elmo@WCSC26

瀧澤 誠

開発コンセプト

- 評価関数の部分だけに絞って開発中
  - どのプログラムにも簡単に移植出来るのが理想。
  - 今後評価関数以外をやらないかどうかは未定。
- KPPファミリの評価値を独自手法で算出
  - Thompson Samplingで強化学習
    - 各特徴量の評価値は学習時の勝敗数から直接算出。

ボナメソ＋強化学習を考えてみたけど大変だ

- (①)Bonanzaメソッドで評価関数の初期値を与え、後は(②)コンピュータ自身で学習するよう作りたいが、思いのほか難しい。
  - 目的関数が異なるので①と②が馴染まない(②は通常、正則化項を使わない)。
    - 学習サンプルが多い特徴も少ない特徴も同じように扱われ、①の結果が大体無視される。
    - ①実行後、①に②の観点を組み込んだサンプルを含めて再学習すれば何とかなりそうだが、結局棋譜(①)から学習しているに過ぎない。
    - 評価値を使って評価値を更新する手法は(迷走する気がするので)嫌。
    - ②は自己対局として実現したい。強くなりそうだから。
    - 学習履歴(分散/確信度)を考慮して更新する手法はあるはず。

→勝敗を報酬とした強化学習＋過去の学習履歴を考慮する手法がイイ。

学習履歴を考慮する手法は
色々あるけど
妥当だと思ったやつで

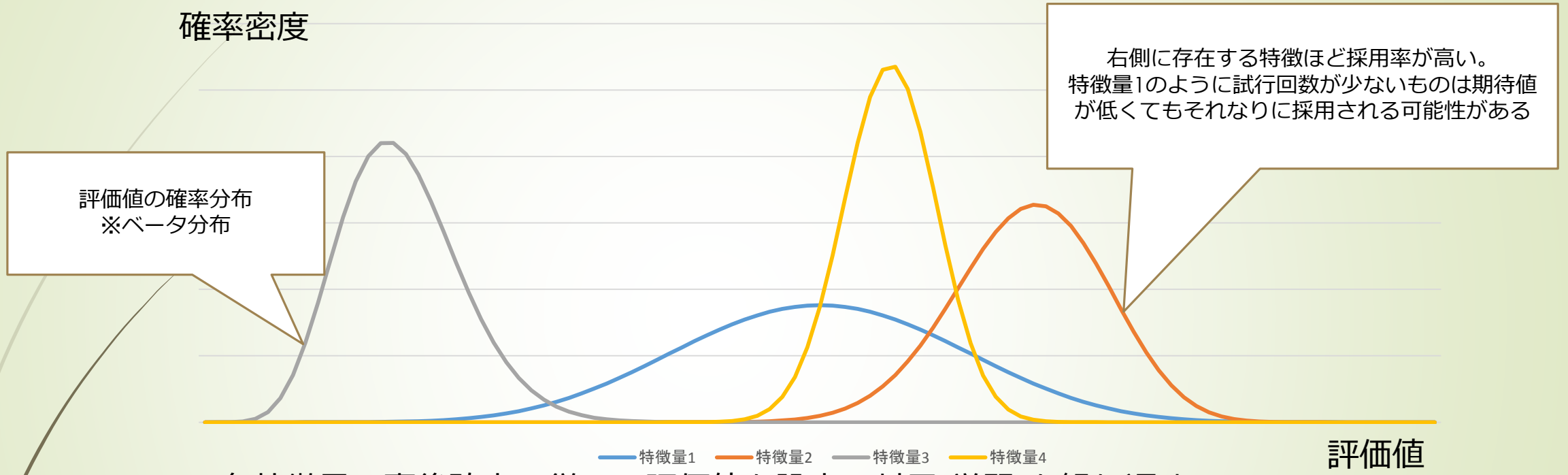
イメージ通りのやつ

Thompson Sampling

- 多腕バンディット問題で用いられる手法の1つ
  - 探索(まだ学習が足りない特徴を試す)と活用(最も良いと思われる特徴を使う)の選択を行う。
    - 学習では探索(上記の意味)が重要だが、勝敗という報酬の性質上(150手等で1つ)、評価を適切に行うために活用も重要※。
  - 行動による報酬を期待値だけではなく、分布として捉え、更新する。

※多腕バンディット問題と強化学習では目的が異なるため必ずしも適切な選択ではないが、上記報酬の性質や、学習時に確信度(confidence)を考慮する必要が無い(サンプリングで実現)などメリットがある。

Thompson Sampling



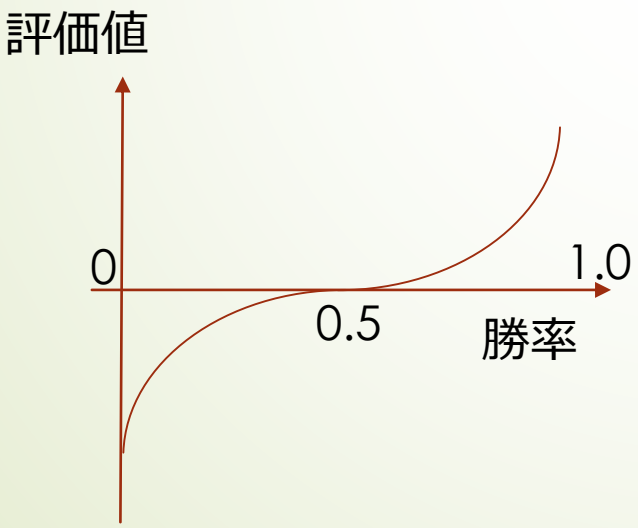
各特徴量の事後確率に従って評価値を設定→対局(学習)を繰り返す。
(対局の度に確率分布を更新する)

※ Thompson Samplingでは概ね評価値の高いものから順に採用されるため、
(上図とは異なり)評価値の高い方からグラフが尖っていく(評価値が確定していく)
※グラフが尖っている＝分散が小さい＝確信度が高い＝試行回数が多い

勝てば評価値が右にシフトし、負ければ左にシフトする。

勝率/評価値変換

- 勝率から評価値を求める。
  - 勝率→評価値の変換は事前に定義したロジット関数。
  - Thompson Samplingの事前確率はボナメソの結果を流用。



あとはひたすら学習

自己対局による評価値のブラッシュアップ

- 自己対局を行い、勝利数と敗北数を更新
  - 平手からの対局
  - 棋譜の途中局面から対局
  - 自動生成された局面から対局

したい。

最後に

- まだ頑張って作ってます。
  - 手法の試行錯誤とバグで…
  - うまくいかないかも
- 元々、勝率と評価値を同一視する手法にしたかった。
  - 実績を出してみるとどの特徴も勝率50%に漸近する(よく考えると当たり前)。
  - ↑3/27に気付いて手法を変更。計算上勝率を補正している(レーティングに近い)。
- 直近の学習をより重視する予定(WCSC以降)
- 強化学習分かりません。 誰か教えて

- 謝辞
  - Aperyを公開していただいた平岡氏をはじめ、その他関係者皆様に感謝いたします。

// 5/1追記 ここから

残念ながら

超弱かったです。

でも良くあること

分かったこと

- 150手指して報酬1つ(勝ち負け)は少な過ぎ。
  - 使用した特徴量の0.01%が間違っていた程度の差で巻き添え多数。
  - 現実的なところまでサンプル数を稼げない。
- ボナメソは優秀。
  - とにかく学習する数が多い。
- Thompson Samplingを使うまでも無く手は分散される。
  - Aperyを使ってる分には多分必要無い。

強化学習は魅力的だが、KPPの特徴空間が巨大過ぎて厳しい。

十分なサンプルを与えるか、KPPよりコンパクトな特徴量が必要と思われる。正直今のままだとしんどい。

取り合えずの出場ネタ※開発者向け

- Aperyを見ていて気になったところを変えてます。
  - r_ee, ee, xee, yeeが特徴量的に不要と思い、削除。※ 計算量も削減
  - e (effect)がDirecMisc(駒の関係がその他(利きが無い位置関係))の時、影響を無視(元々、最大となっていた) ※ 計算量も削減
  - ボナメソ学習時に更に浅い探索を入れて指し手が教師と一致するようについでに学習。
    - NDFの簡易版

でもApery twigに勝率4割くらいです。

オリジナルのバイナリが強くて追いつけないです。

踏み台仕様

- KPPは落として出場予定。
  - KK,KKP(KP含む)で考えています。
- ノートPC(2コア)で参加予定。

// ここまで