エンジニアリングデザイン演習
G2 最終レポート
通販サイトにおけるレコメンドシステム

195705J　多和田真助
195707E　豊見山裕太
195710E　宮城響大
195717B　藤原敦貴

提出日 2022 年 1 月 25 日

目次

1　テーマ・目標

現代社会ではインターネットが広く普及し、通販サイトを用いて買い物することも多くなっている。通販サイトを利用しているとおすすめの商品があげられたりする。グループ 2 では、レコメンドシステムの仕組みを理解し運用するために、通販サイトの売買履歴のデータセットを用いて機械学習を行い、通販サイトユーザーに対するおすすめを出力するプログラムを作成する。Competition に参加し高い評価を得て、より良いレコメンドシステムを作成することを目標とする。

2　アプローチの全体像

まず機械学習に用いるデータセットを用意し、そのデータセットのデータをどういう風に学習データと検証データに分けるか、どの学習モデルを用いるのが良いかを話し合い決定した。コードの書き方、学習モデルの評価指標などについて調べて意味を理解しつつ、全員でコードを書いた。Competition に参加し、コードを実行した結果を投稿し評価を得て、そこから評価を上げるために、データセットの学習データと検証データの割合を考え直して修正した。

3　予定していた実験計画

ひとまず機械学習を行い、おすすめ商品を出力するプログラムを作成し、それから Competition に参加することで作成したレコメンドシステムがどれくらい優れているかを評価してもらい、データセットの学習データ・検証データの分け方を変えたり、学習モデルのパラメータ調整を行うなどして、その評価を上げることで作成したレコメンドシステムの質を上げることを計画していた。実際のところ、ひとまずレコメンドシステムの作成と Competition への参加、データセットの学習データ・検証データの分け方を変えることまでは進めることができたが、まだ学習モデルの細かいパラメータ調整を行うことができずにいる。

4　実験方法

4.1　実験目的

SIGNATE のデータセットを用いて、通販サイトの行動履歴から最適な商品を推薦するための機械学習を行う。Competition に参加し高い評価を得て、より良いレコメンドシステムを作成することを目標とする。

4.2 データセット構築

https://signate.jp/competitions/268/data
上記 URL のデータセットを使用した。train_A.tsv、train_B.tsv、train_C.tsv、train_D.tsv の 4 つのデータセットがあり、それぞれ人材、旅行、不動産、アパレルの購買履歴のデータである。全部で 5 つの特徴量が存在し、user_id はユーザーの id を示しており、product_id は商品の id を示している。event_type はユーザーが商品に対して取った行動について、変数種別に整数型で記載されており、3 が購入、2 が商品をクリック、1 が商品を閲覧、0 が商品をカートに入れるという行動となっている。ad は商品の購入が広告を経由したものかどうかを表した特徴量であり、変数種別は整数型で、そもそも商品を購入していない場合を-1、広告経由で商品を購入した場合を 1、広告経由なしに商品を購入した場合を 0 としている。time_stamp は yyyy-mm-dd hh:mm:ss.sss の形で記載されており、ユーザーがその行動を起こした時の時刻 (○年○月○日○時○分○秒) を表している。説明変数を event_type、ad、time_stamp とし、目的変数を user_id と product_id としているが、目的変数は user_id のみのパターン、product_id のみのパターン、user_id と produc_id の 2 つを使用したパターンの 3 つが存在する。test.tsv はテスト用のデータセットであり、特徴量は user_id のみである。この user_id に対してどの商品を推薦するかを決めるための特徴量である

4.3 モデル選定 (学習器選定理由)

本実験で作成した目的変数データは関連度であり、整数値をとるが、それぞれ”購入した”、”詳細ページを閲覧した”など、意味を持ち、カテゴリ変数とみることができる。つまり、間に値が存在しない。一方で、”詳細ページを閲覧した”¡”購入した”という具合にそれぞれに順序は存在し、一般的な数値 (実数) の性質も一部持っている。一般的に機械学習において、目的変数がカテゴリ変数の場合は分類問題、実数の場合は回帰問題を解くことに対応しており、学習の仕方が異なるが、今回はそれぞれの性質を有する問題と言える。このような問題に対応する学習方法として、ランク学習というものがあり、今回はランク学習によりモデルを学習した。その為、モデルは lightgbm を用いた。lightgbm は分類や回帰のみならず、ランク学習にも対応できるのが特徴の一つである。まず、lightgbm ライブラリをインポートし、LGBMRanker によってモデルを作成した。

4.4 パラメータ調整

lambda_l1 の値はデフォルト値の 0 であり、L1 正則化項の係数である。lambda_l2 の値はデフォルト値の 0 であり、L2 正則化項の係数である。num_leaves の値はデフォルト値の 31 であり、1 本の木の最大の葉の枚数を表している。feature_fraction の値はデフォルト値の 1.0 であり、各決定木においてランダムに抽出される列の割合のことである。bagging_fraction の値はデフォルト値の 1.0 であり、各決定木においてランダムに抽出される標本の割合のことである。bagging_freq はデフォルト値の 0 であり、この指定したイテレーション毎にバギングを実施する。min_data_in_leaf

の値はデフォルト値の 20 であり、1 枚の葉に含まれる最小のデータ数のことである。調整したパラメータとして、eval_at が挙げられる。ランク学習では ndcg という評価指標がよく使われており、LightGBM でもサポートされている。ndcg には検索結果リストの上位何件を評価に用いるかというパラメータが存在し、eval_at=(1,2,3,4,5,6,7,8,9,10) のように指定することによって、上位 10 件までを評価に用いるようにした。

5　実験結果

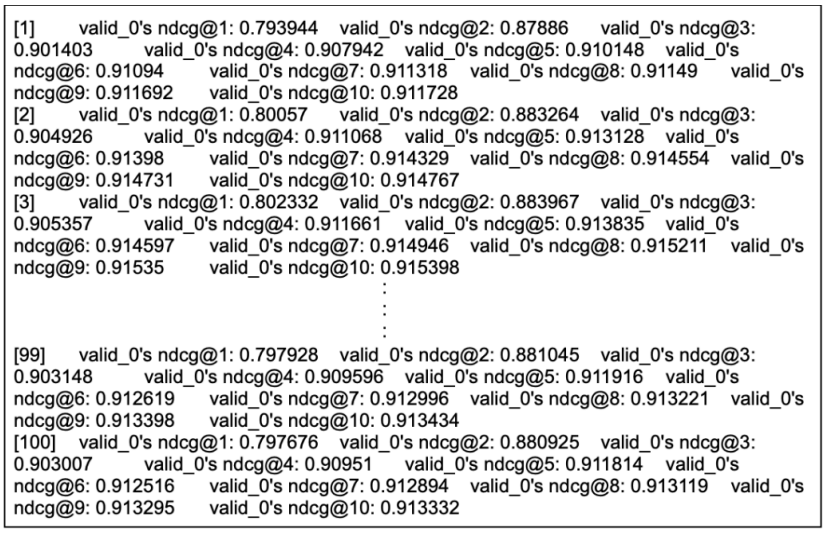

図 1　学習の様子

精度評価の様子は上記のようになっている。今回は精度評価に ndcg を用いた。

$$\mathrm{nDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{IDCG@}k}$$

$$\mathrm{DCG@}k = \sum_{j=1}^{\min(k,n)} \frac{2^{r_j} - 1}{\log(j+1)}$$

図 2　ndcg

ndcg は 0 1 の間の値をとり、値が高ければ高いほど精度が高いことを示す。直感的には、IDCG で表される理想の並び順 (実際の関連度が大きい順) の値に対して DCG で表される予測した並び順の値の割合を表し、上位に関連度が高い商品が並んでいる割合が多いほど精度が高くなる。rj は j 番目の商品の関連度で、k によって予測数の上限を決める。

さらに学習の精度の評価を得るために、Competition に参加した。学習用に使うデータと検証用に使うデータの割合を 5:5 にして提出してみたところ、評価結果は以下のようになった。

図 3　5：5 結果

学習用に使うデータと検証用に使うデータの割合を 8:2 にして提出してみたところ、評価結果は以下のようになった。

図 4　8：2 結果

この評価は SIGNATE によって出された評価であり、1 が MAX の値となっている。

6　考察

本実験では関連度をユーザーの行動パターンを基にして作成した。その関連度を目的変数として用い、関連度は整数値でありカテゴリ変数ともみることができ、ランク学習として機械学習させるために lightgbm を活用した。SIGNATE の Competition に参加した際の評価についてだが、学習データと検証データの割合を 5:5 にすると評価値が 0.1118、8:2 にすると 0.0786 となった。学習データの割合を増やしたことによって、評価が下がってしまったことから、過学習となってし

まったのではないかと考えられる。最適な学習データと検証データの割合はどれくらいなのか考える必要がある。当初はデータの分け方として、単純に 5:5、8:2 というように分けるのではなく、時系列順に前半と後半を 1:1 になるように分けて学習を行ったが、前半部分だけにいるユーザーなどテストデータに反映できないユーザーが出てきたため、SIGNATE の評価を得られずエラーを出されてしまった。そのため、それぞれのユーザーに対して訓練データと検証データを 5:5 や 8:2 に分けることによって、学習させた。

7　自己評価や振り返り

考察で記述したように、最終的に当初予定していたデータの分け方とは違ったデータの分け方で学習を行うこととなった。当初の予定でうまくいかなかったという点に関しては、あまり良いことではないが、新しくデータの分け方を考えついた点や修正できたことに関してはよかった点として挙げることができると考える。

8　時間の都合上省いた項目

細かいパラメータ調整をすることができなかったため、それを行うことによってさらに精度を向上させることが可能であると考える。学習データと検証データの割合を 5:5、8:2 のみで学習を行い、8:2 の場合は過学習であると考えられたため、他のデータの割合で学習してみたり、ただユーザーを割合で分けるのではなく何か他の方法で分けることができないか模索してみたいと考える。

参考文献

[1] レポートのテンプレート,https://github.com/naltoma/info3dm-report-template/blob/master/template.pdf, 2021 参照

[2] 機械学習コンペで人気の LightGBM を Python で使ってみた,https://watlab-blog.com/2020/04/12/light-gbm/ 2021 参照

[3] LightGBM でサクッとランク学習やってみる https://www.szdrblog.info/entry/2018/12/05/001732 2021 参照