Reinforcement Learning
Contextual Bandits

Thomas Bonald
Institut Polytechnique de Paris

Claire Vernade
DeepMind

Till Wohlfarth
Veepee

2020 – 2021

This note is an introduction to contextual bandits, a class of multi-armed bandits where an agent must take sequential actions at time $t = 1, 2, \dots$ based on observed rewards that are supposed to depend on some unknown parameter θ (the context). In recommander systems for instance, the parameter θ is supposed to characterize the user. This parameter is learnt based on the feedback provided by the user for each proposed item. Each item corresponds to an action of the agent, who must learn the best actions, i.e., the items providing the best rewards. We mainly focus on so-called *linear bandits*, where expected rewards are linear functions of the actions, then present the extension to *logistic bandits*.

1 Model

The rewards of each action are supposed to be i.i.d. with mean equal to some function f_θ of the action, for some unknown parameter θ. In linear bandits, this function is linear and the reward r of action a satisfies:

$$\mathrm{E}(r|a) = \theta^T a.$$

Observe that the parameter θ and the action a live in the same space, say $\mathbb{R}^d$. Since the set of actions is possibly infinite, we assume that the agent must select at time t an action in some finite set $A_t \subset \mathbb{R}^d$. In recommander systems for instance, each action $a \in \mathbb{R}^d$ corresponds to an item, described by d features. The parameter $\theta \in \mathbb{R}^d$ characterizes the user and can be interpreted as the sensitivity of the user to each feature. The action set A_t is the set of items that are proposed to the user at time t.

We denote by $a_t \in A_t$ the action selected by the agent at time t and by r_t the corresponding reward. Note that a_t can only depend on previous actions $a_1, \dots, a_{t-1}$ and rewards $r_1, \dots, r_{t-1}$.

2 Linear regression

Since the parameter θ is unknown, it is necessary to estimate it. A natural method is the regularized least squares. The estimate of θ at time t is then given by:

$$\hat{\theta}_t = \arg\min_\theta L_t(\theta)$$

for the loss function:

$$L_t(\theta) = \frac{1}{2}\left(\sum_{s=1}^{t}(\theta^T a_s - r_s)^2 + \lambda||\theta||^2\right).$$

Here $\lambda > 0$ is some regularization parameter. Taking the gradient gives:

$$\nabla L_t(\theta) = \lambda\theta + \sum_{s=1}^{t}(\theta^T a_s - r_s)a_s,$$

so that:

$$\hat{\theta}_t = \left(\lambda I + \sum_{s=1}^{t} a_s a_s^T\right)^{-1} \left(\sum_{s=1}^{t} r_s a_s\right). \tag{1}$$

This expression involves the inversion of a matrix, which turns out to be the Hessian of the loss function L_t:

$$H_t = \nabla(\nabla L_t(\theta)) = \lambda I + \sum_{s=1}^{t} a_s a_s^T.$$

Observe that the Hessian does *not* depend on θ; it only depends on the selected actions.

Online estimation. Interestingly, the inverse of the Hessian can be computed sequentially through simple matrix-vector multiplications using the Sherman-Morrison formula [2] (see the appendix). Letting $\Gamma_t = H_t^{-1}$, we get:

$$\Gamma_{t+1} = \Gamma_t - \frac{\Gamma_t a_t a_t^T \Gamma_t}{1 + a_t^T \Gamma_t a_t}, \tag{2}$$

with $\Gamma_0 = I/\lambda$. The corresponding online estimation of θ is given as Algorithm 1.

Initialize:

- $\Gamma \leftarrow I/\lambda$
- $S \leftarrow 0$

For $t = 1, 2, \ldots$

- $a \leftarrow$ action
- $r \leftarrow$ reward
- $\Gamma \leftarrow \Gamma - \frac{\Gamma a a^T \Gamma}{1 + a^T \Gamma a}$
- $S \leftarrow S + ra$
- $\hat{\theta} \leftarrow \Gamma S$

Algorithm 1: Online estimation of θ.

The Bayesian view. Assume that the reward r of action a has a Gaussian distribution with mean $\theta^T a$ and unit variance. The likelihood of $r_1, \ldots, r_t$ satisfies:

$$p(r_1, \ldots, r_t | \theta) \propto \exp\left(-\frac{1}{2} \sum_{s=1}^{t} (\theta^T a_s - r_s)^2\right).$$

Now assume that θ is itself random, drawn from a centered Gaussian distribution with covariance matrix I/λ. We get:

$$p(\theta | r_1, \ldots, r_t) \propto p(\theta) p(r_1, \ldots, r_t | \theta) \propto \exp\left(-\frac{1}{2}\left(\sum_{s=1}^{t} (\theta^T a_s - r_s)^2 + \lambda ||\theta||^2\right)\right).$$

Thus the negative log-likelihood of θ is, up to some additive constant, equal to the loss function $L_t(\theta)$. In particular, $\hat{\theta}_t$ is the maximum-likelihood estimator for this model (Gaussian rewards, Gaussian prior).

Observing that

$$p(\theta|r_1,\dots,r_t) \propto \exp\left(-\frac{1}{2}\left(\theta^T\left(\lambda I + \sum_{s=1}^{t} a_s a_s^T\right)\theta - 2\theta^T\left(\sum_{s=1}^{t} r_s a_s\right)\right)\right),$$

we deduce from (1) that the posterior distribution of θ at time t is Gaussian with mean $\hat{\theta}_t$ and covariance matrix Γ_t,

$$\theta \sim \mathcal{N}(\hat{\theta}_t, \Gamma_t).$$

3 Greedy policy

The greedy policy consists in selecting at time t:

$$a_t = \arg\max_{a \in A_t} \hat{\theta}_t^T a,$$

where $\hat{\theta}_t$ is given by (1). Unlike usual multi-armed bandits, this is not necessarily a bad policy in the sense that the estimate $\hat{\theta}_t$ might converge to the true parameter θ and the agent eventually take the right actions. The rate of convergence might be slow, however, especially in high dimension d. Exploration can be enforced by the ϵ-greedy policy, to improve the estimation of θ. The parameter ϵ may be time-dependent (typically decreasing to 0).

4 UCB policy

The UCB (Upper Confidence Bound) policy consists in giving a bonus to less explored actions, following the "optimism in face of uncertainty" principle. Specifically, the UCB policy selects at time t:

$$a_t = \arg\max_{a \in A_t} (\hat{\theta}_t^T a + \alpha_t \sqrt{a^T \Gamma_t a}),$$

where $\hat{\theta}_t$ is given by (1) and α_t is some slowly increasing constant, like $c\sqrt{\log t}$ (see [1] for some specific expression of α_t providing confidence bounds on θ). Each action a is given a bonus proportional to $\sqrt{a^T \Gamma_t a}$. Observe that, with the Bayesian view, the expected reward of action a conditionned on θ has a Gaussian distribution with mean $\hat{\theta}_t^T a$ and variance $a^T \Gamma_t a$ at time t. The bonus assigned to action a is thus proportional to the standard deviation of the expected reward, given the posterior distribution of θ: the less confident the estimation, the higher the bonus.

5 Thompson sampling

The TS (Thompson Sampling) policy is based on the Bayesian view of the problem. The principle is to sample some parameter $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \Gamma_t)$ from the posterior distribution of θ at time t, and to apply the greedy policy using this parameter,

$$a_t = \arg\max_{a \in A_t} \tilde{\theta}_t^T a.$$

The exploration-exploitation trade-off is directly controlled by the estimation of θ: the less confident the estimation, the more variant the sample $\tilde{\theta}_t$ and thus the more diverse the selected actions.

6 Logistic regression

For binary rewards, it is natural to assume that the reward of action a has a Bernoulli distribution with parameter:

$$p = \frac{1}{1 + e^{-\theta^T a}}.$$

The estimation of the parameter θ follows from the regularized logistic regression:

$$\hat{\theta}_t = \arg\min_{\theta} L_t(\theta) \tag{3}$$

with

$$L_t(\theta) = -\sum_{s=1}^{t} \left(r_s \log p_s(\theta) + (1 - r_s) \log(1 - p_s(\theta))\right) + \frac{\lambda}{2}||\theta||^2,$$

and

$$p_t(\theta) = \frac{1}{1 + e^{-\theta^T a_t}}.$$

Let:

$$\sigma_t(\theta) = p_t(\theta)(1 - p_t(\theta)) = \frac{1}{(1 + e^{-\theta^T a_t})(1 + e^{\theta^T a_t})}.$$

The solution to the optimization problem (3) is no longer explicit, but can be approximated iteratively using Newton's method[1]. We obtain:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - H_t^{-1}(\hat{\theta}_t)\nabla L_t(\hat{\theta}_t),$$

where

$$\nabla L_t(\theta) = \lambda\theta + \sum_{s=1}^{t}(p_s(\theta) - r_s)a_s$$

and

$$H_t(\theta) = \nabla(\nabla L_t(\theta)) = \lambda I + \sum_{s=1}^{t} \sigma_s(\theta) a_s a_s^T,$$

This can be written is a similar form as in (1) as follows:

$$\hat{\theta}_{t+1} = \left(\lambda I + \sum_{s=1}^{t} \sigma_s(\hat{\theta}_t) a_s a_s^T\right)^{-1} \left(\sum_{s=1}^{t} (\sigma_s(\hat{\theta}_t)\hat{\theta}_t^T a_s + r_s - \hat{p}_s(\hat{\theta}_t))a_s\right). \tag{4}$$

Online estimation. To get an online estimate, we approximate $p_s(\hat{\theta}_t)$ by its estimate $\hat{p}_s = p_s(\hat{\theta}_s)$ at time s, for all $s = 1, \ldots, t$. We define $\hat{\sigma}_s = \hat{p}_s(1 - \hat{p}_s)$ accordingly. Denoting by H_t the corresponding estimate of the Hessian at time t, we get:

$$H_{t+1} = H_t + \hat{\sigma}_t a_t a_t^T.$$

The inverse of the Hessian, $\Gamma_t = H_t^{-1}$, can then be computed sequentially by the Sherman-Morrison formula:

$$\Gamma_{t+1} = \Gamma_t - \frac{\hat{\sigma}_t \Gamma_t a_t a_t^T \Gamma_t}{1 + \hat{\sigma}_t a_t^T \Gamma_t a_t},$$

with $\Gamma_0 = I/\lambda$. The corresponding online estimation of θ is given as Algorithm 2.

[1] Observe that this method gives the exact solution for the linear regression.

Initialize:

- $\Gamma \leftarrow I/\lambda$
- $S \leftarrow 0$
- $\hat{\theta} \leftarrow 0$

For $t = 1, 2, \ldots$

- $a \leftarrow$ action
- $r \leftarrow$ reward
- $\hat{p} \leftarrow \frac{1}{1+e^{-\hat{\theta}^T a}}$
- $\hat{\sigma} \leftarrow \hat{p}(1-\hat{p})$
- $\Gamma \leftarrow \Gamma - \frac{\hat{\sigma}\Gamma a a^T \Gamma}{1+\hat{\sigma} a^T \Gamma a}$
- $S \leftarrow S + (\hat{\sigma}\hat{\theta}^T a - \hat{p} + r)a$
- $\hat{\theta} \leftarrow \Gamma S$

Algorithm 2: Online estimation of θ for binary rewards.

The Bayesian view. The likelihood of $r_1, \ldots, r_t$ is given by:

$$p(r_1, \ldots, r_t|\theta) \prod_{s=1}^{t} p_s(\theta)^{r_s}(1-p_s(\theta))^{1-r_s}.$$

Assuming that θ is drawn from a centered Gaussian distribution with covariance matrix I/λ, we get:

$$p(\theta|r_1, \ldots, r_t) \propto p(\theta)p(r_1, \ldots, r_t|\theta) \propto \exp\left(\sum_{s=1}^{t}(r_s \log p_s(\theta) + (1-r_s)\log(1-p_s(\theta))) - \frac{\lambda}{2}||\theta||^2\right).$$

Thus the negative log-likelihood of θ is, up to some additive constant, equal to the loss function $L_t(\theta)$. In particular, $\hat{\theta}_t$ is the maximum-likelihood estimator for this model. Note that the posterior distribution of θ is no longer Gaussian.

Bandit algorithms. The greedy, UCB and TS policies can be applied accordingly (for TS, pretending that the posterior distribution of θ remains Gaussian with mean $\hat{\theta}_t$ and covariance matrix Γ_t, even if that is not the case).

Appendix

Sherman-Morrison formula

For any invertible matrix A and vectors u, v, the matrix $A + uv^T$ is invertible if and only if $v^T A^{-1} u + 1 \neq 0$, in which case

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}.$$

References

[1] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[2] J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.