応用データ解析　期末課題

担当教員：　米倉

課題１

1. 次の関数からN=1000点のデータをサンプリングし、そのデータを回帰するDNNモデルを作成せよ。ただしεは正規分布に従うノイズとする。

モデルは添付ファイル参照

1. 以下の項目をそれぞれ変えて回帰曲線がどう変わるかを比較せよ。
   1. ノイズεの大きさ

回帰曲線はノイズ大きくすると、上記関数から遠ざかり、小さくすると上記関数に近づく。

* 1. ノード数

回帰曲線はノード数を増やしても、それほど変わらないが、わずかに上記関数に近づき、減らすとわずかに上記関数から遠ざかる。

* 1. 層数

回帰曲線は層数を増やすと、荒くなり、層数を減らすとより直線っぽくなる。

* 1. エポック数

回帰曲線はエポック数減らすと、上記関数とからわずかに遠かったが滑らかになり、エポック数増やすと遠ざかるとともに荒くなった。

* 1. Dropout ratio

回帰曲線はDropout ratioを増やすとより上記関数から遠ざかる

* 1. サンプル数N

回帰曲線はサンプル数を増やすとより上記関数に近づきより滑らかになり、減らすと上記関数から遠ざかり荒くなる。

1. サンプリング点数をN=10点にして、回帰モデルを作成せよ。このとき②に挙がっているパラメータを変えることで過学習したモデルと過学習しないモデルを作成し、どのような場合に過学習するかを述べよ。

エポック数、ノード数、層数増やしすぎる、Dropout ratioが小さいと過学習する

課題２

　Boston housing dataを使っても良いし、各自のデータを使用しても良い。

Boston housing dataを使用する場合、あなたが不動産会社の人間だとして、商品（不動産）の賃貸価格を分析するためにMEDVの回帰モデルを作成せよ。各自のデータを使用する場合、回帰したい変数を１つないし複数設定して回帰モデルを作成せよ。

そのうえで、以下の二つの点について答えよ。ただしどの機械学習モデルを使用しても構わない（SVR, RF, DNN, その他）。

SVRを用いた

モデルは添付ファイル参照

1. 予測精度を適切に評価せよ。ただし学習データと評価データを適切に設定して、適切な交差検証を行なうこと。（DNNの場合は交差検証はしなくても構わない。）

結果は以下になった。

Best score on validation set: 0.8369685773153027

Best parameters: {'gamma': 0.1, 'C': 100, 'epsilon': 1, 'kernel': 'rbf'}

Test set score with best parameters: 0.8167188845982147

r2: 0.8167188845982147

rmse: 3.906532223844641

mae: 2.4905342136961295

1. 中間課題では変数の間で相関の大きいものを調べた。学習モデルを作成するときに、相関の大きい変数を取り除いた場合に予測精度がどのように変化するかを考察せよ。

RMとLSTATを除いた場合

交差検証を行おうとしたが数時間かけて何度試しても交差検証が終了しなかったため、①でのBest parametersをそのまま用いて比較する

test\_score: 0.5467896120481599

r2: 0.5467896120481599

rmse: 6.143027779801729

mae: 3.8826104075440346

上記のように、rmse, maeの値は上がり、r２の値は減少したため予測精度はさがった、今回は交差検証が終わらなかったため最良のパラメーターを利用できなかったが、相関が大きい変数を取り除いた場合、目的関数を表すことが難しくなることから、どう選択肢から選んだ最良パラメータにおいて行ったとしても予測精度はRMとLSTATも含まれているものに敵わないと考えられる。

注）使用するPCによっては稀にpytorchやscikit learnのインストールに手間取る場合があるようです。その場合は米倉までメールで質問してください。

Eメール：　yonekura@struct.t.u-tokyo.ac.jp

以上