Artificial Intelligence: Reinforcement Learning in Python

Multi Armed Bandit Problem: Choose a slot machine in a casino with the best win rate.

Dilemma: You need to collect lots of data for your estimates to be accurate. But when collecting lots of data, a lot of time is spent playing on suboptimal machines. Then it is necessary to find a balance between **Explore** (collect lots of data) and **Exploit** (playing "best-so-far" machine)

Epsilon greedy strategy (This course's algorithm to make a trade-of between exploration and exploitation)

Choose a small number epsilon as probability of exploration. Typical values: 5%, 10%

Pseudo code:

```
> p = random()
> if p < ϵ:
    pull random arm      (this would be explore)
> else:
    pull current-best arm    (this would be exploit)
```

In the long run, epsilon-greedy allows us to explore each arm an infinite number of times. The problem is that we get to a point where we explore when we don't need to. For epsilon = 10%, we will continue to spend 10% of the time doing suboptimal things.

What's the best way of keeping track of the rewards? General method: the mean

Sample mean = Sum of samples values / total number of samples

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

However, you need to keep track of all the samples which can be computationally expensive. An alternative to make a more efficient calculation is to use the mean of N-1 sample:

$$\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{N-1} x_i + \frac{1}{N}X_N = \frac{N-1}{N}\overline{X}_{N-1} + \frac{1}{N}X_N$$

$$\overline{X}_N = \left(1 - \frac{1}{N}\right)\overline{X}_{N-1} + \frac{1}{N}X_N$$

Optimystical initial value method (another way of solving the explore-exploit dilema)

Suppose we know the true mean of a bandit is << 10. The idea is to pick a high ceiling as the initial value for the bandit mean estimate and then do the updates based on that. We say this is optimistic because the initial sample mean is "to good to be true". Since it is too good to be true the only will happen is for the value to go down.
Then we just need to do the greedy part. This will imply exploration as well because when the rewards decrease, the algorithm will choose bandits that haven't been explored yet and therefore have high mean reward still.

UCB1 (another way of solving explore-exploit)
Idea: Confidence bounds.
If you take the mean from 10 samples or 1000 samples, the confidence bounds in the first case are much wider than in the second.

Chernoff-Hoeffding bound:

$P\left\{\left|\overline{X}-\mu\right| \geq \varepsilon\right\} \leq 2exp\left\{-2\varepsilon^2 N\right\}$: Our confidence bounds change exponentially with the number of samples we collect (N).

$\overline{X}$ = Sample mean, μ = true mean, ε = arbitrary small number

Algorithm: we take the upper confidence bound to be the regular sample mean
$X_{UCBj} = \overline{X}_j + \sqrt{2\frac{lnN}{Nj}}$

This is equivalent to taking $\varepsilon = \sqrt{2\frac{lnN}{Nj}}$

$\overline{X}_j$ is the sample mean of the j-bandit, N = total number of times you played all the bandits, Nj = total number of times you played the j-bandit.

The idea is similar to the optimistic initial value method. We use the greedy strategy only but we are greedy not only with respect to the sample mean but also to the upper confidence bound of the sample mean.
The confidence bound is the ratio of ln(N) and Nj. If you play sub other(?) bandits many times but Nj is small, then the ratio is large, which makes the upper confidence bound higher, which will cause you to select this bandit to play. At the same time, if Nj is large, then the ratio is small, at that point the confidence bound will shrink.
The top key of this ratio is that the upper part grows slower than the lower part. That means that as you play all bandits more and more, all upper bounds will shrink and you will be only using

the samples means, which is ok since you have collected lots of data. And so we converge to a purely greedy strategy.

Bayesian method or Thompson sampling and sub sources (another way to solve the explore-exploit dilemma)

To motivate the idea behind the Bayesian method we want to look at confidence intervals again. We know intuitively that a sample mean calculated from 10 samples is not that accurate as the sample mean calculated from 1000 samples.
We can use the Central Limit Theorem* to say that the sample mean is approximately Gaussian with true mean = to the expected value of the random variable and variance = to the original variance/N. The reason that $\overline{X}$ has a distribution is because is essentially the sum of random variables and any function of random variables is also a random variable.

$$\overline{X} \sim Normal(\mu,\ \sigma^2/N)$$
$$X \sim Normal(\mu,\ \sigma^2)$$

* Central limit theorem: In probability theory, the CLT establishes that, in most situations, when independent random variables are added, their properly normalized sum tends towards a normal distribution even if the original variables themselves are not normally distributed.

The Bayesian paradigm takes this a step further. Instead of mu being fixed, mu becomes a random variable too. The basic idea behind the Bayesian paradigm is that data is fixed and the parameters are random. In the Bayesian paradigm parameters have distributions and what we would like to know is the distribution of the parameters given the data: Given some data I should be able to come up with a distribution based on the data that is more accurate that if I didn't have the data. We call this the *Posterior distribution* $p(\Theta \mid X)$.

In the Bayesian paradigm, the key tool is Bayes rule* so we can flip the posterior to find it in terms of the likelihood and prior.

$$p(\Theta \mid X) = \frac{p(X \mid \Theta)p(\Theta)}{p(X)} = \frac{p(X \mid \Theta)p(\Theta)}{\int p(X \mid \Theta)p(\Theta)d\Theta} \propto p(X \mid \Theta)p(\Theta) \ = \ likelihood \ \times prior$$

Bayes rule: describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

$$P\ (A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where A and B are events and P(B) is not 0.
$P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
$P(B \mid A)$ is also a condition probability: the likelihood of event B occurring given that A is true.
P(A) and P(B) are the probabilities of observing A dn B independently of each other; this is known as the marginal probability.
The evidence P(B) is usually considered a scaling term. Bayes theorem also states that is is equal to:

$p(B) = \int p(B \mid A)p(A)dA$ This is known as marginalization.

One disadvantage of the Bayesian method is that we have to choose the prior ourselves, and this can sometimes have a non negligible effect on the posterior.
There are some problems with this formula. One is that is very general, in the sense that a differential equation is very general. The integral in the denominator can't always be found so we need to find mathematical tricks in order to make it work out and it doesn't always work out.

Luckily, there are special sets of (likelihood, prior) pairs of distributions such that the posterior will have the same distribution than the prior. These are called *Conjugate priors.*

Example: How can the conjugate prior method be applied to click-through rates?

We know that click-through rates are like coin tosses so they have a Bernoulli likelihood. The conjugate prior for Bernoulli likelihood is the Beta distribution.