DS 102 Discussion 8

Wednesday, October 28th, 2020

1 Concentration inequality for Poisson distribution

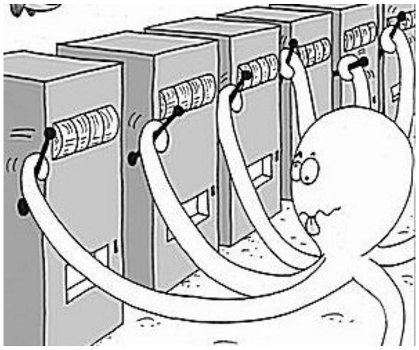
Let X be the sum of 20 i.i.d. Poisson random variables $X_1, \ldots, X_{20}$ with $\mathbb{E}[X_1] = 1$. Use the following techniques to upper bound $\Pr(X \geq 26)$.

1. Markov's Inequality
2. Chebyshev's Inequality
3. Chernoff Bound

(Hint: if X, Y are independent Poisson random variables with parameter λ_1, λ_2, then $Z = X + Y$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.).

2 Multi-Armed Bandits: UCB and Hoeffding's inequality

In the Tuesday's lecture, we began talking about multi-armed bandits. In the multi-armed bandits setting, we consider a decision-maker who is given K options to choose from. We refer to these options as "arms". Associated with each arm is a probability distribution over rewards. Initially, this distribution is unknown to the decision-maker. The decision-maker chooses an arm, usually referred to as pulling an arm, and receives a reward sampled from the corresponding reward distribution. This process is repeated over and over again.



The problem we want to solve is to decide which arm to pull at each time step. One possible algorithm for deciding which arm to pull is the Upper Confidence Bound (UCB) algorithm presented in lecture. If we assume that the reward of each arm is bounded (e.g. the slot machine returns between \$0 and \$100), then the UCB algorithm has bounded regret over time.

In this discussion, we study the derivation of the UCB algorithm using the Hoeffding bound. We will assume that the reward of each arm is bounded.

1. We first set up the framework of a multi-armed bandit problem. Suppose you have a set of K "arms", $\mathcal{A} = \{1, 2, ..., K\}$. Each arm $a \in \mathcal{A}$ has its own reward distribution $X_a \sim \mathbb{P}_a$ with mean $\mu_a = \mathbb{E}[X_a]$. Define the number of times arm a has been pulled up to and including time t as $T_a(t)$. 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.

   We will now derive the upper confidence bound that yields the UCB algorithm. The general formula for constructing an upper confidence bound for the true mean μ_a of an "arm" a, given $T_a(t)$ samples $X_a^{(1)}, ..., X_a^{(T_a(t))}$, is to find a value of $C_a(T_a(t), \delta)$ such that:

$$P(\mu_a < \hat{\mu}_{a,T_a(t)} + C_a(T_a(t), \delta)) > 1 - \delta \tag{1}$$

where $\hat{\mu}_{a,T_a(t)}$ is the sample mean of the reward from arm a, given by $\hat{\mu}_{a,T_a(t)} = \frac{1}{T_a(t)} \sum_{i=1}^{T_a(t)} X_a^{(i)}$. In words, Equation (1) says that with probability at least $1 - \delta$, the true mean μ_a is less than the estimated mean $\hat{\mu}_{a,T_a(t)}$ plus an upper confidence bound $C_a(T_a(t), \delta)$.

(a) Suppose that you know that the reward of any arm is between 0 and 1. That is:

$$X_a \in [0, 1]$$

Construct an upper confidence bound $C_a(T_a(t), \delta)$ for the mean of arm a, after observing $T_a(t)$ samples from arm a.

(b) Suppose we set $\delta = \frac{1}{t^3}$. This controls the probability that the true mean μ_a is greater than our upper confidence bound $C_a(T_a(t), \delta)$ on the estimated mean $\hat{\mu}_{a,T_a(t)}$. What rule does the UCB algorithm use to choose an arm A_t at each iteration t?