

Thompson sampling

From Wikipedia, the free encyclopedia

In artificial intelligence, Thompson sampling,^[1] named after William R. Thompson, is a heuristic for choosing actions that addresses the exploration-exploitation dilemma in the multi-armed bandit problem. It consists in choosing the action that maximizes the expected reward with respect to a randomly drawn belief.

Contents

■ 1 Description

■ 2 History

■ 3 Relationship to other approaches

■ 3.1 Probability matching

■ 3.2 Bayesian control rule

■ 4 References

Description

Consider a set of contexts \mathcal{X} , a set of actions \mathcal{A} , and rewards in \mathbb{R} . In each round, the player obtains a context $x \in \mathcal{X}$, plays an action $a \in \mathcal{A}$ and receives a reward $r \in \mathbb{R}$ following a distribution that depends on the context and the issued action. The aim of the player is to play actions such as to maximize the cumulative rewards.

The elements of Thompson sampling are as follows:

1. a likelihood function $P(r|\theta, a, x)$;
2. a set Θ of parameters θ of the distribution of r ;
3. a prior distribution $P(\theta)$ on these parameters;
4. past observations triplets $\mathcal{D} = \{(x; a; r)\}$;
5. a posterior distribution $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$, where $P(\mathcal{D}|\theta)$ is the likelihood function.

Thompson sampling consists in playing the action $a^* \in \mathcal{A}$ according to the probability that it maximizes the expected reward, i.e.

$$\int \mathbb{I}[\mathbb{E}(r|a^*, x, \theta) = \max_{a'} \mathbb{E}(r|a', x, \theta)] P(\theta|\mathcal{D}) \, d\theta,$$

where \mathbb{I} is the indicator function.

In practice, the rule is implemented by sampling, in each round, a parameter θ^* from the posterior $P(\theta|\mathcal{D})$, and choosing the action a^* that maximizes $\mathbb{E}[r|\theta^*, a^*, x]$, i.e. the expected reward given the parameter, the action and the current context. Conceptually, this means that the player instantiates his or her beliefs randomly in each round, and then acts optimally according to them.

History

Thompson sampling was originally described in an article by Thompson from 1933^[1] but has been largely ignored by the artificial intelligence community. It was subsequently rediscovered numerous times independently in the context of reinforcement learning.^{[2][3][4][5][6][7]} A first proof of convergence for the bandit case has been shown in 1997.^[2] The first application to Markov decision processes was in 2000.^[4] A related approach (see Bayesian control rule) was published in 2010.^[3] In 2010 it was also shown that Thompson sampling is *instantaneously self-correcting*.^[7] Asymptotic convergence results for contextual bandits were published in 2011.^[5] Nowadays, Thompson Sampling has been widely used in many online learning problems: Thompson sampling has also been applied to A/B testing in website design and online advertising;^[8] Thompson sampling has formed the basis for accelerated learning in decentralized decision making;^[9] a Double Thompson Sampling (D-TS)^[10] algorithm has been proposed for dueling bandits, a variant of traditional MAB, where feedbacks come in the format of pairwise comparison.

Relationship to other approaches

Probability matching

Probability matching is a decision strategy in which predictions of class membership are proportional to the class base rates. Thus, if in the training set positive examples are observed 60% of the time, and negative examples are observed 40% of the time, the observer using a probability-matching strategy will predict (for unlabeled examples) a class label of "positive" on 60% of instances, and a class label of "negative" on 40% of instances.

Bayesian control rule

A generalization of Thompson sampling to arbitrary dynamical environments and causal structures, known as Bayesian control rule, has

been shown to be the optimal solution to the adaptive coding problem with actions and observations.^[3] In this formulation, an agent is conceptualized as a mixture over a set of behaviours. As the agent interacts with its environment, it learns the causal properties and adopts the behaviour that minimizes the relative entropy to the behaviour with the best prediction of the environment's behaviour. If these behaviours have been chosen according to the maximum expected utility principle, then the asymptotic behaviour of the Bayesian control rule matches the asymptotic behaviour of the perfectly rational agent.

The setup is as follows. Let $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_T$ be the actions issued by an agent up to time T , and let $\boldsymbol{o}_1, \boldsymbol{o}_2, \dots, \boldsymbol{o}_T$ be the observations gathered by the agent up to time T . Then, the agent issues the action \boldsymbol{a}_{T+1} with probability:^[3]

$$P(\boldsymbol{a}_{T+1} | \hat{\boldsymbol{a}}_{1:T}, \boldsymbol{o}_{1:T}),$$


where the "hat"-notation $\hat{\boldsymbol{a}}_t$ denotes the fact that \boldsymbol{a}_t is a causal intervention (see Causality), and not an ordinary observation. If the agent holds beliefs $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ over its behaviors, then the Bayesian control rule becomes

$$P(\boldsymbol{a}_{T+1} | \hat{\boldsymbol{a}}_{1:T}, \boldsymbol{o}_{1:T}) = \int_{\boldsymbol{\Theta}} P(\boldsymbol{a}_{T+1} | \boldsymbol{\theta}, \hat{\boldsymbol{a}}_{1:T}, \boldsymbol{o}_{1:T}) P(\boldsymbol{\theta} | \hat{\boldsymbol{a}}_{1:T}, \boldsymbol{o}_{1:T}) d\boldsymbol{\theta},$$

where $P(\boldsymbol{\theta} | \hat{\boldsymbol{a}}_{1:T}, \boldsymbol{o}_{1:T})$ is the posterior distribution over the parameter $\boldsymbol{\theta}$ given actions $\boldsymbol{a}_{1:T}$ and observations $\boldsymbol{o}_{1:T}$.

In practice, the Bayesian control amounts to sampling, in each time step, a parameter $\boldsymbol{\theta}^*$ from the posterior distribution $P(\boldsymbol{\theta} | \hat{\boldsymbol{a}}_{1:T}, \boldsymbol{o}_{1:T})$, where the posterior distribution is computed using Bayes' rule by only considering the (causal) likelihoods of the observations $\boldsymbol{o}_1, \boldsymbol{o}_2, \dots, \boldsymbol{o}_T$ and ignoring the (causal) likelihoods of the actions $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_T$, and then by sampling the action \boldsymbol{a}_{T+1}^* from the action distribution $P(\boldsymbol{a}_{T+1} | \boldsymbol{\theta}^*, \hat{\boldsymbol{a}}_{1:T}, \boldsymbol{o}_{1:T})$.

References

- Thompson, William R. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples" (<https://www.dropbox.com/s/yhn9prnr5bz0156/1933-thompson.pdf>). *Biometrika*, 25(3–4):285–294, 1933.
- J. Wyatt. *Exploration and Inference in Learning from Reinforcement*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh. March 1997.
- P. A. Ortega and D. A. Braun. "A Minimum Relative Entropy Principle for Learning and Acting", *Journal of Artificial Intelligence Research*, 38, pages 475–511, 2010.
- M. J. A. Strens. "A Bayesian Framework for Reinforcement Learning", *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford University, California, June 29–July 2, 2000, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.1701>
- B. C. May, B. C., N. Korda, A. Lee, and D. S. Leslie. "Optimistic Bayesian sampling in contextual-bandit problems". Technical report, Statistics Group, Department of Mathematics, University of Bristol, 2011.
- Chapelle O. and Li, L. "An Empirical Evaluation of Thompson Sampling". NIPS, 2011.
- O.-C. Granmo. "Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton", *International Journal of Intelligent Computing and Cybernetics*, 3 (2), 2010, 207-234.
- Ian Clarke. "Proportionate A/B testing", September 22nd, 2011, <http://blog.locut.us/2011/09/22/proportionate-ab-testing/>
- Granmo, O. C.; Glimsdal, S. (2012). "Accelerated Bayesian learning for decentralized two-armed bandit based decision making with applications to the Goore Game". *Applied Intelligence*. doi:10.1007/s10489-012-0346-z (<https://doi.org/10.1007%2Fs10489-012-0346-z>).
- Wu, Huasen; Liu, Xin; Srikant, R (2016), *Double Thompson Sampling for Dueling Bandits*, arXiv:1604.07101 (<https://arxiv.org/abs/1604.07101>)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Thompson_sampling&oldid=783384719"

-
- This page was last edited on 1 June 2017, at 23:10.
 - Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.