Adapting support vector machine methods for horserace odds prediction

清水　豪士

富山県立大学　情報基盤工学講座
t715038@st.pu-toyama.ac.jp

August 5, 2021

はじめに

背景

- SVM は入力の次元が観測値の数に比べて大きい場合，一般化可能なモデルを生成でき，入力変数の数が多い競馬の分野では適切な分析手法である．
- 金融市場と同様に，競馬市場やその他のベッティングマーケットの結果を予測する問題に大きな注目が集まっている．

目的

- オーストラリアのレースデータのサンプルを適用し，サンプル外のレースでテストする．

競馬のオッズ予測

金融市場で一般的に使用されているさまざま「ファンダメンタル」変数や指標と同様に，競馬の世界でも検討可能な変数は豊富にある．

競馬で扱う変数候補

- 繁殖，年齢，調教，レースの履歴，騎手や調教師の履歴
- 天候，馬場，枠順などの補助的な変数

現在のレースと前のレースに関するいくつかの変数と，最も重要なオッズマーケットの情報を含めることで，馬の勝利の可能性に関する信頼性の高い予測モデルを構築することができる．

予測からベッティングへ

「ウィニングネス」指数

- 馬の着順を線形関数として回帰し，変数から算出されるもの．
- 結果変数を勝ち馬の場合は１，それ以外の場合は０として，レースで層別されたマルチノミアル・ロジット回帰の入力として使用される．

- ウィニングネス・インデックスの予測が可能であれば，残りのステップは確率正則化とステーキングの２つの部分に分けられる．
- 本研究では，SVM を適用してウィニングネス・インデックスを生成する．

ウィニングネス指数

ウィニングネス指数を実際に適用できるベットに変換するための方法論

u_{ij} はレース i の馬 j がレースに勝ったら 1，勝っていないと 0 を示す 2 値変数を表し，i はサンプルで分析されたレースの数，j はレース i の馬の数を表している．$\hat{y}_{ij}$ は採用されているモデルによって生成されたウィニングネス指数の予測を示す．

「正規化」パラメータ α の関数として，以下の尤度が最大化される．

$$L(\alpha) = \prod_{i=1}^{N_{Races}} \frac{\prod_{j=1}^{n_i} \exp(u_{ij}\hat{y}_{ij}\alpha)}{\sum_{j'=1}^{n_i} \exp(\hat{y}_{ij'}\alpha)}$$

したがって，確率予測 $\hat{p}_{ij}$ は以下のように表せる．

$$\hat{p}_{ij} = \frac{\exp(\hat{y}_{ij}\hat{\alpha})}{\sum_{j'=1}^{n_i} \exp(\hat{y}_{ij'}\hat{\alpha})},$$

ここで，$\hat{\alpha}$ は前回の最尤法の結果

ウィニングネス指数

ウィニングネス指数を実際に適用できるベットに変換するための方法論

このステップが実行されたと仮定して「ケリー基準」を適用する.
ここでは，各レース i について，なベットのセット $b_{i1}, ..., b_{in_i}; b_{ij} \geq 0$ を解く.

$$\sum_{j=1}^{n_i} \hat{p}_{ij} \log \left\{ 1 + b_{ij} \times t_{ij} - \left(\sum_{j=1}^{n_i} b_{ij} \right) \right\}$$

上記の式が最大となり，t_{ij} は馬 j が勝った場合のグロスリターンを表している.
この最適化は制約付非線形最適化のための多くのルーチンのいずれかによって実行することができる.

サポートベクターマシーン

サポートベクターマシーン

- サポートベクターマシンは，入力変数の次元数が高くてもうまく機能することが知られているので，フィーチャーマップの使用が可能である．
- 入力ベクトルの特徴マップとは，入力を別の（実際にはもっと大きな）「特徴」空間にベクトル値で変換したもので，通常はサポートベクターマシンの手法で次元を大幅に削減する前に行われます．

$$x \longrightarrow h(x)$$

サポートベクターマシーン

サポートベクターマシーン

- ガウスカーネルを用いた放射状基底関数特徴マップを選択し，様々なデータポイントを中心とし，「leave-one-out」クロスバリデーション基準に従ってバンド幅を選択する．

$$K(x; x_j) = \exp\left(-\frac{1}{2\gamma^2}\|x - x_j\|^2\right)$$

- もしカーネルの中心が実際に入力変数のセットそのものであるとすると，この特徴マップは，各入力観測値について，サンプル内の他のすべての変数に対する「特徴の近さ」インデックスのようなものになる．
- このバージョンのサポート・ベクトル・マシン法を，ここでは競馬のコンテキストで採用する．

サポートベクターマシーン

サポートベクターマシーン

勝敗結果の結果変数は，形式的に次のように定義される．

$$y_{ij} = \frac{fp_{ij}}{n_i + 1} - .5, \quad j = 1, 2, \ldots, n_i$$

ここで，fp_{ij} は i^{th} の順位を表し，i^{th} レースには n_i の出走馬がいる．
ここでノーマルスコアによる順位を採用する．

$$y_{ij} = \Phi^{-1}\left(\frac{fp_{ij}}{n_i + 1}\right),$$

Φ^{-1} はガウス分布の逆関数を表す．

$$\int_{-\infty}^{\Phi^{-1}(\alpha)} \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} dz \equiv \alpha$$

サポートベクターマシーン

サポートベクターマシーン

線形の場合は係数ベクトル w を最小化するように選択する

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij}-(w\cdot x_{ij})-b_i)^2,$$

サポートベクターアプローチは，まず ε に依存しない全絶対誤差を最小化して，予測値 y^{ij} を生成する．

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}|y_{ij}-(w\cdot h(x_{ij}))-b_i|_{\varepsilon},$$

次に，最尤法による「強度係数」β の推定により，y から確率的な予測値を生成する最終的な「正則化」ステップが行われる．

$$p_{ij}=\frac{\exp((\hat{y}_{ij}\cdot\beta))}{\exp((\hat{y}_{i1}\cdot\beta))+\exp((\hat{y}_{i2}\cdot\beta))+\cdots+\exp((\hat{y}_{in_i}\cdot\beta))}.$$

モデルのベッティングの有効性を判断するためには，ベッティングルールを適用し，パラメータ・フィッティングに使用したサンプルとホールドアウト・サンプルの両方で収益性を評価することが重要である．

分析結果

分析対象

- 1995 年にオーストラリアで行われたメトロポリタンレースの結果

- 12 個の入力変数で 200 レースの結果を使用し，100 レースのホールドアウトサンプルで結果をテストする．

分析結果

使用する変数

- 前レースのゴール位置
- 前レースのブックメーカーのオッズ
- 前走時の賞金総額
- 前レースの最終コール位置
- 前レースのインジケーター，ノーコールポジション
- 前レースの距離
- 前走オッズ × 賞金額
- 前走の重量
- 前走からの経過日数
- 今回のレースの賞金総額
- 今回のレースでの重量
- 今回のレースの距離

使用する変数

- ベッターやブックメーカーがオッズを設定する方法は，時間の経過とともにあまり安定していない可能性があるため，「現在のレースのオッズ」を使わない
- 基本的な「ファンダメンタル」モデルをフィットさせ，「正規化」ステップで最新の市場情報と組み合わせる．

- 入力から得られた SVM は，結果変数として調整済みのゴールポジションを用いて，トレーニングサンプルでは適合値と予測値の間に 48% の相関関係，ホールドアウトセットでは 45% の相関関係が得られた．

- モデルの最終段階として，訓練サンプルデータを用いて log-1 + Odds と Fundamental 変数 (すなわち，SVM 回帰からフィットした y) の関数として Win(0-1) の Multinomial Logit 回帰を実行する．

- 回帰係数は 0.85 と 1.45 で，それぞれの t 検定は 6.4 と 4.0 となり，各変数に明確な限界予測値があることがわかった．
- また，尤度 R 二乗は 19.2% で，ブックメーカーのオッズのみのモデルの 17.3% よりも高く，ベッティングアドバンテージの可能性を示唆している．

- 完全なモデルは，ケリーベッティングを使用して，100 レースのホールドアウトセットに適用され，例えば，1 つのレースに 1 回賭ける場合のベッティング量 b_i は次のように与えられる．

$$b_i = \frac{p_i(1+o_i)-1}{o_i},$$

ここで，p_i は予想確率，o_i はブックメーカーのオッズである．

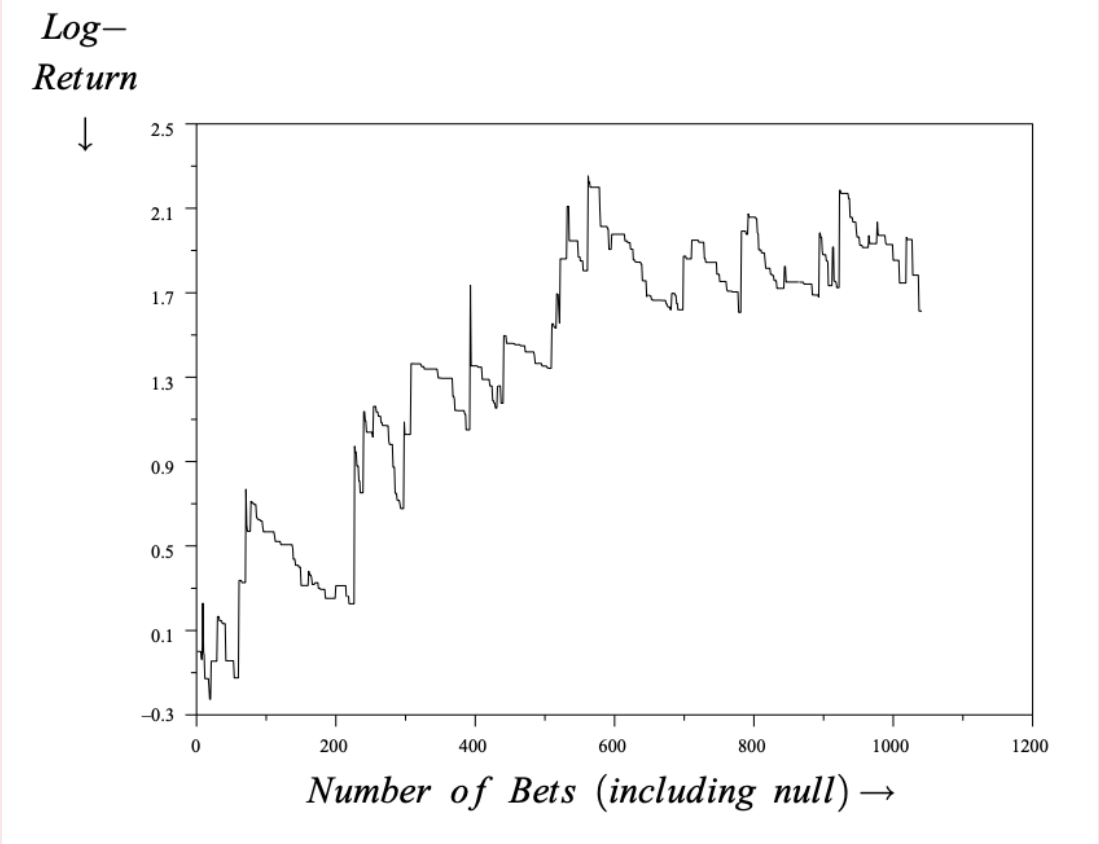

図 1: 対数累積リターン (ホールドアウト・サンプル)

- 図 1 より，ベットの回数を増やせば増やすほど Log-Return（収益率）が増加していることがわかる．

おわりに

まとめ

- SVM の使用は，トレーニングサンプルの数に比べて多数の入力変数でノンパラメトリックモデルをフィットさせる方法として成功した.,
- しかし，本来なら提示した変数以外にも前走以上の情報，レースクラスの情報，騎手情報など他にも含めるべき変数がたくさんある.

課題

- SVM アプローチをより詳細な入力変数に適用し，より大きなデータセットで学習させる.
- SVM 以外のカーネル手法との比較