競馬入門

数学と統計学の視点から

July 14, 2022

はじめに

- 数学と統計学の視点から競馬について語ります。
- 競馬で儲かる夢の投資アルゴリズムに少しでも近づくことはできるのか？
- １０年以上前に私が勉強していた内容を要約したものです。

馬券の種類

- 単勝馬券: 馬を 1 頭指定する。その馬が 1 着になれば当たり。
- 馬単馬券: 馬を 2 頭順序ありで指定する。それらの馬が 1 着と 2 着になれば当たり。
- 3 連単馬券: 馬を 3 頭順序ありで指定する。それらの馬が 1 着、2 着、3 着になれば当たり。

馬券の種類

- 複勝馬券: 馬を 1 頭指定する。その馬が 3 着以内になれば当たり。
- ほかにも、馬連馬券、3 連複馬券等がある。

オッズ

オッズとは、賭けた金が何倍になって払い戻されるかを表す数値。
例: オッズ 3.5 の馬券を 200 円分購入したとする。的中すれば、払戻金は 700 円、はずれの場合、払戻金は 0 円である。

オッズの決め方

Parimutuel 方式 馬券の総売り上げから競馬場が一定割合 (2 割〜3 割) を差し引き、残りの金額を的中馬券に配当する方式。
日本では、オッズは Parimutuel 方式で定まる。オッズの具体的な計算式は競馬法という法律によって定められている。

ブックメーカー方式 馬券購入時にオッズが指定されている方式。

裁定機会

思考実験: 出走馬が 2 頭だけのレースを考える。資金が 10000 円あるとき、あなたならどのように単勝馬券を購入しますか？あるいはお金を賭けない?

馬番	馬名	単勝オッズ
1	エフフォーリア	2.0
2	クロノジェネシス	3.0

裁定機会

私ならば、1 番に 6000 円, 2 番に 4000 円賭けます。

馬番	単勝オッズ	掛け金	払戻金
1	2.0	6000 円	12000 円
2	3.0	4000 円	12000 円

理由は、どちらの馬が 1 着になろうとも、12000 円の払戻金があるからです。このように、損失の可能性が 0 でお金を稼げる機会を裁定機会と言います。

裁定機会

状況を少し一般化します。

馬番	単勝オッズ	掛け金	払戻金
1	$O_1 + 1$	$\frac{1}{O_1+1}u$	u
2	$O_2 + 1$	$\frac{1}{O_2+1}u$	u

ここで $u = \left(\frac{1}{O_1+1} + \frac{1}{O_2+1}\right)^{-1}$。
掛け金の合計は 1 である。
$u > 1$ つまり $\frac{1}{O_1+1} + \frac{1}{O_2+1} < 1$ なので裁定機会がある。

裁定機会

状況をさらに一般化して、n 頭の単勝馬券がオッズ $O_i + 1$, $i = 1, 2, \ldots, n$ で販売されているとする。
$\sum_{i=1}^{n} \frac{1}{O_i+1} < 1$ のとき、各馬 j にそれぞれ $\frac{1}{O_j+1} \left(\sum_{i=1}^{n} \frac{1}{O_i+1} \right)^{-1}$ を賭けると、払戻金は $\left(\sum_{i=1}^{n} \frac{1}{O_i+1} \right)^{-1} > 1$ となる。
裁定機会が存在するための必要十分条件は、

$$\sum_{i=1}^{n} \frac{1}{O_i + 1} < 1$$

である。

裁定機会

馬番	馬名	単勝オッズ
1	エフフォーリア	2.0
2	クロノジェネシス	3.0
3	ステラヴェローチェ	4.0

上記の 3 頭のレースでは、$\frac{1}{2}+\frac{1}{3}+\frac{1}{4}=\frac{13}{12}\geq 1$ となり裁定機会は存在しない。

裁定機会

実際のレースでも、単勝馬券による裁定機会は存在しない。
(ただし、単勝馬券以外の馬券を組合わせる時の裁定機会を否定しているわけではない。)
$\rightarrow$ 確率や統計学を用いた考えを用いる。

投資アルゴリズムの構成

- レース予測
  各レースの状況を予測する。
  - 単勝確率
    各馬が 1 着になる確率を問題にする。
  - 順序確率
    2 着, 3 着がどのような状況になるのかを問題にする。
- ポートフォリオ構成
  どの馬券をどれだけ購入すれば良いのかを問題にする。

単勝確率

各馬 i が 1 着になる確率を π_i とおく。
確率 π_i をどうやって推定すればよいのか？
ちょっと、皆さんも考えてみて下さい。

単勝確率

最もシンプルなモデルとして、

$$\pi_i = P_i \quad (i = 1, 2, \ldots, n)$$

がある。ここで P_i は単勝馬券総売上げ額に対する単勝馬券 i の総売上げ額の割合を表す。
馬券の正式名称は「勝馬投票券」。P_i は馬 i への支持率と言い換えることもできる。

単勝確率

$\pi_i = P_i$ の改良版として、

$$\pi_i = \frac{P_i^\beta}{\sum_{j=1}^n P_j^\beta}, \quad (i = 1, 2, \ldots, n).$$

がある。ここで β はパラメータを表し、値はおよそ 1.1 になる。
$\beta = 1$ のとき、$\pi_i = P_i$ が成立し最もシンプルなモデルとなる。

単勝確率

ロジスティック回帰, サポートベクターマシン, ランダムフォレスト, 勾配ブースティングで単勝確率を推定する前に、知っておいて欲しい統計モデルがあります。
それは, Conditional Logit モデルです。

単勝確率

Conditional Logit モデルを用いると, 単勝確率が以下の形式で推定できる.

$$\pi_i = \frac{\exp(\beta^T x_i)}{\sum_{j=1}^{n} \exp(\beta^T x_j)}$$

ここで、
π_i: 馬 i が 1 着になる確率。
x_i: 馬 i を表す説明変数。
β: パラメータ。

単勝確率

具体例として、次の 4 頭によるレースを考える。

馬番	変数 1	変数 2	変数 3
1	5	20	3.3
2	4	25	2.4
3	1	24	5.2
4	4	22	3.5

パラメータが $\beta = (9, 4, 11)^T$ のとき馬 2 が 1 着になる確率を求める。

単勝確率

各馬のスコアを求める。

馬番	$\beta^T x_i$ (スコア)
1	$9 \cdot 5 + 4 \cdot 20 + 11 \cdot 3.3 = 161.3$
2	$9 \cdot 4 + 4 \cdot 25 + 11 \cdot 2.4 = 162.4$
3	$9 \cdot 1 + 4 \cdot 24 + 11 \cdot 5.2 = 162.2$
4	$9 \cdot 4 + 4 \cdot 22 + 11 \cdot 3.5 = 162.5$

馬 2 が 1 着になる確率は

$$\begin{aligned}\pi_2 &= \frac{\exp(162.4)}{\exp(161.3) + \exp(162.4) + \exp(162.2) + \exp(162.5)} \\ &= 0.3071\end{aligned}$$

となる。

順序確率

レースにおける 2 着、3 着がどのような状況になるのかを問題にする。
馬 i が 1 着かつ馬 j が 2 着である確率を π_{ij} とおく。さらに馬 i が 1 着かつ馬 j が 2 着かつ馬 k が 3 着である確率を π_{ijk} とおく。
確率 π_{ij}, π_{ijk} をどうやって推定すればよいのか？
ちょっと、皆さんも考えてみて下さい。

順序確率

Harville は次のシンプルなモデルを提案した。

$$
\pi_{ij} = \pi_i \frac{\pi_j}{1-\pi_i},
$$
$$
\pi_{ijk} = \pi_i \frac{\pi_j}{1-\pi_i} \frac{\pi_k}{1-\pi_i-\pi_j}.
$$

順序確率

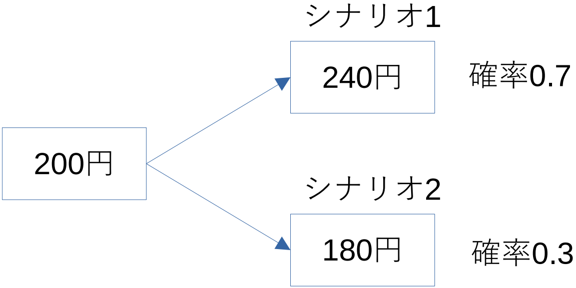
Harville モデルを包含する次のモデルも提案されている。

$$\pi_{ijk} = \pi_i \frac{\pi_j^\lambda}{\sum_{s \neq i} \pi_s^\lambda} \frac{\pi_k^\tau}{\sum_{t \neq i,j} \pi_t^\tau}.$$

λ と τ はパラメータを表す。このモデルはディスカウントモデルと呼ばれている。$\lambda = \tau = 1$ のときディスカウントモデルは Harville のモデルになる。パラメータ λ, τ の一例として, $\lambda = 0.76, \tau = 0.62$ を挙げておく。

株価の評価

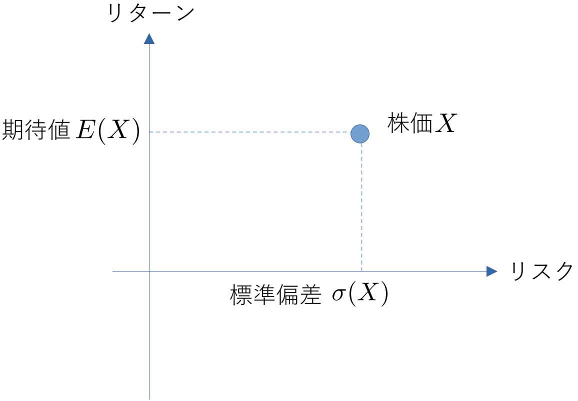
次の株価の変動を考える。



この株価をリターンとリスクの視点から評価したい。
どうすれば良いのか？皆さんも考えてみて下さい。

株価の評価

将来変化する株価を確率変数 X で表す。
株価のリターンは期待値 $E(X)$、
株価のリスクは標準偏差 $\sigma(X)$ で評価する。
この評価方法を平均分散分析 (Mean Variance Analysis) という。



株価の評価

Table: 確率変数 X の確率分布

シナリオ	利益率	確率
1	0.2	0.7
2	-0.1	0.3

具体的な数値計算。

$E(X) = 0.7 \cdot 0.2 + 0.3 \cdot (-0.1) = 0.11$

$V(X) = 0.7 \cdot (0.2 - 0.11)^2 + 0.3 \cdot (-0.1 - 0.11)^2 = 0.0189$

$\sigma(X) = \sqrt{V(X)} = 0.137$

株の合成

2 つの株 X, Y を合成したときの効果を見る。
$E(X) = 0.05$, $\sigma(X) = 0.08$
$E(Y) = 0.10$, $\sigma(Y) = 0.14$
相関係数 $\rho(X, Y) = -0.8$

X はリターン小、リスク小の株。
Y はリターン大、リスク大の株。

株の合成

2 つの株 X, Y を組合せて、新たな株 Z を合成する。
$Z = w_1 X + w_2 Y$ とするとき、
$E(Z) = w_1 E(X) + w_2 E(Y)$
$V(Z) = w_1^2 \sigma(X)^2 + 2 w_1 w_2 \sigma(X) \sigma(Y) \rho(X, Y) + w_2^2 \sigma(Y)^2$
が成立する。

$Z = 0.2X + 0.8Y$ のとき
$E(Z) = 0.09$, $\sigma(Z) = 0.097$
Z はリターン中、リスク中の株。

$W = 0.5X + 0.5Y$ のとき
$E(W) = 0.075$, $\sigma(W) = 0.0449$
W はリターン中、リスク極小の株。

株 (馬券) を組合せることで、リスクを抑えられる。

ポートフォリオ構成

馬券のポートフォリオ構成は、Kelly 基準が用いられることが多い。単勝馬券のみを購入する場合、次の連続最適化問題の最適解が Kelly 基準となる。

$$
\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{n} p_i \log(u_i f_i + w) \\
\text{subject to} \quad & \sum_{i=1}^{n} f_i + w = 1, \\
& f_i \geq 0, \quad i = 1, 2, \ldots, n, \\
& w \geq 0
\end{aligned} \tag{1}
$$

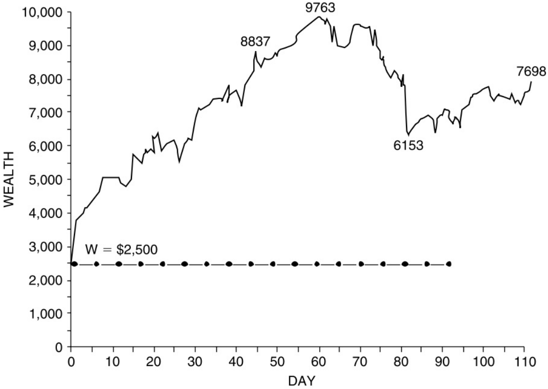
f_i: 所持金のうち馬券 i を購入する割合。
w: 所持金のうち現金として残す割合。
p_i: 馬 i が 1 着になる確率。
u_i: 馬 i の払戻率 (単勝オッズ)。

ポートフォリオ構成

上記の連続最適化問題は解析解を持つ。
なぜ Kelly 基準を用いるのか？
Kelly 基準の長所短所は文献MacLean et al. (2011) に要約されている。しかし、読んでも内容を全く理解できなかった。

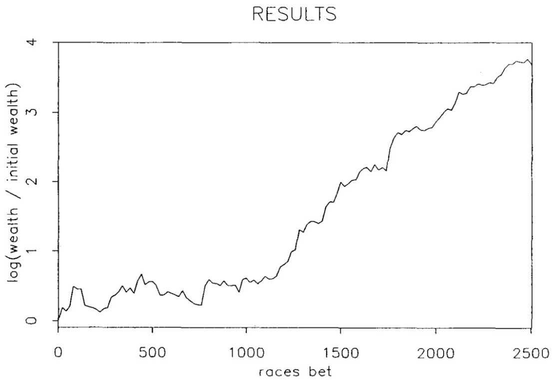
先行研究

文献Ziemba (2008)
単勝確率: 単勝馬券の支持率
順序確率:Harville モデル
ポートフォリオ構成:Kelly 基準
複勝馬券を実際に購入



先行研究

文献Benter (2008)
単勝確率:Conditional Logit モデル
順序確率: ディスカウントモデル
ポートフォリオ構成:Kelly 基準
実際に香港で馬券を購入

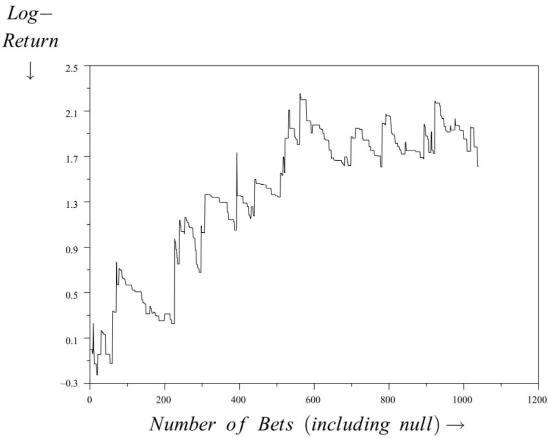


機械学習

近年では機械学習を用いた研究が行われています。私のお気に入りの論文は次の 4 つです。
Edelman (2007)
Lessmann et al. (2009)
Lessmann et al. (2010)
Lessmann et al. (2012)

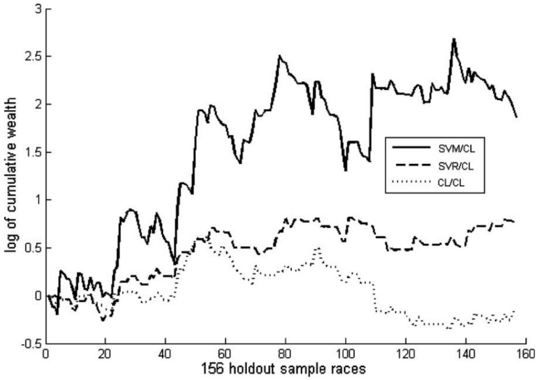
機械学習

文献Edelman (2007)
単勝確率: サポートベクターマシン
ポートフォリオ構成:Kelly 基準
単勝馬券購入のシミュレーション



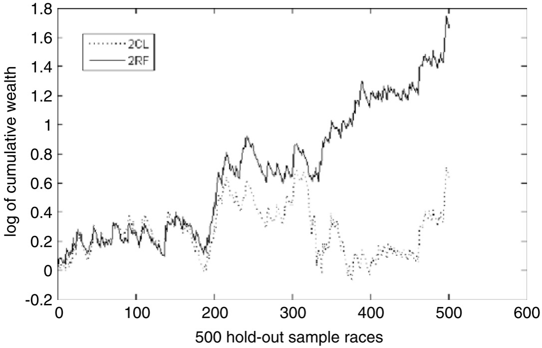
機械学習

文献Lessmann et al. (2009)
単勝確率: サポートベクターマシン
ポートフォリオ構成:Kelly 基準
単勝馬券購入のシミュレーション



機械学習

文献Lessmann et al. (2010)
単勝確率: ランダムフォレスト
ポートフォリオ構成:Kelly 基準
単勝馬券購入のシミュレーション



機械学習

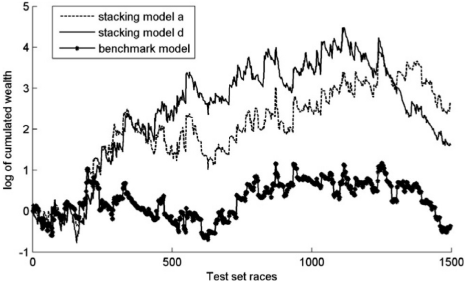
文献Lessmann et al. (2012)

単勝確率: 以下のモデルのアンサンブル学習

Conditional Logit モデル, 線形回帰モデル, ロジスティック回帰, サポートベクターマシン, ランダムフォレスト, 勾配ブースティング, Ada ブースト

ポートフォリオ構成:Kelly 基準

単勝馬券購入のシミュレーション



まとめ

- 競馬を数学と統計学の視点から見た。
- 競馬に勝つための夢の投資アルゴリズムにほんの少しだけ近づいた。
- ただし、機械学習を駆使しても Parimutuel 方式で勝つのは難しそうだ。

参考文献 I

Benter, W. (2008). Computer based horse race handicapping and wagering systems: a report. *Efficiency of racetrack betting markets* (pp. 183–198). World Scientific.

Edelman, D. (2007). Adapting support vector machine methods for horserace odds prediction. *Annals of Operations Research*, *151*(1), 325–336.

Lessmann, S., Sung, M.-C., & Johnson, J. E. (2009). Identifying winners of competitive events: A SVM-based classification model for horserace prediction. *European Journal of Operational Research*, *196*(2), 569–577.

Lessmann, S., Sung, M.-C., & Johnson, J. E. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. *International Journal of Forecasting*, *26*(3), 518–536.

参考文献 II

- Lessmann, S., Sung, M.-C., Johnson, J. E., & Ma, T. (2012). A new methodology for generating and combining statistical forecasting models to enhance competitive event prediction. *European Journal of Operational Research, 218*(1), 163–174.
- MacLean, L. C., Thorp, E. O., & Ziemba, W. T. (2011). Good and Bad Properties of the Kelly Criterion. *The kelly capital growth investment criterion: Theory and practice* (pp. 563–572). World Scientific.
- Ziemba, W. T. (2008). Efficiency of racing, sports, and lottery betting markets. *Handbook of sports and lottery markets* (pp. 183–222). Elsevier.