DS102 - Discussion 8

Wednesday, 30th October, 2019

1. In this question we will look at what happens when we construct an upper confidence bound algorithm using the Chebyshev bound (as opposed to the Chernoff Bound).

   We first set up the framework of a multi-armed bandit (MAB) problem. Suppose you have a set of K 'arms', $\mathcal{A} = \{1, 2, ..., K\}$, each with its own reward distribution $X_a \sim \mathbb{P}_a$ with mean $\mu_a = \mathbb{E}[X_a]$ and variance $\sigma_a^2 = \text{Var}(X_a)$. In these problems we do not know μ_a but we would like to efficiently find the arm with the maximum mean by creating an algorithm that balances *exploration* of the arms with *exploitation* of the best possible arm. The efficiency of the algorithm is measured by a theoretical quantity known as regret, which measures how well the algorithm performs in expectation against an 'oracle' that knows the means of all the arms and always pulls the arm with highest mean.

   From last lecture we saw that an efficient algorithm for doing this makes use of upper confidence bounds. The general formula was for constructing an upper confidence bound for the true mean μ_a of an 'arm' a, given n samples $X_{a,1}, ..., X_{n,1}$, was to find a value of $C_a(\delta, n)$ such that:

   $$P(\mu_a < \hat{\mu}_a + C_a(\delta, n)) > 1 - \delta,$$

   where $\hat{\mu}_a$ is the sample mean given by $\hat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n} X_{a,i}$.

   (a) Suppose that you only knew that the maximum variance of any arm is upper bounded by σ^2. That is:

   $$\max_{a \in \mathcal{A}} \sigma_a^2 < \sigma^2$$

   Construct an upper confidence bound for the mean of arm a, after observing n samples from arm a.

   Solution: Exactly as we have done before:

   $$P(\mu_a < \hat{\mu}_a + C_a(\delta, n)) > 1 - \delta$$
   $$P(\hat{\mu}_a - \mu_a > -C_a(\delta, n)) > 1 - \delta$$

   Now we know that we can use the Chebyshev bound on:

   $$P(\hat{\mu}_a - \mu_a < -C_a(\delta, n)) \leq \delta$$

   $$P(\hat{\mu}_a - \mu_a < -C_a(\delta, n)) \leq \frac{\sigma_a^2}{nC_a(\delta, n)^2}$$

   $$P(\hat{\mu}_a - \mu_a < -C_a(\delta, n)) \leq \frac{\sigma^2}{nC_a(\delta, n)^2}$$

solving gives:

$$C_a(\delta, n) = \sqrt{\frac{\sigma^2}{n\delta}}$$

(b) For each arm $a \in \mathcal{A}$, define the number of times arm a has been pulled up to and including time t as $T_a(t)$. For a given δ, what rule does the modified upper confidence bound algorithm use to choose an arm A_t at each iteration t?

Solution:

$$A_t = \underset{a \in \mathcal{A}}{\text{argmax}} \quad \hat{\mu}_a + \sqrt{\frac{\sigma^2}{T_a(t-1)\delta}}$$

(c) Should this modified UCB algorithm out-perform (meaning have lower regret that) the original UCB algorithm which was based around Chernoff/Hoeffding Bounds when compared? Explain intuitively why or why not. Recall that the original UCB algorithm had a confidence bound given by:

$$C_a\left(n, \frac{1}{N^2}\right) = \sqrt{\frac{4}{n}\log N},$$

where N was the time horizon of the problem. This resulted in a regret that grew at a sublinear rate of $\log n$. We will investigate this more in the next Lab.

Solution: This should not out-perform the original UCB, since by the same heuristic argument the Chebyshev version of UCB does not gurantess a sublinear rate.

2. For this question we will define some basic information theoretic quantities, furthermore we will prove the chain rule of mutual information since it will be useful for future lectures. If you would like to get an intuitive understanding of these quantities the blog post at this link provides a great explanation, although it isn't necessary for the course: https://colah.github.io/posts/2015-09-Visual-Information/.

   Recall that the entropy of a random variable X is defined as

$$H(X) = -\mathbb{E}[\log \mathbb{P}(X)].$$

   Remember that from our lectures on Logistic Regression, the entropy captures a notion of the amount of 'randomness' in a random variable.

   We can also define a notion of the conditional entropy of Y given X as defined by:

$$H(Y|X) = -\mathbb{E}[\log \mathbb{P}(Y|X)],$$

and the joint entropy of Y and X is

$$H(X, Y) = -\mathbb{E}[\log \mathbb{P}(X, Y)].$$

Prove the chain rule of mutual information, which states that

$$H(X, Y) = H(X) + H(Y|X)$$

Solution:

$$\begin{aligned} H(X, Y) &= -\mathbb{E}[\log \mathbb{P}(X, Y)] \\ &= -\mathbb{E}\left[\log\left(\mathbb{P}(Y|X)\mathbb{P}(X)\right)\right] \\ &= -\mathbb{E}[\log \mathbb{P}(Y|X)] - \mathbb{E}[\log \mathbb{P}(X)] \\ &= H(Y|X) + H(X) \end{aligned}$$