Introduction to Data Science

Course 094201

Lab 3:

Clustering: K-Means

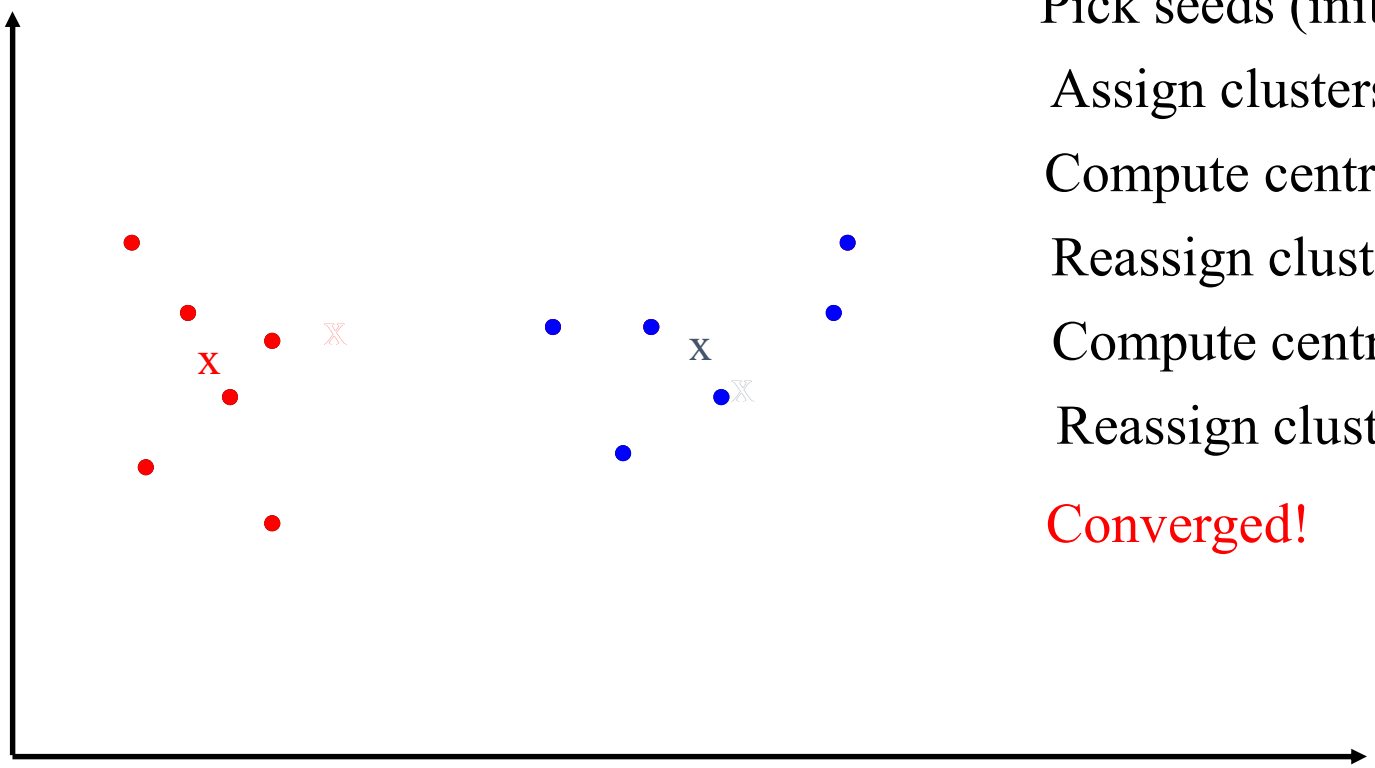
Spring 2019

K-Means

- Partitional clustering method
- Items are represented as points in a space. (we use points/items terminology interchangeably)
- We divide items to K clusters iteratively, until we find a partition that doesn't change
  - Each cluster is associated with a centroid (the arithmetic mean of clusters' items)
  - Each item is assigned to the cluster with the closest centroid
  - After the assignment the centroids are not necessarily correct, thus updated
  - Next we assign again the items to the updated centroids and so on

```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

2D Example



Pick seeds (initial centroids)

Assign clusters

Compute centroids

Reassign clusters

Compute centroids

Reassign clusters

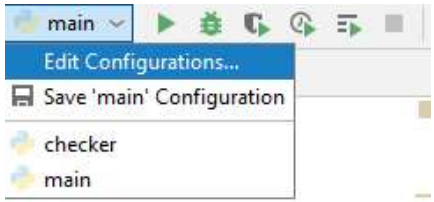
Converged!

The dataset and the code

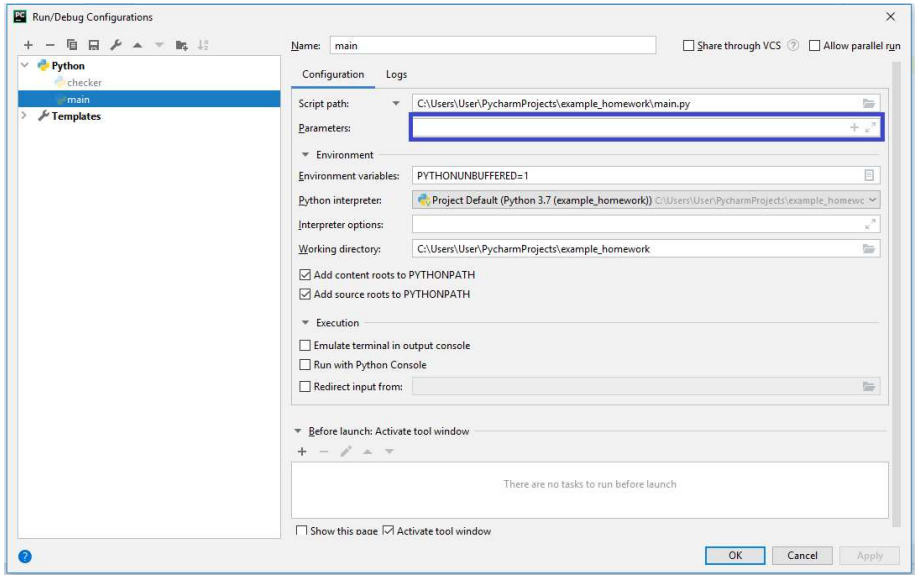
- Download lab 3 content from Moodle and unzip the code: lab3_for_students.zip
- The files **in1 and in2** in the datasets folder are very simple, small inputs to the algorithm.
- The file **colors.txt** contains color names and their RGB values. The dataset was created by showing colors to people and requesting to name the color. We expect that clustering the data based on RGB values will produce clusters of similar colors. The clusters can be used to disambiguate the common color name based on its RGB value or to find synonyms for different color names.

Assignment – Let's start

- What are the arguments for main and how do we set them in PyCharm?



- Here you can edit the arguments for your program (see next slide)



K-Means – Arguments

- The arguments for the program are:
  - K : number of clusters.
  - max_iterations : the limit for the number of iterations.
  - input : path for input file.
- Example: 5 10 ./datasets/in1
- The main class of the project is the KMeans class
- Try to identify which methods in the class KMeans are responsible for which steps in the algorithm

K-Means – add random_seed argument

- We need to add ability to provide different random seeds to the k-means

K-Means – Add loss computation

- Implement the method compute_loss for Cluster which computes total error(sum of distances to the centroid)
- Print average loss in print_results()