

Conditional Variational Autoencoder に基づく分析モデルの構築とその応用

1X16C046-2 川上達也
指導教員 後藤正幸

1. 研究背景・目的

近年、EC サイトを通じた商品の売買が広く行われ、膨大なページ閲覧履歴データや購買履歴データが蓄積されるようになった。ページ閲覧履歴データには、各ユーザが EC サイトにアクセスし、そのサイトから離れるまで (以下、セッション) の閲覧ページの推移や商品購買の有無などが記録されている。各セッションにおいて、ユーザは商品検索や閲覧、購買等を行いながらページを遷移する。また、閲覧履歴データ中には、購買に至るセッション (以下、購買セッション) と至らないセッション (以下、非購買セッション) が混在しており、両セッション間に異なる特徴を発見することができれば、マーケティング上の有用な知見が得られる。

一方、ページ閲覧履歴と購買の関係性の分析には、セッションの閲覧ページを入力として、購買の有無を予測する分類モデルを構築する方法が知られている [1]。上記の分類モデルにより、購買セッションと非購買セッションの閲覧行動の特徴の差異を明らかにすることができる。このような分析は、サイト設計などのユーザ全体の購買行動に関する改善に有用である。

しかし、EC サイトの閲覧を行っている各ユーザに対してカスタマイズされた施策をリアルタイムに打つことを考えた場合、上記のようなユーザ全体に対する分析を行う分類モデルの活用は難しい。この場合、そのセッションでは「非購買」が予測されるユーザに対し、最小の介入で購買の可能性を高めるような施策 (特定ページの閲覧を促す施策など) を導くためのモデルが望まれる。そのためには、「閲覧ページの系列」を入力として「購買/非購買」を出力とする分類モデルではなく、「購買/非購買」から「閲覧ページの系列」を出力する、データの推論が可能なモデルが有用である。このタイプの推論が可能なモデルとして、深層学習の分野では Conditional Variational Autoencoder [2] (以下、CVAE) が提案されており、本研究で対象とする閲覧行動と購買の関係性の分析にも有効であると考えられる。

そこで本研究では、購買の有無をラベルとして閲覧履歴データに CVAE を適用し、各ユーザの購買に寄与する重要な閲覧ページの分析を行うモデルを提案する。具体的には、非購買セッションのラベル以外の入力データの特徴を保持しつつ、ラベルを「購買」に変化させた際の閲覧ページを生成し、入力データと出力データの比較を行う。また、株式会社ヴァリユーズ提供の実閲覧履歴データに提案手法を適用し、購買要因分析を行うことで提案手法の有効性を示す。

2. Conditional Variational Autoencoder

2.1. モデル概要

CVAE はニューラルネットワークに基づく確率的生成モデルの一種であり、エンコーダとデコーダから構成されている (図 1)。エンコーダは入力データとラベルから潜在表現の

推論を行い、デコーダは推論された潜在表現とラベルを用いて出力データを生成する。このときの潜在表現は入力データのラベルに依存しない特徴 (以下、一般特徴) を表しており、出力データは入力データを復元したものとなっている。一方、エンコーダに入力したラベルとは異なるラベルをデコーダに入力した場合、出力データは入力データの一般特徴を保持したまま異なるラベルの特徴を持つデータとなる。したがって、CVAE を用いることで、入力データの一般特徴を保存した状態でラベル変換を行うことができる。

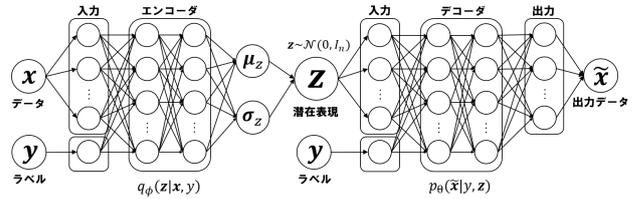


図 1: CVAE のモデル概要

2.2. CVAE の定式化

M 次元の入力データを $\mathbf{x}_i = (x_{i1}, \dots, x_{iM})^T$ 、カテゴリ数 L のラベルを $y_i \in \{1, \dots, L\}$ とすると、 N 件のデータ集合は $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}_i, y_i) : 1 \leq i \leq N\}$ の形で観測される。また、出力データを $\hat{\mathbf{x}}_i = (\hat{x}_{i1}, \dots, \hat{x}_{iM})^T$ 、入力データ \mathbf{x}_i に対応する K 次元の潜在表現を $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})^T$ とし、その事前分布 $p(\mathbf{z}_i)$ を K 次元標準正規分布とする。

このとき、エンコーダでは入力データ \mathbf{x}_i 、ラベル y_i を入力とし、潜在表現 \mathbf{z}_i のパラメータである平均 $\boldsymbol{\mu}_z(\mathbf{x}_i, y_i)$ と分散 $\boldsymbol{\sigma}_z^2(\mathbf{x}_i, y_i)$ を出力する。これらのパラメータを持つ式 (1) の分布を、入力データ \mathbf{x}_i とラベル y_i における潜在表現 \mathbf{z}_i の分布 $q_\phi(\mathbf{z}_i|\mathbf{x}_i, y_i)$ とする。

$$q_\phi(\mathbf{z}_i|\mathbf{x}_i, y_i) = \mathcal{N}(\boldsymbol{\mu}_z(\mathbf{x}_i, y_i), \boldsymbol{\sigma}_z^2(\mathbf{x}_i, y_i)) \quad (1)$$

一方、デコーダではエンコーダによって推論された潜在表現 \mathbf{z}_i とラベル y_i を入力とし、出力データ $\hat{\mathbf{x}}_i$ のパラメータ θ を出力する。このとき、潜在表現 \mathbf{z}_i とラベル y_i における出力データ $\hat{\mathbf{x}}_i$ の分布を $p_\theta(\hat{\mathbf{x}}_i|y_i, \mathbf{z}_i)$ とする。

また、CVAE におけるネットワークのパラメータは、データ (\mathbf{x}_i, y_i) を用いて式 (2) で表される損失関数が最小となるように誤差逆伝播法により学習する。

$$\mathcal{L}(\mathbf{x}_i, y_i) = -\mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{x}_i, y_i)} [\log p_\theta(\mathbf{x}_i|y_i, \mathbf{z}_i) + \log p_\theta(y_i) + \log p(\mathbf{z}_i) - q_\phi(\mathbf{z}_i|\mathbf{x}_i, y_i)] \quad (2)$$

3. 提案モデル

3.1. 概要

本研究では、非購買の可能性が高いセッションに対して、閲覧すると購買確率が高まると期待される重要なページを発見するために、CVAE によるデータ生成を用いて重要なページに関する分析を行うモデルを提案する。具体的には、非購買セッションにおける閲覧ページデータを CVAE に入力

し、エンコーダに「購買」のラベルを入力する。これにより、ユーザの閲覧に関する一般特徴を保存した状態で、そのユーザが購買を行うことを仮定した場合のセッションにおける閲覧ページを生成することが可能となる。そして、入力したデータと生成されたデータの差分により、購買に至る重要な閲覧ページを検出する。

3.2. 変数の定義

ページのカテゴリ数を M 、全セッション数を N 、各セッションにおける閲覧回数ベクトル $\mathbf{x}_i = (x_{i1}, \dots, x_{iM})^T$ を入力データ、セッション内における購買の有無をラベル $y_i \in \{0, 1\}$ によって表す。ただし、非購買セッションを $y_i = 0$ 、購買セッションを $y_i = 1$ とする。

一方、 \mathbf{x}_i における総ページ閲覧回数を s_i 、 $\tilde{\mathbf{x}}_i$ の各ページカテゴリの閲覧確率を $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{iM})^T$ とする。このとき、デコーダは $\boldsymbol{\pi}_i$ を出力し、潜在表現 \mathbf{z}_i とラベル y_i における出力データ $\tilde{\mathbf{x}}_i$ の分布 $p_\theta(\tilde{\mathbf{x}}_i | y_i, \mathbf{z}_i)$ は式 (3) に示す多項分布を仮定する。

$$p_\theta(\tilde{\mathbf{x}}_i | \mathbf{z}_i, y_i) = \text{Multi}(\tilde{\mathbf{x}}_i | \boldsymbol{\pi}_i, s_i) \quad (3)$$

3.3. 提案モデルの手順

提案モデルを用いた購買要因分析の手順を以下に示す。

Step1) CVAE の学習

出力データ $\tilde{\mathbf{x}}_i$ が入力データ \mathbf{x}_i を復元し、入力データ \mathbf{x}_i における潜在表現 \mathbf{z}_i の分布 $q_\phi(\mathbf{z}_i | \mathbf{x}_i, y_i)$ と潜在表現の事前分布 $p(\mathbf{z}_i)$ の差が小さくなるように CVAE のパラメータを学習する。

Step2) 非購買セッションにおける潜在表現の推論

入力データ \mathbf{x}_i とラベル $y_i = 0$ を学習済みモデルのエンコーダに入力することで潜在表現 \mathbf{z}_i を推論する。

Step3) 非購買セッションの購買セッションへの変換

Step2 で推論した潜在表現 \mathbf{z}_i とラベル $y_i = 1$ を学習済みモデルのデコーダに入力することにより、該当ユーザが購買すると仮定した場合の各ページの閲覧回数ベクトル $\tilde{\mathbf{x}}_i$ を得る。このとき、非購買セッション ($\mathbf{x}_i, y_i = 0$) の一般特徴を保存した状態でラベルを「購買」 ($y_i = 1$) に変換したときの出力が $\tilde{\mathbf{x}}_i$ となる。

Step4) 購買要因分析

Step2 で入力した非購買セッションの閲覧回数ベクトル \mathbf{x}_i と、Step3 で生成した閲覧回数ベクトル $\tilde{\mathbf{x}}_i$ を比較することにより、各非購買セッションに対する購買に寄与する閲覧ページを分析する。

4. 実データ分析

提案手法の有効性を示すため、提案手法を実データに適用して購買要因分析を行う。

4.1. 分析データ

株式会社ヴァリューズにより提供された楽天市場サイト上の閲覧履歴データを用いる。データ収集期間は 2017 年 8 月 1 日から 2017 年 10 月 31 日であり、総ユーザ数は 766 人、総セッション数は 37,304 件、総購買数は 1,079 件である。また、ユニークページ数は 53,462 ページであり、これらのユニークページを 35 カテゴリに集約した。

4.2. 分析条件

本実験では、購買に寄与する重要なページの分析を行うため、購買セッションにおいて出現頻度が高いことが明白であるカートページ、購買手続ページ、購買完了ページの 3 つについては入力データから除外し、32 種類のページを用いた。また、閲覧ページ数が 5 件以上のセッションのみを用いて実験を行った。CVAE において、入力データの次元数を 32、潜在表現の次元数を 20 とし、中間層についてはエンコーダ、デコーダそれぞれ、ユニット数を 100、層数を 2 とした。

4.3. 分析結果と考察

提案手法により、ある非購買セッションを購買セッションに変換したときの各ページの閲覧確率を表 1 に示す。このとき、表中の観測データは実際非購買セッションにおける各ページカテゴリの閲覧割合とする。また、復元データは、デコーダに入力するラベルを $y_i = 0$ とし、非購買セッションを指定して生成した各ページカテゴリの閲覧確率 $\boldsymbol{\pi}_i$ 、変換データは提案モデルにより変換した購買セッションにおける閲覧確率 $\boldsymbol{\pi}_i$ を示している。

表 1: ある非購買セッションのページごとの閲覧確率 (%)

ページ内容	セール商品 検索ページ	イベント ページ	商品詳細 ページ	トップ ページ	...	商品価格 比較ページ
観測データ ($\mathbf{x}_i, y_i = 0$)	58.33	19.44	13.89	5.56	...	0.00
復元データ ($\tilde{\mathbf{x}}_i, y_i = 0$)	49.02	15.83	12.59	9.47	...	0.61
変換データ ($\tilde{\mathbf{x}}_i, y_i = 1$)	9.03	12.39	34.69	9.40	...	12.68

表 1 より、観測データでは、ある非購買セッションにおいてユーザはセール商品検索ページ、イベントページを多く閲覧している。このユーザのラベルを「購買」に変更した場合のページ閲覧確率を表す変換データでは、観測データに比べて商品価格比較ページと商品詳細ページの閲覧確率が大きく上昇している。そのため、この非購買セッションを購買セッションに変えるためには、商品詳細ページと商品価格比較ページに誘導するような施策を打つことが効果的であり、売上向上に結びつく可能性がある。

5. まとめと今後の課題

本研究では、CVAE を閲覧履歴に適用し、非購買セッションの閲覧ページから購買セッションの閲覧ページを生成することで、購買に寄与する重要なページを発見する分析モデルを提案した。また、実際の閲覧履歴データを用いてその有効性を示した。

今後の課題として、CVAE に時系列性を導入し、閲覧順序を考慮した分析モデルの構築などが挙げられる。

参考文献

- [1] M.Hotoda, H.Mizuochi, G.Kumoi, M.Goto, "Analytical Model of Customer Purchase Behavior Considering Page Transitions on EC Site," *Total Quality Science*, Vol.5, No.1, pp.23-33, 2019.
- [2] D.P.Kingma, S.Mohamed, D.J.Rezende, and M.Welling, "Semi-supervised learning with deep generative models," *Neural Information Processing Systems*, 2014.