統計的機械学習 レポート課題

所属：総合文化研究科広域科学専攻相関基礎科学系
学籍番号：31-217907　氏名：崎下雄稀

2021 年 7 月 28 日

1　概要

本レポートでは，オープンソース機械学習 scikit-learn で提供されているデータセット「California Housing dataset」を用いて住宅価格を予想するタスクを Kernel Ridge 法および Random Forest 法を用いて行い，ハイパーパラメータやデータ量による性能の変化を評価した．

2　手法

「California Housing dataset」を用いて Kernel Ridge 法および Random Forest 法で予測を行い，K-fold 法で性能の評価を行った．学習，性能評価は Python ライブラリ scikit-learn (©2007 - 2020, scikit-learn developers (BSD License)) を用いて行った．

2.1　Data Set

本レポートで使用するデータセットは，住宅価格を含む 9 つの経済的指標が地区ごとに集計されたものである (R. K. Pace and R. Barry (1997))．各変数を表 1 に示す．

表 1　California Housing dataset の内容

目的変数	住宅価格の中央値
特徴量	• 収入の中央値
	• 住宅の築年数の中央値
	• 住宅の部屋数の平均値
	• 住宅の寝室の数の平均値
	• 地区の人口
	• 住宅の占有率
	• 地区の経度
	• 地区の緯度

変数は全て実数，サンプル数は 20640 で，欠損値は含まれない．本データセットは，トイデータセットとして提供されている「Boston house prices dataset」などと比べサンプル数が多く，実世界データセットとして提供されている．

2.2 Kernel Ridge

Kernel Ridge 法は Ridge 回帰をカーネル空間で行うものである．

本実験では RBF カーネルを使用する．ハイパーパラメータは正則化係数 `alpha` と，カーネル変数 `gamma` である．`alpha` は大きいほど正則化が強まるため，経験誤差は大きくなり汎化誤差との差は小さくなることが期待される． `gamma` はカーネル関数であるガウシアン RBF 関数の幅に影響する．大きいほど表現力が上がるため，一定の`alpha` のもとでは相対的に正則化が弱まり過学習の傾向が強まることが予想される．学習は入力データを正規化して行う．

2.3 Random Forest

Random Forest 法はバイアスが小さくバリアンスが大きい決定木学習をアンサンブル学習によりバリアンスの減少を図る手法である．

本実験ではハイパーパラメータとして `n_estimator` と `max_features` を設定する．`n_estimator` はアンサンブルを行う木の数であり，多いほど性能が上がることが期待される．`max_features` は各決定木で用いる特徴量の最大数である．特徴量の数の制限は正則化と捉えることができ，小さいほど経験誤差は大きくなり汎化誤差との差は小さくなることが期待される．Kernel Ridge 法と異なり，パラメータ間の大きさの違いは学習に影響を与えないため正規化は行わない．

2.4 性能の評価

予測モデルの性能の評価は K-fold 法を用いて行った．すなわち，データセットを K 個に分割してそのうち 1 つを検証用データに，それ以外を学習データとして学習，テストを行いうことを K 通り行い，それぞれの結果の平均値を見る方法である．

性能の評価は目的変数と予測値の R^2 値で行った．以下学習データにおける R^2 値を学習性能，検証用データにおける R^2 値を検証性能と呼ぶ．

計算時間の都合により，Kernel Ridge 法では $K = 3$ ，Random Forest 法では $K = 5$ とした．また，学習サンプル数を変化させる場合は各分割のの後に一部を抽出した．

3 結果

3.1 Kernel Ridge

図 1 は Kernel Ridge 法でサンプル数を 13 から 13760 まで変化させたときの学習性能・検証性能の推移を表したものである．これは全データから検証用データを除いた 2/3 の 0.1% から 100% まで変えたものである．

サンプル数が 5000 以下と少ない領域では極端にスコアの悪いものが含まれるためエラー領域が大きく上下しているが，概ねサンプル数を増やすと学習性能が下がり検証性能は上昇するという傾向が見られる．

図 2 は `alpha` を 10^{-4} から 1 まで変化させたときの性能の推移である．`alpha` が大きくなるにつれ正則化が強まり，学習性能が低下する反面経験誤差と汎化誤差の差が小さくなり 10^{-1} 程度までは検証性能が上昇していることがわかる．

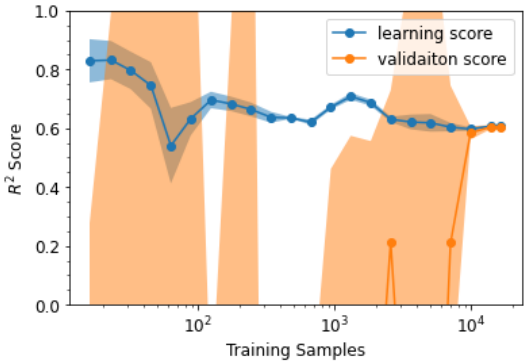

図 1　学習曲線（Kernel Ridge）.

図 3 は gamma を 10^{-3} から 10 まで変化させたときの性能の推移である． gamma が大きくなるほど学習性能が 1 に近づく一方検証性能は gamma $= 0.5$ 程度を境に減少し始め，過学習していることがわかる．

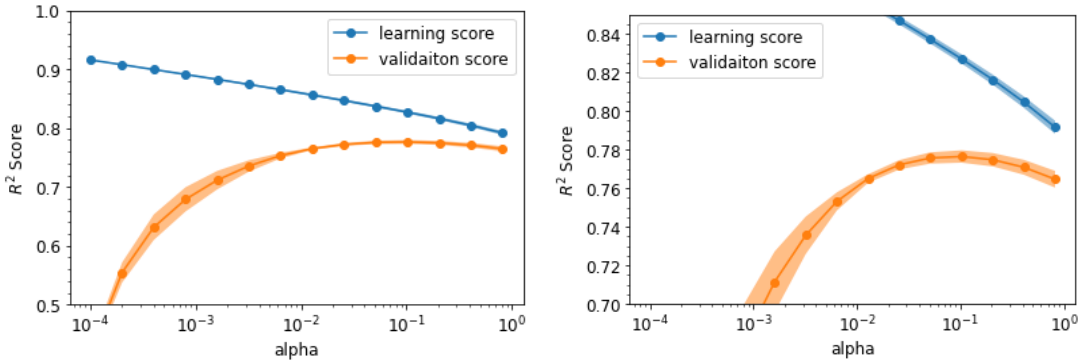

図 2　alphaの検証曲線（Kernel Ridge）．右は R^2score が 0.7–0.85 の拡大図.

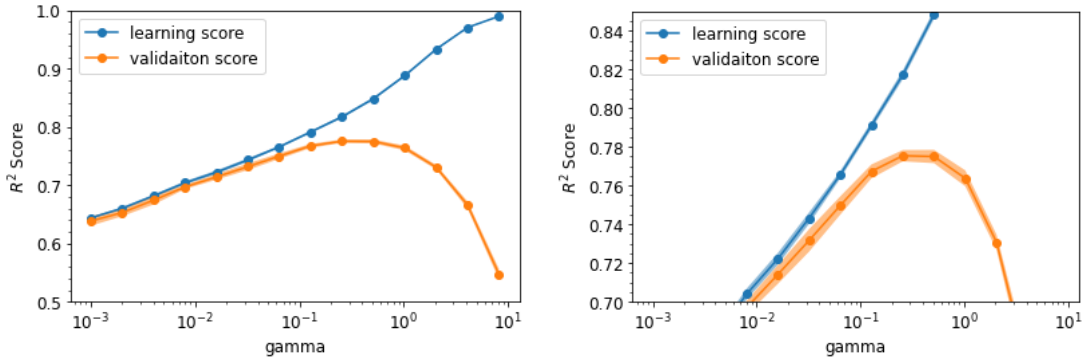

図 3　gammaの検証曲線（Kernel Ridge）．右は R^2score が 0.7–0.85 の拡大図.

3.2 Random Forest

図 4 は Random Forest 法でサンプル数を 16 から 16512 まで変化させたときの性能の推移を表したものである．概ねサンプル数を増やすと学習性能が下がり検証性能は上昇するという傾向が見られる．

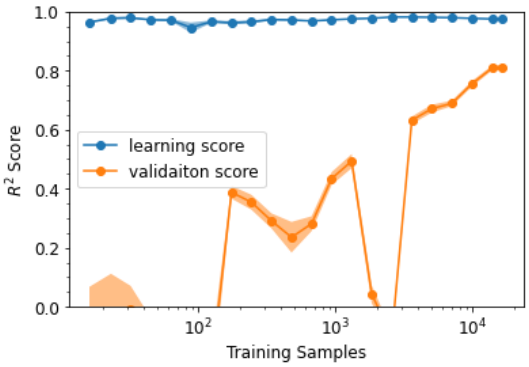

図 4　学習曲線（Random Forest）．

図 5 は `n_estimator` を 10 から 1000 まで変化させた時の性能の推移である．`n_estimator` を大きくすると学習性能・検証性能ともに上昇するが `n_estimator`＝ 100 程度で飽和しているという結果が得られた．

図 6 は `max_features` を 1 から 8 まで変化させた時の性能の推移である．`max_features`＝ 2 でピークを持つ形となった．これは scikit-learn でのデフォルト値が特徴量数の平方根に設定されることと整合的である．

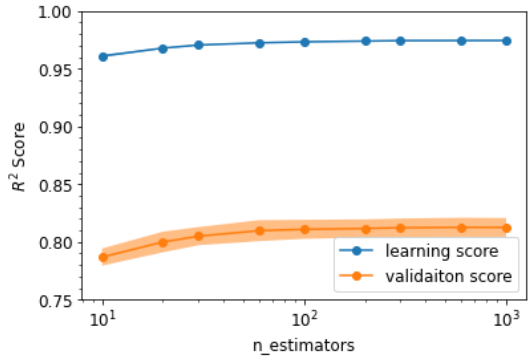

図 5　`n_estimator`の検証曲線（Random Forest）．

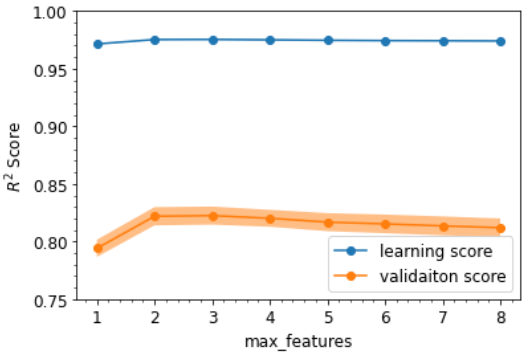

図 6　`max_features`の検証曲線（Random Forest）．

4　まとめ

2 つの手法を用いて性能の検証を行ったが，事前の予想と概ね一致する傾向が得られた．今回の結果では Random Forest 法は Kernel Ridge 法に比べハイパーパラメータによる性能の変化が小さく，また全体的にも性能が高かった．これは特徴量から目的変数への写像 が Kernel Ridge 法で仮定されるようななめらかな連続関数となっておらず，十分にフィッティングできなかった一方，決定木をベースとしている Random Forest 法では不連続的な変化も学習できるため対応できたのではないかと考えられる．