Partial and Delayed Feedback

Third Thematic Programme

Months 12-18

Nicolò Cesa-Bianchi

Università degli Studi di Milano

Yishay Mansour (Tel-Aviv) and Michele Sebag (CNRS)

Thanks to Giorgio Valentini (Milano)

Motivation

- Learning models (statistical, on-line) are based on idealized protocols in which the learner receives feedback indicating the optimal action (e.g., the correct label of an instance)

Motivation

- Learning models (statistical, on-line) are based on idealized protocols in which the learner receives feedback indicating the optimal action (e.g., the correct label of an instance)
- In real-world cognitive applications, this feedback is typically partial, delayed, or indirect

Motivation

- Learning models (statistical, on-line) are based on idealized protocols in which the learner receives feedback indicating the optimal action (e.g., the correct label of an instance)
- In real-world cognitive applications, this feedback is typically partial, delayed, or indirect
- Active or semi-supervised learning: correct labels obtained only for a small subset of the dataset (extreme case: clustering)

Motivation

- Learning models (statistical, on-line) are based on idealized protocols in which the learner receives feedback indicating the optimal action (e.g., the correct label of an instance)
- In real-world cognitive applications, this feedback is typically partial, delayed, or indirect
- Active or semi-supervised learning: correct labels obtained only for a small subset of the dataset (extreme case: clustering)
- Reinforcement learning: utility of actions is known after a variable number of time steps

Motivation

- Learning models (statistical, on-line) are based on idealized protocols in which the learner receives feedback indicating the optimal action (e.g., the correct label of an instance)
- In real-world cognitive applications, this feedback is typically partial, delayed, or indirect
- Active or semi-supervised learning: correct labels obtained only for a small subset of the dataset (extreme case: clustering)
- Reinforcement learning: utility of actions is known after a variable number of time steps
- Common themes:
  1. Exploration vs. exploitation trade-off
  2. Use of unlabeled data to regularize solutions

Bandit problems

. . .

K possible actions

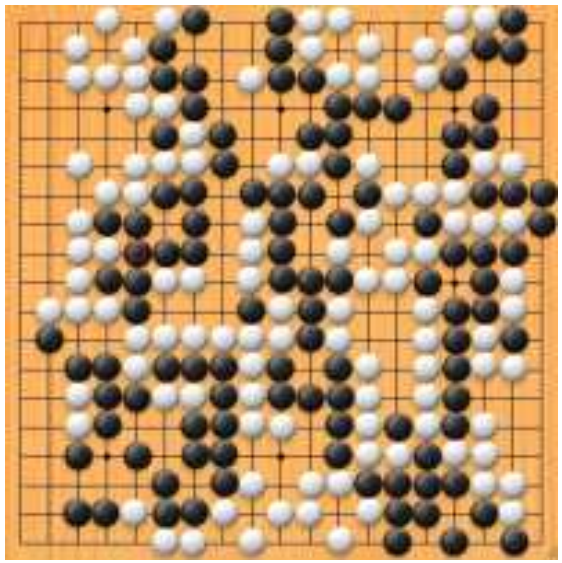
- A sequential decision problem in which an agent must, at each step, choose one among K actions, possibly based on side information (multivariate bandits)
- Feedback is obtained only for the chosen option – utility of other actions remains unknown
- Multiclass classification: actions = classes, feedback = Yes/No. The true class of an instance remains unknown on mistakes

Some applications in PASCAL

- Adaptive content management of websites (PASCAL challenge in 2006 sponsored by Omniture/TouchClarity).
- Direct commercial application of multivariate bandit technology

Some applications in PASCAL

- Adaptive content management of websites (PASCAL challenge in 2006 sponsored by Omniture/TouchClarity).
- Direct commercial application of multivariate bandit technology
- Programs to play GO: bandit-based program *MoGo* is the highest rated program on the Computer GO Online Server
  [Gelly, Wang, Munos and Teytaud, 2006]



Lightweight reinforcement learning models

- *MoGo* is a convincing example of a practical game-playing algorithm based on principled bandit algorithms (*UCB, UCT*)

Lightweight reinforcement learning models

- *MoGo* is a convincing example of a practical game-playing algorithm based on principled bandit algorithms (*UCB, UCT*)
- Goal: Definition of RL models in between the bandit problem and the full MDP problem that allow for the design and analysis of practical algorithms in reactive environments

Lightweight reinforcement learning models

- *MoGo* is a convincing example of a practical game-playing algorithm based on principled bandit algorithms (*UCB, UCT*)
- Goal: Definition of RL models in between the bandit problem and the full MDP problem that allow for the design and analysis of practical algorithms in reactive environments
- Apprenticeship learning: learn a good policy on a given MDP by observing the trajectory of an unknown expert policy
  [Abbeel and Ng, 2004; Syed and Schapire, 2008]

Game-theoretic models

- Partial monitoring: dynamic-pricing, revealing actions
  [C-B, Lugosi and Stoltz, 2006]
- Budgeted learning: an overall budget on the number of attributes that can be accessed
  [Lizotte, Madani and Greiner, 1993]
- Incentive-based learning: feedback as a result of a negotiation between decision maker and environment (which is neither oblivious nor adversarial)
- Goal: Develop unified models of game-theoretic learning to address realistic cognitive problems

Connection with Multicomponent Learning

- Networks of bandit players [Awerbuch and Kleinberg, 2005]
- Multitask learning [Evgeniou, Micchelli and Pontil, 2005]
- Maximization of overall utility when adaptive components are selfish (e.g., [Blum, Even-Dar and Ligett, 2006])

Clustering

The mother of all ill-posed problems. . .

- Principled and efficient algorithms
- Stability
- Other criteria for consistency of clustering
- Assessment of validity/reliability of solutions (e.g., model order selection, reliability of each discovered cluster)
- Discovery of multiple structures
- Scaling to high-dimensional, large size and/or large number of clusters
- Integration of multiple data sources

Preliminary list of events

- Joint workshop with multicomponent thematic programme
- Practical reinforcement learning
- Game-theoretic and utility-based approaches to learning
- Challenging unsupervised learning problems in bioinformatics