Multi-Arm Bandit

Team Members: Tristan Caputo, Paul Lenharth, River Ludington

Project Description

- Artificial Neural Networks have found a wide range of uses, and their learning algorithms determine their usefulness.
- Much is known about the theoretical performance of learning algorithms [1][4]. Empirical evidence suggests, however, that naïve approaches often outperform more theoretically sound ones [2][3].
- Robbins multi-arm bandit problem offers a useful framework for evaluating desirable performance characteristics of learning algorithms.
- The goal of this project is to inform algorithm choice based on empirical evaluations of short-term vs. long-term performance characteristics.

Scientific Challenges

- Industry and Academia are seeking strategies to optimize Neural Network performance

- With widely varying application, the appropriate choice of algorithm is dependent on the algorithms tradeoff between fast-learning and optimal choice convergence.

Potential Applications

- This work has a large span of applications in ad-targeting, stock market prediction, recommendation algorithms, data acquisition, and data processing.

The Multi-Arm Bandit

- Each slot machine i has a different unknown distribution with an unknown expectation μ_i
- Each turn, the algorithm selects an arm
- The normalized regret each round is the difference between the mean reward of the optimal arm, and the reward of the arm chosen [1] divided by the mean reward of the optimal arm.

$$R = \frac{\mu^* - \mu_{j(t)}}{\mu^*}$$

Where R is the regret, μ^* is the expected reward from the best arm, $\mu_{j(t)}$ is the reward from the arm j chosen at round t

Figure 1. Diagram of slot machines colloquially known as “single-arm bandits”

Methodology

1. We coded the Multi-arm bandit problem to accommodate:
   - Bernoulli or Gaussian Distributions
   - Different numbers of arms
   - Different means and variances

   Coded common learning algorithms:
   - Thompson Sampling
   - Softmax (Boltzmann)
   - Epsilon-Greedy

   Coded a custom algorithm:
   - Vary-Greedy
2. At every turn, when the algorithm chose an arm to play, the regret was measured.
3. Parameters for the Multi-Arm Bandit were varied and each algorithm was tested
4. The average was taken over 100 runs

Results

1. We observe that EG (w/ exp. Decay) and Softmax generally outperform the more complex algorithms such as VG and TS

2. Optimal Algorithm choice is dependent on the type of distribution and need for short-term vs long-term performance.

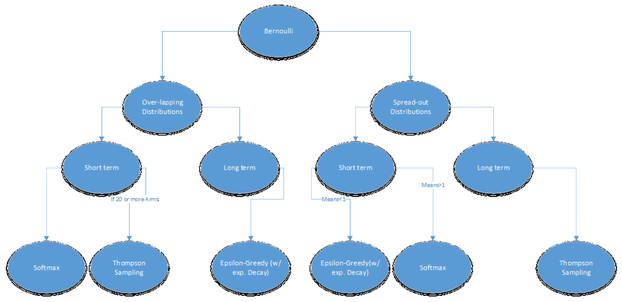

Figure 2. Tree diagram of optimal algorithms for Bernoulli Distributions

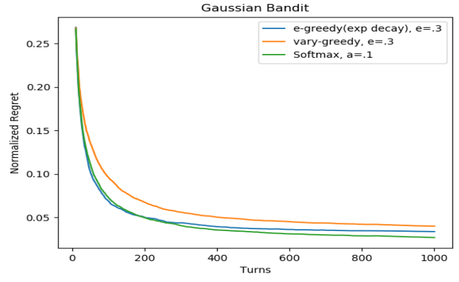

Figure 3. Sample results plot comparing three algorithms against a Gaussian Distribution

Glossary of Technical Terms

Artificial Neural Network: A computing system loosely based on biological neural networks, through the use of nodes and weighted connections.
Optimal Choice Convergence: An algorithm’s ability to converge to zero regret per iteration.
Spread-out Distributions: Distributions with large differences in means compared to Standard Deviation.
Over-lapping Distributions: Distributions with small differences in means compared to Standard Deviation.

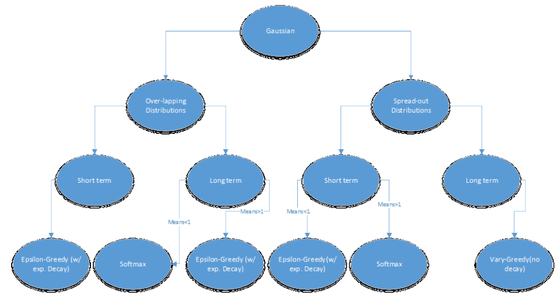

Figure 4. Tree diagram of optimal algorithms for Gaussian Distributions

References

1. Russo, Daniel, et al. "A Tutorial on Thompson Sampling.", 2017.
2. Kuleshov, Volodymyr, and Doina Precup. "Algorithms for Multi-Armed Bandit Problems.", 2014.
3. Vermorel, Joannès, and Mehryar Mohri. "Multi-Armed Bandit Algorithms and Empirical Evaluation." vol. 3720, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
4. Agrawal, Shipra, and Navin Goyal. "Analysis of Thompson Sampling for the Multi-Armed Bandit Problem." *Journal of Machine Learning Research*, vol. 23, 2012, pp. 39.26.

Acknowledgments

This project was mentored by Joseph Gibney, whose help is acknowledged with great appreciation.