Homework 3

CSE 541: Interactive Learning
Instructor: Kevin Jamieson
Due 11:59 PM on May 29, 2024 (late homework not accepted)

Contextual Bandits

In this exercise we will implement several contextual bandit algorithms. We will "fake" a contextual bandit problem with multi-class classification dataset where each example is context, and the learner chooses an "action" among the available class labels, and receives a reward of 1 if the guess was correct, and 0 otherwise. However, keeping with bandit feedback, we assume the learner only knows the reward of the action played, not all actions.

We will use the MNIST dataset[1]. The MNIST dataset contains 28x28 images of handwritten digits from 0-9. Download this dataset and use the python-mnist library[2] to load it into Python. Rather than using the full images, you may run PCA on the data to come up with a lower dimensional representation of each image. You will have to experiment with what dimension, d, to use. Scale all images so that they are norm 1.

Let the d dimensional representation of the tth image in the dataset, c_t, be our "context." Our action set $\mathcal{A} = \{0, 1, \ldots, 9\}$ has 10 actions associated with each label. For each $i \in \mathcal{A} = \{0, 1, \ldots, 9\}$ define the feature map $\phi(c, i) = \text{vec}(c\mathbf{e}_i^\top) \in \mathbb{R}^{10d}$. If $v(c, a)$ is the expected reward of playing action $a \in \mathcal{A}$ in response to context c, then let us "model the world" with the simple linear model so that $v(c, a) \approx \langle \theta_*, \phi(c, a) \rangle$ for some unknown $\theta_* \in \mathbb{R}^{10d}$. Of course, when actually playing the game we will observe image features c_t as the context, choose an "action" $a_t \in \{0, \ldots, 9\}$, and receive reward $r_t = \mathbf{1}\{a_t = y_t\}$ where y_t is the true label of the image c_t and a_t is the action played.

Implement the Explore-Then-Commit algorithms, Follow-The-Leader, LinUCB, and Thompson Sampling algorithms for this problem. You can use just the training set of $T = 50000$ examples. The training set is class balanced meaning that there are 5000 examples of each digit. Important: randomly shuffle the dataset so the probability of any particular class showing up at any given time is 1/10. The algorithms work as follows:

- **Explore-Then-Commit** ("Model the world"): Fix $\tau \in [T]$. For the first τ steps, select each action $a \in \mathcal{A}$ uniformly at random. Compute $\widehat{\theta} = \arg\min_\theta \sum_{t=1}^{\tau} (r_t - \langle \phi(c_t, a_t), \theta \rangle)^2$. For $t > \tau$ play $a_t = \arg\max_{a \in \mathcal{A}} \langle \phi(c_t, a), \widehat{\theta} \rangle$. Make one plot showing performance for this method alone with a number of choices of τ (from unreasonably small to unreasonably large). Identify one of the best performing values and plot that performance curve against all the other methods (using their best hyperparameters) on the same plot.

- **Explore-Then-Commit** ("Model the bias"): Fix $\tau \in [T]$. For the first τ steps, select each action $a \in \mathcal{A}$ uniformly at random. Our goal is to identify a policy $\widehat{\pi} : \mathcal{C} \to \mathcal{A}$ using the dataset $\{(c_t, a_t, p_t, r_t)\}_{t \le \tau}$ such that

$$\begin{aligned}
\widehat{\pi} &= \arg\max_{\pi \in \Pi} \sum_{t=1}^{\tau} \frac{r_t \mathbf{1}\{\pi(c_t) = a_t\}}{p_t} \\
&= \arg\min_{\pi \in \Pi} \sum_{t=1}^{\tau} \frac{r_t \mathbf{1}\{\pi(c_t) \neq a_t\}}{p_t} \\
&= \arg\min_{\pi \in \Pi} \sum_{t \in [\tau] : r_t = 1} \mathbf{1}\{\pi(c_t) \neq a_t\}
\end{aligned}$$

  where the last line uses the fact that $p_t = 1/10$ due to uniform exploration and the definition of r_t. Note that this is just a multi-class classification problem on dataset $\{(c_t, a_t)\}_{t \in [\tau] : r_t = 1}$ where one is trying to identify a classifier $\widehat{\pi} : \mathcal{C} \to \mathcal{A}$ that predicts label a_t from features c_t. Train a 10-class linear logistic classifier[3] $\widehat{\pi}$ on the data up to time $[\tau]$ and then for $t > \tau$ play $a_t = \arg\max_{a \in \{0,\ldots,9\}} \widehat{\pi}(c_t)$. Make

[1] http://yann.lecun.com/exdb/mnist/

[2] https://pypi.org/project/python-mnist/

[3] Please feel free to use an off-the-shelf method to train logistic regression such as https://scikit-learn.org/stable/auto_examples/linear_model/plot_iris_logistic.html#sphx-glr-auto-examples-linear-model-plot-iris-logistic-py

one plot showing performance for this method alone with a number of choices of τ (from unreasonably small to unreasonably large). Identify one of the best performing values and plot that performance curve against all the other methods (using their best hyperparameters) on the same plot.

- **Follow-The-Leader**: Fix $\tau \in [T]$. For the first τ steps, select each action $a \in \mathcal{A}$ uniformly at random. For $t > \tau$ play $a_t = \arg\max_{a\in\mathcal{A}} \langle \phi(c_t, a), \widehat{\theta}_{t-1} \rangle$ where $\widehat{\theta}_t = \arg\min_\theta \sum_{s=1}^t (r_s - \langle \phi(c_s, a_s), \theta \rangle)^2$. Make one plot showing performance for this method alone with a number of choices of τ (from unreasonably small to unreasonably large). Identify one of the best performing values and plot that performance curve against all the other methods (using their best hyperparameters) on the same plot.
- **LinUCB** Using Ridge regression with an appropriate regularization parameter $\gamma > 0$ ($\gamma = 1$ may be okay) construct the confidence set $\mathcal{C}_t$ derived in class (and in the book). At each time $t \in [T]$ play $a_t = \arg\max_{a\in\mathcal{A}} \max_{\theta\in\mathcal{C}_t} \langle \theta, \phi(c_t, a) \rangle$.
- **Thompson Sampling** Fix $\gamma > 0$ ($\gamma = 1$ may be okay). At time $t \in [T]$ draw $\widetilde{\theta}_t \sim \mathcal{N}(\widehat{\theta}_{t-1}, V_{t-1}^{-1})$ and play $a_t = \arg\max_{a\in\mathcal{A}} \langle \widetilde{\theta}_t, \phi(c_t, a) \rangle$ where $\widehat{\theta}_t = \arg\min_\theta \sum_{s=1}^t (r_s - \langle \theta, \phi(c_s, a_s) \rangle)^2$ and $V_t = \gamma I + \sum_{s=1}^t \phi(c_s, a_s)\phi(c_s, a_s)^\top$.

Implement each of these algorithms and show a plot of the regret (all algorithms on one plot with their best hyperparameters) when run on MNIST for good choices of τ, γ. Hint, for computing V_t^{-1} efficiently see `https://en.wikipedia.org/wiki/Sherman%E2%80%93Morrison_formula`.

Modeling and solving real-world problems

In this course you have learned a lot of technical skills and algorithms for abstract math problems. Now its time to apply those skills to solve real problems. Below, a number of scenarios are described that you will need to model mathematically and propose a solution for. Your solution should address the following questions (with justification):

- Is the task a pure-exploration (i.e., best-arm identification) or reward accumulation (i.e., regret minimization) task?
- What are the *actions*, *rewards*, and (if applicable) *contexts* or *states*?
- What assumptions are you making about the setting? Propose a sanity check experiment to justify the assumptions.
- Is the problem best described as a k-armed multi-armed bandit, linear bandit, non-parametric bandit[4], contextual bandit, or finite horizon Markov Decision Process?
- What is a specific algorithm you will use to address the problem (e.g., Linear-UCB or propose something custom if necessary)?
- If your solution relies on features, how will you construct them or where will you get them from?
- What is a very simple baseline that you will compare your algorithm to as a sanity check?
- If your proposed solution appears to not be performing much better than the baseline, what might you try changing?

Several sentences are expected per scenario with succinct answers to the above. Be **specific** with your answers. Another engineer familiar with the concepts of this course should be able to take your proposal and implement it.

[4] For example, the Lipschitz reward function like in problem 4.1 of HW2.

1. NASA is designing a new Mars rover, which harvests fuel from chemicals stored in Martian rocks. The scientists have no idea which rocks store these chemicals, but they do have instruments on board the rover that can give them some information about each rock. NASA plans to send this rover with an initial supply of fuel, which it will use to investigate promising rocks. As the rover traverses its environment, it analyzes nearby rocks with its instruments. After analyzing each rock, the rover can choose whether to consume the rock (which takes fuel, but may give the robot its much-needed chemicals), or to pass over the rock. That is, the robot has no agency in where it goes next, it only decides whether to consume the rock or not when presented with a sequence of rocks. NASA has asked you to design an algorithm for determining which rocks in the sequence to consume.

2. Consider the same NASA Mars rover setting as above. But after many years, the rover's solar panels have been covered with dust and therefore only allows for a short amount of time to explore each day. Communication has also been cut so the rover must make its own decisions on where to move the rover each day before returning to its primary charging location. The goal is to consume as many chemicals as possible over time.

3. The Department of Energy (DoE) is looking for new ways to turn spent nuclear fuel into glass (so-called "vitrification" techniques). There are three primary inputs to a vitrification process: temperature, pressure, and ratio of silica to nuclear waste. The success of a process is measured by the fraction of waste that converts to glass. The DoE would like to find the most effective combination of inputs. Further, they are willing to assume that the relationship between the inputs and the effectiveness of the process is dominated by simple, low-order effects[5]. How should the DoE decide which combinations of the inputs to test in order to come up with a good solution using as few measurements as possible?

4. Your office mate has suddenly acquired a strong love for canned tuna and buttery popcorn, and also a new puppy they won't stop showing you videos of. In response, you have decided to avoid the office and work at the several cafes on the Ave all day. You don't want to be in a nearly empty cafe, but you also don't want it to be so busy you cannot focus. The issue is that the amount of people throughout the day fluctuates requiring you to move every so often to maintain your happy state. Luckily, the traffic pattern at any given cafe throughout the day is pretty consistent. Without relying on external information (e.g., the business of a cafe based on Google traffic) how will you learn when and where to work at throughout the day?

5. Love Canal, located in New York State, was one of the worst environmental disasters in US history. The 70-acre site was used as a landfill for chemical waste in the 1940's, subsequently developed into a school and neighborhood in the 1950's, and started leaking toxic waste in the 1970's. Cleanup took over twenty years and cost $400 million. Suppose it's 1970 and you work for the Environmental Protection Agency, which managed the cleanup. Your goal is to clean up the most polluted acre of Love Canal. In order to find this most polluted acre, you can analyze soil samples from across the site of your choosing. Note that there is some inherent uncertainty when measuring parts-per-million concentrations of chemicals - two soil samples from the same spot can be expected to vary slightly in reported contamination. You know that since all the chemicals have had two or three decades to mix underground, soil pollution readings will be similar for nearby samples, but there is no reason to expect there is a single connected region of high pollution. How will you find this most contaminated region?

6. Your friend is starting a tulip farm in the Skagit Valley, and she needs your advice to grow the most bountiful crop. Her farm covers a wide variety of land - some is shaded, some is in full sun, some fields are mostly clay, others are mostly sand, etc. - and she has mapped this all in great detail. She wants to grow tulips on the entire farm, but knows that the watering and fertilizer needs of the tulips will vary as the conditions change day to day. She is willing to apply water and fertilizer to each part of the farm in whatever way you specify each day, and will keep you on for several growing seasons as a consultant. She is paid by the crop yield and wishes to make as much money as possible. What's your plan?

[5] For example, the true response is described by a function such as $f(T, P, R) = \alpha_0 + \alpha_1 T + \alpha_2 P + \alpha_3 R + \beta_1 TP + \beta_2 TR + \beta_3 PR$ for some unknown coefficients α_i, β_i

7. Your friend is a committed runner, and participates in a five mile race every Saturday. At the end of the year, he adds his race times together and compares them with the other runners in his running club. He knows that his race time is influenced by many factors, like what he ate the previous day, the training exercises he did that week, and how much water he drank before the race. He wants to improve his year-end total time, and wants you to choose a new set of actions each week that will help him do this. How will you help?

8. Chemists are interested in developing techniques to turn methane gas into more stable or useful compounds (e.g. methanol), but existing industrial processes in this space are characterized by high energy consumption and the use of dangerous or expensive reagents. In the search for new chemical pathways for these reactions, chemists are seeking inspiration from proteins found in methane-eating (methanotropic) bacteria. Suppose scientists have identified a new strain of methanotropic bacteria, and have isolated a thousand protein-coding genes from its genome. They want to understand which single gene (and its associated protein) is most important for methane conversion. To do this, they can run the following experiment: remove gene i from the bacteria, place the edited bacteria in a sealed container full of methane, and measure the methane concentration after five days. A higher methane content indicates a less efficient conversion, and implies that the removed gene was important to the methane conversion process. Of course, this experiment is also subject to experimental noise, so the scientists are prepared to replicate the experiment as many time as they need in order to find the best gene. How should they model this problem?

9. You run a successful ice cream shop with 32 flavors and take pride in your ability of finding the best flavor for each customer. Upon entering the ice cream shop, the customer is given a free sample of ice cream of your choosing on a small wooden spoon. If the customer likes it, they purchase a pint of that flavor. If the customer does not like it, you keep offering samples of your choosing until you find one that the customer likes. You are pretty good at this task and usually find a flavor that the customer likes within a few samples. Unfortunately, lumber costs have been skyrocketing lately and the wooden spoons used for the samples are destroying your profit margin. You have decided that it is critical that you identify the ice cream flavor that is favored by the majority of the population. That way you can suggest this single ice cream flavor to everyone on the first sample and hopefully avoid a follow-up sample. You hire a chemist who provides you with a $d = 128$ dimensional feature vector of each ice cream that you are assured perfectly aligns with a linear model reflecting preferences (i.e., there exists a vector $\theta_* \in \mathbb{R}^d$ such that the probability a customer likes the ith ice cream is Bernoulli$(\frac{1}{2} + \langle x_i, \theta_* \rangle)$). What do you do to minimize the number of customers necessary to identify the majority preferred ice cream?