NAME	NIA	GRADE

Mining of Massive Datasets (2019-2020)

FINAL EXAM

WRITE YOUR ANSWERS <u>CLEARLY</u> IN THE BLANK SPACES. Please write clearly, as if you were trying to communicate something to another person who needs to understand what you write to be able to evaluate you properly. If an answer requires intermediate steps, please mark clearly your final response with a rectangle. If you answer with text, please underline the key words or phrases of your answer. If absolutely necessary, you can attach an extra sheet to your exam, indicating that the solution can be found in the extra sheet.

Problem 1 *1 point*

Suppose you are doing reservoir sampling with a reservoir of size 3 from the stream $a, b, c, d, e, \ldots$. Prove that $p_5(a) = p_5(b) = p_5(c) = p_5(d) = p_5(e) = 3/5$, with $p_5(u)$ being the probability that element u is in the sample after seeing 5 elements.

•$p_5(a) =$

•$p_5(b) =$

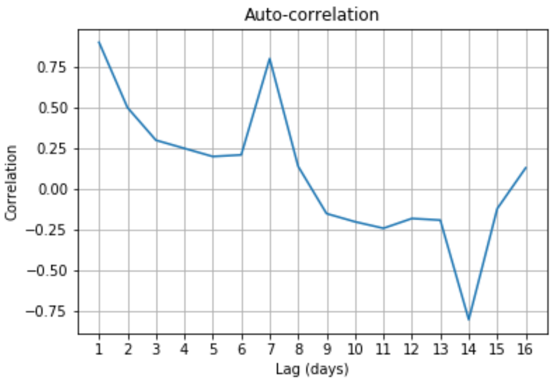
•$p_5(c) =$

•$p_5(d) =$

•$p_5(e) =$

Problem 2 *1 point*

Suppose the series x_t is seasonal and its auto-regressive plot is as shown on the right. Assume auto-correlation has a small absolute value for lags larger than what is shown in the figure.



1. What would be the lags you would use to create an auto-regressive model in this series, if you can use 3 lags? Justify briefly your answer.
   - Lags:
   - Justification:
2. How would you remove the weekly (7 days) seasonality of this series? Be precise and clear, using formulas when appropriate. Assume function $DayOfWeek(t)$ returns the day of the week $(1, 2, \ldots 7)$ of day t.

Problem 3 *2 points*

We would like to implement a Wi-Fi access point having the capability of denying access to known disruptive users, whose identifiers are assumed to be stored in an array W having N addresses. Suppose we will use a Bloom filter of M bits with two hash functions: h_1 and h_2.

1. Describe in pseudocode two functions: $m = \textit{initialize}(W)$ that creates and initializes bloom filter m, and $\textit{check}(W, m, x)$ that tests whether the address x is in the vector W, using the bloom filter m to accelerate the process. Assume you have function $\textit{lookup}(x, W)$ that indicates if x is in W.

   $\textit{initialize}(W)$: $\textit{check}(W, m, x)$:

2a. What is the probability of hash function h_1 not mapping any of the N input addresses to the first bit of the table?

2b. What is the probability that none of h_1, h_2 maps any of the N elements to the first bit on the table?

2c. What is fraction of bits that will be set to 1 in the table after $\textit{initialize}()$.

Problem 4 *1 point*

In non-negative matrix factorization (NMF) we decompose a matrix D into a user-matrix U and an item-matrix V such that $D \approx UV^T$. Suppose there are n users, m items, and we will use ℓ latent factors. Suppose set Ω contains the indices of the positions of D that are known, i.e., $(i, j) \in \Omega \iff$ user i has rated item j.

1. Write the objective function that is minimized by NMF.

2a. Describe what U_i (row i of matrix U) and V_j (row j of matrix V) contain.

- U_i contains

- V_j contains

2b. Indicate, with the formula and a brief justification, how to compute the predicted rating of user i over item j.

- $\tilde{D}_{i,j} =$

- Justification:

Problem 5 *2 points*

K-means (seen in the ML course) consists of P passes over the data. Items are first assigned random centroids, and on each pass: each centroid is calculated as the mean of the items associated to it, and each item is then assigned to its nearest centroid. Describe the general idea and in pseudocode a variant of the k-means algorithm that in P passes over the data, in addition to clustering elements $X = \{x_1, x_2, \ldots, x_N\}$ into C clusters, returns $M < N$ outliers. In this algorithm, possible outliers should be identified at each iteration of k-means, and centroids should not consider possible outliers in their computation. Name $cluster[x]$ the cluster to which x must belong, with $NULL$ indicating x is an outlier, $centroid[c]$ the centroid of cluster c. Assume you have function $CalcCentroid(S)$ that computes the centroid of set S.

- General idea:

- Pseucode of $KMeansVariant(P, X, C, M)$:

Problem 6 *1 point*

Consider the following transactions:

Transaction	Items
T1	a b c
T2	a b
T3	a b c d
T4	b d
T5	a c

1. What is the confidence and lift of the rule $a, b \Rightarrow c$?

- Confidence =
- Lift =

2. What is the confidence and lift of the rule $a \Rightarrow c$?

- Confidence =
- Lift =

Problem 7 *2 points*

We would like to develop a system to compute how similar are two people based on their musical taste. The input will be a dataset in which each person X_i with $i = 1, 2, \dots, n$ is represented a set of tracks they like $X_i = \{t_{i,1}, t_{i,2}, ..., t_{i,n_i}\}$, with $t_{i,j} \in T$ (the set of all tracks), and $n_i \geq 1$ the number of tracks liked by person i.

Let $p(t) = |\{i : t \in X_i\}|/n$ be the popularity of track t.

1. Provide the formula for a measure of similarity $sim(X_u, X_v) \in [0, 1]$ between two people X_u, X_v that makes more similar people who share tracks that are less popular. Justify briefly your formula.
   - $sim(X_u, X_v) =$
   - Justification:
2. Measure the similarity between users $X_1 = \{a, b\}$ and $X_2 = \{b, c\}$, and between X_2 and $X_3 = \{c, d\}$ if the popularities of tracks are $p(a) = 0.01$, $p(b) = 0.12$, $p(c) = 0.05$, $p(d) = 0.04$.
   - $sim(X_1, X_2) =$
   - $sim(X_2, X_3) =$