Playground Scoring*

!Optimizer

Sharif Optimization and Applications Laboratory
Department of Mathematical Sciences
Sharif University of Technology

May 2022

1 File Formats, Submissions and Scoring

The !Optimizer competition for 2022 consists of four different rounds, each concentrating on a prefixed setup for multi-manifold clustering in four possibly different scenarios and datasets. Descriptions and details of the questions for each round as well as what is expected to be submitted by participating teams are explained in Section 2. For each one of the rounds, all four datasets will be made available to all teams at the beginning of that specific round of the competition and each team may have as many submissions as they wish for each dataset, until the time for that specific competition round is up (the timetable of competition rounds will be announced before the competition starts in July 2022. For input/output file formats see Section 1.1). During each competition round, only the highest score for each dataset and for each one of the participating teams will be displayed according to their submissions on the competition leader-board that will be updated based on latest submissions in real-time, visible to all participants (for competition's scoring policy see Section 1.3).

1.1 File formats

All required input files will be made available to all participating teams in compressed text formats via the official website according to the announced timeline (in addition to the compressed text format, when necessary, input files may also be made available to participants in some other formats that may facilitate reading data). All submissions (i.e. output files) must be uploaded,

Playground scoring procedure for *Optimizer Competition* 2022. For problem description please go to (http://optimizer.math.sharif.edu/).

in a text format, to the announced portal for the online judge system (details of the timeline for data releases and submission deadlines will be announced before the competition starts in July 2022. For input and output file formats and some samples see below).

The first line of the input file contains integers d, n, m, k, and ρ, respectively, separated by one space, if each one of these parameters are specified as part of the given data for that specific dataset (otherwise there is a symbol "_" in the file, indicating that the corresponding parameter must be evaluated and submitted in the output file as part of the teams response to that specific dataset, according to the description of the round in Section 2). The next line of the input file contains m integers, indicating k_i's (if m is not specified in the input, this line is left empty.). After that, each one of the next n lines contains d real numbers specifying coordinates of the corresponding input vector.

Each submission of teams, as an output file, should follow the strict specifications explained below as the output file format (note that any format error by the judge will result in exclusion of the corresponding submission! For samples see Section 1.2).

The output file starts with a line containing the integers n and m, respectively, separated by one space. After this there is a list of m records, each containing the information of the ith manifold for $1 \leq i \leq m$, described as follows.

- The first line of the record for the ith manifold contains parameters d_i, k_i, and t_i, respectively. The parameters d_i, k_i are integers defined in Section 2. The parameter t_i is a word representing the type of the manifold which is either "Shpere" for spherical manifolds (**note that this type must be used for both balls and spheres**), "Affine" for affine subspaces, or "Complex" for more complex manifolds.

- The next part of the record is the geometric specification of the ith manifold that depends on the type of the manifold, specified by t_i. Here are the geometric specifications for different values of t_i:

  - t_i = Affine:
    The specification of an affine d_i-dimensional affine subspace of the ambient space must be provided by $d - d_i + 1$ lines containing the necessary information characterizing $d - d_i$ equations $\langle a_j, x\rangle = b_j$ with $x \in \mathbb{R}^d$ and $1 \leq j \leq d - d_i$ for an **orthonormal** set of vectors $\{a_j \in \mathbb{R}^d \mid 1 \leq j \leq d - d_i\}$, in such a way that, the jth line contains the coordinates of the vector a_j, and proceeding these $d - d_i$ lines, there must be a new line containing $d - d_i$ real numbers b_j, separated by one space. For the case $d = d_i$ just leave a blank line.

  - t_i = Sphere:
    The specifications of a spherical manifold residing in an affine d_i-dimensional subspace, starts with specifications of the affine subspace in $d - d_i + 1$ lines (if $d = d_i$ then leave a blank line), followed by a line containing the coordinates of the center as a point in $\mathbb{R}^d$, separated

by one space, as well as and the value of the radius of the sphere (all in one line).

- ⋆ t_i = Complex:
  In this case there is no specification for the manifold and the record contains no line for this case.

- Information of k_i clusters follows the specifications of the ith manifold with the following format. Information of the jth cluster (for $1 \leq j \leq k_i$) appears in a separate line, starting with the number of vectors (in the jth cluster) followed by the *index* of input vectors assigned to this cluster. Note that the *index* of an input vector is its order of appearance in the input file, starting from 1 (i.e. the index of the first vector in the input file).

- The last record is dedicated to the information of outliers. This record starts with an integer indicating the number of outliers ρ, followed by the indices of the input vectors that are determined as outliers, all separated by one space. If you do not find any outlier (i.e. $\rho = 0$), this record just contains a single "0". (Note that each index of datapoints must appear exactly once in the union of your clustering and your outlier set.)

Please refer to Section 1.2 for some examples of input and output files.

1.2 Input/Output samples

Note that sample input and outputs presented here are only provided to show the formatting of the input and outputs. So following sample inputs may not be compatible with requirements of the rounds of the competition and also following sample outputs may represent not good solutions for the problem specified in the corresponding sample input.

Input #1:	Description of input
`2 5 2 _ 0`	$d = 2$, $n = 5$, $m = 2$, k is not specified, and $\rho = 0$.
`_ _`	Values k_i are not specified.
`0 0`	Vector #1:
`0 1`	Vector #2:
`10 10`	Vector #3:
`1 0`	Vector #4:
`1 1`	Vector #5:

Output #1:	Description of output
`5 2`	$n = 5$ and $m = 2$.
`2 1 Affine`	An affine manifold in $\mathbb{R}^2$ having $k_1 = 1$ cluster.
`4 1 2 4 5`	This cluster contains 4 vectors: 1, 2, 4, and 5.
`0 1 Sphere`	A sphere manifold in $\mathbb{R}^0$ having $k_2 = 1$ cluster.
`1 1`	$a_{2,1} = (1, 1)$
`1 -1`	$a_{2,2} = (1, -1)$
`20 0`	$b_{2,1} = 20$ and $b_{2,2} = 0$.
`10 10 0`	Center of sphere is $(10, 10)$, and its radius is 0.
`1 3`	This cluster contains 1 vector: 3.
`0`	There is no outlier ($\rho = 0$).

Input #2:	Description:
`2 6 _ _ 1`	$d = 2$, $n = 6$, m and k are not specified, and $\rho = 1$.
	Values k_i are not specified.
`0 0`	Vector #1:
`0 1`	Vector #2:
`10 10`	Vector #3:
`1 0`	Vector #4:
`10 11`	Vector #5:
`11 10`	Vector #6:

Output #2:	Description:

6 2	$n = 6$ and $m = 2$.
2 3 Complex	A complex manifold in $\mathbb{R}^2$ having $k_1 = 3$ cluster.
1 1	This cluster contains 1 vectors: 1.
1 2	This cluster contains 1 vectors: 2.
1 4	This cluster contains 1 vectors: 4.
2 1 Complex	A complex manifold in $\mathbb{R}^2$ having $k_1 = 1$ cluster.
2 3 5	This cluster contains 2 vectors: 3, and 5.
1 6	There is one outlier ($\rho = 1$): vector 6.

1.3 Scoring

The scoring procedure in each case contains the following three steps.

1) **Format verification:**

   The format of the submitted file is verified to match what has already been specified in Section 1. In case of any kind of format error, the total score for the submission will be set to **zero**.

2) **Input data verification:**

   Then, the submission is verified to contain and match the information provided in the input file as the input dataset and what is asked for within the specific round. **The total score of a submission that is not compatible with this information is set to be equal to zero**. In particular, any submission must be a hard/crisp clustering of the dataset, i.e. each data point must appear exactly in one of the parts of a clustering (including the outlier set if there is any).

3) **Scoring:**

   In the final stage, the score of a submission is evaluated according to the sum of scores related to *parameter compatibility* and *cluster compatibility* of the submission, compared to the *ground truth clustering* which is available to the judge. In this scoring procedure, which is described explicitly in what follows, the constants ε, σ, ω_d, ω_I, ω_M, ω_C, ω_k, and ω_O are determined by the judge for each dataset.

   For two given numbers a and b, and two given subsets C and C', let us define

$$s(a,b) \stackrel{\text{def}}{=} e^{-\sigma\left(\frac{a-b}{\max(\varepsilon,|a|)}\right)^2},$$

$$s^+(a,b) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } b < 0, \\ b/a & \text{if } b < a, \\ s(a,b) & \text{o.w.} \end{cases}$$

   and

$$J(C,C') \stackrel{\text{def}}{=} \frac{|C \cap C'|}{|C \cup C'|},$$

which is the Jaccard index for the clusters C and C'. Also, for a given manifold M of dimension d, specific parameters I (see below for the definition) and a ground-truth clustering

$$\mathcal{C} = (C_1, C_2, \ldots, C_k),$$

the score of a submitted clustering

$$\mathcal{C}' = (C'_1, C'_2, \ldots, C'_{k'})$$

with the submitted dimension d' and specific parameters I', is defined as

$$s(M, M') \stackrel{\text{def}}{=} \omega_I \pi_{M,M'} + \omega_d s^+(d, d') + \omega_k s^+(k, k') + \omega_C \max_{\mathcal{P}} \sum_{(i,j)\in\mathcal{P}} \omega_{C_i} J(C_i, C'_j)$$

in which $\mathcal{P}$ is a maximum pairing (i.e. matching) in $\{1, \ldots, k\} \times \{1, \ldots, k'\}$ and the *parameter-score*, $\pi_{M,M'}$, is defined below.

Now, let

$$\mathcal{C} = (C_{1,1}, C_{1,2}, \ldots, C_{1,k_1}, C_{2,1}, C_{2,2}, \ldots, C_{2,k_2}, \ldots, C_{m,1}, \ldots, C_{m,k_m})$$

be the ground truth clustering in which

$$\mathcal{C}_i = (C_{i,1}, C_{i,2}, \ldots, C_{i,k_i})$$

is the induced clustering on M_i having k_i clusters. Then, the total score of a submitted clustering $\mathcal{C}'$ on a list of manifolds M'_i of size k'_i's for $1 \le i \le m'$ with dimensions d'_i's is defined to be

$$s(\mathcal{C}, \mathcal{C}') \stackrel{\text{def}}{=} \omega_M \max_{\mu} \sum_{(i,j)\in\mu} \omega_{M_i} \, s(M_i, M'_j) + \omega_O J(O, O'),$$

in which μ is a maximum pairing (i.e. matching) in $\{1, \ldots, m\} \times \{1, \ldots, m'\}$, and O and O' are the sets of points indicated as outliers in the ground truth dataset and in the submitted solution, respectively.

The parameter score $\pi_{M,M'}$ depends on the type t of the manifold M (from the ground truth) and type t' of the manifold M' (submitted to the judge) and is defined as follows:

- $t = t' =$ Affine:
  For an affine subspace determined by the equation $Ax = b$ (where rows of A are chosen to be orthonormal), we define $I \stackrel{\text{def}}{=} A^T[A \quad b]$, where $[A \quad b]$ is a row block matrix. Hence, again, for two given affine subspaces we similarly let

  $$\pi_{M,M'}(I, I') \stackrel{\text{def}}{=} \frac{1}{N} \sum_{(i,j)} s(I_{i,j}, I'_{i,j}),$$

  where N is the number of entries of the matrix I.

- If $t = t' =$ Sphere:

  For a sphere residing on an affine subspace having the parameter matrix I, let $\tilde{I}$ be the $d+1$ dimensional vector whose first coordinate is the radius of the sphere and the rest of the coordinates are the coordinates of the center of the sphere in $S = \mathbb{R}^d$. Then, for two given spheres we define

$$\pi_{M,M'}(I, \tilde{I}, I', \tilde{I}') \stackrel{\text{def}}{=} \left(\frac{1}{d+1} \sum_{i=1}^{d+1} s(\tilde{I}_i, \tilde{I}'_i) + \frac{1}{N} \sum_{(i,j)} s(I_{i,j}, I'_{i,j}) \right).$$

- $t = t' =$ Complex: $\pi_{M,M'} \stackrel{\text{def}}{=} 1$.
- Otherwise: $\pi_{M,M'} \stackrel{\text{def}}{=} 0$.

!Optimizer

Sharif Optimization and Applications Laboratory,
Department of Mathematical Sciences,
Sharif University of Technology,
Tehran, Iran.
optimizer.math.sharif.edu