?????????? ML

- 00_Общая информация, pytorch
- 01_Ранжирование

00_????? ???????????, pytorch

Тензор - многомерный массив.

Направления, где может использоваться ML:

- Если есть большие списки правил
- Постоянно меняющиеся правила
- Большой массив данных

Менее подходящие сферы:

- Если требуется объяснение, почему получен данный ответ
- Если можно сделать при помощи алгоритмической системы
- Если ошибки недопустимы
- Если мало данных

01_??????????

Ранжирование - упорядочивание объектов в соответствии с некоторой мерой, т е создание частично упорядоченного множества. Может быть указана зависимость для пар объектов. Следовательно, некоторые пары могут быть не связаны соотношением, т к относятся к разным множествам зависимостей.

Матчинг (соответствие) - процесс сопоставления объектов на основе сравнения и расчета некоторой меры схожести. Подзадача ранжирования.

Learning to rank - класс задач ML с учителем (с частичным привлечением учителя) поиска модели наилучшего приближения и обобщения способа ранжирования. Пример: псевдолейблинг. Небольшое количество данных с разметкой, затем предсказания на огромном объеме данных. Предсказания становятся источником для обучения.

Мера релевантности - степень соответствия между запросом и набором документов.

SKU - идентификатор товарной позиции, идентификатор сущности (не обязательно физический товар).

TP (True Positives) — верно предсказанные положительные случаи
FP (False Positives) — ложноположительные (модель сказала “да”, но это ошибка)
TN (True Negatives) — верно предсказанные отрицательные случаи
FN (False Negatives) — ложноотрицательные (модель сказала “нет”, но это ошибка)

Качество ранжирования

Критерии репрезентативности выборки:

- соответствие структуры выборки структуре реальных данных
- Отсутствие систематического смещения (bias) Данные не должны быть перекошены в сторону одной группы.
- Достаточный объем
- Случайный или контролируемый отбор

Критерии качества ранжирования

- Качество / точность
- Эффективность (скорость предоставления ответа, объем ресурсов)
- Удобство использования

Методология оценки Кранфилда: оценка релевантности моделей на основе фиксированных репрезентативных наборов документов и запросов.

Метрики рассчитываются по топу документов, обозначается metric@k. Например recall@5 это полнота среди 5 документов.

Метрика точность

Из всех объектов, которые модель назвала положительными, какая доля действительно положительная?

Precision = TP / (TP + FP)

Precision = кол-во найденных релевантных документов среди выданных / кол-во выданных

Метрика полнота

Из всех реально положительных объектов сколько мы нашли?

Recall = TP / (TP + FN)

Recall = кол-во найденных положительных релевантных документов среди выданных / кол-во положительных релевантных документов

Fb-мера

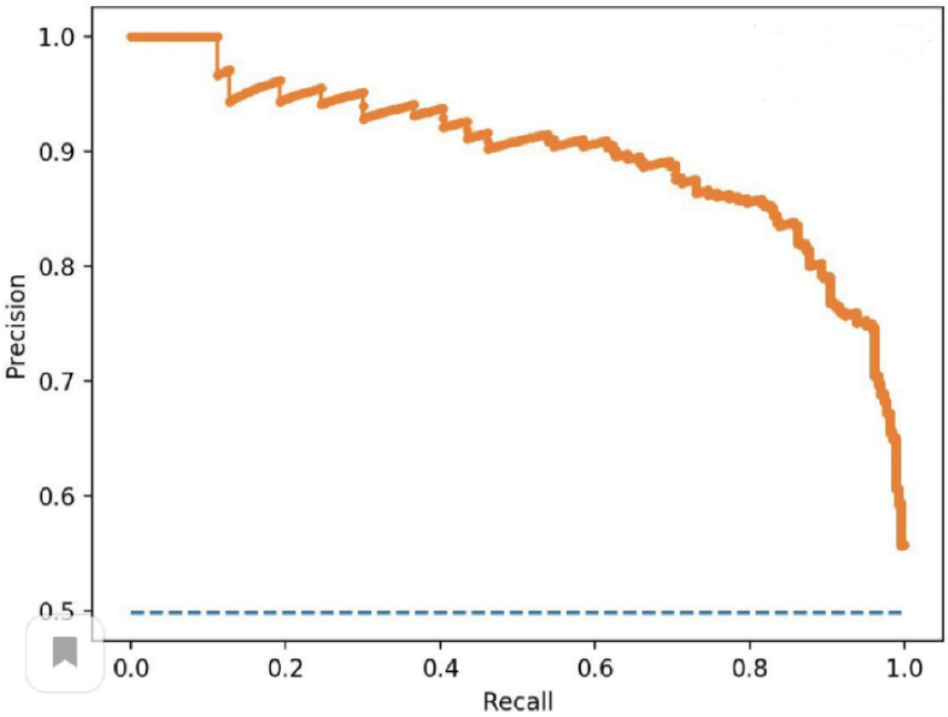
Агрегированный критерий качества, b - вес точности в метрике. Способ объединения двух метрик.

Fb = (1 + b^2) precision * recall / (b^2 * precision + recall)

PR-кривая

- сортируем предсказания по убыванию релевантности
- считаем значение точности и полноты в первой паре
- снижаем значение порога, чтобы попало две пары
- повторяем, пока не попали все элементы

Метрика - площадь под PR кривой



Average Precision

Сколько релевантных объектов сконцентрировано среди самых высоко оцененных.

$$AP = \sum_{K}(Recall@k - Recall@[k-1]) \cdot Precision@k$$

Main avarage precision

MAP = AP / Q

Качество многоуровневого ранжирования

Cumulative gain - сумма рангов

Discounted cumulative gain - сумма, каждый последующий делится на логарифм по основанию 2 от номера позиции. DCG@k

IdealDCG@k считается для случая идеальной выдачи

Normalized DCG = DCG@k/IdealDCG@k

Обучение моделей ранжирования

1. Pointwise (поточечный) функция ошибки по конкретному объекту минимизируется
$$\sum_{q,j} \ell(f(\mathbf{x}_j^q), r_j^q) \to \min$$

2. Pairwise (попарный) функция ошибки по паре объектов минимизируется RankNet
$$\sum_q \sum_{i,j:r_i^q > r_j^q} \ell(f(\mathbf{x}_i^q) - f(\mathbf{x}_j^q)) \to \min$$
3. Listwise (списочный) функция ошибки на всем списке документов ListNet
$$\ell\left(\{f(\mathbf{x}_j^q)\}_{j=1}^{m_q}, \{r_j^q\}_{j=1}^{m_q}\right) \to \min$$

Поточечные методы

BM25:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$