scikit-learn を用いた機械学習演習

離散数理分野

2022 年 6 月

1　概要

化学グラフ推定プロジェクトにおける機械学習実験の流れを体験してもらう．

ここでは**回帰問題**を取り扱う．回帰問題では**訓練集合** (training set) $D_{\text{train}} = \{(x_1, a_1), \ldots, (x_m, a_m)\} \subseteq \mathbb{R}^K \times \mathbb{R}$ ($x_i \in \mathbb{R}^K$, $a_i \in \mathbb{R}$, $i = 1, 2, \ldots, m$, K は特徴ベクトルの次元数) が与えられ，これを手がかりに，未知の特徴ベクトル $x \in \mathbb{R}^k$ の物性値 $a \in \mathbb{R}$ を予測する関数 $f : \mathbb{R}^K \to \mathbb{R}$ を訓練 (学習) することが求められる．学習された f の質 (= 学習アルゴリズムの質) は，**試験集合** (test set) $D_{\text{test}} \subseteq \mathbb{R}^K \times \mathbb{R}$ に対する**決定係数** (coefficient of determination; R^2) などを用いて評価される．実際には与えられたデータ $D \subseteq \mathbb{R}^K \times \mathbb{R}$ を D_{train} と D_{test} に分割して評価を行うのが一般的である．この分割に基づいた評価を行うための枠組みとして**交叉検定** (cross validation; CV) が知られる．整数値 $k \geq 2$ に対する k-交叉検定 (k-fold cross validation; k-fold CV) の手順をまとめておく．

k-交叉検定

1. 与えられたデータ D を $S_1, S_2, \ldots, S_k$ に (おおよそ) 等分割する．

2. $t = 1, 2, \ldots, k$ について以下を行う．

2-1. $D_{\text{train}} := D \setminus S_t$, $D_{\text{test}} := S_t$ とする．

2-2. D_{train} から予測関数 $f : \mathbb{R}^K \to \mathbb{R}$ を学習する．

2-3. 予測関数 f の D_{test} に対する決定係数 r_t を求める．

3. $r_1, r_2, \ldots, r_k$ の中央値をもって学習アルゴリズムの評価値とする．学習アルゴリズムの

当研究室では通常 $k = 5$ を用いる．また $T \geq 2$ 回の k-fold CV を行い (T 回を通じてデータの分割を変える)，kT 個の決定係数の中央値を用いて評価している．

2　学習実験の手順

学習実験は**予備実験**と**評価実験**の大きく二つに分けられる．

予備実験． ハイパーパラメータ[1]をチューニングするための実験である．一般に，学習モデル[2]は多くのハイパーパラメータを持つ．プロジェクトでは様々な物性を取り扱っており，**それぞれの物性に対して適切なハイパーパラメータの値を定める**必要がある．

与えられた学習モデルに対し, ハイパーパラメータの取りうる値 (もしくはベクトル) の集合を Θ とする．ハイパーパラメータの値 $\theta \in \Theta$ に対して学習された予測関数を $f_\theta : \mathbb{R}^K \to \mathbb{R}$ とする．

一般的な予備実験の手順は以下のとおりである．

1. 試行するハイパーパラメータの集合を $\tilde{\Theta} \subseteq \Theta$ を定める．

2. 各 $\theta \in \tilde{\Theta}$ を評価する．評価の手段として 5-fold CV を複数回繰り返すとよい．

3. **2** において最良の評価を達成した θ を, $\theta^* := \theta$ として選択．

評価実験． 評価実験は論文などに報告する数値を記録するための実験である．その手順は以下のとおりである．予備実験で定められたパラメータ θ^* を 10 回の 5-fold CV で評価. 10 回の CV では, 予備実験で用いたデータの分割を用いてはならない．

3　演習課題

従来の学習実験であまり性能の出なかった (R^2 が 0.8 を超えていない) 物性について，これまで試してこなかった学習モデルを試し，あわよくば従来のベスト値を超えてもらいたい．

物性．

物性	従来の最良 $\mathbf{R}^2$	最良値を達成した学習モデル
BHL (3 元素)	0.60	ANN
Fp (3 元素)	0.74	ANN
Vp (3 元素)	0.75	決定木

各物性は, 特徴ベクトルの記述された csv ファイルと，物性値の記述された txt ファイルの二つから成る．

[1]学習に先立って値を決めておかなければならないパラメータ．たとえばニューラルネットワークでは，ネットワークの構造や素子における活性化関数はハイパーパラメータである．

[2]線形回帰, 決定木, 人工ニューラルネットワーク (Artificial Neutal Network; ANN) など．

学習モデル.

1. 多次特徴を用いた線形回帰. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html
2. ランダムフォレスト. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
3. k-近傍法[3]. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html
4. サポートベクター回帰. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

このうち 1 は現在プロジェクトで検討中である．2 はプロジェクトで取扱可能だが, まだ研究に着手していない. 3 はプロジェクトで取扱可能かもしれない. 4 はプロジェクトで取扱困難である (逆問題の定式化が難しいから).

サンプルプログラム. Lasso 線形回帰を用いた学習実験用のサンプルスクリプトを渡す．詳細は別途説明する.

[3] この k は交叉検定のパラメータとは異なる.