edX

> **Audit Access Expires May 11, 2020**
> You lose all access to this course, including your progress, on May 11, 2020.

2. K-means

K-means

1.0/1.0 point (graded)
For this part of the project you will compare clustering obtained via K-means to the (soft) clustering induced by EM. In order to do so, our K-means algorithm will differ a bit from the one you learned. Here, the means are estimated exactly as before but the algorithm returns additional information. More specifically, we use the resulting clusters of points to estimate a Gaussian model for each cluster. Thus, our K-means algorithm actually returns a mixture model where the means of the component Gaussians are the K centroids computed by the K-means algorithm. This is to make it such that we can now directly plot and compare solutions returned by the two algorithms as if they were both estimating mixtures.

Read a 2D toy dataset using `X = np.loadtxt('toy_data.txt')`. Your task is to run the K-means algorithm on this data using the implementation we have provided in `kmeans.py` Initialize K-means using `common.init(X, K, seed)`, where K is the number of clusters and `seed` is the random seed used to randomly initialize the parameters.

Note that `init(X,K)` returns a K-component mixture model with means, variances and mixing proportions. The K-means algorithm will only care about the means, however, and returns a mixture that is retrofitted based on the K-means solution.

Try $K = [1, 2, 3, 4]$ on this data, plotting each solution using our `common.plot`

function. Since the initialization is random, please use seeds $0, 1, 2, 3, 4$ to and select the one that minimizes the total cost. Save the associated plots (best solution for each K). The code for this task can be written in `main.py` .

Report the lowest cost for each K:

$\text{Cost}\big|_{K=1} =$ 5462.297452340001   ✔ **Answer: 5462.2974**

$\text{Cost}\big|_{K=2} =$ 1684.9079502962372   ✔ **Answer: 1684.9079**

$\text{Cost}\big|_{K=3} =$ 1329.5948671544297   ✔ **Answer: 1329.5948**

$\text{Cost}\big|_{K=4} =$ 1035.4998265394659   ✔ **Answer: 1035.4998**
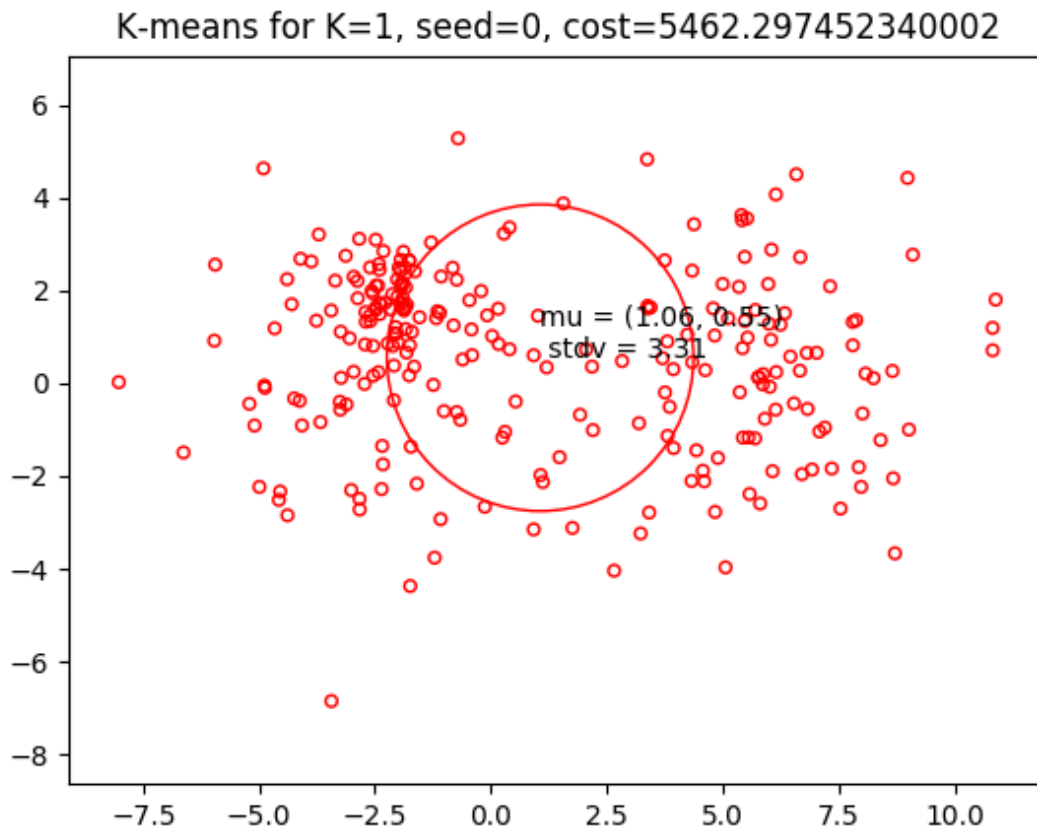
Solution:

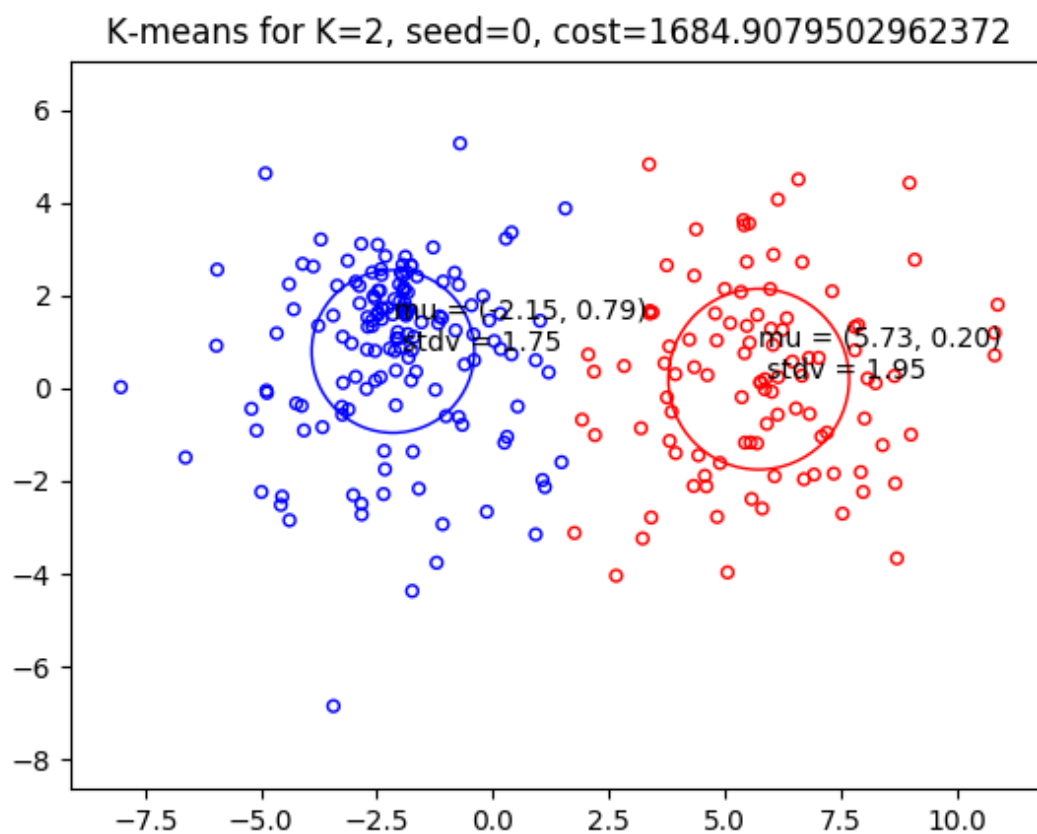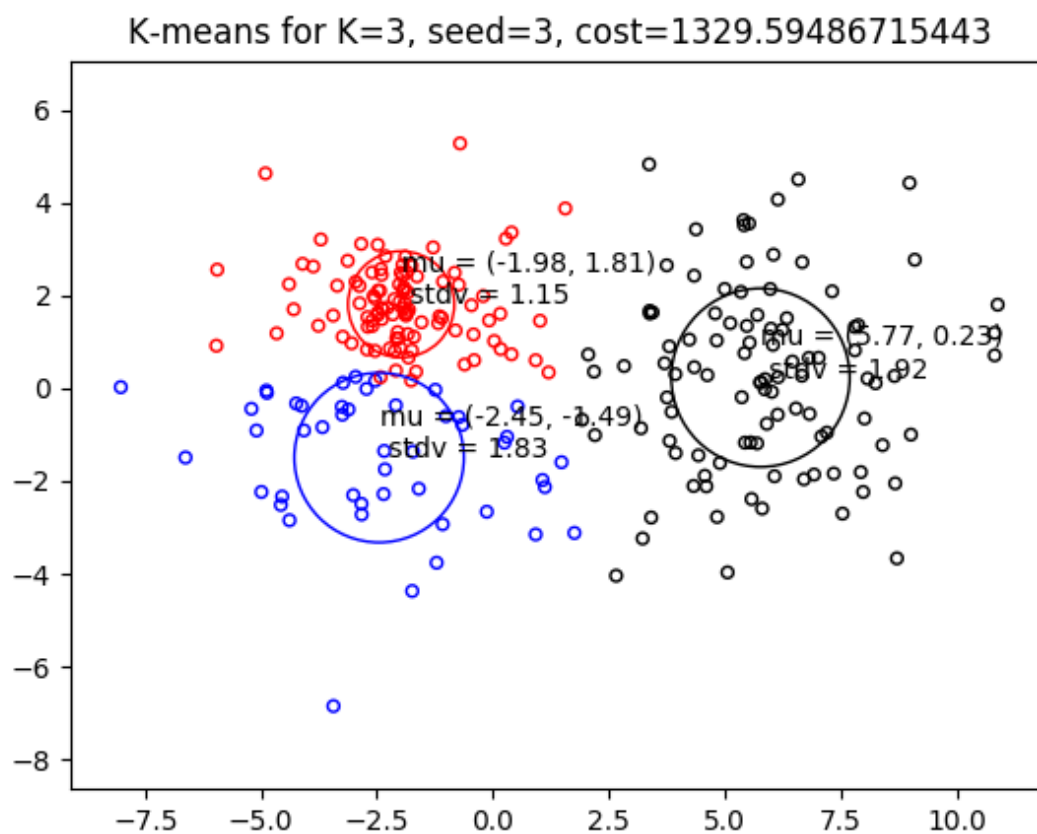Code:

```
def run_kmeans():
    for K in range(1, 5):
        min_cost = None
        best_seed = None
        for seed in range(0, 5):
            mixture, post = common.init(X, K, seed)
            mixture, post, cost = kmeans.run(X, mixture, post)
            if min_cost is None or cost < min_cost:
                min_cost = cost
                best_seed = seed

        mixture, post = common.init(X, K, best_seed)
        mixture, post, cost = kmeans.run(X, mixture, post)
        title = "K-means for K=, seed=, cost=".format(K, best_seed,
 min_cost)
        print(title)
        common.plot(X, mixture, post, title)
```
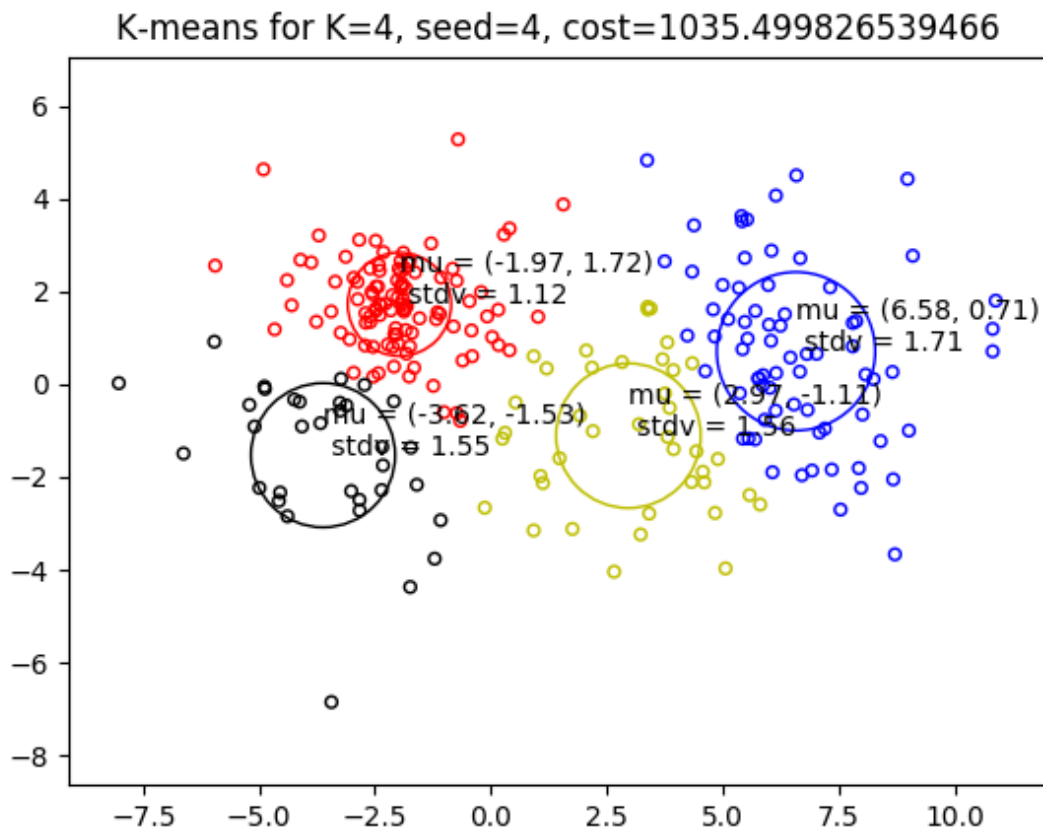
Plots:

K-means for K=1, seed=0, cost=5462.297452340002

K-means for K=2, seed=0, cost=1684.9079502962372

K-means for K=3, seed=3, cost=1329.59486715443

K-means for K=4, seed=4, cost=1035.499826539466

Submit   You have used 1 of 25 attempts

ℹ   Answers are displayed within the problem

Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Project 4:
Collaborative Filtering via Gaussian Mixtures / 2. K-means

Hide Discussion

Add a Post

Show all posts	by recent activity

💬 **seed**   5

how to set seed having five values of seed for only 4 values of k. Won't it influence the result ...

☑ **Need Help - Lowest Cost?**

Assuming we are using the cost function contained in the code - which "lowest cost" is the gr...

1 new

💬 **namedTuple**

This block of code is throwing an error inside common.py class GaussianMixture(NamedTupl...

4

💬 **Am I the only one that requires github posts to even try to understand what to do or the relation between videos and a script????**

Am I the only one that requires github posts to even try to understand what to do or the relat...

5

☑ **Strange Python function header**

I've noticed a kinda strange function headers in means.py, such as `def estep(X: np.ndarray, ...

2

💬 **Docstring in kmeans.run() is missing an argument**

The docstring in kmeans.run reads as: Args: X: (n, d) array holding the data post: (n, K) array h...

👤 Community TA

2

❓ **What is the use of 'post' return from init?**

It is not used in estep, and m step uses the one from estep, so why is 'post' returned from init?

2

💬 **Just a gentle feedback..**

The problem could have been stated in a much simpler way and the code could be way simpl...

2

☑ **"post" parameter in common.plot?**

Any hints on what this refers to? Thanks! JAC

3

💬 **kmeans attribute mu**

I get an error when trying to call kmeans.estep: AttributeError: 'tuple' object has no attribute '...

2

☑ **Cost function?**

What cost function should we use? Euclidean?

4

❓ Reference on Python def code