

# Машинное обучение

Лекция 15

Ранжирование

## На прошлых лекциях

- Дано: матрица «объекты-признаки»  $X$  и ответы  $y$
- Модель должна выдавать прогнозы, близкие к истинным ответам

# На прошлых лекциях

- Методы обучения с учителем: линейные модели, решающие деревья, случайные леса, ...
- Дано: матрица «объекты-признаки»  $X$  и ответы  $y$
- Найти: модель  $a(x)$
- Модель должна выдавать прогнозы, близкие к истинным ответам

**W** **Машинное обучение** — Википедия[ru.wikipedia.org](https://ru.wikipedia.org) > Машинное обучение ▾

**Машинное обучение** (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

**л** **Что такое машинное обучение и почему оно может...**[lifehacker.ru](https://lifehacker.ru) > Лайфхакер > ...-mashinnoe-obuchenie ▾

**Машинное обучение** избавляет программиста от необходимости подробно объяснять компьютеру, как именно решать проблему.

**▶** **Курс «Машинное обучение» 2014 - YouTube**[youtube.com](https://youtube.com) > playlist?list=...\_b9zqEQiiBtC ▾

Курс "**Машинное обучение**" является одним из основных курсов Школы, поэтому он является обязательным для всех студентов ШАД.

**P** **Машинист электропоезда - обучение** | Про профессии.ру[proprof.ru](https://proprof.ru) > Машинист электропоезда ▾

**Машинист электропоезда - обучение.** И метрополитен, и РЖД приглашают на **обучение** в собственные учебно-производственные центры.

**т** **Обучение - машина - Большая Энциклопедия Нефти...**[ngpedia.ru](https://ngpedia.ru) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

w **Машинное обучение** — Википедия[ru.wikipedia.org](https://ru.wikipedia.org) > Машинное обучение ▾

**Машинное обучение** (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

л **Что такое машинное обучение и почему оно может...**[lifehacker.ru](https://lifehacker.ru) > Лайфхакер > ...-mashinnoe-obuchenie ▾

**Машинное обучение** избавляет программиста от необходимости подробно объяснять компьютеру, как именно решать проблему.

v **Курс «Машинное обучение» 2014 - YouTube**[youtube.com](https://youtube.com) > playlist?list=...\_b9zqEQiiBtC ▾

Курс "**Машинное обучение**" является одним из основных курсов Школы, поэтому он является обязательным для всех студентов ШАД.

P **Машинист электропоезда - обучение** | Про профессии.ру[proprof.ru](https://proprof.ru) > Машинист электропоезда ▾

**Машинист электропоезда - обучение.** И метрополитен, и РЖД приглашают на **обучение** в собственные учебно-производственные центры.

т **Обучение - машина - Большая Энциклопедия Нефти...**[ngpedia.ru](https://ngpedia.ru) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...



# Ранжирование

- Дан набор запросов  $\{q_1, \dots, q_m\}$
- Дан набор документов  $\{d_1, \dots, d_n\}$
- Нужно для каждого запроса правильно упорядочить документы
- Что такое «правильно»?

# Ранжирование

- Дан набор запросов  $\{q_1, \dots, q_m\}$
- Дан набор документов  $\{d_1, \dots, d_n\}$
- Рассматриваем пары «запрос-документ»  $(q, d)$
- Для некоторых троек  $(q, d_1, d_2)$  известно, что для запроса  $q$  документ  $d_1$  должен стоять раньше, чем  $d_2$
- Обозначение:  $R$  — множество троек  $(q, d_1, d_2)$ , для которых известен такой порядок

# Ранжирование

- Раньше: строим модель  $a(x)$ , которая приближает ответы
- Сейчас: строим модель  $a(q, d)$ , которая правильно упорядочивает документы для запросов

$$(q, d_1, d_2) \in R \Rightarrow a(q, d_1) > a(q, d_2)$$



# Пример

- Для запроса  $q$  известны пары  $(d_3, d_1)$ ,  $(d_3, d_2)$ ,  $(d_1, d_4)$
- Какие наборы прогнозов модели лучше?
- $(3, 2, 4, 1)$
- $(2, 3, 4, 1)$
- $(3, 4, 2, 1)$
- $(13, 10, 20, 7)$

# Пример

- Для запроса  $q$  известны пары  $(d_3, d_1)$ ,  $(d_3, d_2)$ ,  $(d_1, d_4)$
- Какие наборы прогнозов модели лучше?
- **(3, 2, 4, 1)**
- **(2, 3, 4, 1)**
- **(3, 4, 2, 1)**
- **(13, 10, 20, 7)**
- Важен порядок, а не абсолютные значения!

# Метрики качества ранжирования

# Целевая переменная

- Определение задачи через пары — правильно, но сложно
- Упростим постановку:
  - Объекты — пары «запрос-документ»  $x_i = (q, d)$
  - Ответы — числа  $y_i$
  - Требование — если есть объекты  $(q, d_1)$  и  $(q, d_2)$ , такие что  $y_1 > y_2$ , то должно быть  $a(q, d_1) > a(q, d_2)$

## Целевая переменная, пример

- $(q_1, d_1), 1$
- $(q_1, d_2), 0.7$
- $(q_1, d_3), 0$
- $(q_2, d_1), 0$
- $(q_2, d_2), 1$
- Для  $q_1$  должны получить ранжирование  $(d_1, d_2, d_3)$
- Для  $q_2$  должны получить ранжирование  $(d_2, d_1)$

# Качество ранжирования

## Машинное обучение — Википедия

[ru.wikipedia.org](http://ru.wikipedia.org) > Машинное обучение ▾

**Машинное обучение** (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

## Обучение машиниста бурильно-крановых машин — АНО...

[ccrp.ru](http://ccrp.ru) > rabochie/mashinist\_burilno-kranovoy... ▾

**Обучение машиниста** бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

## Обучение - машина - Большая Энциклопедия Нефти...

[ngpedia.ru](http://ngpedia.ru) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

## Обучение машиниста бурильно-крановых машин — АНО...

[ccrp.ru](http://ccrp.ru) > rabochie/mashinist\_burilno-kranovoy... ▾

**Обучение машиниста** бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

## Обучение - машина - Большая Энциклопедия Нефти...

[ngpedia.ru](http://ngpedia.ru) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

## Машинное обучение — Википедия

[ru.wikipedia.org](http://ru.wikipedia.org) > Машинное обучение ▾

**Машинное обучение** (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

## Обучение машиниста бурильно-крановых машин — АНО...

[ccrp.ru](http://ccrp.ru) > rabochie/mashinist\_burilno-kranovoy... ▾

**Обучение машиниста** бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

## Машинное обучение — Википедия

[ru.wikipedia.org](http://ru.wikipedia.org) > Машинное обучение ▾

**Машинное обучение** (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

## Обучение - машина - Большая Энциклопедия Нефти...

[ngpedia.ru](http://ngpedia.ru) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

- Какое ранжирование лучше?
- Какое хуже всех?

# Метрики классификации

$$\text{precision} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{retrieved}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{relevant}\}|}$$

# Метрики классификации

- Рассмотрим precision
- На самом деле нас интересуют документы, которые оказались вверху выдачи
- Поэтому рассматривают precision@k: метрика, вычисленная на k первых выдачах
- (!) Мы никак не учитываем порядок внутри первых k позиций

$$AP@k(q) = \sum_{i=1}^k \frac{y_{(i)}}{\sum_{j=1}^k y_{(j)}} \text{precision}@i(q),$$



## DCG (Discounted cumulative gain)

$$\text{DCG}@k(q) = \sum_{i=1}^k \frac{2^{y_i} - 1}{\log(i + 1)}$$

- Вычисляется по первым  $k$  документам из выдачи для запроса  $q$
- $y_i$  — истинный ответ для документа на  $i$ -й позиции
- Чтобы получить итоговую оценку, DCG усредняется по всем запросам

# DCG (Discounted cumulative gain)

W **Машинное обучение** — Википедия

[ru.wikipedia.org](http://ru.wikipedia.org) > Машинное обучение ▾

**Машинное обучение** (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

 **Обучение машиниста бурильно-крановых машин** — АНО...

[ccrp.ru](http://ccrp.ru) > [rabochie/mashinist\\_burilno-kranovoy...](http://rabochie/mashinist_burilno-kranovoy...) ▾


**Обучение машиниста** бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

 **Обучение - машина** - Большая Энциклопедия Нефти...

[ngpedia.ru](http://ngpedia.ru) > [id201843p1.html](http://id201843p1.html) ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

$$DCG = \frac{2^1 - 1}{\log(2)} + \frac{2^0 - 1}{\log(3)} + \frac{2^0 - 1}{\log(4)} \approx 1.44$$

 **Обучение машиниста бурильно-крановых машин** — АНО...

[ccrp.ru](http://ccrp.ru) > [rabochie/mashinist\\_burilno-kranovoy...](http://rabochie/mashinist_burilno-kranovoy...) ▾

**Обучение машиниста** бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

W **Машинное обучение** — Википедия

[ru.wikipedia.org](http://ru.wikipedia.org) > Машинное обучение ▾

**Машинное обучение** (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

 **Обучение - машина** - Большая Энциклопедия Нефти...

[ngpedia.ru](http://ngpedia.ru) > [id201843p1.html](http://id201843p1.html) ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

$$DCG = \frac{2^0 - 1}{\log(2)} + \frac{2^1 - 1}{\log(3)} + \frac{2^0 - 1}{\log(4)} \approx 0.91$$

Доля дефектных пар

$$\text{DP}@k(q) = \frac{2}{k(k-1)} \sum_{i < j}^k [y_i < y_j]$$

- Число инверсий порядка среди первых  $k$  документов

# pFound

- Вероятностная модель поведения пользователя
- При неуспехе с очередным документом выдачи пользователь разочаруется и уйдет с вероятностью  $P_{out}$
- $P_i$  — вероятность дойти до  $i$ -ого документа,  $y_i$  — вероятность того, что пользователь удовлетворится  $i$ -ым документом

$$P_1 = 1, \quad P_{i+1} = P_i(1 - y_i)(1 - P_{out})$$

$$\text{pFound}@k(q) = \sum_{i=1}^k P_i y_i$$

# pFound



<https://habr.com/ru/company/yandex/blog/197838/>

# Разнообразие поисковой выдачи

- Неоднозначные запросы
- Пример: «ягуар»
  - Животное?
  - Марка автомобиля?
  - Танк? (немецкий или китайский?)
  - Напиток?

# Разнообразие поисковой выдачи

- Неоднозначные запросы
- С точки зрения обычных метрик, весь топ выдачи нужно замостить одинаковыми релевантными документами
- Разнообразие позволяет собрать разнородную выдачу, чтобы удовлетворить в среднем всех

# Wide pFound

- Предполагается, что пользователь, делая запрос, мог иметь в виду один из интенгов  $I = \{I_1, \dots, I_m\}$
- Примеры интенгов: автомобили, картинки, новости, животные, ...
- Каждый интенг имеет некоторую вероятность  $p(I_i)$  и порождает собственное распределение релевантностей на документах

$$\text{wide pFound} = \sum_{i=1}^m p(I_i) \text{pFound}(I_i)$$



# Wide pFound

- Как вычислить вероятности интенгов?

# Wide pFound

- Как вычислить вероятности интенгов?
- Интент пользователя определяется по продолжениям введенного запроса
- Продолжения классифицируются по различным тематикам
- Тематики являются интенгами
- Вероятности определяются по частоте соответствующих продолжений запросов

# Качество ранжирования

- Также можно сформулировать задачу классификации ( $Y = \{0, 1\}$ ):

$$\text{precision} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{retrieved}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{relevant}\}|}$$

# Методы ранжирования

# Поточечный (pointwise) подход

- Обучим модель  $a(q, d)$ , чтобы она как можно точнее приближала ответы  $y_i$
- Например, линейная регрессия:

$$\sum_{(q,d,y) \in R} (\langle w, x(q, d) \rangle - y_i)^2 \rightarrow \min_w$$

- $x(q, d)$  — признаки для пары «запрос-документ»

# Поточечный (pointwise) подход

- Простой в реализации
- Можно использовать любую из известных моделей (линейные, деревья, случайные леса, нейронные сети...)
- Восстанавливает точные значения  $y_i$ , хотя нас интересует порядок

## Попарный (pairwise) подход

- В ранжировании требуется правильно располагать пары документов — формализуем это

$$\sum_{(q, d_i, d_j) \in R} [a(q, d_i) - a(q, d_j) < 0]$$

- Штрафуем, если второй документ из пары оказался раньше

## Попарный (pairwise) подход

- Получили разрывный функционал — сложно оптимизировать
- Перейдём к гладкой верхней оценке (как в линейных классификаторах):

$$\sum_{(q, d_i, d_j) \in R} [a(q, x_i) - a(q, x_j) < 0] \leq \sum_{(q, d_i, d_j) \in R} L(a(q, x_i) - a(q, x_j))$$

- Пример:  $L(z) = \log(1 + e^{-z})$



## Попарный (pairwise) подход

- Сложнее поточечного (больше слагаемых в функционале)
- Обычно даёт качество выше, чем поточечный
- Реализации:  $SVM^{light}$ , xgboost (rank:pairwise)

# Признаки в задачах ранжирования

# Типы признаков

- Запросные
  - Популярность запроса
  - Тип запроса (навигационный, товарный и т.д.)
- Статические — зависят только от документа
  - Популярность документа
  - Тематика
  - Распределение слов
- Динамические — зависят от документа и от запроса
  - Расстояния между запросом и документом

# Признаки ранжирования Google

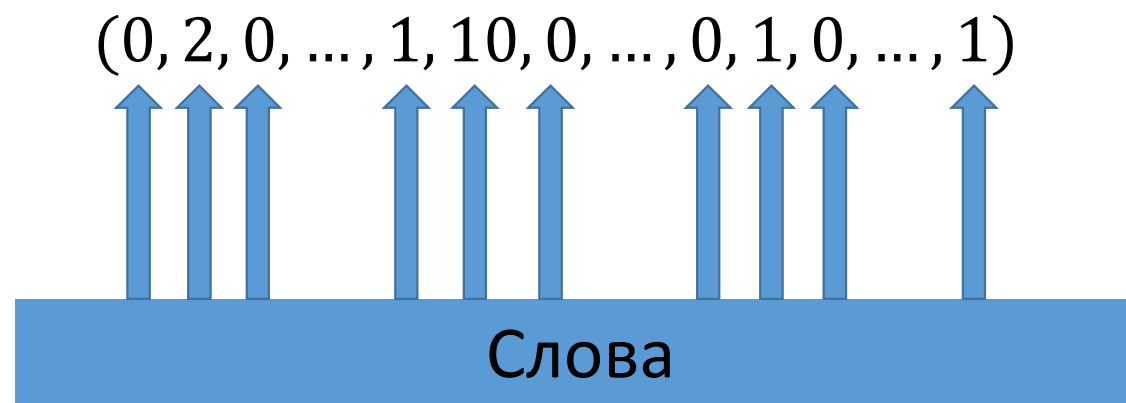
- <https://backlinko.com/google-ranking-factors>

# Мешок слов

- $v(\text{большое}) = (1, 0, 0, 0, \dots, 0)$
- $v(\text{спасибо}) = (0, 1, 0, 0, \dots, 0)$
- $v(\text{минус}) = (0, 0, 1, 0, \dots, 0)$
- $v(\text{зарубежный}) = (0, 0, 0, 1, \dots, 0)$
- ...
- $v(\text{инквизиция}) = (0, 0, 0, 0, \dots, 1)$

# Мешок слов

- Текст — это вектор  $x$ , содержащий счётчики слов



# Косинусное расстояние

- Пусть  $\vec{q}$  — вектор запроса,  $\vec{d}$  — вектор документа
- Мера сходства:

$$s(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^n q_i d_i}{\|\vec{q}\| \|\vec{d}\|}$$

- Чем больше, тем сильнее тексты похожи по долям слов

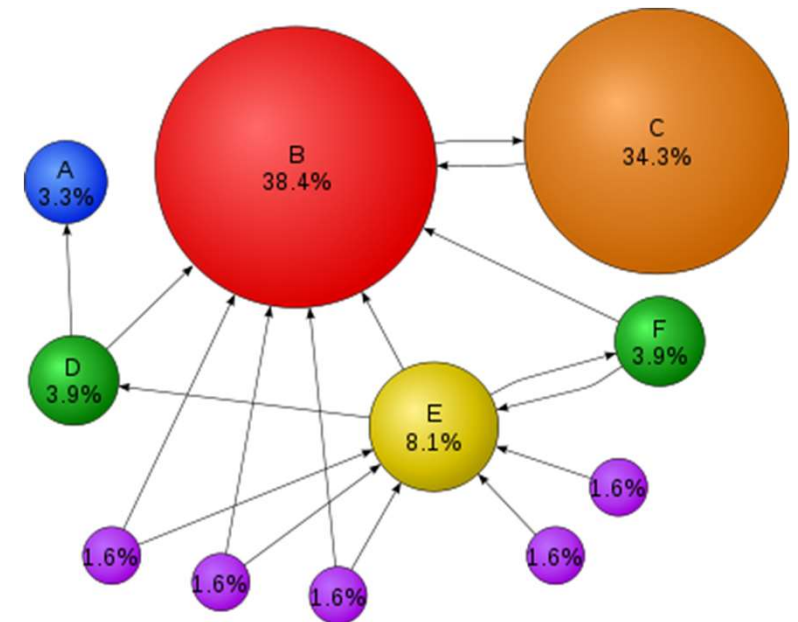
## Продвинутое расстояние: BM25

$$\text{BM25}(q, d) = \sum_{i=1}^n \text{IDF}(q_i) \frac{\text{tf}(q_i, d)(k_1 + 1)}{\text{tf}(q_i, d) + k_1 \left(1 - b + b \frac{|D|}{\bar{n}_d}\right)}$$

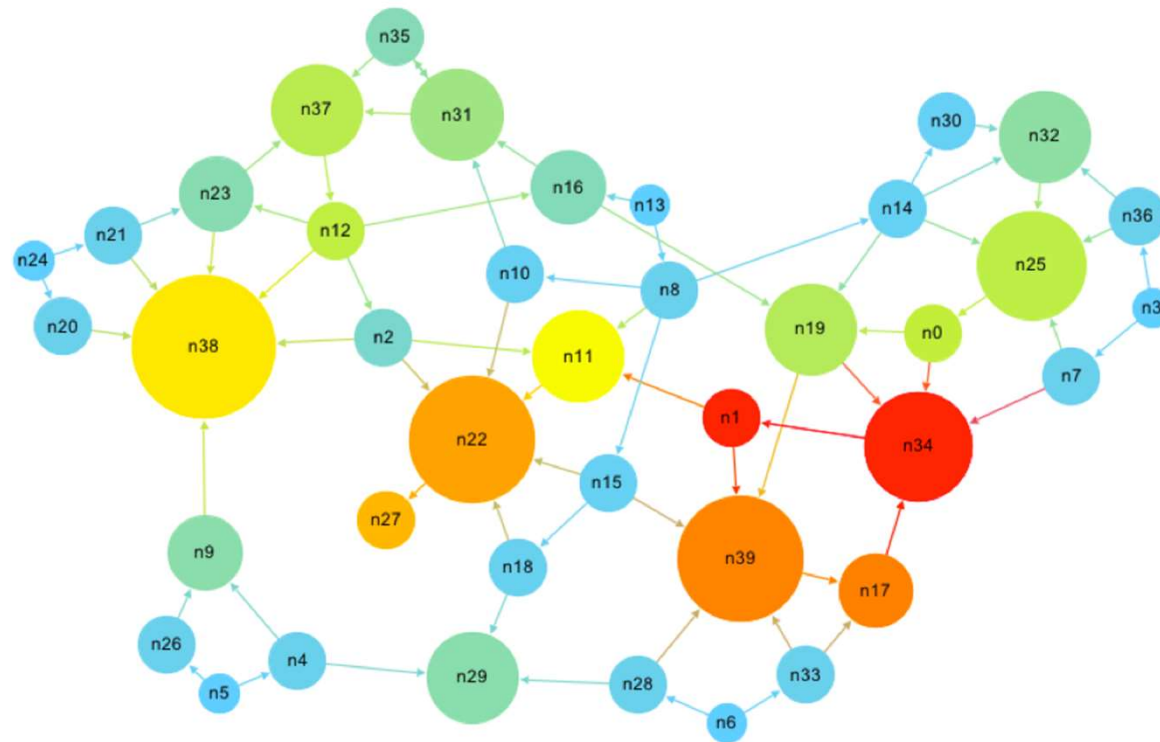


# PageRank

- Документы в сети ссылаются друг на друга
- Если документ A ссылается на документ B, то он «голосует» за B
- Чем меньше голосов отдаёт A, тем сильнее его голос
- Документ B важен, если за него отдано много сильных голосов



# PageRank



# PageRank

- Пусть пользователь бродит по сети
- Стартует из случайного документа
- С вероятностью  $(1 - \delta)$  переходит по одной из ссылок с равными вероятностями
- С вероятностью  $\delta$  переходит на случайный документ из всей сети
- PageRank — вероятность при таком случайном блуждании попасть в данный документ

# PageRank

- PageRank страницы  $u$  зависит от PageRank страниц  $v$  из множества  $B_u$  (страниц, которые ссылаются на  $u$ ), деленного на число исходящих ссылок  $L(v)$  из страницы  $v$ :

$$\text{PR}(u) = \sum_{v \in B_u} \frac{\text{PR}(v)}{L(v)}$$

# PageRank

- Учтем, что пользователь может остановиться в какой-то момент
- Установим damping factor (фактор затухания) – обычно  $d \approx 0.85$
- $N$  – число рассматриваемых страниц

$$\text{PR}(u) = \frac{1 - d}{N} + d \sum_{v \in B_u} \frac{\text{PR}(v)}{L(v)}$$

# Резюме

- Ранжирование — задача сортировки документов по релевантности
- Метрика должна учитывать позиции, а не абсолютные значения прогнозов — например, DCG
- Поточечный и попарный подходы
- Отдельная задача — разработка признаков