Python で始める機械学習

6 アルゴリズムチェーンとパイプライン

6.1 前処理を行う際のパラメータ選択

15T4032G 芝山直希

2018 年 6 月 29 日

アルゴリズムチェーン？

機械学習を活用するアプリケーションは、様々な処理と複数の機械学習アルゴリズムを連鎖的に実行する必要がある

- この連鎖的構造をアルゴリズムチェーンという
- scikit-learn では Pipeline クラスを使用することで簡単に構築できる
- Pipeline では任意個数の処理とモデル 1 つを連結し、使用法を変えることなく連鎖的に実行できる
- この章では Pipeline と、GridSearchCV を用いたパイプライン内の処理の最適化について触れていく

活用例：前処理を行う際のパラメータ選択

スケール変換をしたデータセットに対し、交差検証 (GridSearchCV) を用いて SVC の最適なパラメータを選びたい

- スケール変換、SVC の交差検証の順にプログラムに書いてしまうと訓練データセットをスケール変換してから分割するため、検証用データの情報が一部漏洩してしまう
  - 適切な選択ができない、楽観的すぎる結果が出る等の問題が起こる可能性がある
- 前処理の前に交差検証用の分割を行うことで解決できる (Pipeline を用いて実現可能)