機械学習による株価上昇下落予測モデルの構築

インディ・パ株式会社

2017 年 4 月

はじめに

株価上昇下落予測モデルを構築することを考えます。正解率の高い株価上昇下落予測モデルを構築することが出来れば、直ちに投資やトレードで良い結果が得られるというわけではありませんが、予測モデルから得られる知見は、投資やトレードに大変有益であると考えられます。今回は機械学習の手法を使用して株価上昇下落予測モデルを構築してみます。機械学習の手法として、今回は、特に決定木と k 近傍法を考えます。決定木と k 近傍法の構築には python の scikit-learn を利用してみます。

決定木による株価上昇下落予測モデルの構築

決定木 (decision tree) は、データを分類する質問をノードとし、分類結果をリーフとする木構造で概念を表現するものです。決定木識別器は、一連の質問に対する答えに基づいて決断を下すという方法により、データを識別するモデルであると考えることができます。

決定木モデルでは、学習データセットの特徴量に基づいて一連の質問を学習し、サンプルのクラスラベルを推測します。質問は、学習データセットの分割であると考えることが出来ます。学習データセットの分割は以下のように行われます。決定木のルートから始めて、情報利得 (information gain) が最大となる特徴量でデータを分割します。木のリーフが純粋になる (分割されたデータのばらつきの減少がなくなる) まで、この分割を子ノードごとに繰り返すことができます。リーフが純粋になるというには、各リーフのサンプルがすべて同じクラスに属するということを意味します。実際には、リーフが純粋になるまで分割を繰り返すと多くのノードを持つ非常に深い決定木になることがあり、過学習に陥りやすくなるため、通常は決定木の最大の深さに制限を設けます。

python の scikit-learn の決定木には、回帰を行うものと、識別を行うものがありますが、今回は識別を行うものを使い識別を行ってみます。[当日の終値 – 当日の始値] が 0 以上であるものを正例、負であるものを負例と考えて、正例、負例を予測するモデルを決定木識別器により構築してみます。分割条件で

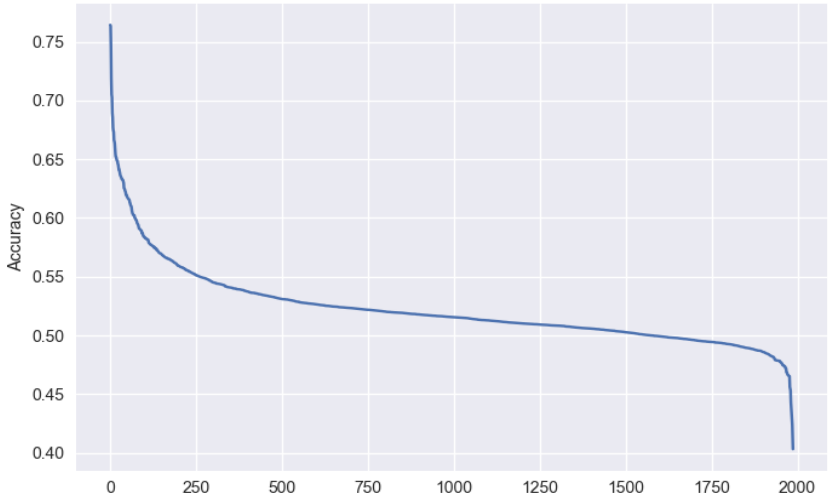

図 1: 正解率 (決定木)

よく使用される不純度の指標または分割条件は、エントロピー (entropy)、ジニ不純度 (Gini impurity)、分類誤差 (classification error) 等がありますが、今回はジニ不純度を使用してみます。

表 1: 識別予測結果 (決定木)

銘柄	期間	Accuracy	Precision	Recall
A	1993 年 2 月 25 日–2017 年 2 月 10 日	0.764	0.772	0.983
B	1992 年 1 月 7 日–2017 年 2 月 10 日	0.610	0.624	0.922
C	1992 年 1 月 7 日–2017 年 2 月 10 日	0.475	0.490	0.606

10-分割交差検証を行い、テストデータの正例、負例の予測を行ってみた結果を表 1 に示します。学習データとしては対数収益を与え、ツリーの最大深さは 4 に設定し、決定木を構築しました。表 1 の期間は学習・テストデータの期間を表し、Accuracy、Precision、Recall はテストデータにおける正解率、適合率、再現率を表しています。1987 銘柄の日本株に適用してみましたが、表 1 には正解率が非常に高いもの、正解率が高いもの、正解率が低いものの 3 銘柄のみを選んで示しています。図 1 には調べた 1987 銘柄に関して、正解率を縦軸に示し、降順に並べたものを示しています。

k 近傍法による株価上昇下落予測モデルの構築

k 近傍法 (k-nearest neighbor) は、以下のようにして識別、回帰モデルを構築します。

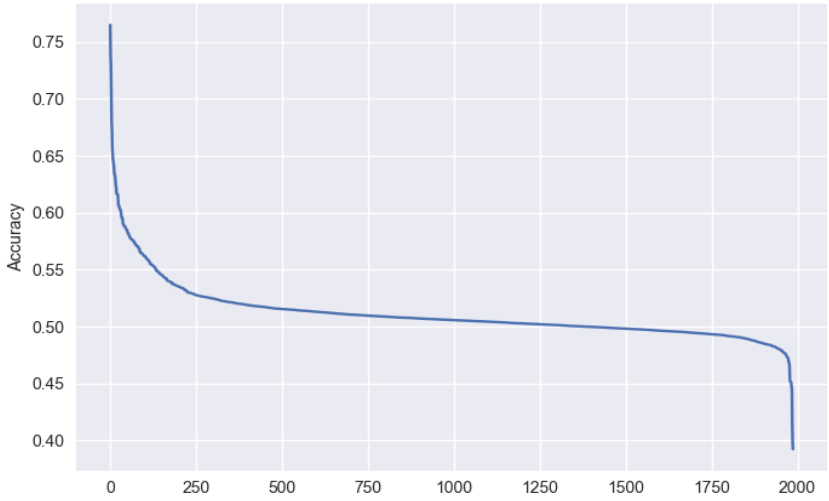

図 2: 正解率 (k 近傍法)

1. k の値を決め、距離尺度を定義する。

2. 定義した距離尺度に基づいてサンプルから k 個の最近傍データ点を見つけ出す。

3. k 個のデータ点の加重平均、多数決等により、クラスラベル、数値を割当てる。

python の scikit-learn の k 近傍法には、回帰を行うものと、識別を行うものがありますが、今回は回帰を行うものを使ってみます。決定木の場合と同様に、[当日の終値 − 当日の始値] が 0 以上であるものを正例、負であるものを負例と考えて、正例、負例を予測するモデルを構築してみます。k 近傍法の設定は、$k = 5$, 距離をマンハッタン距離、加重平均を一様加重平均としてみます。

表 2: 識別予測結果 (k 近傍法)

銘柄	期間	Accuracy	Precision	Recall
A	1993 年 2 月 25 日–2017 年 2 月 10 日	0.738	0.777	0.924
B	1992 年 1 月 7 日–2017 年 2 月 10 日	0.580	0.626	0.783
C	1992 年 1 月 7 日–2017 年 2 月 10 日	0.509	0.520	0.524

10-分割交差検証を行い、テストデータの正例、負例の予測を行ってみた結果を表 1 に示します。学習データとしては対数収益を与えました。表 2 には、比較のために表 1 と同じ銘柄に関して、正解率、適合率、再現率を示しています。図 2 には調べた 1987 銘柄に関して、正解率を縦軸に示し、降順に並べたものを示しています。

まとめ

決定木、k 近傍法により株価上昇下落を予測するモデルを構築してみました。日本株 1987 銘柄に関して、10-分割交差検証により、正解率、適合率、再現率を計算してみました。高い正解率を示す銘柄もそれなりの割合で存在しますが、それらの銘柄のモデルの予測能力に関しては更に調査が必要かもしれません。決定木、k 近傍法ともに、予測の性能の全体的な傾向は似通っていると言えます。