SI100B, Spring 2018
Homework 2: Line Clustering via RANSAC

Due: April 4th, 23:59 p.m.

1 Description

The key concept behind this homework is Random Sample Consensus (RANSAC), which is covered in class. You can refer to the details of this algorithm on Wikipedia. With this algorithm, you can estimate the parameters of outliers and inliers of data points in a given space. In this homework, the data points you will be given are in $\mathbb{R}^2$. The inliers can be considered to be on the same line, and the outliers can be seen as noise points. Running RANSAC on the target dataset will do line clustering. This homework is split into 3 parts. The first two parts let you implement two different ways to compare the output inliers with the ground truth inliers, and the third part is the RANSAC alrogithm itself.

Part 1 In the first part, you need to design a function to estimate geometric error between two lines. The lines are represented by the line parameter vector, which is basically `[a, b, c]` in $ax + by + c = 0$. The geometric error is the euclidean distance between the two parameter vectors.

Part 2 In the second part, you should design a function to estimate clustering error between two sets of labeled points. The two sets have the same length, and each element in the set is an integer label, which represents the cluster ID. You should traverse all the permutations of the labels to find the minimum error between two sets.

Part 3 In the last part, you have to implement the RANSAC algorithm. Your program will be given some input parameters and a set of points in $\mathbb{R}^2$, and it should output the best fitted lines following the requirements of the parameters. You are recommended to implement this part step by step. First, you can design a function to calculate the distance from a point to a line. Then, you can write a function for a single run of RANSAC. After all these steps, you can implement the recursive version of RANSAC algorithm and output the line parameters you've found accordingly.

2 Requirements

Part 1 You will be given two lines, each line is represented by the line parameters `[a, b, c]` in the line equation $ax + by + c = 0$. The input has two file lines, each file line contains line parameters of the two lines. Your program should output the geometric error between these two lines. The result must be an floating point number with 8 decimal digits. An example input/output format is as following:

Input:
1 2 3
4 0 6

Output:
4.69041576

The two lines, $x + 2y + 3b = 0$ and $4x + 6 = 0$ has the geometric error 4.69041576.

Part 2 You will be given two sets of labeled elements. Each element is labeled by an integer, which represents the cluster that this element belongs to. In the same set, the same integer represents the same cluster. But this does not apply across different sets. The quantity of clusters in each set are the same. You need to output the (minimum) clustering error between the two sets. An example input/output format is as following:

Input:
1 1 1 1 2 2 2 3
0 0 0 0 3 3 2 2

Output:
2

The two lines denote two clusterings: $(0,1,2,3),(4,5,6),(7)$ and $(0,1,2,3),(4,5),(6,7)$. To calculate the difference, we traverse all the permutations. We find that $(4,5,6)$ and $(4,5)$ and 1 element difference, and (7) has 1 element difference with $(6,7)$. So the total error is 2, which is minimum.

Part 3 In the first line, you will be given a set of parameters `[n, eps, T, p, w]`. The first parameter `n` is the run times of RANSAC, which means you should find `n` sets of inliers. The second parameter `eps` is the threshold with which you can determine whether a data point is an inlier or not. The parameter `T` is the maximum iteration times in a single run of RANSAC. The parameter `p` and `w` are the probability of success and inliers quantity percentage. If these two parameters are not zero, you should calculate an optimal iteration times `k` of RANSAC, and take the smaller one between `k` and `T`. From the second line, you will be given the coordinate of a data point on each line, until the end of file. Output the normalized line parameters of your estimated inliers on line by line, and keep 8 digits. "Normalized" means that $a^2 + b^2 + c^2 = 1$. You should make sure that the first parameter is non-negative. If it is zero, keep the second parameter non-negative. Then, sort the results in ascending order, by the value priority of the first, second and the last parameter. An example input/output format is as following (partial):

Input:
1 1 300 0.99 0.75
901.67 60943.72
...

Output:

```
0.06591805 0.68489159 -0.72565716
```

You can refer to the detailed information of this part on the recitation and TA session.

3 Submit

Submit this homework on the online judging system. RANSAC is a random algorithm, though we have let you keep 8 digits to avoid minor difference between your output and the ground truth, it is still possible that you occasionally fail some test cases in part 3 even your program is correct. Try to resubmit for a new judge if you are in this situation.