

MALSS: 機械学習支援ツール

MALSS: Machine Learning Support System

日立製作所

鴨志田 亮太

背景と課題

背景: データサイエンティストの需要が高まり、人材不足のため未習熟者が機械学習によるデータ分析に従事
課題: 経験・知識不足から不十分な分析となり、想定した成果を得られない

手法・ツールの開発による解決

アプローチ: 自動化による分析支援、分析レポートによる知識習得支援を行うツールを提案
開発: 機械学習支援ツールMALSSをオープンソースPythonライブラリ*として開発
 *<http://pypi.python.org/pypi/malss/>

データ分析の流れとMALSSの機能

データの前処理

データの種類に応じた前処理

- ・欠損値補間
- ・シャッフル
- ・ダミー変数化 (カテゴリ変数)
- ・標準化

アルゴリズム選択

データの
 ・サンプル数
 ・次元数
 ・分析タスク (回帰/分類)
 に応じて適切なアルゴリズムの候補を複数選択

分析/評価

・パラメータチューニング (グリッドサーチ)
 ・汎化性能評価 (交差検証) により
 予測モデルを自動作成

考察

・分析結果
 ・手順/用語解説
 ・分析指針
 を記載した分析レポートを作成
 分析、および知識習得を支援

使用例

MALSSを利用して回帰分析を行う例

```
from malss import MALSS

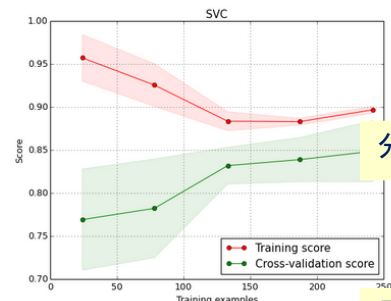
# タスクの設定
clf = MALSS('regression')

# 分析
# repdir: レポート出力先
clf.fit(X, y, 'repdir')

# 未知データの予測
pred = clf.predict(X_test)
```

分析レポート(一部)

学習曲線 (Learning curve)



学習曲線 (Learning curve)

- ・学習曲線はデータサイズを変えた時の訓練データでのスコア、交差検証スコア
- ・学習曲線が以下のような場合、モデルはハイバリエンス(オーバーフィッティング)
 - ・学習データが増加に伴って交差検証のスコアの改善が飽和しない
 - ・訓練データのスコアと交差検証のスコアの差が大きいの
- ・学習曲線が以下のような場合、モデルはハイバイアス(アンダーフィッティング)
 - ・訓練データのスコアでもスコアが低い(誤差が大きい)
 - ・訓練データのスコアと交差検証のスコアの差が小さい

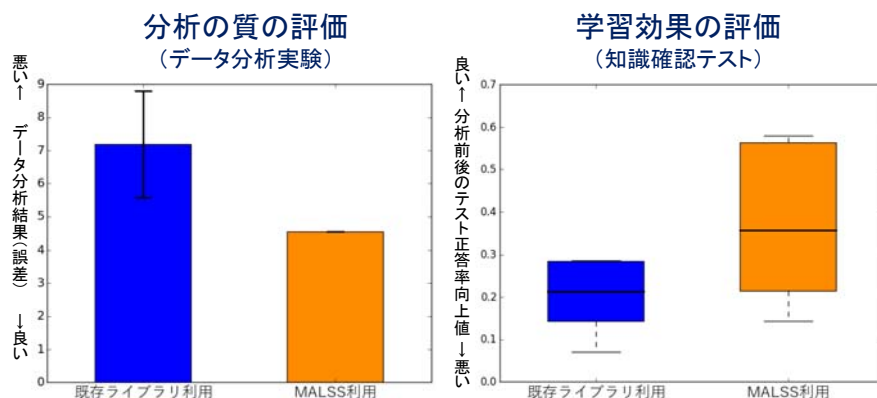
ハイバリエンス(High variance)への対策:

- ・特徴量選択や次元削減により特徴量の数を減らす。
- ・データ量を増やす。

分析指針

評価

既存ライブラリ+参考資料を利用して分析した場合と比較



MALSSを利用することで、必要な知識を身に付けながら質の高い分析を行うことが可能