Exercises
Data Mining: Learning from Large Data Sets
HS 2016

Series 6, Dec 15th, 2016 (Bandits)

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.ethz.ch/teaching/dm-f16/`
Email questions to:
Johannes Kirschner, jkirschner@inf.ethz.ch

It is not mandatory to submit solutions and sample solutions will be published after one week. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise6` containing a PDF (LaTeX or scan) to `jkirschner@inf.ethz.ch` until Thursday, Dec 22th 2016.

Problem 1 (Analysis of UCB1):

In this exercise, we will prove a regret bound of the UCB1 algorithm, under the assumption that we know the total number of rounds T beforehand. We assume that there are k arms with random payoffs in $[0, 1]$ and means $\mu_1, \mu_2, \ldots, \mu_k$. We denote the optimal mean by $\mu^* = \max_{i=1}^k \mu_i$. Furthermore, let $\hat{\mu}_i^t$ be the empirical estimate of the mean μ_i at time t and denote by $\Delta_i = \mu^* - \mu_i$ the sub-optimality gaps. The full algorithm is given below.

Algorithm 1 UCB1 Policy for k-armed bandits with fixed T

```
function UCB1(T)
  Initialize: μ̂_i^0 = 0, n_i^0 = 0 for each i = 1, 2, ... k
  Play each arm once for initialization purpose and update μ̂_i^t and n_i^t
  for t = k + 1, ..., T do
    pick arm j ← argmax_i μ̂_i^t + sqrt(ln T / n_i^t)
    update count n_j^{t+1} ← n_j^t + 1 and mean estimate μ̂_j^{t+1} ← μ̂_j^t + (y^t − μ̂_j^t) / n_j^t
  end for
```

1. We will prove that the expected regret $\mathbb{E}[R_T] = T\mu^* - \mathbb{E}[\sum_{t=1}^T y_t]$ of UCB1 after T rounds is at most

$$\mathbb{E}[R_T] \leq 4 \sum_{\Delta_i > 0} \frac{\ln(T)}{\Delta_i} + 5 \sum_{\Delta_i > 0} \Delta_i = O\left(\frac{k \ln(T)}{\min_i \Delta_i}\right) . \tag{1}$$

   (a) Denote by n_i^t the number of times arm i has been played until round t (note that this is a random variable). Show that the total expected regret can be written as $\mathbb{E}[R_T] = \sum_{i=1}^k \mathbb{E}[n_i^T]\Delta_i$.

   (b) Next, we define a confidence set $\mathcal{C}_i^t = \{\mu : |\mu - \hat{\mu}_i^t| \leq \sqrt{\frac{\ln(T)}{n_i^t}}\}$ for each arm i. Note that the UCB1 policy plays the arm with the largest upper bound of the confidence set. Use Hoeffding's inequality to show that

$$\mathbb{P}[\mu_i \notin \mathcal{C}_i^t] \leq \frac{2}{T^2} \quad \text{for any } t = 1, \ldots, T. \tag{2}$$

   (c) Let i^* denote the index of an optimal arm, ie $\mu_{i^*} = \mu^*$. Consider any suboptimal arm i and show that if $\mu_i \in \mathcal{C}_i^t$ and $\mu_{i^*} \in \mathcal{C}_{i^*}^t$ for all $t = 1, \ldots, T$, then $n_i^T \leq \frac{4\ln(T)}{\Delta_i^2} + 1$.

   (d) Use the probabilistic bounds above to bound the expected number of times $\mathbb{E}[n_i^T]$ a suboptimal arm i is played, and put everything together to obtain the desired regret bound.

2. The bound we derived in the first part of the exercise is called an *instance dependent* regret bound, as it contains the sub-optimality gaps Δ_i. In particular the bound degrades as $\Delta_i \to 0$. Use the regret decomposition and that $\mathbb{E}[n_i^T] \in O\left(\frac{\log(T)}{\Delta^2}\right)$ to prove the *worst-case* regret bound $\mathbb{E}[R_T] = O(\sqrt{kT\ln(T)})$.

Solution 1 (Analysis of UCB1):

1. (a) This is a simple observation. If we denote by a_t the arm chosen at step t and by y_t the observed payoff, we have that $\sum_{i=1}^k \mathbb{1}(a_t = i) = 1$ and and $n_i^T = \sum_{t=1}^T \mathbb{1}(a_t = i)$. Using the law of iterated expectation it follows that

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}\left[\sum_{t=1}^{T}(\mu^* - y_t)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{k}\sum_{t=1}^{T}\mathbb{1}(a_t = i)(\mu^* - y_t)\right] \\
&= \sum_{i=1}^{k}\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}[\mathbb{1}(a_t = i)(\mu^* - y_t)|a_t]] \\
&= \sum_{i=1}^{k}\sum_{t=1}^{T}\mathbb{E}[\mathbb{1}(a_t = i)(\mu^* - \mu_i)] \\
&= \sum_{i=1}^{k}\mathbb{E}[n_i^T]\Delta_i
\end{aligned}$$

(b) Here we show that the $\mathcal{C}_i^t$ are indeed confidence sets, such that the true parameter μ_i is contained at least with probability $1 - 1/T^2$. A direct application of Hoeffding's inequality gives

$$\mathbb{P}[\mu_i \notin C_i^t] = \mathbb{P}\left[|\mu_i - \hat{\mu}_i^t| > \sqrt{\frac{\ln(T)}{n_i^t}}\right] \leq 2\exp\left(-2n_i^t\frac{\ln(T)}{n_i^t}\right) = \frac{2}{T^2}$$

(c) A suboptimal arm i is only selected over an optimal arm i^* if

$$\hat{\mu}_i^t + \sqrt{\frac{\ln(T)}{n_i^t}} \geq \hat{\mu}_{i^*}^t + \sqrt{\frac{\ln(T)}{n_{i^*}^t}}$$

Now, if we assume that $\mu_i \in C_i^t$ and $\mu_{i^*} \in C_{i^*}^t$ for all t, we find that

$$\mu_i + 2\sqrt{\frac{\ln(T)}{n_i^t}} \geq \mu_{i^*}$$

Using the definition $\Delta_i = \mu^* - \mu_i$ and solving for n_i^t we find that whenever the arm i is selected, $n_i^t \leq 4\frac{\ln(T)}{\Delta_i^2}$. It follows that $n_i^T \leq 4\frac{\ln(T)}{\Delta_i^2} + 1$.

(d) Let A_i denote the event that $\mu_i \in \cap_{t=1}^T \mathcal{C}_i^t$ and $\mu_{i^*} \in \cap_{t=1}^T \mathcal{C}_{i^*}^t$. With the result from 1c), we find that

$$\mathbb{E}[n_i^T|A_i] \leq \frac{4\ln(T)}{\Delta_i^2} + 1 \tag{3}$$

Using the union bound on $\mathbb{P}[A_i{}^c] = \mathbb{P}[\mu_i \in \cup_t \mathcal{C}_i^{t\,c}$ or $\mu_{i^*} \in \cup_t \mathcal{C}_{i^*}^{t\;c}] \leq \sum_t \mathbb{P}[\mu_i \notin \mathcal{C}_i^t] + \sum_t \mathbb{P}[\mu_{i^*} \notin \mathcal{C}_{i^*}^t]$ and with 1b) we find that $\mathbb{P}[A_i{}^c] \leq 2T \cdot \frac{2}{T^2} = \frac{4}{T}$. By the fact that $n_i^T \leq T$ and the law of iterated expectations we find

$$\mathbb{E}[n_i^T] = \mathbb{E}[n_i^T|A_i]\mathbb{P}[A_i] + \mathbb{E}[n_i^T|A_i{}^c]\mathbb{P}[A_i{}^c] \leq \frac{4\ln(T)}{\Delta_i^2} + 5 \tag{4}$$

Finally we use the regret decomposition found in 1a) to prove the final result

$$\mathbb{E}[R_T] = \sum_{\Delta_i > 0} \Delta_i \mathbb{E}[n_i^T] \leq 4 \sum_{\Delta_i > 0} \frac{\ln(T)}{\Delta_i} + 5 \sum_{\Delta_i > 0} \Delta_i \tag{5}$$

2. Let $\Delta > 0$ to be chosen later. Using the regret decomposition from part 1, we find that for some constant $C > 0$,

$$\begin{aligned}
\mathbb{E}[R_T] &= \sum_{\Delta_i < \Delta} \Delta_i \mathbb{E}[n_i^T] + \sum_{\Delta_i \geq \Delta} \Delta_i \mathbb{E}[n_i^T] && (6)\\
&\leq T\Delta + C \sum_{\Delta_i \geq \Delta} \frac{\log(T)}{\Delta_i} && (7)\\
&\leq T\Delta + C\, k \frac{\log(T)}{\Delta} && (8)
\end{aligned}$$

Choosing $\Delta = \sqrt{\frac{k \log(T)}{T}}$ completes the proof.