

乱択化フーリエ潜在変数モデルによるユーザー選好の埋め込みと分析

竹原 一彰^{†,††} 持橋 大地^{†††}

† 総合研究大学院大学 〒240-0193 神奈川県三浦郡葉山町 湘南国際村
†† mercari R4D 〒106-6118 東京都港区六本木 6-10-1 六本木ヒルズ森タワー
††† 統計数理研究所 〒190-8562 東京都立川市緑町 10-3
E-mail: †{kazuaki,daichi}@ism.ac.jp

あらまし アイテムに対するユーザーの選好（好き・嫌いの度合い）を可視化・分析するためのモデルを提案する。提案モデルは、高次元なアイテム特徴量と、多数のユーザーによる評価ラベルを入力として、視覚可能な低次元潜在空間における、アイテムの座標と、ユーザーの選好関数を表す重みベクトル（ユーザー埋め込み）を出力する。自由度が制限された低次元潜在空間に、様々な情報を埋め込むという課題に、乱択化フーリエ特徴量を用いた解決を試みる。公開されている実データセットを用いた評価実験にて、良好な結果を確認した。

キーワード ガウス過程, 乱択化フーリエ特徴量 (Random Fourier features), 潜在変数モデル

Random Fourier feature latent variable model for user preference embedding and analysis

Kazuaki TAKEHARA^{†,††} and Daichi MOCHIHASHI^{†††}

† The Graduate University for Advanced Studies Shonan Village, Hayama, Kanagawa 240-0193, Japan
†† mercari R4D Roppongi Hills Mori Tower, 6-10-1 Roppongi, Minato-ku, Tokyo 106-6118, Japan
††† The Institute of Statistical Mathematics 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
E-mail: †{kazuaki,daichi}@ism.ac.jp

Abstract We propose a model to visualize and analyze user preferences (degrees of liking or disliking) for items. The proposed model takes high-dimensional item features and rating labels from a large number of users as input and outputs item coordinates in a visually interpretable low-dimensional latent space and weight vectors (user embeddings) representing user preference functions. To address the challenge of embedding various information into a constrained low-dimensional latent space, we adopt a solution using Random Fourier features. In evaluation experiments using publicly available real-world datasets, we confirmed favorable results.

Key words Gaussian processes, Random Fourier features, Latent variable models

1. はじめに

ユーザーの選好（好き・嫌いの度合い）の可視化・分析は、Webサイトのユーザー、運営者、双方にとって重要である。ユーザーは、自身の意外な選好に気付いたり、視覚化された選好分布を指針に、探索を進めることができる。運営サイドとしては、どのような選好を持つユーザーが、どのくらい存在するのか分析することで、サイト全体の改善につなげることができる。

アイテムは、一般に高次元の特徴量を有しているため、PCA [1], t-SNE [2] などの次元削減手法を用いて可視化される。これらは教師なし（非識別的）モデルであり、アイテムの特徴量と、ユーザー評価ラベルなど、種別の異なる情報を同時に、潜在空間へ

反映することができない。

本研究では、高次元な特徴量を持つアイテムと、ユーザーの評価ラベルを入力として、視覚可能な低次元潜在空間における、アイテムの座標と、ユーザーの選好関数を表す重みベクトル（ユーザー埋め込み）を、出力するモデルを提案する。視覚化のために自由度（ ≤ 3 次元）が制限された潜在空間に、種別が異なる様々な情報を同時に反映させるという課題に対して、乱択化フーリエ特徴量を用いた潜在変数モデルを提案する。

提案モデルでは、複雑な非線形関数であるユーザーの選好が、ユーザー埋め込みベクトルとして表現される。このユーザー埋め込みベクトルは、内積計算などの演算が適用でき、可視化・分析に有用である。

2. 関連研究

2.1 ガウス過程 (Gaussian Processes, \mathcal{GP})

入力 $X = \{x_n \in \mathbb{R}^D\}_{n=1}^N \in \mathbb{R}^{N \times D}$ に対する, 出力 $y = \{y_n\}_{n=1}^N \in \mathbb{R}^N$ が, 平均 $\mu \in \mathbb{R}^N$, 共分散行列 $K_X = (k(x_i, x_j; \theta))_{i,j} \in \mathbb{R}^{N \times N}, i, j \in \{1, \dots, N\}$ のガウス分布に従うとき, y はガウス過程に従うといい, $y \sim \mathcal{GP}(\mu, k)$ と書く [3], [4]. ここで, $k(x_i, x_j; \theta) \in \mathbb{R}$ は, 2つの入力 $x_i, x_j \in \mathbb{R}^D$ を引数とした, パラメータ θ のカーネル関数である. カーネル関数として, 例えば, RBF (Radial Basis Function) カーネル

$$k_{RBF}(x_i, x_j; \tau, \sigma, \epsilon) = \tau \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma}\right) + \epsilon \delta(i, j)$$

などが用いられる. $\delta(i, j)$ は $i = j$ ならば 1, それ以外は 0 を返す関数である.

2.2 乱択化フーリエ特徴量 (Random Fourier features, RFF)

ガウス過程には, データ数 N に対して, 計算量が $O(N^3)$ であるという課題がある. 本研究では, 乱択化フーリエ特徴量 (Random Fourier features, RFF) [5] により, 計算量を低減する. $x_n \in \mathbb{R}^D$ に対応する RFF $\phi_n \in \mathbb{R}^M$ は, 次のように表される.

$$\phi_n = \varphi(x_n) = \sqrt{\frac{2}{M}} \begin{pmatrix} \cos(\omega_1^\top x_n + b_1) \\ \cos(\omega_2^\top x_n + b_2) \\ \vdots \\ \cos(\omega_M^\top x_n + b_M) \end{pmatrix} \in \mathbb{R}^M$$

$\varphi: \mathbb{R}^D \rightarrow \mathbb{R}^M$ を RFF の特徴写像という. 導出は 付録 1. を参照. $\omega_m \sim \mathcal{N}(\omega_m | 0, I_D)$, $b_m \sim \text{Unif}(b_m | 0, 2\pi)$ のとき, $\phi_i^\top \phi_j \approx k_{RBF}(x_i, x_j)$ となる. $\mathcal{N}(x | \mu, \Sigma)$ は, 平均 μ , 共分散行列 Σ の正規分布, $\text{Unif}(x | a, b)$ は, 区間 $[a, b]$ における一様分布, I_D は, D 次の単位行列を表す.

N 個のデータ点を用いたカーネル回帰関数 f を, M 個の RFF を用いて, 次のように近似する.

$$f(x_n) = \sum_{i=1}^N \alpha_i k(x_i, x_n) \approx \sum_{i=1}^N \alpha_i \phi_i^\top \phi_n = w^\top \phi_n$$

$w = \sum_{i=1}^N \alpha_i \phi_i \in \mathbb{R}^M$ とおいた. 単純な線形結合 $w^\top \phi_n$ で, 非線形関数 $f(x_n)$ を近似しているのが重要である. この手法により, $N \times N$ のカーネル行列の直接計算を避けることができ, 計算量を $O(NM^2)$, $M \ll N$ へと低減することができる.

2.3 ガウス過程潜在変数モデル (Gaussian Process Latent Variable Model, GPLVM)

本研究では, ガウス過程潜在変数モデル (Gaussian Process Latent Variable Model, GPLVM) [6]~[8] をベースに考える. このモデルは, 非線形なデータの関係性を捉えることができる.

観測 $X = \{x_n \in \mathbb{R}^D\} \in \mathbb{R}^{N \times D}$ に, 対応する潜在変数を $Z = \{z_n \in \mathbb{R}^Q\} \in \mathbb{R}^{N \times Q}$, $Q \ll D$ とする. GPLVM は, 観測の各次元 $X_{:,d} \in \mathbb{R}^N$, $d \in \{1, \dots, D\}$ が, ガウス過程 $\mathcal{GP}(0, K_Z)$ に従う確率的生成モデルである.

次の対数尤度を最大化することで, 潜在変数 Z を推定する.

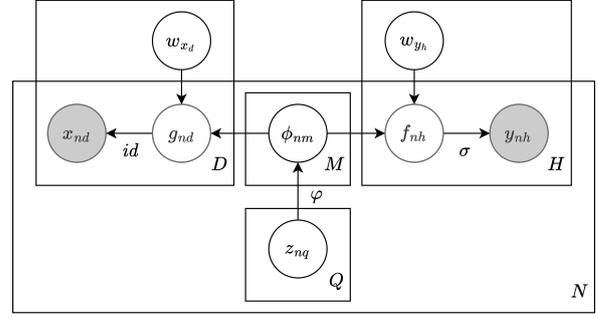


図1 提案モデルのグラフィカルモデル表現

$$\mathcal{L}(Z) = \log p(X|Z)p(Z)$$

$$\propto -\frac{D}{2} \log |K_Z| - \frac{1}{2} \text{tr}(K_Z^{-1} X X^\top) - \frac{1}{2} \text{tr}(Z Z^\top)$$

アイテムの特徴量 X だけでなく, ユーザーの評価ラベル Y など, 別の情報も, 潜在変数 Z に反映する識別的なモデル [9]~[11] も提案されている. [11] は, 各アイテムへのクラスラベル付与が, 前提になっているという課題がある. [9], [10] は, クラスラベル付与などの前提はないが, ユーザーの嗜好を表す重みが積分消去されているため, 分析の際に利用したいユーザー埋め込みが得られない. また, 多数のユーザーによる評価情報を, 低次元潜在空間に反映するためのモデルの表現力も, 課題として挙げられている. [12], [13] では, RFF に基づく GPLVM が提案されている. しかし, 種別の異なる特徴量は, 入力として想定していない. 本研究では, これらの課題を解決するモデルを提案する.

3. 提案モデル

D 次元の特徴量を持つ N 個のアイテムを $X = \{x_n \in \mathbb{R}^D\}_{n=1}^N \in \mathbb{R}^{N \times D}$ とする. H 人のユーザーによる評価を $Y \in \{0, 1, -1\}^{N \times H}$ とする. ユーザー h による, アイテム n の評価を, y_{nh} で表す. ポジティブな評価 (正例) は $y_{nh} = 1$, ネガティブな評価 (負例) は $y_{nh} = 0$, 未評価 (欠測値) は $y_{nh} = -1$ とする.

提案モデルのグラフィカルモデル表現を 図 1 に示す. グラフィカルモデル表現と対応する, 次のデータ生成プロセスを仮定する.

$$\Omega = \{\omega_m \sim \mathcal{N}(\omega_m | 0, I_D)\}_{m=1}^M \in \mathbb{R}^{D \times M}$$

$$b = \{b_m \sim \text{Unif}(b_m | 0, 2\pi)\}_{m=1}^M \in \mathbb{R}^M$$

for each item $n = 1, 2, \dots, N$

$$\phi_n = \varphi(z_n; \Omega, b), z_n \sim \mathcal{N}(z_n | 0, I_Q)$$

for each user $h = 1, 2, \dots, H$

$$f_{nh} = w_{yh}^\top \phi_n, w_{yh} \sim \mathcal{N}(w_{yh} | 0, I_M)$$

$$y_{nh} \sim \text{Bern}(y_{nh} | \sigma(f_{nh}))$$

for each item feature dimension $d = 1, 2, \dots, D$

$$g_{nd} = w_{xd}^\top \phi_n, w_{xd} \sim \mathcal{N}(w_{xd} | 0, I_M)$$

$$x_{nd} = \text{id}(g_{nd})$$

ここで, $\sigma(x) = 1/(1 + e^{-x})$, $\text{id}(x) = x$, $\text{Bern}(x|p)$ は, パラ

Algorithm 1 推論アルゴリズム

```
1: Given  $T, M, Q, \lambda_X, \lambda_Y, \eta$ 
2: Initialize  $Z, W_Y, W_X, \varphi$ 
3: Training
4:  $\Theta_0 = \{z_n \in Z, w_{y_n} \in W_Y, w_{x_d} \in W_X\}$ 
5: for  $t = 1, \dots, T$  do
6:    $\Theta_t = \text{ADAM}(\Theta_{t-1}; \frac{\partial L(\Theta)}{\partial \Theta}, \eta)$ 
7: end for
8: return  $\Theta_t$ 
```

メータ $0 \leq p \leq 1$ のベルヌーイ分布である。

データ生成プロセスでは、まず、RFF の特徴写像 φ のパラメータ Ω, b が初期化されている。次に、視覚化の対象となるアイテム n の潜在座標 $z_n \in \mathbb{R}^Q$ が初期化され、特徴写像 $\varphi(z_n; \Omega, b)$ によって、RFF ϕ_n に変換されている。ユーザー h には、重み $w_{y_h} \in \mathbb{R}^M$ が割り当てられ、 $f_{nh} = w_{y_h}^\top \phi_n$ 、 $y_{nh} \sim \text{Bern}(y_{nh} | \sigma(f_{nh}))$ によって、評価ラベルが生成されている。アイテム特徴量は、各次元 d ごとに、重み w_{x_d} が与えられ、 $g_{nd} = w_{x_d}^\top \phi_n$ 、 $x_{nd} = \text{id}(g_{nd})$ によって、生成されている。

視覚化可能な次元 $Q \leq 3$ に、アイテムの潜在座標 z_n の自由度を制限しつつも、特徴写像 φ によって、より高次元な再生核ヒルベルト空間 (Reproducing Kernel Hilbert Space, RKHS) 上の回帰問題とすることで、制限を緩和している。また、 Z だけに、様々な情報を埋め込むのではなく、ユーザーごとの重み $\{w_{y_h}\}_{h=1}^H$ 、アイテムの特徴次元ごとの重み $\{w_{x_d}\}_{d=1}^D$ と協調させることで、モデルの表現力を高めている。

4. 推論アルゴリズム

提案モデルの潜在変数を、データから推論するアルゴリズムを、Algorithm-1 に示す。イテレーション数 T 、RFF の次元 M 、潜在変数の次元 Q 、正則化係数 λ_X, λ_Y 、学習率 η は事前に与えるハイパーパラメータである。 Z は、PCA で初期化する。 W_Y, W_X は、各要素を $\mathcal{N}(0, I_M)$ で初期化する。事前に $\Omega = \{\omega_m \sim \mathcal{N}(\omega_m | 0, I_D)\}_{m=1}^M$ 、 $b = \{b_m \sim \text{Unif}(b_m | 0, 2\pi)\}_{m=1}^M$ を、サンプリングし、RFF の特徴写像は $\varphi(\cdot; \Omega, b)$ で固定する。

次の損失関数 $L(\Theta)$ 、 $\Theta = \{Z, W_Y, W_X\}$ を、勾配法 (ADAM [14]) によって、最小化する。

$$L(\Theta) = L_X(Z, W_X) + \gamma L_Y(Z, W_Y)$$

$$L_X(Z, W_X) = \sum_{n=1}^N \sum_{d=1}^D (x_{nd} - g_{nd})^2 + \lambda_X \sum_{d=1}^D \|w_{x_d}\|_2^2$$

$$L_Y(Z, W_Y) =$$

$$\begin{aligned} & - \sum_{n, h \in \mathcal{I}} \{\kappa_h y_{nh} \log(\sigma(f_{nh})) + (1 - y_{nh}) \log(1 - \sigma(f_{nh}))\} \\ & + \lambda_Y \sum_{h=1}^H \|w_{y_h}\|_2^2 \end{aligned}$$

L_X がカーネルリッジ回帰、 L_Y がカーネルロジスティックリッジ回帰に対応している。 γ は、2つの損失関数 L_Y, L_X のバランスを調整する係数である。 $\mathcal{I} = \{(n, h) | y_{nh} \neq -1, \forall n, h\}$

は、観測された Y のインデックス集合を表す。 $\kappa_h = |\{y_{nh} | y_{nh} = 0, \forall n\}| / |\{y_{nh} | y_{nh} = 1, \forall n\}|$ は、ユーザー h による評価の正例・負例の不均衡性を調整する係数である。

実装においては、jax [15]、optax [16] を用いた。勾配 $\partial L(\Theta) / \partial \Theta$ は、自動微分で求めた。

5. 評価実験

5.1 データセット

公開されている実データセット、MIND¹ [17]、ANIME²、BEER³ [18], [19] を加工して、評価用のデータセットを構築した。加工の詳細は付録 2. にまとめた。MIND はニュースの閲覧履歴、ANIME はアニメの評価、BEER はビールの評価データセットである。

構築したデータセットの統計情報を表 1 に示す。ユーザー評価 Y は欠測が多くスパースなデータであること、正例・負例が不均衡であることがわかる。

評価実験にあたって、イテレーション数 $T = 2,000$ 、可視化の次元 $Q = 2$ 、RFF の次元 $M = 128$ 、オプティマイザの学習率 $\eta = 0.001$ 、正則化係数 $\lambda_X = \lambda_Y = 0.01$ 、損失関数の調整係数 $\gamma = 10$ とした。プロセッサは、GeForce RTX 3090 (1.70 GHz, 24 GB) を用いた。

5.2 ユーザー評価 Y の予測

推論した Z を特徴量に、テスト用の Y を予測することで、ユーザー評価の反映度を評価した。観測値 Y の 20% を、テスト用に分離してから、推論アルゴリズムを実行した。

比較対象として、ロジスティック回帰 (LR, 線形モデル)、XGBoost (XGB, 非線形モデル, ツリー系) [20]、多層パーセプトロン (MLP, 非線形モデル, ニューラル系)、ガウス過程分類器 (GPC, 非線形モデル, カーネル法) を採用した。ロジスティック回帰、多層パーセプトロン、ガウス過程分類器の実装は [21], [22] を用いた。XGBoost⁴ は、[20] を用いた。比較対象の LR, MLP, XGB, GPC は、次元削減前の特徴量 X からテスト用の Y を予測している。

二値分類かつ不均衡データであることを考慮して、評価指標は ROC-AUC, F1-Score を採用した。提案モデルの数値は、 $T = 2,000$ イテレーション中の最良値を採用した。結果を表 2 に示す。ROC-AUC, F1-Score とともに、提案モデルが良好な結果を示しており、ユーザーの評価情報 Y を反映していることがわかる。

5.3 アイテム特徴量 X の復元

推論した Z から、アイテムの特徴量を復元し、誤差を測定することで、アイテム特徴量の反映の程度を評価した。 x_{nd} は観測値、 $\hat{x}_{nd} = w_{x_d}^\top \varphi(z_n)$ は予測値を表すとする。指標は、平均絶対誤差 (Mean Absolute Error, MAE)、平均絶対パーセント誤差 (Mean Absolute Percentage Error, MAPE) を採用した。

(注1) : <https://msnews.github.io/>

(注2) : <https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset>

(注3) : https://cseweb.ucsd.edu/~jmcauley/datasets.html#multi_aspect

(注4) : <https://github.com/dmlc/xgboost>

表1 評価データセットの統計情報. ユーザー評価 Y の内訳からデータセットがスパースであること, 観測値の内訳から正例・負例が不均衡であることがわかる.

データセット	アイテム数 N	特徴量の次元 D	ユーザー数 H	ユーザー評価 Y の内訳		観測値の内訳	
				観測	欠測	正例	負例
MIND	14,595	384	4,131	1,187,381	59,104,564	62,473	1,124,908
ANIME	10,714	384	5,000	948,851	52,621,149	257,919	690,932
BEER	24,003	384	235	24,003	5,616,702	5,414	18,589

表2 ユーザー評価ラベル Y の予測精度. スコアの平均値 \pm 標準偏差の形式で記載. ROC-AUC, F1-Score とともに大きいほど良い. 提案モデルが良好な結果を示している.

Model	MIND		ANIME		BEER	
	ROC-AUC	F1-Score	ROC-AUC	F1-Score	ROC-AUC	F1-Score
LR	0.631 \pm 0.184	0.000 \pm 0.000	0.636 \pm 0.128	0.112 \pm 0.228	0.549 \pm 0.201	0.067 \pm 0.193
MLP	0.577 \pm 0.171	0.046 \pm 0.137	0.620 \pm 0.115	0.313 \pm 0.232	0.545 \pm 0.195	0.283 \pm 0.251
XGB	0.557 \pm 0.194	0.003 \pm 0.025	0.591 \pm 0.126	0.238 \pm 0.236	0.531 \pm 0.200	0.251 \pm 0.230
GPC	0.630 \pm 0.191	0.000 \pm 0.000	0.637 \pm 0.127	0.117 \pm 0.234	0.565 \pm 0.184	0.077 \pm 0.198
提案モデル	0.674 \pm 0.158	0.165 \pm 0.120	0.726 \pm 0.145	0.510 \pm 0.158	0.554 \pm 0.190	0.331 \pm 0.186

表3 特徴量 X の再構成誤差

	MIND	ANIME	BEER
MAE	0.191	0.194	0.193
MAPE	4.737	4.62	4.82

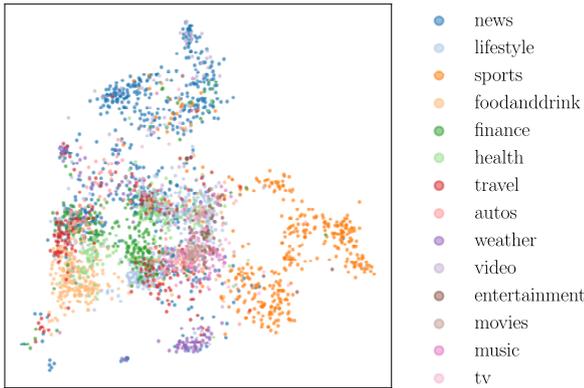


図2 潜在空間 Z の散布図. 同色のノードが近傍に配置されているということは, X の情報が Z に反映されていることを示す.

$$\text{MAE} = \frac{1}{ND} \sum_{n,d} |x_{nd} - \hat{x}_{nd}|, \text{MAPE} = \frac{100}{ND} \sum_{n,d} \left| \frac{x_{nd} - \hat{x}_{nd}}{x_{nd}} \right|$$

結果を表3に示す. 各データセットに対するMAEは約0.19, MAPE 4.6~4.8%であった. 定性的にも確認するため, MINDデータセットから推論された Z を3,000個サンプリングし, 散布図を描いた図2. 各点の色は, 各アイテムに付与されているカテゴリラベルに対応している. 同色のノードが近傍に存在するようになっており, 元の特徴量 X が反映されていることが視覚的にも確認できる.

5.4 RFFの次元 M の影響

$M = \{2, 4, 8, 16, 32, 64, 128, 256, 512\}$ と変化させ, ROC-AUCの変化を調べた. MINDデータセットを用いた結果を図3に示す. 意外であるが, M の大小は, ROC-AUC にさほど影響がな

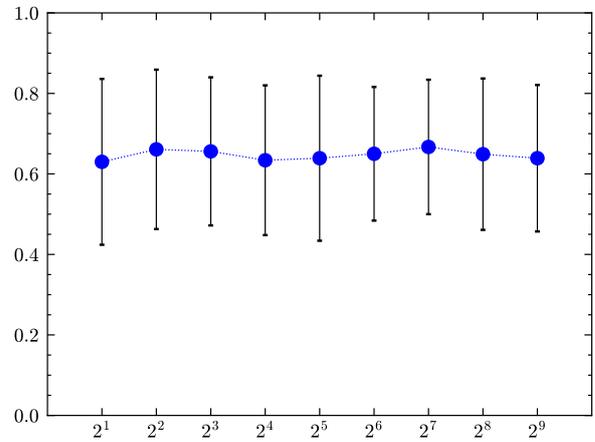


図3 RFFの次元 M (x軸) とROC-AUC (y軸) の関係. 青点が平均, 縦黒線が標準偏差を表す. M によって, ROC-AUCは, さほど変化していない.

かった. しかし, 得られる視覚化の様子はかなり異なる図4. 上段は, アイテムに付与されたカテゴリラベルで色付けした, Z の散布図である. 下段は, あるユーザーが評価したアイテム (赤: 正例, 青: 負例, 未評価アイテムは非表示) と, 等高線で示した選好分布である. $M = \{2, 8, 32\}$ は, アイテムの潜在座標が近傍に集中しており, かつ, ユーザーの選好分布は, 周期性が目立ち解釈性が低い. 視覚化という目的を考慮すると, アイテムの潜在座標が適度な広がりを持ちつつ, なめらかなユーザー選好分布が得られる $M = 128$ 以上が, 解釈性が高く望ましかった.

6. 可視化・分析への応用

ユーザーの選好は $f_{nh} = w_{yh}^T \varphi(z_n)$ で表される. 複雑な非線形関数であるユーザーの選好は, M 次元のベクトル w_{yh} (ユーザー埋め込み) で表現されているというのが重要なポイントである. 本章では, アイテムの潜在座標 Z と, ユーザー埋め込み $w_{yh} \in W_Y$ を用いた可視化・分析の実例を示す. 図表は, すべ

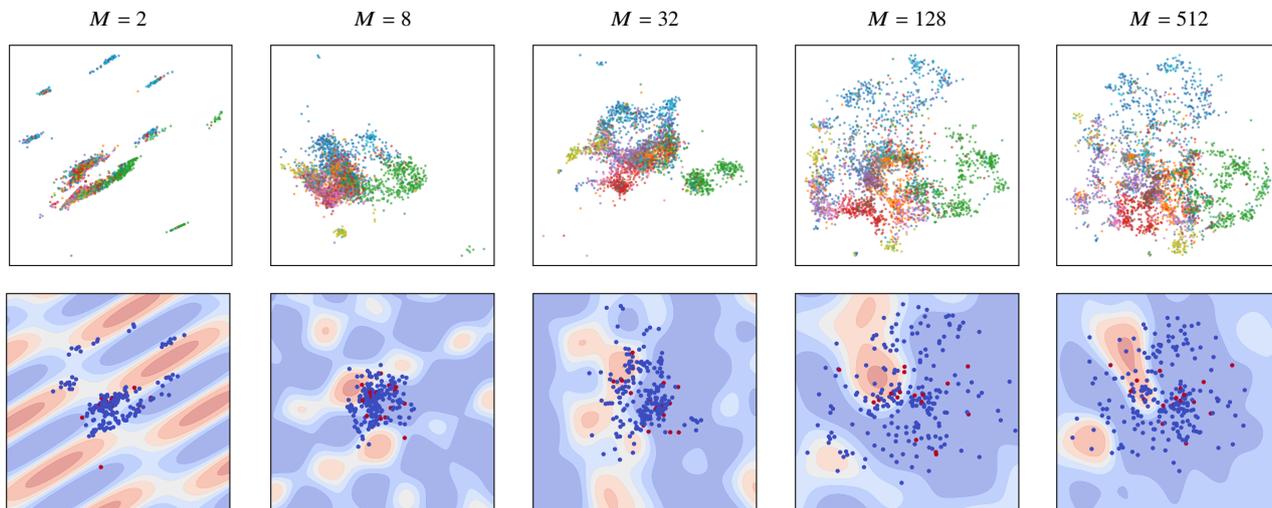


図4 M による視覚化の変化. 上段が Z の散布図. 下段がユーザーの選好分布. $M \geq 128$ では, Z が適度にばらつき, なめらかな選好分布が得られ, 視覚的な解釈性が高い.

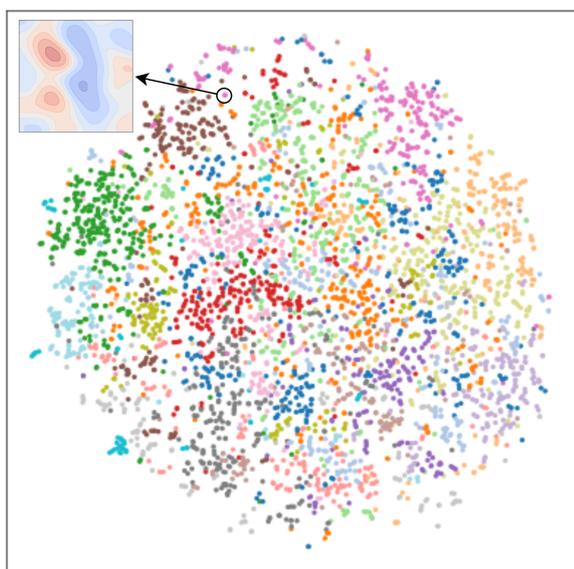


図5 ユーザー埋め込みの散布図. 各点はユーザー埋め込み w_{y_h} に対応し, 色は階層型クラスタリングによって付与した番号を表す. 左上のように, 各点はユーザーの選好分布に対応している.

て, MIND データセットから得られたものである.

6.1 ユーザー埋め込み W_Y の可視化

ユーザー埋め込み $\{w_{y_h}\}_{h=1}^H \in W_Y$, $H = 4,131$ の散布図を示す図5. 描画にあたっては, t-SNE を用いて, $M = 128$ 次元から, 2次元へ削減した. 各点の色は, クラスタ数を20に固定した階層型クラスタリングによって求めたクラスタ番号に対応している. 図5中の一点が, 左上に示すような, ひとつの選好分布に対応している.

6.2 ユーザーのクラスタリング

Web サイトを改善する場合に, どのような選好のユーザーが, どのくらい存在しているのか把握したいというニーズがある. W_Y をクラスタリングすることにより, このニーズに応えることができる.

W_Y から100人分をサンプリングし, 階層型クラスタリングによるデンドログラムを描いた図6. 横軸がユーザー番号, 縦軸がユーザー埋め込み間の距離である. 水平点線は, クラスタリングに用いた距離の閾値 (この図では6.0) である. 同一のクラスタに所属する場合, デンドログラムの枝が同色となる. 各クラスタの下部に, 代表ユーザーの選好分布を表示してある. 図6から, クラスタの大小, クラスタ間の距離を把握することができる.

6.3 類似ユーザーの抽出

選好の類似性に基づいて, アイテムを推薦したいなどのニーズがある. 例えば, ユーザー埋め込み間の内積値から, 類似ユーザーを抽出できる. 図7では, 左端のユーザーに対する, 類似ユーザーの選好分布が表示されている. 各ユーザーが実際に閲覧したアイテム10件を, 表4に示す.

W_Y から類似度が算出できることを利用して, 例えば, 選好が類似しているユーザーが実際に閲覧したアイテムを, 推薦することで, 協調フィルタリングのような推薦アルゴリズムを実現することができる.

7. おわりに

本研究では, アイテムに対するユーザーの選好 (好き・嫌いの度合い) を可視化・分析するためのモデルを提案した. 視覚化のために自由度 (≤ 3 次元) が制限された潜在空間に, アイテムの特徴量や多数のユーザによる評価という, 種別が異なる情報を同時に反映させるという課題を, 乱択化フーリエ特徴量により, RKHS 上の問題とすることで対処した.

アイテムの特徴量とユーザーの評価ラベルを, 同時に反映するための損失関数を定義し, 自動微分を用いた勾配法による推論アルゴリズムを実装した. 推論アルゴリズムは, 非線形な関係性を捉えつつも, データ数 N に対して線形オーダーの計算量であり, データの規模に対してロバストである.

実データを用いた評価実験を行い, アイテムの特徴量とユー

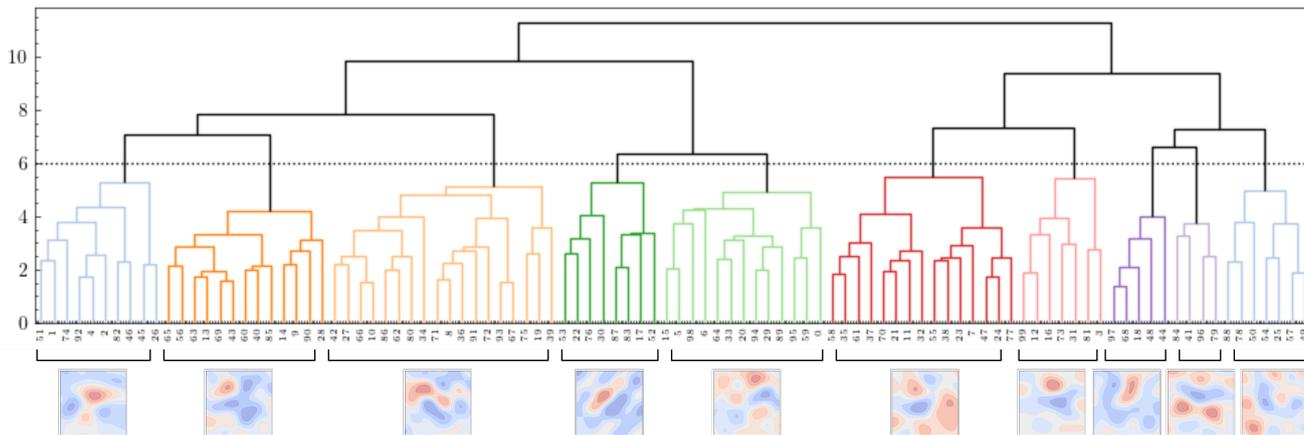


図6 ユーザー埋め込み W_Y のデンドログラム. 枝の形状と色 (クラスタ番号を表す) から, どういう選好分布を持つユーザーが, どのように存在するのか把握することができる.

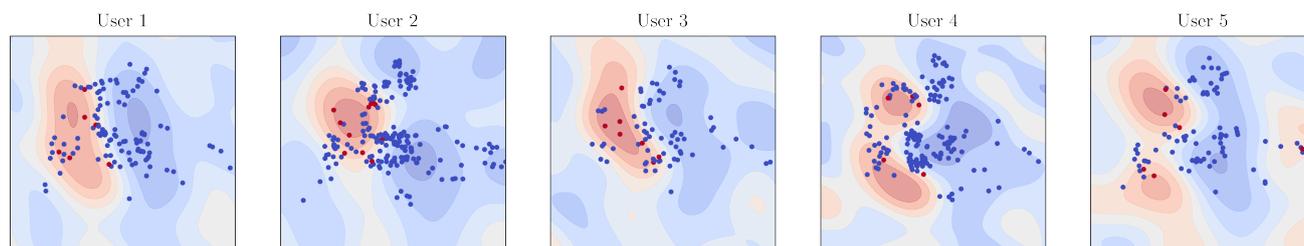


図7 類似ユーザーの選好分布. User 1 の類似ユーザー (User 2 ~ User 5) が表示されている. 選好の形状は類似しつつも, アイテムの閲覧・評価パターンは異なっている.

ザーの評価ラベル, 双方を反映した潜在空間を学習できていることを確認した. また, 獲得したユーザー埋め込みを用いた可視化・分析の実例を示した.

本モデルは, 映画, 書籍, レストラン, 不動産, 自動車など, アイテムに対するユーザーの評価という形式でデータが蓄積される, 幅広い領域への応用が期待できる.

謝 辞

本研究は, 社会人博士支援制度「mercari R4D PhD Support Program」の支援により実施しています.

文 献

- [1] H. Hotelling, “Analysis of a complex of statistical variables into principal components.,” *Journal of educational psychology*, vol.24, no.6, p.417, 1933.
- [2] L. van derMaaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol.9, no.86, pp.2579–2605, 2008. <http://jmlr.org/papers/v9/vandermaaten08a.html>
- [3] C.K. Williams and C.E. Rasmussen, *Gaussian processes for machine learning*, vol.2, MIT press Cambridge, MA, 2006.
- [4] 持橋大地, 大羽成征, *ガウス過程と機械学習*, 講談社, 2019.
- [5] A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” *Proceedings of the 20th International Conference on Neural Information Processing Systems*, p.1177–1184, NIPS’07, Curran Associates Inc., Red Hook, NY, USA, 2007.
- [6] N.D. Lawrence, “Gaussian process latent variable models for visualisation of high dimensional data,” *Proceedings of the 16th International Conference on Neural Information Processing Systems*, p.329–336, NIPS’03, MIT Press, Cambridge, MA, USA, 2003.
- [7] M. Titsias and N.D. Lawrence, “Bayesian gaussian process latent variable model,” *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, pp.844–851 2010.
- [8] V. Lalchand, A. Ravuri, and N.D. Lawrence, “Generalised gplvm with stochastic variational inference,” *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, eds. by G. Camps-Valls, F.J.R. Ruiz, and I. Valera, vol.151, pp.7841–7864, *Proceedings of Machine Learning Research*, PMLR, 28–30 Mar 2022. <https://proceedings.mlr.press/v151/lalchand22a.html>
- [9] 竹原一彰, 持橋大地, “推薦の解釈を目的とした識別的なベイズ離散ガウス過程潜在空間モデル,” *研究報告 数理モデル化と問題解決研究会 (MPS)*, vol.2022, no.32, pp.1–6, 2022.
- [10] 竹原一彰, 持橋大地, “ユーザー選好可視化のための識別的な離散ベイズガウス過程潜在変数モデル,” *人工知能学会全国大会論文集*, vol.JSAI2023, pp.4L2GS403–4L2GS403, 2023.
- [11] R. Urtasun and T. Darrell, “Discriminative gaussian process latent variable model for classification,” *Proceedings of the 24th international conference on Machine learning*, pp.927–934, 2007.
- [12] G. Gundersen, M. Zhang, and B. Engelhardt, “Latent variable modeling with random features,” *International Conference on Artificial Intelligence and Statistics*, PMLR, pp.1333–1341 2021.
- [13] M.M. Zhang, G.W. Gundersen, and B.E. Engelhardt, “Bayesian non-linear latent variable modeling via random fourier features,” 2023.
- [14] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.
- [15] J. Bradbury, R. Frostig, P. Hawkins, M.J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, “JAX: composable transformations of Python+NumPy programs,” 2018. <http://github.com/google/jax>
- [16] DeepMind, I. Babuschkin, K. Baumli, A. Bell, S. Bhupatiraju, J. Bruce, P. Buchlovsky, D. Budden, T. Cai, A. Clark, I. Danihelka, A. Dedieu, C. Fantacci, J. Godwin, C. Jones, R. Hemsley, T. Hennigan, M. Hessel, S. Hou, S. Kapturovski, T. Keck, I. Kemaev, M. King, M. Kunesch, L. Martens, H. Merzic, V. Mikulik, T. Norman, G. Papamakarios, J. Quan, R. Ring, F. Ruiz, A. Sanchez, L. Sartran, R.

表4 図7のユーザーが閲覧したアイテムのリスト. 共通して news, finance カテゴリの頻度が高いが, ユーザーそれぞれ, 異なるアイテムを閲覧している.

ユーザー	カテゴリ	タイトル
User 1	finance	No. 1 milk company declares bankruptcy amid drop in demand
	news	FBI investigating killing of US women and children in Mexico
	news	Fast food cashier and manager fired for refusing to serve police officer
	lifestyle	A master suite with an 'open-concept' bathroom that has no doors or privacy is asking for \$1,200/month in rent
	video	Creepy fish with 'human-like face' found swimming in lake
	news	Iraq protests death toll rises to 305 with nearly 15,000 injured
	news	Louisville police: TARC bus, car collision results in TARC passenger's death, two injured
	news	Multistate police pursuit ends on Cane Run Road
	foodanddrink	Snow crab sells for record price in Japan
	finance	A \$30,000 handbag. A disastrous wine spill. And now, a country club is suing its own waiter.
User 2	finance	'Priceless' finds that turned out to be worthless
	finance	No. 1 milk company declares bankruptcy amid drop in demand
	travel	Do you have the right to recline your airline airplane seat? No, and here's why
	finance	One of America's biggest solar panel makers quits manufacturing
	news	Experts crack mystery of ancient Egypt's sacred bird mummies
	travel	Video shows a Boeing 737 plane carrying 196 people burst into flames just after landing at an Egyptian airport
	news	College gymnast dies following training accident in Connecticut
	finance	The cost to become a US citizen is going up 61 percent
	news	Missing California hiker found dead at top of glacier just weeks before baby was due
	lifestyle	Kendall Jenner Wore the Tiniest Dress to Go Jewelry Shopping
User 3	music	Top 100 country songs of all time
	finance	Supreme Court refuses to block lawsuit against gun manufacturer brought by Sandy Hook families
	finance	One of America's biggest solar panel makers quits manufacturing
	news	Fast food cashier and manager fired for refusing to serve police officer
	autos	Japan Wants You To Stop Buying All Its Skyline GT-Rs
	health	This 'new' strain of HIV is actually a good thing
	finance	Dean Foods files for bankruptcy
	news	A rock in Omaha became an overnight attraction thanks to the cars that can't seem to avoid it
	autos	These 7 sports cars crushed their horsepower ratings in our dyno tests
	video	Dramatic Footage Shows Plane Skidding Off Runway at O'Hare Airport in Chicago
User 4	news	Mother of missing Florida girl charged; human remains found
	lifestyle	18 Quiet Dogs That Don't Bark (Too Much)
	news	A woman was found dead with an ax in her head. 37 years later, her husband faces charges
	news	Family of woman missing since 2006 refocuses search after new tip
	news	Nurse took in disabled teen, then tortured and killed her, police say.
	news	Missing California hiker found dead at top of glacier just weeks before baby was due
	foodanddrink	When to Use Your Oven's Convection Function
	news	Porsche launches into second story of New Jersey building, killing 2
	music	Broadway Star Laurel Griggs Suffered Asthma Attack Before She Died at Age 13
	finance	Are Stores Open on Veterans Day? Target, Aldi, Walmart and More
User 5	health	40 Restaurant Meals You Should Never Eat After 40
	travel	50 must-see stops on the East Coast
	sports	The Geno Smith coin toss conspiracy is the dumbest thing in Seahawks' win over 49ers
	news	College gymnast dies following training accident in Connecticut
	news	Photos of the Day
	news	Officials: At least 13 dead in Slovakia bus crash
	foodanddrink	Thanksgiving side dishes: 11 unique recipes to try this season
	sports	Christian McCaffrey on pace to threaten NFL record for scrimmage yards
	news	3 castaway cows swam to North Carolina's Outer Banks to escape Hurricane Dorian
	news	'One in a million' deer captured on camera in Michigan woods

- Ecosystem,” 2020. <http://github.com/google-deeppmind>
- [17] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, et al., “Mind: A large-scale dataset for news recommendation,” Proceedings of the 58th annual meeting of the association for computational linguistics, pp.3597–3606, 2020.
- [18] J. McAuley, J. Leskovec, and D. Jurafsky, “Learning attitudes and attributes from multi-aspect reviews,” 2012 IEEE 12th International Conference on Data Mining ICDM, pp.1020–1025 2012.
- [19] J.J. McAuley and J. Leskovec, “From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews,” Proceedings of the 22nd international conference on World Wide Web, pp.897–908, 2013.
- [20] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp.785–794, 2016.
- [21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol.12, pp.2825–2830, 2011.
- [22] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp.108–122, 2013.
- [23] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), eds. by K. Inui, J. Jiang, V. Ng, and X. Wan, pp.3982–3992, Association for Computational Linguistics, Hong Kong, China, Nov. 2019. <https://aclanthology.org/D19-1410>

付 録

1. RFF の導出

$k(x, y) = k(x - y)$ であるとすると、Bochner の定理より [5],

$$k(x - y) = \int_{\mathbb{R}^D} p(\omega) e^{i\omega^\top(x-y)} d\omega$$

オイラーの公式 $e^{i\theta} = \cos \theta + i \sin \theta$ を用いて,

$$= \int_{\mathbb{R}^D} p(\omega) \{ \cos(\omega^\top(x-y)) + i \sin(\omega^\top(x-y)) \} d\omega$$

実数の項のみに着目して,

$$\begin{aligned} &= \int_{\mathbb{R}^D} p(\omega) \cos(\omega^\top(x-y)) d\omega \\ &= \int_{\mathbb{R}^D} p(\omega) \cos(\omega^\top x - \omega^\top y) d\omega \end{aligned}$$

$b \sim \text{Unif}(b|0, 2\pi)$ を導入して,

$$= \int_{\mathbb{R}^D} \int_0^{2\pi} p(b)p(\omega) \underbrace{\cos(\omega^\top x + b)}_{\alpha} \underbrace{\cos(\omega^\top y + b)}_{\beta} db d\omega$$

加法定理 $\cos(\alpha - \beta) = 2 \cos(\alpha) \cos(\beta) - \cos(\alpha + \beta)$ を用いて,

$$\begin{aligned} &= \int_{\mathbb{R}^D} \int_0^{2\pi} p(b)p(\omega) \{ 2 \cos(\omega^\top x + b) \cos(\omega^\top y + b) \\ &\quad - \cos(\omega^\top(x+y) + 2b) \} db d\omega \end{aligned}$$

$\int_0^{2\pi} p(b) \cos(\omega^\top(x+y) + 2b) db = 0$ となるので,

$$\begin{aligned} &= \int_{\mathbb{R}^D} \int_0^{2\pi} p(b)p(\omega) \{ 2 \cos(\omega^\top x + b) \cos(\omega^\top y + b) \} db d\omega \\ &\approx \frac{1}{M} \sum_{m=1}^M 2 \cos(\omega_m^\top x + b_m) \cos(\omega_m^\top y + b_m) \\ &= \sum_{m=1}^M \sqrt{\frac{2}{M}} \cos(\omega_m^\top x + b_m) \sqrt{\frac{2}{M}} \cos(\omega_m^\top y + b_m) \\ &= \sqrt{\frac{2}{M}} \begin{pmatrix} \cos(\omega_1^\top x + b_1) \\ \cos(\omega_2^\top x + b_2) \\ \vdots \\ \cos(\omega_M^\top x + b_M) \end{pmatrix}^\top \sqrt{\frac{2}{M}} \begin{pmatrix} \cos(\omega_1^\top y + b_1) \\ \cos(\omega_2^\top y + b_2) \\ \vdots \\ \cos(\omega_M^\top y + b_M) \end{pmatrix} \\ &\equiv \phi_x^\top \phi_y \\ &\therefore k(x, y) \approx \phi_x^\top \phi_y \end{aligned}$$

$p(\omega)$ の取り方によって、様々なカーネル関数 $k(x, y)$ を近似できる。詳細は [5] を参照。

2. 評価データセットの構築

共通の前処理として、1) 欠損を含む行の削除、2) 同一ユーザーによる同一アイテムへの評価は最新のみを採用した。

2.1 MIND データセット

ユーザー h に対して、アイテム n が表示され、閲覧（クリック）していたら $y_{nh} = 1$ （正例）、していないなら $y_{nh} = 0$ （負例）、表示されていない場合は、 $y_{nh} = -1$ （欠測）とした。正例が 10 件以上あるユーザーのみ抽出し、ユーザー評価 Y を構築した。

上述の条件を満たす H 人のユーザーの中で、1 回以上、閲覧されたことがある、ニュース $n \in \{1, 2, \dots, N\}$ を抽出した。各ニュースに対して、title, abstract カラムのテキストを結合して sentence-transformers⁵ [23] にて 384 次元のベクトルに変換し、アイテムの特徴量 X とした。

2.2 ANIME データセット

評価は、1～5 で付与されている。4.0 以上を正例 1、それ以外を負例 0 とした。正例が 10 件以上あるユーザーを $H = 5,000$ 人抽出し、ユーザー評価 Y を構築した。

H 人のユーザーの中で、1 回以上、評価されたアニメを抽出し、各アニメの Synopsis（概要）カラムを、sentence-transformers で 384 次元のベクトルに変換し、アイテムの特徴量 X とした。

2.3 BEER データセット

評価は、1～10 で付与されている。9.0 以上を正例 1、それ以外を負例 0 とした。正例が 10 件以上あるユーザーを抽出し、 Y を構築した。

H 人のユーザーの中で、1 回以上、評価されたビールを抽出した。各ビールに対する review カラムを、ランダムに最大 10 件サンプリングし、sentence-transformers で 384 次元のベクトルに変換し、アイテムの特徴量 X とした。

(注5) : <https://github.com/UKPLab/sentence-transformers>

(注6) : モデルは all-MiniLM-L6-v2 を利用。