Partner 1
Partner 2
Partner 3

Homework 3

CSCI 1550 / 2540
March 6th, 2025

Due: March 31, 2025
Remember to show your work for each problem to receive full credit.

Problem 1 [40 points]

k-Mean Clustering: Let $S = \{x_1, \ldots, x_n\}$ be a set of points in a high dimension Euclidian space. A k-mean clustering of S is a partition of S into k disjoint sets (clusters), $C_1, \ldots, C_k$, with centers $c_1, \ldots, c_k$,

$$c_j = \frac{1}{|C_j|} \sum_{x \in C_j} x.$$

$||x||^2$ is the standard Euclidean norm.
The cost of k-mean a clustering $C_1, \ldots, C_k$ solution, with centers $c_1, \ldots, c_k$, is

$$Cost(C_1, \ldots, C_k) = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - c_i||^2$$

where $||x||$ is the standard Euclidean norm. (If $x \in R^m$ then $||x||^2 = \sum_{i=1}^{m} x_i^2$.)
The goal is to find a clustering that minimizes the cost function. (In general the problem is NP-complete, but there are efficient approximations.) Here we show that we can speed-up the computation by solving the problem in a lower dimension space.

1. Prove that for any set of points $C \subset R^m$:

$$\sum_{x \in C} \sum_{y \in C} ||x - y||^2 = 2|C| \sum_{x \in C} ||x - c||^2, \quad \text{where} \quad c = \frac{1}{|C|} \sum_{x \in C} x.$$

   [Hint: prove first for dimension 1.]

   Solution:

$$\sum_{x \in C} \sum_{y \in C} ||x - y||^2 = 2|C| \sum_{x \in C} x^2 - 2 \sum_{x \in C} \sum_{y \in C} xy = 2|C| \sum_{x \in C} (x^2 - c^2) = 2|C| \sum_{x \in C} ||x - c||^2$$

2. Apply dimension reduction (JL-Lemma) to show that a cost of a solution for k-mean clustering in a low dimension vector space is within $1 \pm \epsilon$ times the cost of the corresponding solution in the original space. How small can the dimension of the projected space be?

   Solution: From part 1, the cost for k-mean clustering can be written as $1/2|C| \sum_{x \in C} \sum_{y \in C} \|x - y\|^2$. Take $\delta = \sqrt{1 + \epsilon} - 1 \in O(\epsilon)$. By JL-lemma, we can project $\{x_i\}$ down into a subspace of dimension $\Omega(\log n / \epsilon^2)$, $\{y_i\}$, so that

$$(1 - \delta)\|x_i - x_j\| \leq \|y_i - y_j\| \leq (1 + \delta)\|x_i - x_j\|$$

If we square and sum, we have

$$(1-\delta)^2 \sum_{x_i,x_j \in C} \|x_i - x_j\|^2 \leq \sum_{y_i,y_j \in C} \|y_i - y_j\|^2 \leq (1+\delta)^2 \sum_{x_i,x_j \in C} \|x_i - x_j\|^2$$

That is, we immediately get a solution in the projected space that is within a factor of $[(1-\delta)^2, (1+\delta)^2]$ of the original. I claim that this is contained within $[1-\epsilon, 1+\epsilon]$. The upper bound is clear by the choice of δ. It remains to prove that $(1-\delta)^2 \geq 1-\epsilon$. We show that $(1-\delta)^2 - (1-\epsilon) \geq 0$:

$$\begin{aligned}
(1-\delta)^2 - (1-\epsilon) &= (2-\sqrt{1+\epsilon})^2 - (1-\epsilon) \\
&= 5 + \epsilon - 4\sqrt{1+\epsilon} - (1-\epsilon) \\
&= 2(\sqrt{1+\epsilon} - 1)^2 \\
&\geq 0
\end{aligned}$$

as desired. We can check that this is tight from the upper bound, so $\Omega(\log n/\epsilon^2)$ is the smallest possible space we can project into.

Partner 1
Partner 2
Partner 3

Homework 3

CSCI 1550 / 2540
March 6th, 2025

Problem 2 [40 points]

1. Let $X(m, \epsilon)$ be a random variable with a Binomial distribution $B(m, \frac{1}{2} + \epsilon)$, for some $0 < \epsilon \leq \frac{1}{2}$. Prove that for any even $m \geq 2$ there is an $0 \leq \epsilon \leq \frac{1}{2}$, such that $Pr(X(m, \epsilon) \leq m/2) \geq 1/4$. [Hint: It's hard to compute an explicit value of ϵ that satisfies the requirement. Instead, write $Pr(X(m, \epsilon) \leq m/2)$ as a function of ϵ and use simple calculus to show that there is a value of ϵ that satisfies the requirement.]

   Solution: We may write

$$f(\epsilon) = Pr(X(m, \epsilon) \leq m/2) = \sum_{i=0}^{m/2} \binom{m}{i} (\frac{1}{2} + \epsilon)^i (\frac{1}{2} - \epsilon)^{m-i}$$

   Since m is fixed, we can see that this is a polynomial in ϵ. This implies that f is continuous in ϵ: because $f(0) = 1/2$ and $f(1/2) = 0$, we can apply the Intermediate Value Theorem to get there is some $\epsilon \in (0, 1/2]$ for which $f(\epsilon) = 1/4$, as desired.

2. Assume that we have k arms. $k-1$ of the arms give 0 with probability $1/2$ and 1 with probability $1/2$. The k-th arm gives 0 with probability $1/2 - \epsilon$ and 1 with probability $1/2 + \epsilon$. Show that the Explore and Commit has linear expected regret for any m and some choice of ϵ.

   Solution: For each of the first $k-1$ arms, we have $\mu_i = 1/2$ and $\mu_k = 1/2 + \epsilon$. As a result, $\Delta_i = \epsilon$ for $i < k$. The expected regret may now be written as

$$(k-1)m\epsilon + (T - km)\epsilon \cdot \sum_{j=1}^{k-1} \Pr\left(j = \arg\max_{i \in [k]} M_i\right)$$

   Since each of the $k-1$ arms are i.i.d., this can be rewritten as

$$(k-1)m\epsilon + (T - km)\epsilon(k-1) \cdot \Pr(k \neq \arg\max_{i \in [k]} M_i)$$

   To see the regret is exactly linear, we know by definition it is $O(T)$. To see it is $\Theta(T)$, note that

$$\Pr(k \neq \arg\max_{i \in [k]} M_i) \geq \Pr(M_k \leq m/2) \cdot \left(1 - (1/2)^k\right) \geq 1/4 \cdot \left(1 - (1/2)^k\right)$$

   where we are looking at the case that the kth arm is pulled $\leq m/2$ times and at least one of the other arms is pulled at least $m/2$ times. The $1/4$ follows from the choice of ϵ in part 1. This is enough to show the expected regret is bounded below by a linear function in T, which proves it is $\Theta(T)$.

Homework 3

Problem 3 [40 points]

Consider a doubling version of the UCB algorithm:

Algorithm UCB(α):

1. For all $i \in [k]$ activate arm i once, update M_i and set $N_i = 1$.
2. $s = k$
3. For $t = 1, \ldots, \log T$ do:
   (a) Activate 2^t times arm $j = \arg\max_{i \in [k]} \left(M_i + \sqrt{\frac{\alpha \log s}{2N_i}} \right)$.
   (b) Update M_j, and set $N_j = N_j + 2^t$, and $s = s + 2^t$

Prove that the expected regret of this algorithm is no more than:

$$\sum_{i \in [k],\ \Delta_i > 0} \left(\frac{2\Delta_i}{1 - 2^{-(\alpha-1)}} + \frac{4\alpha \log T}{\Delta_i} \right)$$

Solution: From the lecture, we know that if $N_i(t) > \frac{2\alpha \log s}{{\Delta_i}^2}$, we pull arm i at iteration t with probability $\leq 2 \cdot s^{-\alpha}$. Here, we know that at the start of iteration t that $s = k + 2^t - 2 > 2^t$, so the probability is bounded above by $2 \cdot 2^{-\alpha t}$. The total number of pulls until the probability of pulling arm i is $\leq 2 \cdot 2^{-\alpha t}$ is bounded by $2 \cdot \frac{2\alpha \log s}{\Delta_i^2} = \frac{4\alpha \log s}{\Delta_i^2}$, since we are able to pull one more time when $N_i(t) < \frac{2\alpha \log s}{\Delta_i^2}$. Once we are past this point, the expected number of pulls is

$$\sum_{N_i(t) > 2\alpha \log s / \Delta_i^2} 2^t \cdot 2 \cdot 2^{-\alpha t} \leq 2 \sum_{i=0}^{\infty} 2^{-(\alpha-1)t} = \frac{2}{1 - 2^{-(\alpha-1)}}$$

As a result, the expected regret is bounded by

$$\begin{aligned} \sum_{i \in [k], \Delta_i > 0} E[N_i(T)] \cdot \Delta_i &\leq \sum_{i \in [k], \Delta_i > 0} \left(\frac{4\alpha \log s}{\Delta_i^2} + \frac{2}{1 - 2^{-(\alpha-1)}} \right) \cdot \Delta_i \\ &= \sum_{i \in [k], \Delta_i > 0} \left(\frac{4\alpha \log s}{\Delta_i} + \frac{2\Delta_i}{1 - 2^{-(\alpha-1)}} \right) \end{aligned}$$

Finally, because s represents the total number of pulls, we get $s \leq T$ to obtain the bound of

$$\sum_{i \in [k], \Delta_i > 0} \left(\frac{4\alpha \log T}{\Delta_i} + \frac{2\Delta_i}{1 - 2^{-(\alpha-1)}} \right)$$

Homework 3

Problem 4 [25 points]

A parking-lot attendant has mixed up n keys for n cars. The n car owners arrive together. The attendant gives each owner a key according to a permutation chosen uniformly at random from all permutations. If an owner receives the key to their own car, they take it and leave; otherwise, they return the key to the attendant. The attendant now repeats the process with the remaining keys and car owners. This continues until all owners receive the keys to their cars.

Let R be the number of rounds until all car owners receive the keys to their cars. We want to compute $\mathbb{E}[R]$. Let X_i be the number of owners who receive their car keys in the ith round.

1. Prove that
$$Y_i = \sum_{j=1}^{i} (X_j - \mathbb{E}(X_j \mid X_1, \ldots, X_{j-1}))$$
is a martingale with respect to the X_i's.

   Solution: By definition, Y_i is a function of $X_1, ...X_i$.

   At each turn, we can return between 0 and n keys, so $(X_j - E\left(X_j | X_1, \ldots, X_{j-1}\right))$ is bounded above by n. Then $\mathbb{E}[Y_i] \leq \sum_{j=1}^{i} n < \infty$

   We also have that
$$\mathbb{E}[Y_{i+1} - Y_i | X_1, ...X_i] =$$
$$\mathbb{E}[(X_{i+1} - E\left(X_{i+1} | X_1, \ldots, X_i\right)) + \sum_{j=1}^{i} \left(X_j - E\left(X_j | X_1, \ldots, X_{j-1}\right)\right) - Y_i | X_1, ...X_i] =$$
$$\mathbb{E}[(X_{i+1} - E\left(X_{i+1} | X_1, \ldots, X_i\right)) | X_1, \ldots, X_i]$$
   By linearity, we have
$$\mathbb{E}[X_{i+1} | X_1, \ldots, X_i] - \mathbb{E}[\mathbb{E}[X_{i+1} | X_1, \ldots, X_i] | X_1, \ldots, X_i] = 0$$
   We conclude that Y_i is a martingale.

2. Use the martingale stopping theorem to compute $\mathbb{E}[R]$.

   Solution: To apply the MST, it is sufficient to show that $\mathbb{E}[R] < \infty$ and that $|Y_{i+1} - Y_i | X_1, ...X_i] < c$ for some constant c.

   We have that the probability that all keys are returned to their owners in a given round is at least $1/n!$. Consider the geometric r.v. G with parameter $1/n!$. We see that
$$\mathbb{E}[R] \leq \mathbb{E}[G] = n! < \infty$$

In addition, from part a), as $Y_{i+1} - Y_i = X_{i+1} - \mathbb{E}[X_{i+1}|X_1, ..., X_i]$ and we can only return between 0 and n keys,

$$\mathbb{E}[|Y_{i+1} - Y_i||X_1, ...X_i] < c] \leq \mathbb{E}[|X_{i+1}||X_1, ...X_i] + \mathbb{E}[|\mathbb{E}[X_{i+1}|X_1, ..., X_i]||X_1, ...X_i] \leq 2n$$

Therefore, we can apply the MST to show that

$$\mathbb{E}[Y_R] = \mathbb{E}[Y_1] = \mathbb{E}[X_1 - \mathbb{E}[X_1]] = 0$$

Applying linearity to $\mathbb{E}[Y_R]$, we get

$$\mathbb{E}[Y_R] = \sum_{j=1}^{R} \mathbb{E}[X_j] - \mathbb{E}[\sum_{j=1}^{R} \mathbb{E}[X_j|X_1, ..., X_{j-1}]] = 0$$

By definition of our stopping time, $\sum_{j=1}^{R} \mathbb{E}[X_j] = n$. By linearity of expectation, assuming there are m people at round j, each person gets their key with probability $1/m$, so letting M_i be the indicator variable for the ith person getting their key back after the first $j-1$ rounds,

$$\mathbb{E}[X_j|X_1, ..., X_{j-1}] = \mathbb{E}[\sum_{1}^{m} M_i] = 1$$

We conclude that $\mathbb{E}[R] = n$.