Default Prediction Documentation

Introduction to Artificial Intelligence and Machine Learning

December 22, 2017

Professor Tian-Li Yu

電機四 B03901018 楊程皓

Algorithm

我以 Gradient Boosting 作為 ranking 的 algorithm / model. 套件使用的是 xgboost，主要是將多個 boosted trees 的預測值 ensemble 起來作為最終預測的分數。在這之前也有使用 DNN 直接預測每個 instance default 的機率，但在這個問題下如果沒有夠好的 preprocessing, neural network 很難將有用的資訊擷取出來，故我在嘗試過 normalization, one-hot, discretization, 但 public score 仍只能到 0.77 後，轉而使用基礎是 Decision Tree 的 Gradient Boosting。而 objective function 我使用的是 rank:pairwise, 因相較於 entropy, rank:pairwise 是兩兩比較更像於 default 的可能性，比較貼近本題的 evaluation: map@500 的形式。

Parameters

重要參數如下:

- max depth (5): 每個 boosted tree 的最大深度 (測試範圍: 3-10)。
- num of estimators (100): boosted trees 的數量 (測試範圍: 100-1000)。
- learning rate (0.1): boosted trees 的 learning rate, 決定每棵樹的參數更新幅度 (測試範圍: 0.01-0.1)。
- colsample bytree (1.0): 每棵 boosted tree 所用到的 columns 數量上限 (測試範圍: 0.4-1.0)。
  我的更改參數方式是隨機切 1/4 的 training data (1/4 default 0, 1/4 default 1) 作為 validation, 計算 map@500 並重複做十次作為衡量標準。最前面的括號是我最後使用的參數。

Preprocessing

因為我的 model 是以 Decision Tree 為基礎架構，上述三種 normalization, one-hot, discretization 都對他沒什麼用 (我也嘗試過了)，故最後我就沒有使用任何 feature 前處理了。

Packages

xgboost (0.6a2): model 主要使用套件。
sklearn: 方便計算 accuracy。

Script Usage

Default_train.sh data/Train.csv
Default_predict.sh data/Test_Public.csv data/Test_Private.csv