Коллаборативная фильтрация

Машинное Обучение, 20!7

Малютин Евгений Алексеевич

С чего всё началось?

Netflix prize

- 480189 users
- 17770 movies
- 100480507 scores
- 02.10.2006 > 21.09.2009
- 1 000 000 000$
- Task: RMSE to 10% (0.9514 $\rightarrow$ 0.8563)

Определения

- U – множество субъектов (users/пользователи/субъекты)
- R – множество объектов (items/предметы/товары)
- Y – пространство транзакций;

Сырые исходные данные:
$D = (u_i, r_i, y_i)_{i=1}^{m}$ – транзакционные данные;
Агрегированные данные:
$F = |f_{ur}|$– матрица кросс-табуляции размера $|U| \times |R|$, где
$f_{ur} = agg\{(u_i, r_i, y_i) \in D | u_i = u, r_i = r|\}$
Задачи:

- прогнозирование незаполненных ячеек f_{ur};
- оценивание сходства: $\rho(u,u), \rho(r,r), \rho(u,r)$;
- выявление скрытых интересов $p(t|u), q(t|r)$ относительно заданного либо неизвестного набора тем $t = 1, ..., T$.

Рекомендательные системы в Интернете

- U – пользователи Интернет;
- R – ресурсы (сайты, документы, новости, и т.п.);
- f_{ur} = [пользователь u посетил ресурс r];

Основная гипотеза Web Usage Mining:
Действия (посещения) пользователя характеризуют его интересы, вкусы, привычки, возможности.

Задачи персонализации:

- выдать оценку ресурса r для пользователя u;
- выдать пользователю u ранжированный список рекомендуемых ресурсов;
- сгенерировать для ресурса r список близких ресурсов.

Ещё пример

Окуда-то из эпохи ЖЖ

- U – пользователи;
- R – текстовые документы (форумы, блоги);
- K – ключи (ключевые слова или выражения);
- f_{ur} = [пользователь u участвует в r];
- g_{rk} = частота встречаемости ключа k в тексте r;
- h_{uv} = [пользователю u интересен пользователь v].

Некоторые задачи анализа социальной сети:

- рекомендовать пользователю интересные ему блоги,
- найти единомышленников (like-minded people);
- охарактеризовать интересы пользователя ключами;
- найти все блоги по данным или похожим ключам;
- найти все блоги, похожие на данный;
- построить иерархический тематический каталог блогов.

Методы

Корреляционные модели

- хранение всей исходной матрицы данных F;

Латентные модели

Методы

Корреляционные модели

- хранение всей исходной матрицы данных F;
- сходство клиентов — это корреляция строк матрицы F;

Латентные модели

Методы

Корреляционные модели

- хранение всей исходной матрицы данных F;
- сходство клиентов — это корреляция строк матрицы F;
- сходство объектов — это корреляция столбцов матрицы F.

Латентные модели

Методы

Корреляционные модели

- хранение всей исходной матрицы данных F;
- сходство клиентов — это корреляция строк матрицы F;
- сходство объектов — это корреляция столбцов матрицы F.

Латентные модели

- оценивание профилей клиентов и объектов(профиль — это вектор скрытых характеристик);

Методы

Корреляционные модели

- хранение всей исходной матрицы данных F;
- сходство клиентов — это корреляция строк матрицы F;
- сходство объектов — это корреляция столбцов матрицы F.

Латентные модели

- оценивание профилей клиентов и объектов(профиль — это вектор скрытых характеристик);
- хранение профилей вместо хранения F;

Методы

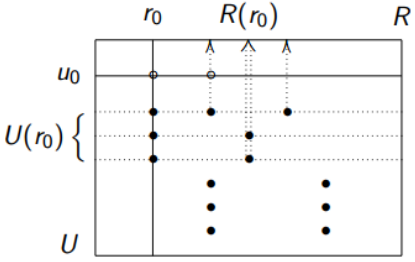
Корреляционные модели

- хранение всей исходной матрицы данных F;
- сходство клиентов — это корреляция строк матрицы F;
- сходство объектов — это корреляция столбцов матрицы F.

Латентные модели

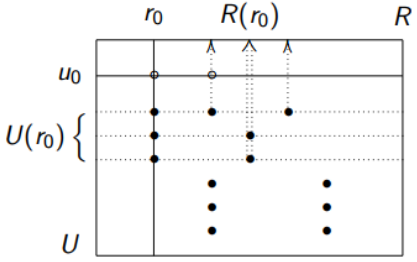
- оценивание профилей клиентов и объектов(профиль — это вектор скрытых характеристик);
- хранение профилей вместо хранения F;
- сходство клиентов и объектов — это сходство их профилей

«клиенты, купившие r_0, также покупали $R(r_0)$» [Amazon.com]



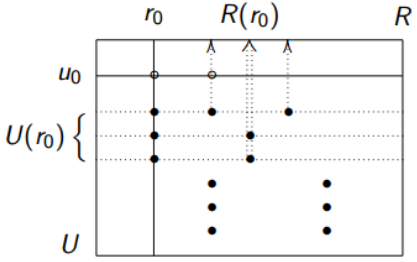
- $U(r_0) = \{u \in U | f_{ur_0} \neq \emptyset, u \neq u_0\}$ – коллаборация
- $R(r_0) = \{r \in R | B(r) = \frac{|U(r_0) \cap U(r)|}{|U(r_0) \cup U(r)|}\}$, B – любая метрика близости
- отсортировать $R(r_0)$ по убыванию $B(r)$, взять $topN$.

«клиенты, купившие r_0, также покупали $R(r_0)$» [Amazon.com]



Проблемы?

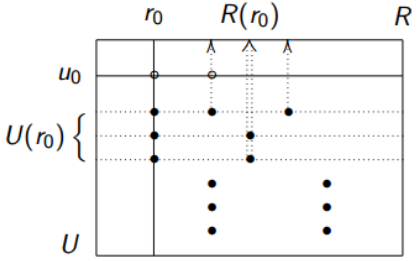
«клиенты, купившие r_0, также покупали $R(r_0)$» [Amazon.com]



Проблемы?

-
- Тривиальные рекомендации

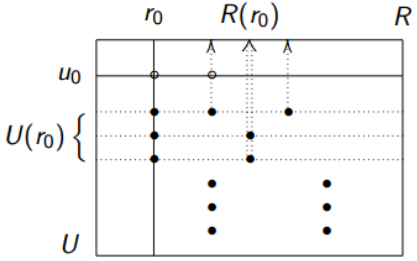
«клиенты, купившие r_0, также покупали $R(r_0)$» [Amazon.com]



Проблемы?

-
- Тривиальные рекомендации
- Не учитывают интересы конкретного пользователя

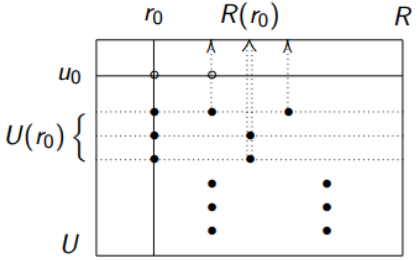
«клиенты, купившие r_0, также покупали $R(r_0)$» [Amazon.com]



Проблемы?

-
- Тривиальные рекомендации
- Не учитывают интересы конкретного пользователя
- Холодный старт

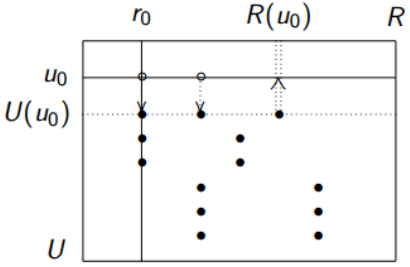
«клиенты, купившие r_0, также покупали $R(r_0)$» [Amazon.com]



Проблемы?

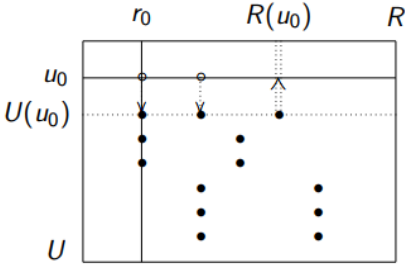
-
- Тривиальные рекомендации
- Не учитывают интересы конкретного пользователя
- Холодный старт
- надо хранить всю матрицу R

«клиенты, похожие на u_0, также покупали $R(u_0)$»



- $U(u_0) = \{u \in U | corr(u, u_0) > \alpha\}$ – коллаборация
- $R(r_0) = \{r \in R | B(r) = \frac{|U(u_0) \cap U(u)|}{|U(u_0) \cup U(u)|}\}$, B – любая метрика близости
- отсортировать $R(r_0)$ по убыванию $B(r)$, взять $topN$.

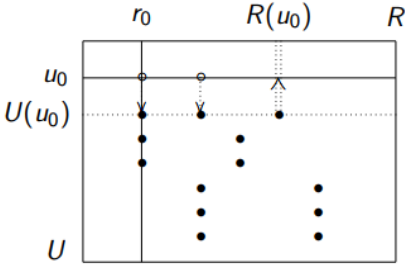
«клиенты, похожие на u_0, также покупали $R(u_0)$»



Прр-роблемы?

- матрица R

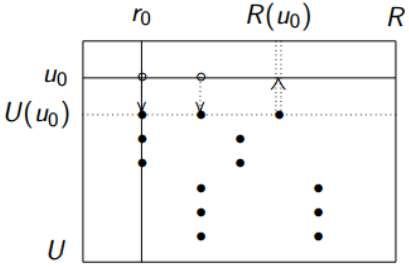
«клиенты, похожие на u_0, также покупали $R(u_0)$»



Прр-роблемы?

- матрица R
- холодный старт

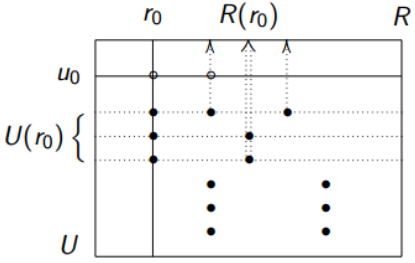
«клиенты, похожие на u_0,
также покупали $R(u_0)$»



Прр-роблемы?

- матрица R
- холодный старт
- новые, нетипичные пользователи

«клиенты, купившие r_0, также покупали $R(r_0)$» [Amazon.com]



- $f(u_0) = \{i \in I | \exists i_0 : r_{u_0,i_0} \neq \emptyset,\ sim(i, i_0) > \alpha\}$ – коллаборация
- $R(r_0) = \{r \in R | B(r) = \frac{|U(r_0) \cap U(r)|}{|U(r_0) \cup U(r)|}\}$, B – любая метрика близости
- отсортировать $R(r_0)$ по убыванию $B(r)$, взять *topN*.

Восстановление пропущенных значений

Непараметрическая регрессия Надарая-Ватсона:

$$\widehat{f}_{ur} = \overline{f_u} + \frac{\sum\limits_{u' \in U_\alpha(u)} K(u,u')(f_{u'r} - \overline{f}_{u'})}{\sum\limits_{u' \in U_\alpha(u)} K(u,u')}$$

, где $\overline{f_u}$ – средний рейтинг пользователя u

- $R(u)$ – множество объектов, которые клиент u оценил,
- $K(u,u')$ — сглаживающее ядро, функция близости u и u'
- $U_\alpha(u)$ – коллаборация, пользователи в α-окрестности пользователя u

Недостатки:

-

Восстановление пропущенных значений

Непараметрическая регрессия Надарая-Ватсона:

$$\widehat{f}_{ur} = \overline{f_u} + \frac{\sum\limits_{u' \in U_\alpha(u)} K(u,u')(f_{u'r} - \overline{f}_{u'})}{\sum\limits_{u' \in U_\alpha(u)} K(u,u')}$$

, где $\overline{f_u}$ – средний рейтинг пользователя u

- $R(u)$ – множество объектов, которые клиент u оценил,
- $K(u,u')$ — сглаживающее ядро, функция близости u и u'
- $U_\alpha(u)$ – коллаборация, пользователи в α-окрестности пользователя u

Недостатки:

-
- Холодный старт

Восстановление пропущенных значений

Непараметрическая регрессия Надарая-Ватсона:

$$\widehat{f}_{ur} = \overline{f_u} + \frac{\sum\limits_{u' \in U_\alpha(u)} K(u,u')(f_{u'r} - \overline{f}_{u'})}{\sum\limits_{u' \in U_\alpha(u)} K(u,u')}$$

, где $\overline{f_u}$ – средний рейтинг пользователя u

- $R(u)$ – множество объектов, которые клиент u оценил,
- $K(u,u')$ — сглаживающее ядро, функция близости u и u'
- $U_\alpha(u)$ – коллаборация, пользователи в α-окрестности пользователя u

Недостатки:

-
- Холодный старт
- Хранить матрицу R

Функции близости

- Корреляция Пирсона(Спирмена, Kendall):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

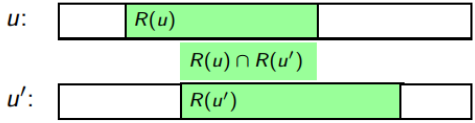
- Косинусная мера:

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$

- Статистические критерии:
  χ^2, точный тест Фишера (для бин. данных)
- Графовые меры: Jaccard, PageRank, ADAMIC-ADAR, etc.

Функции близости на основе точного теста Фишера

Рассмотрим случай бинарных данных, $f_{ur} \in \{0, 1\}$:
Нулевая гипотеза:
клиенты u и u' совершают свой выбор независимо:



Вероятность случайной реализации r совместных выборов:

$$p(r) = P\left\{|R(u) \cap R(u')| = r\right\} = \frac{C^{r}_{|R(U)|} C^{|R(u')|-r}_{|R|-|R(u)|}}{C^{|R(u')|}_{|R(u)|}}$$

Функция близости: $R(u, u') = -log\ p(|R(u) \cap R(u')|)$

Резюме:

Преимущества для бизнес-приложений:

Недостатки:

Резюме:

Преимущества для бизнес-приложений:

- Легко понять.

Недостатки:

Резюме:

Преимущества для бизнес-приложений:

- Легко понять.
- Легко реализовать.

Недостатки:

Резюме:

Преимущества для бизнес-приложений:

- Легко понять.
- Легко реализовать.

Недостатки:

- Не хватает теоретического обоснования:

Резюме:

Преимущества для бизнес-приложений:

- Легко понять.
- Легко реализовать.

Недостатки:

- Не хватает теоретического обоснования:
- придумано много способов оценить сходство...

Резюме:

Преимущества для бизнес-приложений:

- Легко понять.
- Легко реализовать.

Недостатки:

- Не хватает теоретического обоснования:
- придумано много способов оценить сходство...
- придумано много гибридных (item-user-based) методов... ...и не ясно, что лучше;

Резюме:

Преимущества для бизнес-приложений:

- Легко понять.
- Легко реализовать.

Недостатки:

- Не хватает теоретического обоснования:
- придумано много способов оценить сходство...
- придумано много гибридных (item-user-based) методов... ...и не ясно, что лучше;
- Все методы требуют хранения огромной матрицы F.

Резюме:

Преимущества для бизнес-приложений:

- Легко понять.
- Легко реализовать.

Недостатки:

- Не хватает теоретического обоснования:
- придумано много способов оценить сходство...
- придумано много гибридных (item-user-based) методов... ...и не ясно, что лучше;
- Все методы требуют хранения огромной матрицы F.
- Проблема «холодного старта»

Латентные модели

Латентная модель:

по данным D оцениваются векторы:

p_{su}, $s \in S$ – профили клиентов $u \in U$; q_{tr}, $t \in T$ – профили объектов $r \in R$.

Типы латентных моделей (основные идеи):

- Ко-кластеризация:
  - жёсткая:

    p_{su} = [клиент u принадлежит кластеру s];

    q_{tr} = [объект r принадлежит кластеру t];
  - мягкая: p_{su}, q_{tr} – степени принадлежности кластерам.
- Матричная факторизация: $S = T$; по p_{tu}, q_{tr} должны восстанавливаться f_{ur}.
- Вероятностные (байесовские) модели: $S = T$; $p_{tu} = p(t|u), qtr = q(t|r)$.

Бикластеризация

Общие положения:

Пусть f_{ur} – вещественные числа или рейтинги;

$g : U \to G$ — функции кластеризации клиентов ($|G| < \inf$);

$h : R \to H$ – функции кластеризации объектов ($|H| < \inf$);

Модель усреднения по блокам (Block Average):

$\widehat{f}_{ur}(g,h) = \overline{f}_{g(u),h(r)} + (\overline{f}_u - \overline{f}_{g(u)}) + (\overline{f}_r - \overline{f}_{h(r)})$

Функционал качества:

$$\sum_{(u,r)\in D} (\widehat{f}_{ur}(g,h) - f_{ur})^2 \to min$$

Матричные разложения

T – множество тем (интересов): $|T| \ll |U|, |T| \ll |R|$
p_{tu} – неизвестный профиль клиента u; $P = (p_{tu})_{|T| \times |U|}$
q_{tr} – неизвестный профиль объекта r; $P = (q_{tr})_{|T| \times |R|}$
Задача: найти разложение $f_{ur} = \sum_{t \in T} \pi_t p_{tu} q_{tr}$
Матричная запись: $F = P^T \Delta Q$, где $\Delta = diag(\pi_1, ..., \pi_{|T|})$
Вероятностный смысл: $p(u, r) = \sum_{t \in T} p(t) p(u|t) p(r|t)$

Методы решения:

- SVD — сингулярное разложение (плохо интерпретируется (?));
- NNMF — неотрицательное матричное разложение: $p_{tu} > 0, q_{tr} > 0$;
- PLSA — вероятностный латентный семантический анализ

Разреженный SVD (Singular Value Decomposition)

Обычный не-разреженный SVD: $||F - P^T Q||^2 \to \min\limits_{P,Q}$

Разреженный SVD: $\sum\limits_{(u,r)\in D} (f_{ur} - \sum\limits_{t\in T} p_{tu} q_{tr})^2 \to \min\limits_{P,Q}$

Метод стохастического градиента:

перебираем все $(u, r) \in D$ и делаем градиентный шаг: ϵ_{ur}:

$$p_{tu} = p_{tu} + \eta \epsilon_{ur} q_{tr}$$
$$q_{tr} = q_{tr} + \eta \epsilon_{ur} p_{tr}$$

Разреженный SVD

Плюсы

- легко вводится регуляризация:
  $\epsilon_{ur}^2 + \lambda||p_u||^2 + \mu||q||^2 \to min$
- легко вводятся ограничения неотрицательности:
  $p_{tu} \geq 0, q_{tu} \geq 0$ (метод проекции градиента)
- легко вводятся обобщение для ранговых данных:
  $\sum_{(u,r)\in D} (\beta(f_u r) - \sum_{t\in T} p_{tu} q_{tr}) \to \min_{P,Q,\{\beta_t\}}$
- легко реализуются все виды инкрементности: добавление
  - ещё одного клиента u,
  - ещё одного объекта r,
  - ещё одного значения fur
- высокая численная эффективность на больших данных; (?)

NNMF (Non-negative matrix factorization)

Метод чередующихся наименьших квадратов (Alternating Least Squares):

$$D = ||R - \sum_{t \in T} p_t q_t^T||^2 = ||R_t - p_t q_t^t||^2 \to \min_{p_t \geq 0, q_t \geq 0}$$

Идея:

поочерёдно перебирать то строки, то столбцы, считая все остальные фиксированными: $R_t = R - \sum_{s \in T/s} p_s q_s^T$:

$$\frac{\partial D}{p_t} = 0 \Rightarrow (p_t^T q_t - R_t) q_t^T \Rightarrow p_t = (\frac{q_t R_t^T}{q_t q_t^T})_+$$

$$\frac{\partial D}{q_t} = 0 \Rightarrow p_t (p_t^T q_t - R_t) \Rightarrow p_t = (\frac{o_t R_t^T}{p_t p_t^T})_+$$

PLSA – probabilistic lantent semantic analysis

Пусть T – множество тем (интересов); Вероятностная модель посещений:
$p(u,r) = \sum_{t \in T} p(t)p(u|t)p(r|t)$
Задача максимизации правдоподобия по $p(t), p(u|t), q(r|t)$:

$$L(\Delta, P, Q) = \sum_{u,r} f_{ur} ln p(u,r) \to max$$

при ограничениях нормировок:

$$\sum p(t) = 1, \sum p(u|t) = 1, \sum p(r|t) = 1$$

Тематические профили вычисляются по формуле Байеса:

$$p(t|u) = \frac{p(u|t)p(t)}{\sum_{s \in T} p(u|s)p(s)} \qquad q(t|r) = \frac{p(r|t)p(t)}{\sum_{s \in T} p(r|s)p(s)}$$

Сформировать начальные приближения $p(t), p(u|t), q(r|t)$;
Повторять итерации до сходимости:

- **E-шаг:** скрытые переменные H по формуле Байеса:

$$H(t|u,r) = \frac{p(t)p(u|t)q(r|t)}{p(u,r)}$$

- **M-шаг:** аналитическое решение задачи $L(\Delta, P, Q) \to max$:

$$p(t) = \frac{S(t)}{S}, \quad S(t) = \sum_{u,r} f_{ur} H(t|u,r), \quad S = \sum f_{ur}$$

$$p(u|t) = \frac{1}{S(t)} \sum_{t} f_{ur} H(T|u,r)$$

$$q(r|t) = \frac{1}{S(t)} \sum_{r} f_{ur} H(t|u,r)$$

Ещё

- Если f_{ur} – рейтинги, то вместо $p(u, r) = P(f_{ur} \neq \emptyset)$ надо оценивать $(z_{max} - 1)$ вероятностей $p_z(u, r) = P(f_u r, \leq z), z \in Z$
- Иерархические профили: темы разбиваются на подтемы;
- Инкрементные алгоритмы: обработка потока данных D;
- Учёт априорной информации через
  - начальное приближение профилей:
  - тематический каталог объектов;
- соц-дем (анкеты) клиентов;
- унифицированный профиль объектов и клиентов;
- долгосрочный и краткосрочный профили;
- оценивание сходства по частям профиля.

Резюме

- Коллаборативная фильтрация (Collaborative Filtering) – это набор методов для построения рекомендательных систем (Recommender Systems).
- Тематическое моделирование (Topic Modeling) – это набор методов для выявления латентных интересов клиентов или для выявления латентных тем в корпусе текстов.
- Латентные модели обладают рядом преимуществ:
  - тематические профили содержательно интерпретируемы,
  - могут оцениваться по внешним данным,
  - что позволяет решать проблему «холодного старта»
  - и строить тематическую кластеризацию (таксономию);
  - оценки сходства клиентов и объектов более адекватны;
  - резко сокращается объём хранимых данных.

Порассуждать

Factorization machines

$$h(x) = w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j'=j+1}^{p} x_j x_{j'} v_j^T v_j'$$

- $x \in R^p$
- h(x) – предскахание
- w_0 – смещение
- w_0, w, V - параметры
- модель "квадратичной"регрессии

Как граф?

Музыкальная рекомендация

- Группы, песни, жанры, инструменты
- А может построить граф!
- Personalized page rank
- А как же big-data?
- Оказывается можно считать не через собственные числа - а парой блужданий по графу
- И explonation завезти.