2020･08･03　ゼミ資料

Pythonではじめる教師なし学習
9章 半教師あり学習

1116 17 9036
山口真哉

やること

- 半教師あり学習
- クレジットカードデータを使って，教師あり学習，教師なし学習，半教師あり学習 の比較

モチベーション

- 教師あり学習はラベル付けされたものに適していて,
  ラベル付けされていなければ,教師なし学習が必要.
  - 実際は完全に区別されていない.
- ラベル付けされたものだけを使うのはうまくデータを活用できていない.
- ラベル付されたデータの情報を活用しつつ,
  ラベル付されていないものからも効率的に情報を得たい.
- 半教師あり学習は両者のデータをうまく活用する手法.
- 詳細は後ほど

データの準備

- 2, 4, 8章で使用したクレジットカードデータ(PCAされたもの)を使って不正検出をする
- 284,807のうち492が不正データである.
- Class列とTime列を取り除いたデータを標準化しこれを使用する.
- このうち2/3を訓練データ, 残りの1/3をテストデータする.
- 異常スコア(不正っぽさ)を 元のデータと学習したデータの二乗和を
  正規化して[0, 1]に収めたものとする.
- 状況を再現するために訓練データの不正なトランザクションの90%を削除する.

Case 1 (教師あり学習)

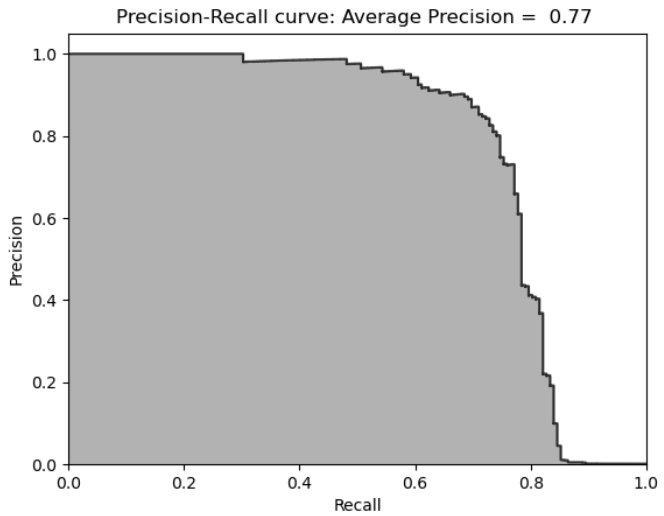
- LightGBMを使用し, 5分割交差検証を行う
- ハイパーパラメータは右の通り(教科書との変更点は赤色)
  - 学習率が大きくてうまく収束できていなかったため
- 半教師あり学習と合わせるため,
  num_boost_round=5000に変更した.
  (勾配ブースティングのイテレーションの回数)

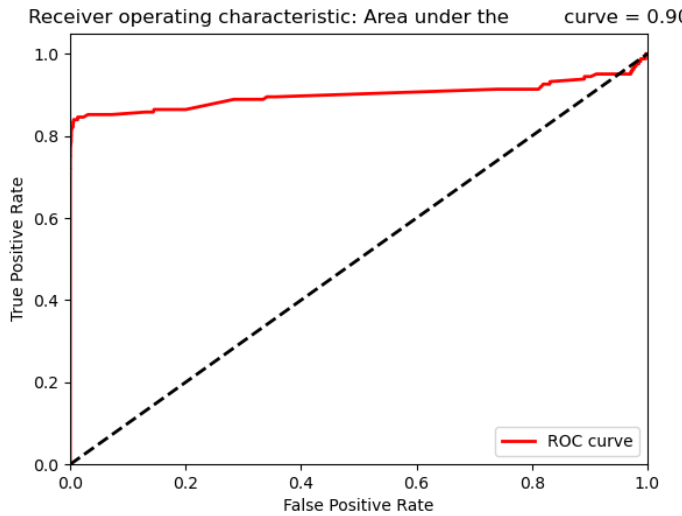
```
params_lightGBM = {
    'task': 'train',
    'application':'binary',
    'num_class': 1,
    'boosting': 'gbdt',
    'objective': 'binary',
    'metric': 'binary_logloss',
    'metric_freq': 50,
    'is_training_metric':False,
    'max_depth': -1,
    'num_leaves': 31,
    'learning_rate': 0.001,
    'feature_fraction': 1.0,
    'bagging_fraction': 1.0,
    'bagging_freq': 0,
    'bagging_seed': 2018,
    'verbose': -1,
    'num_threads':16
}
```

Case 1 result

- 平均適合率は77%, auROCは90%であった.
- 75% の不正を検出をしようとすると, 適合率が0.7485になる. (教科書よりよくなりすぎた)

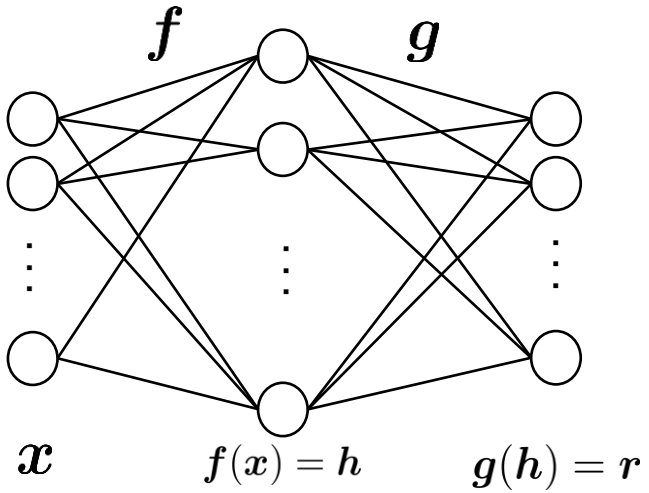
test データに対する適合率-再現率曲線とROC曲線





Case 2 (過完備線形オートエンコーダ)

- 線形活性化関数を用いた2層過完備オートエンコーダを使用する.
- バッチサイズを64, エポック数を100とする.
  - ➢ バッチサイズ大きくして学習速度を速めつつ, エポック数を大幅に増やした. (先週の先生の指摘)
- 10^{-4}のL^1正則化を入れて, 2%をドロップアウトする.
- 不正なものの感度を上げるためにオーバーサンプリングを行う.
  不正なデータを100倍に複製して訓練データに加える.

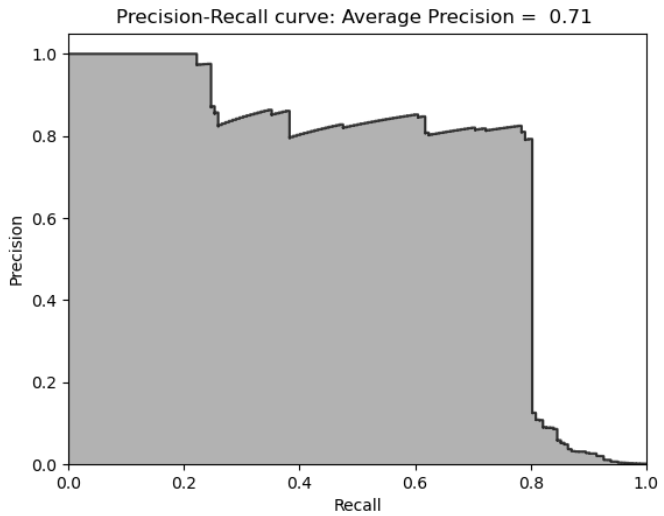


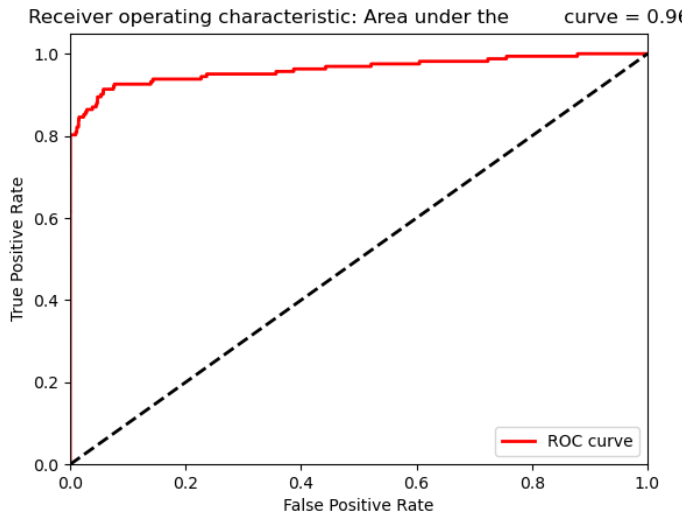
$\boldsymbol{x}, \boldsymbol{r}$ の次元は 29, $\boldsymbol{h}$ の次元は 40

Case 2 result

- 平均適合率は71%, auROCは96%であった.
- 75% の不正を検出をしようとすると, 適合率が0.8188になる. (教科書よりよくなった)

test データに対する適合率-再現率曲線とROC曲線





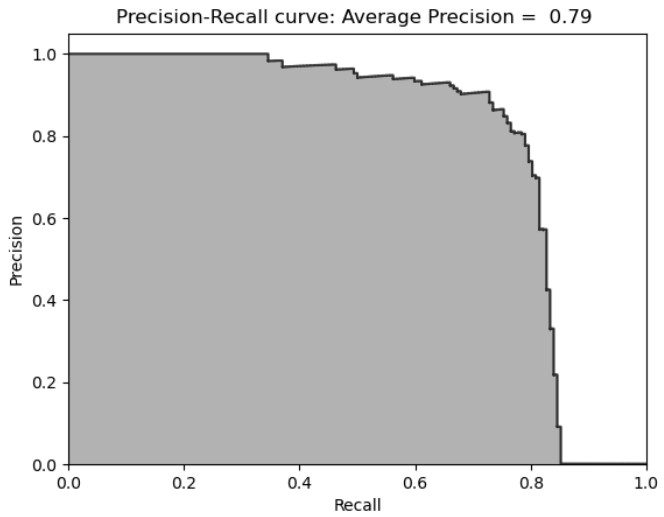
Case 3 (半教師あり学習)

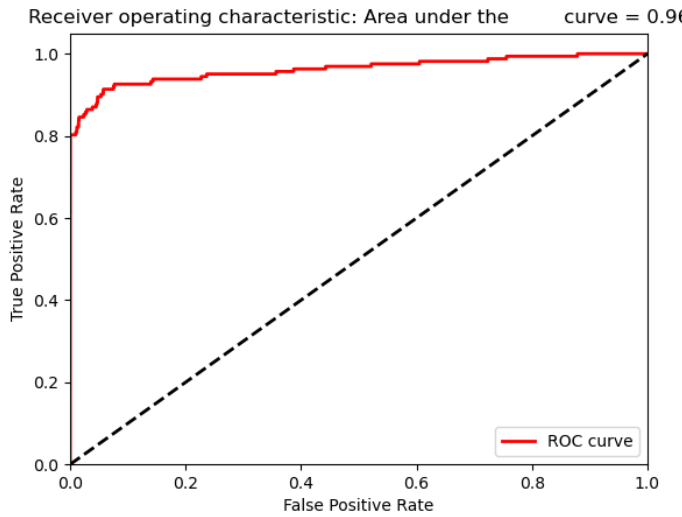
- 線形活性化関数を用いた2層過完備オートエンコーダで得た特徴量を元の訓練データにマージする.
- 29+40個の特徴量もつ訓練セットをLightGBMを使用して学習する.
- 上の一連の作業が半教師あり学習である.

Case 3 result

- 平均適合率は79%, auROCは96%であった.
- 75% の不正を検出をしようとすると, 適合率が0.8652になる.

test データに対する適合率-再現率曲線とROC曲線





Case 3 result

- Feature Importanceを見てみると, 加えられた特徴量も重要であることがわかる.

Feature Importance

V11	0.071236
V26	0.049848
4	0.044701
V15	0.040725
23	0.039089
…	…
19	0.001748
V28	0.001427
3	0.001219
30	0.000689
V10	0.000625

まとめ

- クレジットカードデータを使って，教師あり学習，教師なし学習，半教師あり学習 の比較
- 結果を表でまとめる.

アルゴリズム	平均適合率 (%)	auROC (%)	Precision at 75% recall(%)
教師あり学習	77	90	75
教師なし学習	71	96	82
半教師あり学習	79	96	87

- 微差ながら半教師あり学習が勝っている.
- (教科書は教師ありと教師なしが悪すぎて半教師あり学習が誇張されていた)