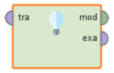
# モデル作成・検証 Cheat Sheet

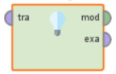
## 代表的なアルゴリズム

**k-NN**




学習データをプロットしておき、新しいデータに対し、そこから距離が近い順に任意のK個を取得し、多数決でデータが属するクラスを推定するクラス判別のアルゴリズムです。

**Naive Bayes**




ベイズの定理をもとにデータが与えられたときの全ての推定の確率を計算し、最も確率の高いものを推定結果として出力します。

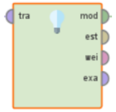
**Decision Tree**

ツリーによって分類を行う手法です。可読性が高く、外れ値にも強い手法ですが、過学習を起こしやすいため、パラメータ調整や枝の刈り込みが必要となります。

**Random Forest**

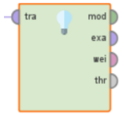
アンサンブル学習をベースに、多数の異なる決定木(Decision Tree)を集めて分析を行う手法です。
決定木で起きやすい過学習を避けることが出来ます。

**SVM**



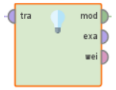

クラスを分類する境界線を引き、データ判別を行う手法です。少ないデータで結果を出すことが出来ます。カーネルトリックにより、非線形データにも対応できます。

**Logistic Regression**



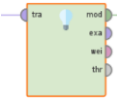

回帰手法を応用して二項分類を行うロジスティック回帰のオペレータです。
2値カテゴリ(Binominal)がlabelとして設定されている必要があります。

**Linear Regression**




線形回帰を行うオペレータです。
数値項目が目的変数(Label)として設定されている必要があります。

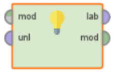
**Generalized Linear Model**




従来の線形回帰を拡張した一般化線形モデルを使用するオペレータです。
正規分布ではないデータに対しても有効に利用できます。

## モデル適用

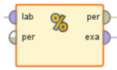
**Apply Model**




unlポートに接続されたラベル無しデータに対し、Modポートに接続されているモデルを適用するオペレータです。

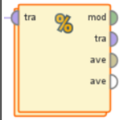
## 検証

**Performance**



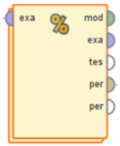

精度検証に用いるオペレータです。
汎用できるものから、回帰や分類に特化したものもありますので、用途に合わせて使い分けをしてください。

**Validation**




分割検証(Split Validation)に使用するオペレータです。
入力されたデータセットを指定した割合で学習データとテストデータに分割し、1回のみ検証を実施します。
内部プロセスの左側で学習モデルの作成、右側で検証を行います。

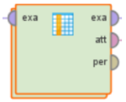
**Cross Validation**




交差検証を行う際に使用します。データセットを指定したk個に分割し、一つをテストデータ、残りを学習データとして検証を行い、組み合わせを変えてk回反復します。
モデルに過剰適合が起きていないかの確認に用います。

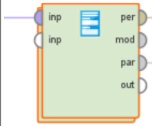
## 精度向上

**Forward Selection**




説明変数が多いとき、説明変数の組み合わせを変えながら、最も高い正答率となる組み合わせを探すときに使用するオペレータです。

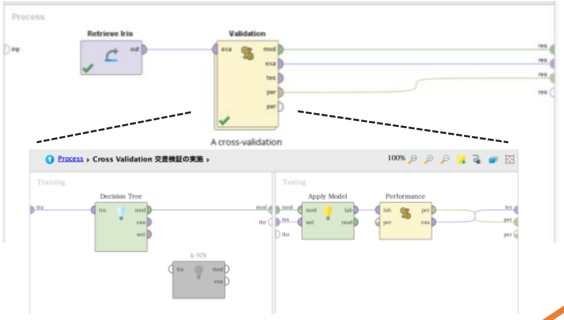
**Optimize Parameters (Grid)**



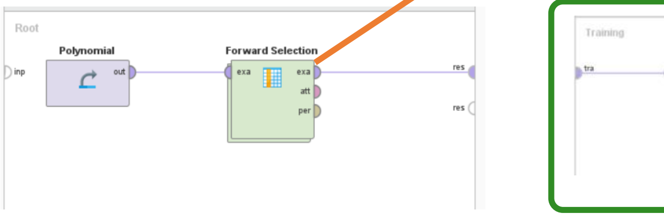

オペレータのハイパーパラメータの最適化をします。
選択された範囲内から最適なパラメータの組み合わせをグリッドサーチによって探索します。

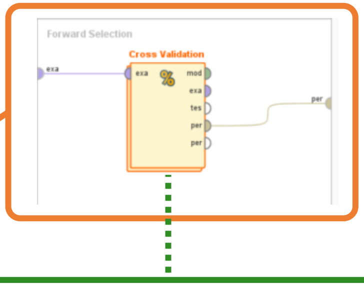
## サンプル

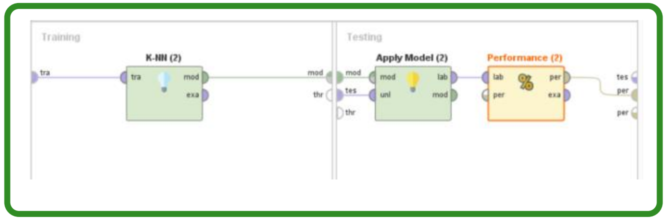
■ Cross Validation使用の一例




■ Forword Selection使用の一例




© 2021 KSK Analytics Inc. All Rights Reserved.