = i\right).$$

Матрица $\mathbf{P} = (p_{ij})_{i,j=1}^{N}$ называется *матрицей переходных вероятностей.* Пусть

$$\pi_j^n = \mathbb{P}\left(X_n = j\right), j = 1, \dots, N, \quad \sum_j \pi_j^n = 1.$$

Вектор $\pi^0 = (\pi_1^0, \dots, \pi_N^0)$ – начальное распределение цепи. Имеет место соотношение

$$\pi^n = \pi^{n-1}\mathbf{P} = \dots = \pi^0\mathbf{P}^n. \tag{1}$$

Bonus-malus system

Бонус-малус (англ. Bonus-Malus System, BMS) – система тарифных коэффициентов, обычно применяемая в страховом бизнесе, которая изменяет размер страховой премии, которую платит клиент (страхователь) страховщику в зависимости от его истории страховых случаев.

По прошествии каждого года страхового договора, клиент попадает в один из классов дисконтирования страховых выплат. Чем выше класс, тем большую скидку на страховые выплаты он получает. Понижение или повышение класса зависит от того, были ли в прошедшем году страховые случаи.

Мы рассмотрели пример с тремя классами: 0, 1 и 2. Если за предыдущий год не было страховых случаев, то класс клиента повышается

(остается вторым, если уже был там). Если был хотя бы один страховой случай, клиент снова попадает в нулевой класс. Оказалось, что после 2 года страхования, распределение по классам перестает меняться. Такие распределения называются *стационарными.*

PageRank

Алгоритм PageRank решает задачу ранжирования поисковой выдачи. Основная идея алгоритма – наиболее релевантными должны быть страницы, имеющие большую «важность», то есть на них часто ссылаются авторитетные источники, которые тоже сами часто посещаются. Алгоритм действует так: мы считаем, что некий средний пользователь Интернета в исходный момент времени находится на одной из страниц. Затем этот пользователь начинает случайным образом начинает перемещаться, переходя на каждой странице по одной из ссылок, которые ведут на другую страницу рассматриваемого множества. На любой странице все допустимые ссылки имеют одинаковую вероятность нажатия. Если с текущей страницы нет исходящих ссылок, то пользователь равновероятно переходит на любую другую.

Так мы задаём цепь Маркова: страницы — это возможные состояния, переходные вероятности задаются ссылками со страницы на страницу (взвешенными таким образом, что на каждой странице все связанные страницы имеют одинаковую вероятность выбора). Такое ранжирование можно сделать, если найти стационарное распределение соответсвующей цепи Маркова и выдавать первыми ссылки на те страницы, вероятность перехода на которые в стационарном распределении самая большая. Однако описанный алгоритм не гарантирует существования стационарного распределения. Для того, чтобы этого добиться, преобразуем матрицу переходных вероятностей:

$$P^* := (1-d)\frac{1}{n} \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} + d \cdot P,$$

где P – исходная матрица, параметр d обычно полагается равным 0.85.

Такая версия алгоритма слишком проста и поддается «взлому». Было показано, как владелец любой страницы может сделать ее PR сколь угодно большим с помощью «фермы» из страниц-клонов.

Что почитать еще?

Общая теория цепей Маркова изложена в классическом учебнике [1] и [2].

Хороший обзор на Bonus-Malus System сделан в учебнике [3]. Там можно почитать на русском языке про задачи, связанные с BMS: например, про рассчет эффективности тарифной системы или «жажду бонуса», где рассчитывается, выгодно ли клиенту заявлять об аварии.

Собственно, статья Сергея Брина и Ларри Пейджа, в которой впервые были описаны поисковая система Google и алгоритм PageRank [4]. Ее перевод на русский – [5].

Хорошая презентация, в которой еще раз (более подробно) разбирается PageRank [6]. Для продвинутых: обзор стратегий типа Sybil обмана алгоритма PageRank [7]. Еще можно просто погуглить PageRank: написано куча статей, где разбираются алгоритмы обмана старых версий, предлагаются улучшения и обсуждаются вычислительные сложности.

Список литературы

[1] Ширяев, Альберт Николаевич. Вероятность. Изд-во МЦНМО, 2011.

[2] Чжун, Кай-лай. Однородные цепи Маркова. Мир, 1964.

[3] `http://insurance-institute.ru/library/kaas/kaas06.pdf`

[4] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine."(1998).

[5] `https://web.archive.org/web/20130625003347/http://wseob.ru/seo/searchengine-anatomy`

[6] `http://statweb.stanford.edu/~tibs/sta306bfiles/pagerank/ryan/01-24-pr.pdf`

[7] Cheng, Alice, and Eric Friedman. "Manipulability of PageRank under sybil strategies."(2006).