基般

04

学生チームによるデータ分析プロセスとメリット

田中一樹（慶應義塾大学大学院）　池田春之介（慶應義塾大学）

チーム構成と役割

私たちは KDD Cup 開始当初，それぞれ個人で分析を行い，基礎集計などを通じてデータ特性の理解を進めた．その後，終了 3 週間前にチームを組み，2 人でアイディアを出し合い，主に田中が実装するという形で分析を行った．分析環境は MacBook Pro 2 台，使用した言語は主に Python，R であり，チームポリシーとして「1 日 1 特徴量作成」を掲げて分析を行った．

アルゴリズム

特徴量エンジニアリング

今回の KDD Cup では，インターネット上で誰でも無料で大学の授業を受けることができる Massive Open Online Course（MOOC）におけるユーザの受講ログや講座の情報などのデータが与えられたがすべて質的変数（講座 ID，セッション内容，セッション時間など）であったので，それらをダミー変数などで数値化し，予測器に入力できる形に前処理を行った．また，特徴量を増やさずに予測器の改良だけを行っても精度が上がることはなかったので，新たに自分たちで特徴量を作成していく必要があった．

基礎特徴量

まず，以下の基礎的な特徴量を作成した．

- カウント変数（その講座を受講したユーザ数，ユーザのセッション回数など約 200 個）
- ダミー変数（セッションの曜日，月，時間，講座 ID，約 100 個）
- 割合変数（各ユーザの受講率や宿題提出率といった行動の割合，ユーザやコースの離反率，約 20 個）
- 時間変数（最初と最後のログイン日時，セッション間隔，それらの平均・分散・最大値―平均，平均―最小値，最大値―最小値，約 50 個）

これらの基礎特徴量を用いることによって性能の良さを表す Area Under the Curve（以下 AUC）0.900 付近まで精度を出すことができた．

効果的だった特徴量

基礎特徴量のみを用いた場合では上位に入賞することはできず，より精度を上げるためにより効果的な特徴量を作成する必要があった．そこで，MOOC ではどのようなことが重要かを考え，ユーザのやる気が大きく関係すると仮定し，各ユニークユーザに注目した特徴量を作成した．具体的には，複数回異なる講座を受講しているユーザに関して，基礎特徴量の一部を集約し，それらの合計，平均，分散を求め特徴量として追加した．その結果，AUC は約 0.903 まで上がり，この特徴量によってユーザ特性をより表現できたと考えられる．

また，講座期間を前期・中期・後期と 3 つに分割し，それぞれに対して同様に基礎特徴量とその集約特徴量を作成したところ，精度が改善した（この 3 分割は手元の交差検証の試行錯誤の末決定），この特徴量では講座の前半はよく受講しているが後半は欠席しがちというユーザや継続的に受講しているユーザを上手く表現できたと考えられる．

さらに，効果的であったのは，今回の離反の定義である各講座の後の 10 日間で，ほかの講座に参加しているか否かという特徴量である．各講座の開講期間を調べてみると開講期間が重複している講座が複数あったため，離反するかを予測する講座の 10 日間にほかの講座を受講していれば，同時にその講座も受講するのではないかという仮説を立てた．そ

して，離反を判別する定義の 10 日間でほかに受講している講座数，ログイン日数などの特徴量を作成し，AUC を約 0.9055 まで伸ばすことができた．

モデル構築

私たちは作成した訓練データを用いて，ロジスティック回帰，Deep Learning，Factorization Machine などさまざまなモデルを使用したが最終的に Gradient Boosting Decision Tree（以下 GBDT）を採用した．今まで述べた AUC の値はすべて 1 つの GBDT を用いた予測結果であったが，上位入賞のためにはより精度を向上させなければならなかった．そこで，作成した約 1,300 個の全特徴量が入った訓練データを用いてモデルパラメータが異なる 7 つの GBDT を作成した．そして，それらの予測結果を単純平均（いわゆるアンサンブル）し，AUC を約 0.9057 に上げることができた．

さらに，KDD Cup 締切が近づくに連れて 10 日間に着目した特徴量を追加しても AUC が頭打ちになる状況が起こったため，一部の特徴量（追加しても AUC が変化しない特徴量）が異なる 7 つの訓練データを作成した．そして，そのそれぞれに対して GBDT を構築し，GBDT の予測結果をアンサンブルした．その結果，AUC を約 0.90599 まで伸ばすことができ，10 位に入賞することができた．これは，より訓練データにバリエーションを与えることができる相互補完の意味で，アンサンブルが成功したからだと考えている．また，興味深いことに部分的に異なる特徴量を使用した予測スコアと全特徴量を使用した予測スコアがシングルモデルではほぼ一致していた（約 0.9055）．

学生参加のメリット

最後に，学生が社会人も参戦するデータ分析コンペに参加することで得られるメリットについて解説する．

最も大きいメリットの 1 つが，教科書や論文などで学んだ種々の手法を実際のデータで試し，それぞれの手法間の違いや利点・欠点など，文字上では得られない知識を得られることである．自分でプログラムを組み利用することで，パラメータの意味やアルゴリズムの理解も進む．また，実際のデータはそのままモデルに入力できない形であることが多いため，データの前処理が必要となる．そういったデータ分析と言われる作業の一連の流れをデータ分析コンペでは学ぶことができ，データが与えられても何から手を付ければよいか分からない！という状況からはすぐに抜け出せるだろう．

さらに，データ分析コンペに関する特徴として，常に結果を表すリーダボードが変動しているため，毎日新たなことを考え実装しなければ上位に入り込むことは困難である．そのため，論理力や忍耐力といった考え抜く力が必要であるが，それらは自然と身に付いていくと思われる．コンペ終了後には上位入賞者の分析手法やそのコードが公開されることもあり，自分と同じ点，異なっていた点を発見し，新たなアイディア・技術を身に付けることができる．

今回の KDD Cup では，上位入賞者は KDD Workshop で発表があり，上位の方々との交流という学生にとって非常に有意義な時間を過ごすことができた．普段出会うことができない世界トップレベルの分析者達とお互いの研究や分析方法を議論でき，新たな視界を拡げることができた．

以上をまとめると，研究では扱うことができないようなデータでさまざまな手法を試しながら，機械学習等に関する実用的な理解を深めることができる．よって，敷居が高いと思われがちなデータ分析コンペだが，気軽にデータをダウンロードし，予測結果を提出することもできるので，時間に余裕のある学生はぜひ一度挑戦してみるべきだと思う．

（2015 年 10 月 28 日受付）

田中一樹 ■ ikki0407@gmail.com
2011 年東葛飾高校卒業，2015 年慶應義塾大学理工学部卒業，2015 年同大学院理工学研究科総合デザイン工学専攻入学．大森研究室所属．

池田春之介 ■ ikeda.shunnosuke@gmail.com
2011 年旭丘高校卒業，2012 年慶應義塾大学理工学部システムデザイン工学科入学．大森研究室所属．