Машинное обучение, осень 2016. Никита Волков

Ранжирование

Семинар

0. Ключевые слова темы

- Задача ранжирования, оценка релевантности (асессорская), функция релевантности (модель)
- Метрики качества: MAP, NDCG, pFound
- RankSVM
- RankNet, его факторизация, LambdaRank, LambdaMART
- PageRank, эргодическая теорема

1. Задача ранжирования

Пусть D — некоторая коллекция текстовых документов и Q — множество запросов. Обозначим D_q неупорядоченный набор документов, потенциально релевантных запросу $q \in Q$. Основная задача ранжирования — упорядочить документы внутри D_q по убыванию степени их релевантности запросу, то есть более релевантные документы должны иметь более высокий ранг.

Пример. Задача ранжирования — основа поисковых систем. В наших обозначениях множество D — база документов, проиндексированных поисковой системой. Пользователи задают поисковой системе запросы из множества Q. По запросу $q \in Q$ поисковая система в своей базе D должна найти все документы D_q, которые соответствуют заросу q. После того, как документы D_q найдены, поисковой системе нужно упорядочить их по убыванию степени их релевантности запросу. Важность последнего этапа поиска состоит в том, что часто на запросы находятся сотни тысяч различных документов, и ни один пользователь не будет просматривать их всех чтобы найти нужный. Скорее всего пользователь просмотрит не более 10 первых документов. Как следствие, поисковая система должна качественно делать ранжирование документов по степени их релевантности запросу.

Отношение порядка на парах запрос-документ. Если документ d_1 релевантнее документа d_2 по запросу $q \in Q$ (то есть $d_1, d_2 \in D_q$), то будем писать $(q, d_1) \triangleright (q, d_2)$. Таким образом определяется порядок на парах запрос-документ. Отметим, что этот порядок определен только внутри одного конкретного запроса.

Специальные люди, называемые асессорами, для некоторых запросов размечают пары запрос-документ, определяя для них *оценку релевантности* $y(q, d)$. Обозначим X — множество пар (d, q), для которых определена оценка релевантности $y(q, d)$, и назовем их асессорской выборкой. Стоит отметить, что в настоящее время процедура оценивания релевантности ассессорами достаточно хорошо формализована в виде списка многочисленных правил.

Задача. Чтобы упорядочить пары запрос-документ, на основе асессорской выборки будем искать *функцию релевантности* $\alpha(q, d)$, которая "как можно лучше" удовлетворяет условию

$$(q, d_1) \triangleright (q, d_2) \quad \Leftrightarrow \quad \alpha(q, d_1) > \alpha(q, d_2).$$

2. Метрики качества

Различные ранжирования пар запрос-документ сравниваются при помощи метрик качества ранжирования. Для измерения качества ранжирования традиционно (традиция задана конференцией TREC) используются меры точности. Приведем описание одной из них.

NDCG

Метрика NDCG не накладывает огрничений на возможные значения оценки релевантности $y(q, d)$.

Зафиксируем запрос q и упорядочим все документы по этому запросу в соответствии со значением $\alpha(q, d)$, получив набор $\{d_q^{(i)}\}_i$. Метрика DCG (Discounted Cumulative Gain, взвешенная сумма выигрышей) определяется как

$$DCG_n(q) = \sum_{i=1}^{n} G_q\left(d_q^{(i)}\right) \; D(i),$$

где величина $G_q(d) = 2^{y(q,d)} - 1$ придает больший вес релевантным документам, а величина $D(i) = 1/\log_2(i+1)$ придает больший вес документам, которые функция ранжирования $\alpha(q, d)$ посчитала более релевантными.

Значения получаемой метрики сильно зависят от запроса. Чтобы снизить эту зависимость, значение DCG нормируют, получая NDGC — нормированный DCG

$$NDCG_n(q) = \frac{DCG_n(q)}{\max DCG_n(q)},$$

где максимум берется по всем возможным ранжированиям. Очевидно, что он достигается, если построенная функция релевантности задает тот же порядок, что и асессорские оценки релевантности.

3. PageRank

История

(Взято с Википедии https://ru.wikipedia.org/wiki/PageRank)

В 1996 году Сергей Брин и Ларри Пейдж, тогда ещё аспиранты Стэнфордского университета, начали работу над исследовательским проектом BackRub — поисковой системой по Интернету, использующей новую тогда идею о том, что веб-страница должна считаться тем «важнее», чем больше на неё ссылается других страниц, и чем более «важными», в

свою очередь, являются эти страницы. Через некоторое время BackRub была переименована в Google. Первая статья с описанием применяющегося в ней метода ранжирования, названного PageRank, появилась в начале 1998 года, за ней следом вышла и статья с описанием архитектуры самой поисковой системы.

Их система значительно превосходила все существовавшие тогда поисковые системы, и Брин с Пейджем, осознав её потенциал, основали в сентябре 1998 года компанию Google Inc., для дальнейшего её развития как коммерческого продукта.

Описание

Введем понятие веб-графа. Ориентированный граф $G = (V, E)$ называется веб-графом, если

- $V = \{url_i\}_{i=1}^n$ — некоторое подмножество страниц в интернете, каждой из которых соответствует адрес url_i.
- Множество E состоит из тех и только тех пар (url_i, url_j), для которых на странице с адресом url_i есть ссылка на url_j.

Рассмотрим следующую модель поведения пользователя. В начальный момент времени он выбирает некоторую страницу из V в соответствии с некоторым распределением $\Pi^{(0)}$. Затем, находясь на некоторой странице, он может либо перейти по какой-то ссылке, которая размещена на этой странице, либо выбрать случайную страницу из V и перейти на нее (damping factor). Считается, что если пользователь выбирает переход по ссылке, то он выбирает равновероятно любую ссылку с данной страницы и переходит по ней. Если же он выбирает переход не по ссылке, то он также выбирает равновероятно любую страницу из V и переходит на ее (в частности может остаться на той же странице). Будем считать, что переход не по ссылке пользователь выбирает с некоторой вероятностью $p \in (0, 1)$. Соответственно, переход по ссылке он выбирает с вероятностью $1 - p$. Если же со страницы нет ни одной ссылки, то будем считать, что пользователь всегда выбирает переход не по ссылке.

Описанная выше модель поведения пользователя называется моделью PageRank. Нетрудно понять, что этой модели соответствует некоторая марковская цепь. Опишем ее.

- Множество состояний: V
- Начальное распределение: $\Pi^{(0)}$
- Переходные вероятности: $P = (p_{ij})$, где

$$p_{ij} = \begin{cases} \frac{1-p}{N_i} I\{(url_i, url_j) \in E\} + \frac{p}{|V|}, & \text{если } N_i > 0; \\ \frac{1}{|V|}, & \text{если } N_i = 0. \end{cases},$$

$$N_i = |\{j \in V \mid (url_i, url_j) \in E\}|$$

Вычисление

Для начала напомним теорему из курса случайных процессов.

Теорема 1 (Эргодическая)**.** *Пусть* $(X_n, n \in \mathbb{Z}_+)$ *— однородная марковская цепь со значениями в* $\mathscr{X} = \{1, ..., N\}$ *и матрицей переходных вероятностей* $P = (p_{ij})$*. Если существует число* n_0 *такое, что все элементы матрицы* P^{n_0} *положительны, то существует* $\Pi = (\pi_1, ..., \pi_N)$ *(назыв. эргодическое распределение), для которого выполнено*

1. $\pi_j > 0$ *для любого* j *и* $\sum_{j=1}^{N} \pi_j = 1$*;*
2. $\lim_{n \to +\infty} p_{ij}^{(n)} = \pi_j$ *для любых* i, j*, где* $p_{ij}^{(n)}$ *— элемент матрицы* P^n *(матрица переходных вероятностей за* n *шагов;*
3. *Распределение* Π *является предельным и единственным стационарным.*

Причем скорость сходимости экспоненциальная

$$\left| p_{ij}^{(n)} - \pi_j \right| < (1 - \varepsilon)^{\lfloor n/n_0 \rfloor},$$

где $\varepsilon = \min_{ij} p_{ij}^{(n_0)} > 0$.

Марковская цепь в модели PageRank является эргодической, поскольку все элементы матрицы переходов положительны, то есть $n_0 = 1$. А это означает, что цепь имеет некоторое эргодическое распределение Π, которое является предельным и единственным стационарным. Данное распределение называется весом PageRank для нашего подмножества интернета.

Как вычислить это распределение Π для данного веб-графа? Обычно для этого используют степенной метод (power iteration), суть которого состоит в следующем. Выбирается некоторое начальное распределение $\Pi^{(0)}$. Далее производится несколько итераций по формуле $\Pi^{(k)} = \Pi^{(k-1)} P$, где P — матрица переходных вероятностей цепи, до тех пор, пока $\left|\left|\Pi^{(k)} - \Pi^{(k-1)}\right|\right| > \varepsilon$. Распределение $\Pi^{(k)}$ считается приближением распредения Π.

Из теоремы так же следует, что абсолютное отклонение вероятностей от предела будет не превосходить $\left(1 - \frac{p}{|V|}\right)^n$.

Пример

Ниже на первой картинке приведен пример реального веб-графа. Это веб-граф сайта кафедры Дискретной математики, который построил студент 494 группы Степан Каргальцев в практическом задании по курсу случайных процессов. Второй рисунок — этот же веб-граф, но размер вершин и их цвет пропорционален его весу RageRank.