推薦システムのアルゴリズム

神嶌　敏弘

研究ゼミ第2回（2022/5/9）

担当者：柳　智也

目次

推薦システムとは何か

Konstan[1]による定義:

Recommenders: Tools to help identify worthwhile stuff

・利用者にとって有用と思われる対象、情報などを選び、利用者の目的に合わせた形で推薦するシステム

なぜ推薦システムが必要になったのか？

1. 大量の情報が発信されるようになった
2. 大量の情報の蓄積・流通が可能になったことで、誰もが大量の情報を獲得できるようになった

→情報が多すぎて欲しい情報を特定できない「情報過多」が起きるようになった

推薦システムの種類

個人化の度合いによって、以下の3段階に分けられる。

1. 非個人化（no personalization）
   全ての利用者に対して、全く同じ推薦をする

2. 一時的個人化（ephemeral personalization）
   システムを利用する中で同じ入力や振る舞いをした利用者には同じ推薦をする

3. 永続的個人化（persistent personalization）
   同じ入力や行動をしている利用者でも、個人情報や過去の利用履歴に応じて異なる推薦をする

推薦システムの種類：具体例

1. 非個人化（売上ランキング）

2. 一時的個人化（一緒に購入されている商品）

3. 永続的個人化（あなたへのおすすめ）

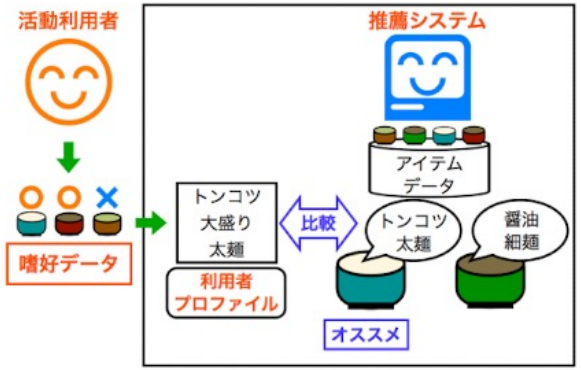
嗜好の予測：内容ベースフィルタリング

推薦システムを作るには、その人が好みそうなアイテムを探す「嗜好の予測」が必要

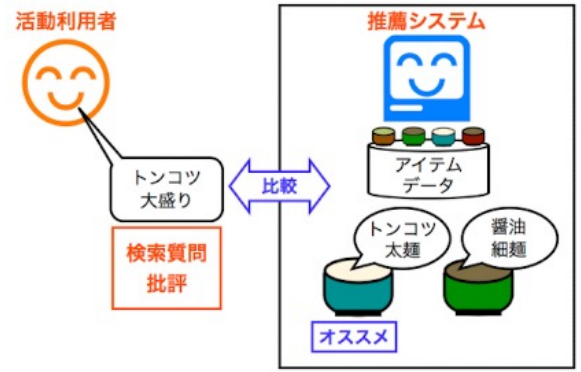
嗜好の予測の実現方法は、大きく2つに分類される

1. 内容ベースフィルタリング

・アイテムの性質と利用者の嗜好パターンを比較して、利用者が好むであろうものを推薦
・アイテムの性質は特徴ベクトルによって表される
例：（スープ＝トンコツ, 麺の太さ＝太麺, 量＝大盛り）
・嗜好データを過去の履歴から貯めておいて利用する「間接指定型」と、質問に答えてもらって集める「直接指定型」がある



(a) 内容ベースフィルタリング（間接指定型）



(b) 内容ベースフィルタリング（直接指定型）

嗜好の予測：協調フィルタリング

推薦システムを作るには、その人が好みそうなアイテムを探す「嗜好の予測」が必要

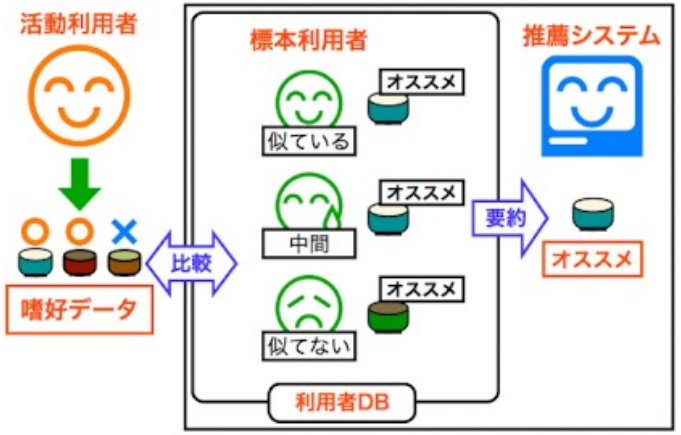
嗜好の予測の実現方法は、大きく2つに分類される

2. 協調フィルタリング

・アイテムの性質は考慮せず、多くの利用者のあらゆるアイテムに対する嗜好データを蓄積しておく

・今までの嗜好パターンが似ている利用者は、これからも同じアイテムを好み、同じアイテムを嫌うだろうという仮説を立てる

・嗜好パターンが似ている利用者を探し、その人が好むものを推薦する



(c) 協調フィルタリング

内容ベースと協調フィルタリングの比較

内容ベースフィルタリングと協調フィルタリングにはそれぞれ短所と長所がある

1. ドメイン知識
内容ベースフィルタリングは各アイテムの特徴ベクトルを持つ必要がある一方で、協調フィルタリングは利用者同士の類似度がわかればいいのでドメイン知識を必要としない

2. 多様性
内容ベースフィルタリングでは、アイテムの類似性をもとに推薦を行うため、多様性が低くなりやすく、推薦したアイテムを受け入れることでより嗜好の偏りが強化される場合がある

3. スタートアップ問題
新たな利用者や新たなアイテムが追加された時、協調フィルタリングでは嗜好パターン・アイテム評価がないため推薦を行うことが難しい

協調フィルタリングによる嗜好の予測

協調フィルタリングによる推薦候補の予測手法は2つに分かれる

1. メモリベース法（memory-based method）
   ・推薦システムが利用される以前には何もせず、利用者データベースを持っている
   ・推薦の際にデータベース内の嗜好データと活動利用者の嗜好データを併せて予測
   ・システム利用時に1から計算するため、推薦時間が長い
2. モデルベース法（model-based method）
   ・推薦システムが利用される以前にモデルを構築する
   （例：佐藤さんが好むものは鈴木さんも好むことが多い）
   ・推薦時間は短いが、利用者やアイテムの削除があると再度モデルを作る必要がある
   （メモリベースは事前に計算をしないため、データベースの変更に影響されない）

メモリベース法：利用者間型

まず、活動利用者と嗜好パターンが似ている利用者を見つけ、彼らが好むものを推薦する「利用者間型メモリベース法」について、代表的なGroupLensの方法[2]を説明する

記号の定義

・n 人の利用者の集合：$\mathrm{X} = \{1, \ldots, n\}$
・m 種類の全アイテムの集合：$Y = \{1, \ldots, m\}$
・利用者 $x \in X$ の、アイテム $y \in Y$ への評価値 r_{xy} を要素とする評価値行列：R
・評価値行列における欠損：$\perp$
・活動利用者（推薦を行う対象）：添字 a
・利用者 x が評価済みのアイテムの集合：$Y_x = \{y | y \in Y, r_{xy} \neq \perp\}$
・利用者 a と x が共通に評価したアイテムの集合：$Y_{ax} (= Y_a \cap Y_x)$

利用者間型メモリベース法：類似度の計算

活動利用者 a と標本利用者 x の類似度は、共通に評価しているアイテムについてのPearson相関係数で測る

$$\rho_{ax} = \frac{\sum_{y \in Y_{ax}} (r_{ay} - \bar{r}_a{}')(r_{xy} - \bar{r}_x{}')}{\sqrt{\sum_{y \in Y_{ax}} (r_{ay} - \bar{r}_a{}')^2} \sqrt{\sum_{y \in Y_{ax}} (r_{xy} - \bar{r}_x{}')^2}} \quad (\text{ただし、} \bar{r}_x{}' = \frac{\sum_{y \in Y_{ax}} r_{xy}}{|Y_{ax}|})$$

共通で評価したアイテムでの、 a, x それぞれの評価の標準偏差

なお、$|Y_{ax}| \leq 1$ 、すなわち共通で評価したアイテムが1つ以下である時、上の式は計算できないので $\rho_{ax} = 0$ とする

利用者間型メモリベース法：アイテムの評価値予測

活動利用者 a のアイテム $y \notin Y_a$ の評価値 $\hat{r}_{ay}$ は、類似度で重みづけを行った、各標本利用者のアイテム y への評価値の加重平均で予測する

$$\hat{r}_{ay} = \bar{r}_a + \frac{\sum_{x \in X_y} \rho_{ax} \left(r_{xy} - \bar{r}_x{}' \right)}{\sum_{x \in X_y} |\rho_{ax}|}$$

・X_y ：アイテム y を評価済みの利用者の集合
・$\bar{r}_x$ ：利用者 x の全評価アイテムに対する平均評価値

$$\bar{r}_x = \frac{\sum_{y \in Y_x} r_{xy}}{|Y_x|}$$

利用者間型メモリベース法：例題

・あるどんぶり専門店の評価値行列が表1で与えられている

・1は嫌い、2は普通、3は好きと数字が大きくなるにつれ評価が高いと考える

・この時、活動利用者を2.田中さんとして、親子丼への評価推定値を予測してみよう

	1.親子丼	2.牛丼	3.海鮮丼	4.カツ丼
1.山田	1	3	⊥	3
2.田中	⊥	1	3	⊥
3.佐藤	2	1	3	1
4.鈴木	1	3	2	⊥

表1：評価値行列 R の例

利用者間型メモリベース法：例題

1.山田、3.佐藤、4,鈴木の3人とも親子丼を評価済みなので $X_1 = \{1, 3, 4\}$ の各利用者との相関係数を求める

・1.山田との相関係数
2.田中と1.山田が共通に評価しているアイテムは1つだけなので、$\rho_{2,1} = 0$

・3.佐藤との相関係数
共通に評価しているのは2.牛丼と3.海鮮丼なので $Y_{2,3} = \{2,3\}$ となる。これらのアイテムについて、人物ごと平均評価値を算出すると、

$$\bar{r}_2{}' = \frac{(\sum_{y=2,3} r_{2,y})}{2} = \frac{1+3}{2} = 2, \qquad \bar{r}_3{}' = \frac{(\sum_{y=2,3} r_{3,y})}{2} = \frac{1+3}{2} = 2$$

利用者間型メモリベース法：例題

2.田中と3.佐藤の相関係数は、

$$\rho_{2,3} = \frac{\sum_{y=2,3}(r_{2,y} - \bar{r}_2')(r_{3,y} - \bar{r}_3')}{\sqrt{\sum_{y=2,3}(r_{2,y} - \bar{r}_2')^2}\sqrt{\sum_{y=2,3}(r_{3,y} - \bar{r}_3')^2}}$$

$$= \frac{(1-2)(1-2) + (3-2)(3-2)}{\sqrt{(1-2)^2 + (3-2)^2}\sqrt{(1-2)^2 + (3-2)^2}}$$

$$= 1$$

2.田中と3.佐藤は非常に嗜好が似ている！

同様に、2.田中と4.鈴木の相関係数を求める

・共通評価アイテム： $Y_{2,4} = \{2, 3\}$

・$\bar{r}_4' = (\sum_{y=2,3} r_{4,y})/2 = (3+2)/2 = 2.5$

$$\rho_{2,4} = \frac{\sum_{y=2,3}(r_{2,y} - \bar{r}_2')(r_{4,y} - \bar{r}_4')}{\sqrt{\sum_{y=2,3}(r_{2,y} - \bar{r}_2')^2}\sqrt{\sum_{y=2,3}(r_{4,y} - \bar{r}_4')^2}}$$

$$= \frac{(1-2)(3-2.5) + (3-2)(2-2.5)}{\sqrt{(1-2)^2 + (3-2)^2}\sqrt{(3-2.5)^2 + (2-2.5)^2}}$$

$$= -1$$

利用者間型メモリベース法：例題

ここから、2.田中の2.親子丼の評価推定値を計算する

まず、2.田中の全評価済みアイテム上の平均評価値を求める

$$\bar{r}_2 = \frac{\sum_{y=2,3} r_{2,y}}{2} = \frac{1+3}{2} = 2$$

最後に、評価推定値を計算すると

$$\begin{aligned}\hat{r}_{2,1} &= \bar{r}_2 + \frac{\sum_{x=1,3,4} \rho_{2,x}\left(r_{x,1} - \bar{r}'_x\right)}{\sum_{x=1,3,4}\left|\rho_{2,x}\right|} \\ &= 2 + \frac{0(1-3) + 1(2-2) + (-1)(1-2.5)}{|0| + |1| + |-1|} \\ &= \underline{2.75}\end{aligned}$$

田中は親子丼が好きであると予測される！

アイテム間型メモリベース法

・利用者間型では評価値行列 R の行ベクトル（利用者方向）でのベクトルの類似度を計算していたが、アイテム間型では列ベクトル（アイテム方向）の類似度を計算する

→同じような評価を受けるアイテムは似ていて、ある利用者が関心を持っているアイテムの類似アイテムにも、その利用者は関心を持つだろうという仮定

・アイテム y と j の類似度 ρ'_{yj} は、2つの列ベクトルのコサイン類似度(Appendix)や相関係数などで計算される

・類似度を計算した後、活動利用者 a のアイテム y への推定評価値 $\hat{r}_{ay}$ を以下式で求める

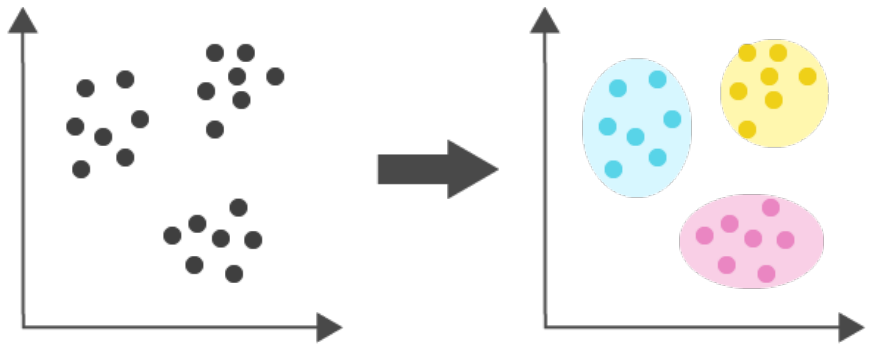
$$\hat{r}_{ay} = \frac{\sum_{j \in Y_a} \rho'_{yj} \, r_{aj}}{\sum_{j \in Y_a} |\rho'_{yj}|}$$

モデルベース法：クラスタモデル

モデルベース法では、活動利用者に推薦をする前にモデルを構築するが、最も直感的なものがクラスタリングモデルである[3]

・クラスタとは対照の集合を分割した部分集合
・同じクラスタ内の対象は違いに似ており、違うクラスタは似ていないという条件を満たす

クラスタリングのイメージ



実際に推薦システムに適用する時は以下の手順で行う

①活動利用者の嗜好と最も似ているクラスタを見つける
②クラスタ内の標本利用者の平均表価値が高いアイテムから順に活動利用者に推薦する

モデルベース法：回帰問題

・類似度や評価値そのものを予測するモデルの獲得として最も簡単な、線形関数による回帰問題への帰着を考える

・アイテム間型メモリベース法の推定評価値の式は、詳細を無視すれば以下の線形モデルとみなせる

$$\hat{r}_{ay} = \sum_{j} w_{yj} r_{aj}$$

・今まで、パラメータ w_{yj} はアイテム y と j の嗜好パターンの類似性を相関係数などで決めた

与えられた評価値の集合を訓練事例として、機械学習の手法を適用すればもっと予測精度の高い関数を獲得できるのではないか？

モデルベース法：回帰問題の設定

・まず、回帰モデルを評価地行列 R の行列分解として考えてみる

・与えられた行列 R では未評価アイテムの部分が欠損しているが、欠損していない完全な評価値行列 R^* を考え、以下のように分解する

$$R^* \approx U^T V$$

・U と V はそれぞれ $K \times n, K \times m$ 行列で、 U の第 x 列ベクトル $\boldsymbol{u}_x$ は利用者 x の特徴、V の第 y 列ベクトル $\boldsymbol{v}_y$ はアイテム y の特徴を表す

利用者 x のアイテム y への評価値 r^*_{xy} は $r^*_{xy} \approx \boldsymbol{u}_x^T \boldsymbol{v}_y$ のようなモデルで表すことができ、 $\boldsymbol{u}_x$ の要素を説明変数、 $\boldsymbol{v}_y$ の要素をパラメータとみなせばp19の線形モデルと同等のモデルである

モデルベース法：行列分解の方法

・ここで、もし K が $\max\{m,n\}$ なら、行列の分解は近似ではなく厳密に $R^* = U^T V$ となるように分解することができるが、これは観測データを書き写しただけにすぎない

・また、実際に観測できる評価値行列は欠損のある R

⇩

嗜好パターンを要約するため、 $K \ll m,n$ に固定し、R の非欠損部分の損失を最小化するように行列分解を行い、モデルを獲得する

・損失関数には残差 $(R - U^T V)$ などが用いられ、勾配降下法などを用いて U,V の各成分の更新を行う

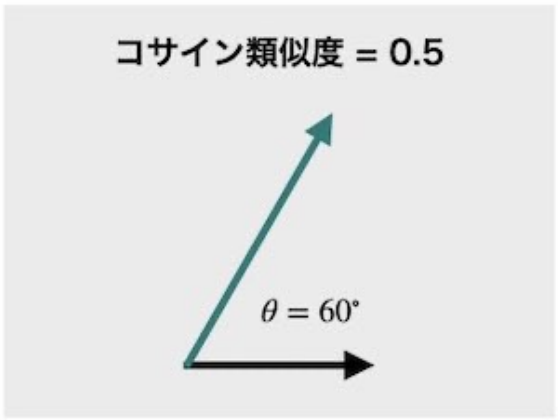
・残差 $(R - U^T V)$ の各成分が正規分布に従うとしてモデル化し、観測された評価値の生成確率が高くなるように U,V を計算する手法[4]もある

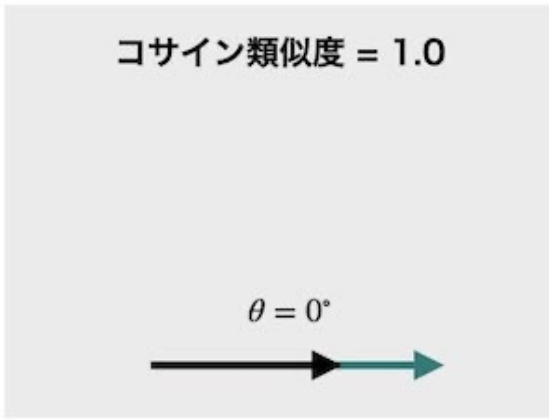
Appendix：コサイン類似度

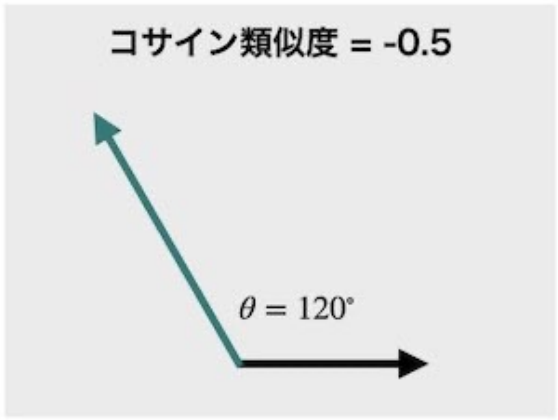
・ベクトル $\boldsymbol{a}, \boldsymbol{b}$ のコサイン類似度は以下の式で計算できる

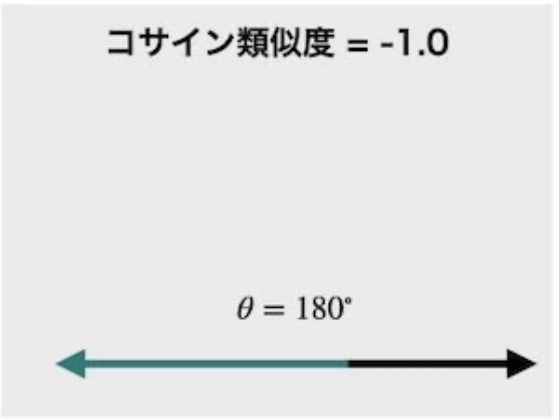
$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{||\boldsymbol{a}|| \, ||\boldsymbol{b}||}$$

・2本のベクトルが似ている時、コサイン類似度は1に近い値を取り、逆に似ていないときは-1に近い値を取る









参考文献

[1] : J.A.KonstanandJ.Riedl.Recommendersystems:Collaboratingincommerce and communities. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems,Tutorial, 2003.

[2] : P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of Netnews. In Proc. of the Conf. on Computer Supported Cooperative Work, pp. 175–186, 1994

[3] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Uncertainty in Artificial Intelligence 14, pp. 43–52, 1998.

[4] J.Canny.Collaborativefilteringwithprivacyviafactoranalysis.InProc.ofthe 25th Annual ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 238–245, 2002.