2020･07･27　ゼミ資料

Pythonではじめる教師なし学習
8章1節～5節

1116 17 9036
山口真哉

やること

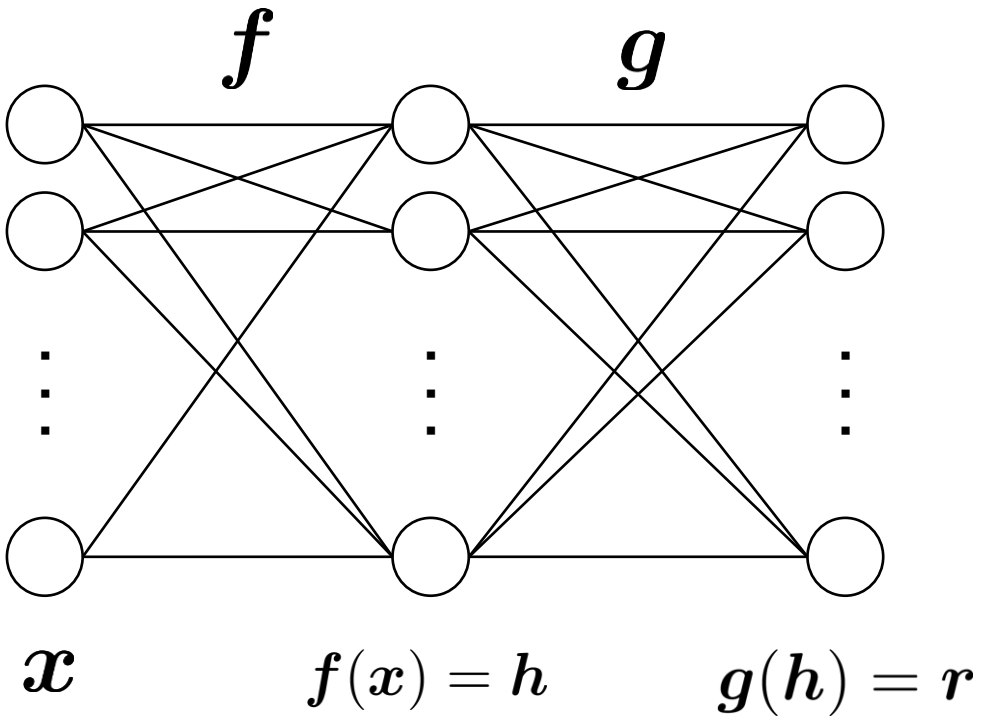
- オートエンコーダを構築する準備.
- KerasのSequentialモデルAPIを使って, オートエンコーダをいろいろ試してみる.

データの準備

- 2章, 4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.
- 教師ありだと平均適合率は82%でPCAでは69%であった.
- 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.

Case 1

- まずはじめに(実験的に)線形活性化関数を用いた2層完備オートエンコーダを実装する.
- バッチサイズを32, エポック数を10とする.
- パラメータ更新はAdamを使う.



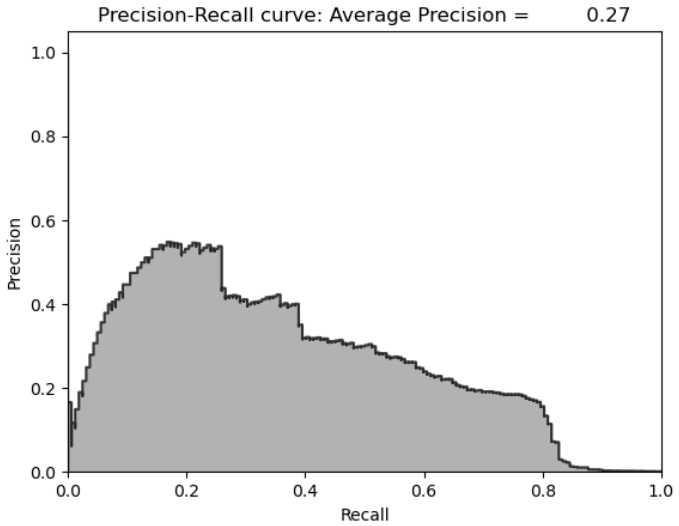
x, h, r の次元は共に 29

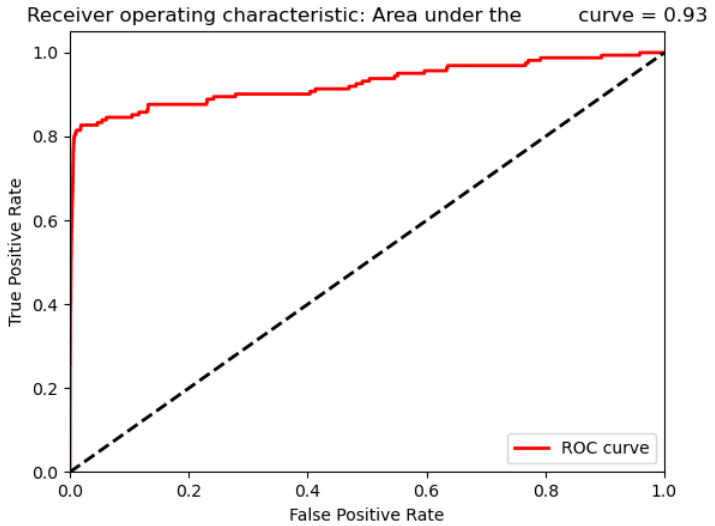
Case 1 result

- 同じ次元から同じ次元に写しているからあまり良い結果が得られない.
(入力をそのまま記憶するイメージ)

- 平均適合率は19%, 変動係数(相対的なばらつきの大きさ)は0.53であった.

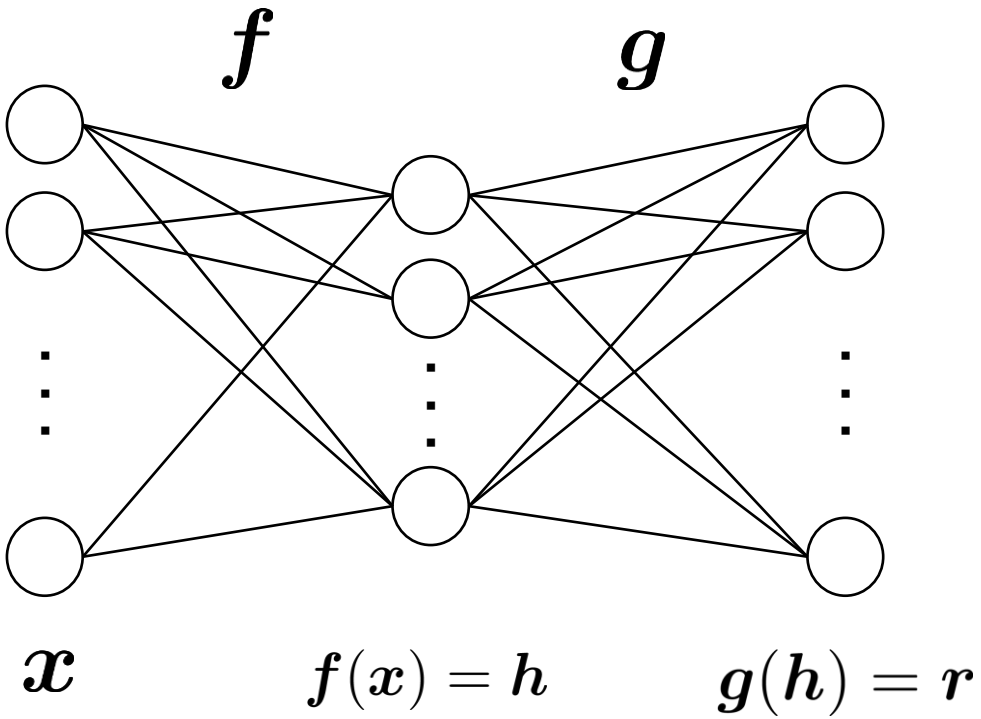
1回目の学習の時の適合率-再現率曲線とauROC





Case 2

- 線形活性化関数を用いた2層未完備オートエンコーダを実装する.
- バッチサイズを32, エポック数を10とする.
- パラメータ更新はAdamを使う.

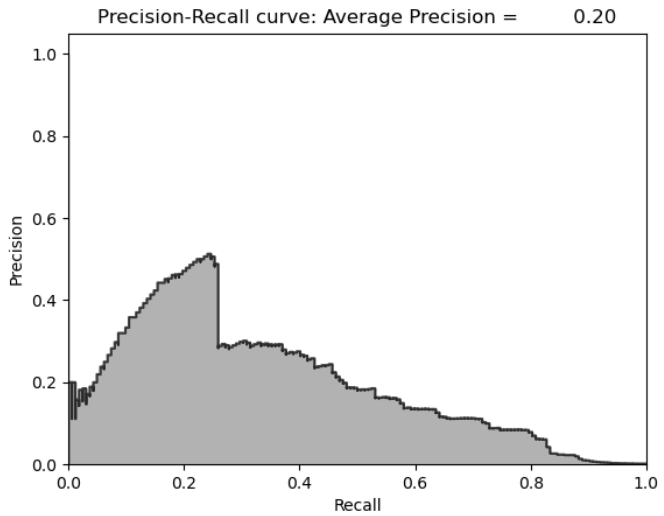


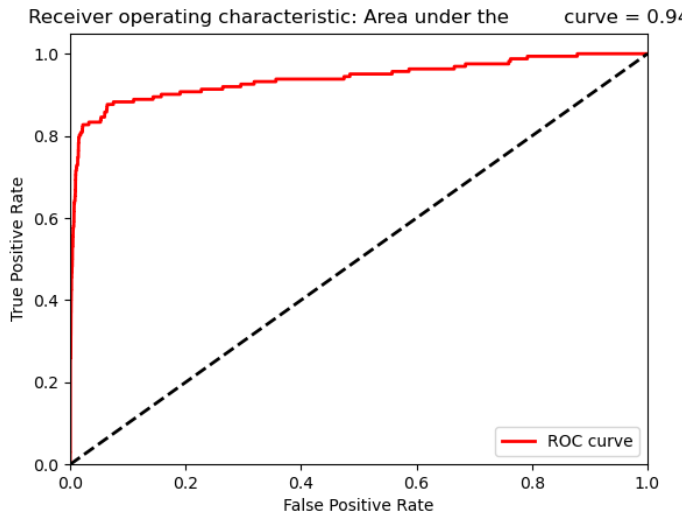
x, r の次元は 29, h の次元は 20

Case 2 result

- 完備のものに比べて要点だけ抑えるイメージ
- 平均適合率は30%, 変動係数は0.034であった.
  → 完備のものより平均適合率が高く, ばらつきも小さい(安定している)

1回目の学習の時の適合率-再現率曲線とauROC





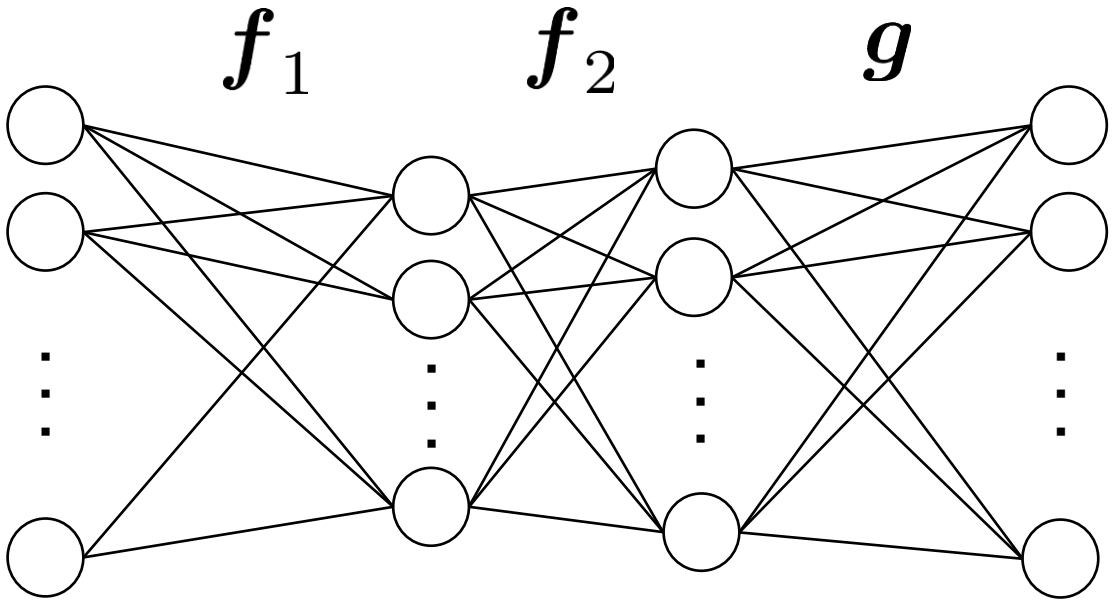
Case 2 補足

- 一般にハイパーパラメータの調整をする必要がある.

- 隠れ層の次元が $dim = 27$ の時, 教科書ではうまくいっているが,
  手元の環境(tensorflow 2.2)だとうまくいかなかった.

Case 3

- 線形活性化関数を用いた3層未完備オートエンコーダを実装する.
- バッチサイズを32, エポック数を10とする.
- パラメータ更新はAdamを使う.

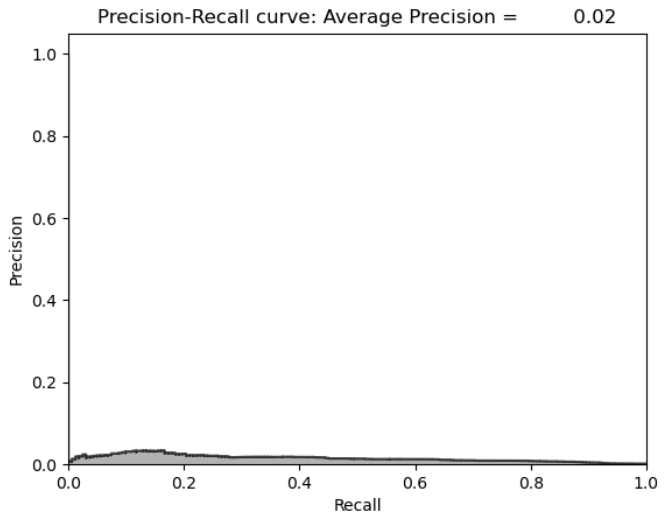


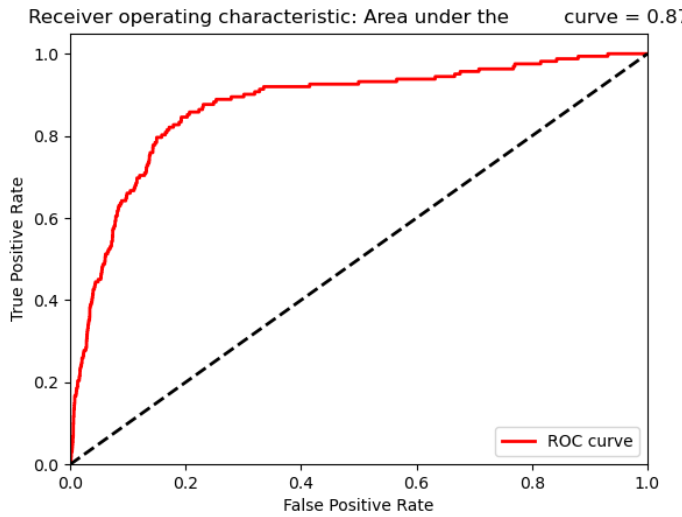
x, r の次元は 29,
h_1 の次元は 27,
h_2 の次元は 28.

Case 3 result

- 線形写像 φ, ψ に対し, $\phi \circ \psi$ も線形写像なのでやってることは無駄 ？
- 平均適合率は26%, 変動係数は1.2であった.
  → Case 2 に比べて変動係数が大きく結果が安定していない.

1回目の学習の時の適合率-再現率曲線とauROC





ここまでのまとめ

- 完備のものと未完備のもののオートエンコーダを見てきた.
- 結果を表でまとめる.

隠れ層の次元	平均適合率 (%)	変動係数
29	19	0.53
20	30	0.034
27 -> 28	26	1.2
(参考) 教師あり	82	N/A
(参考) PCA	69	N/A

- 隠れ層20のものが最善だった（ほかの次元も試す価値がある）
- PCAには勝てていない