Kai Wong

UID: 704451679

CS 145 Fall 2018, Homework 4

**1. Clustering Evaluation**

**Purity:**

N = 20

|  |  |  |
| --- | --- | --- |
| Labels | Predicted IDs | Ground Truth IDs |
| 1 | 10,12,14,16,18 | 3,4,5,13,17 |
| 2 | 1,2,7,8,15,17 | 10,12,14,16,18 |
| 3 | 3,4,5,9,13 | 1,2,7,8,15 |
| 4 | 6,11,19,20 | 6,9,11,19,20 |

Match predicted cluster labels to ground truth cluster labels:

**1 🡪 2, 2 🡪 3, 3 🡪 1, 4 🡪 4**

Let Ci be the predicted clustering results and Wibe the ground truth clustering results.

| C1 W2 | = 5

| C2 W3 | = 5

| C3 W1 | = 4

| C4 W4 | = 4

purity(**C**, **W**) =

**Precision, Recall & F-measure:**

# pairs of data points = = 190

Using a python script, I computed the following:

TP = 32

FP = 9

TN = 141

FN = 8

Precision = TP/(TP + FP) = 32/41 = 0.780

Recall = TP/(TP + FN) = 32/40 = 0.8

F-measure = (2 \* Precision \* Recall) / (Precision + Recall) = 0.790

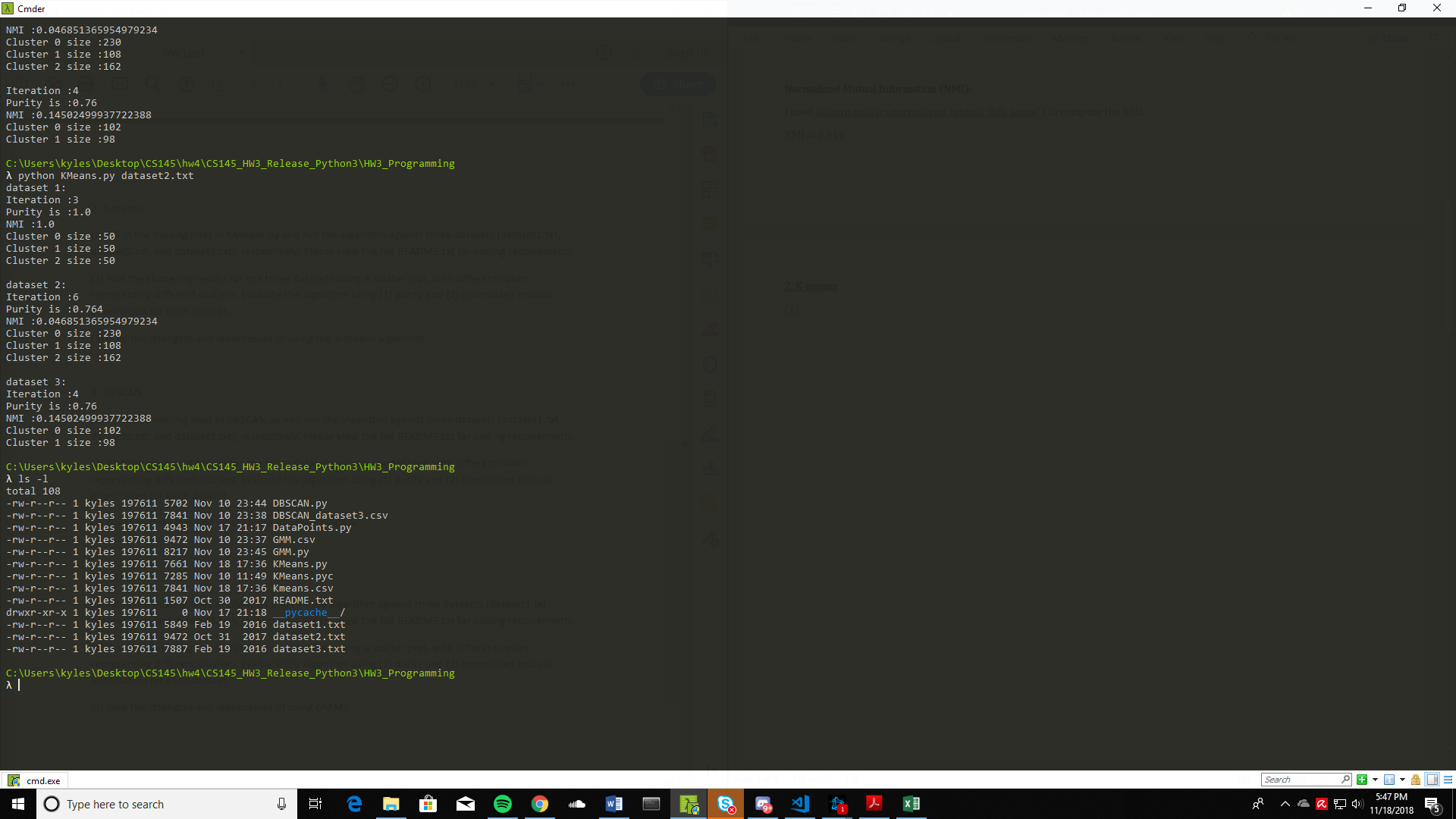
**Normalized Mutual Information (NMI):**

I used sklearn.metrics.normalized\_mutual\_info\_score() to compute the NMI:

NMI = 0.815

**2. K-means**

(1):



(2):

Dataset 1:

Purity: 1.0

NMI: 1.0

Dataset 2:

Purity: 0.764

NMI: 0.046851365954979234

Dataset 3:

Purity: 0.76

NMI: 0.14502499937722388

(3):

**Strengths:**

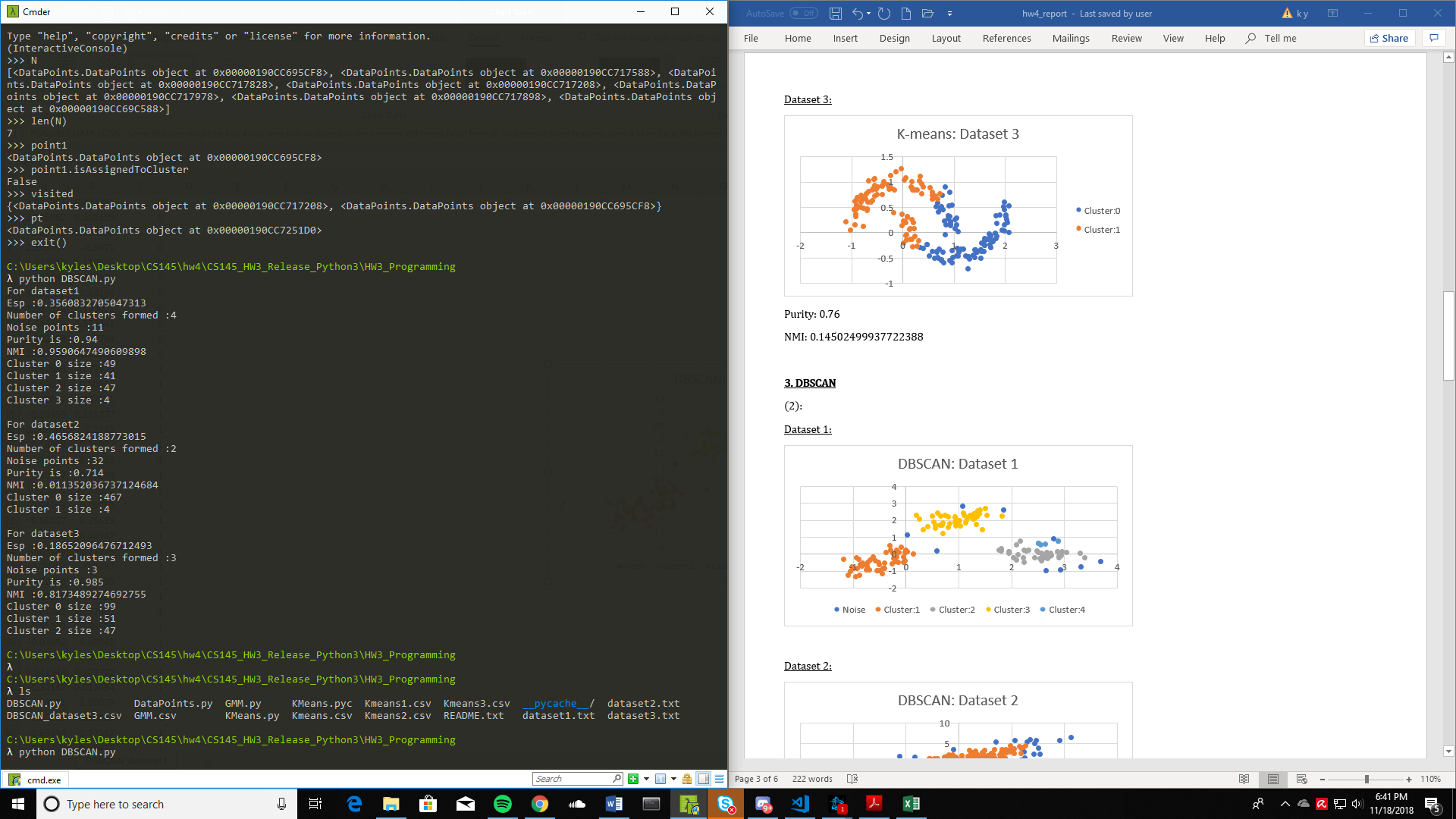
* Efficient: O(tkn) where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.

**Weaknesses:**

* Applicable only to objects in a continuous n-dimensional space.
* Need to specify k, the number of clusters, in advance.
* Sensitive to noisy data and outliers.
* Not suitable for discovering clusters with non-convex shapes.

**3. DBSCAN**

(1):



(2):

Dataset 1:

Purity: 0.94

NMI: 0.9590647490609898

Dataset 2:

Purity: 0.714

NMI: 0.011352036737124684

Dataset 3:

Purity: 0.985

NMI: 0.8173489274692755

(3)

**Strengths:**

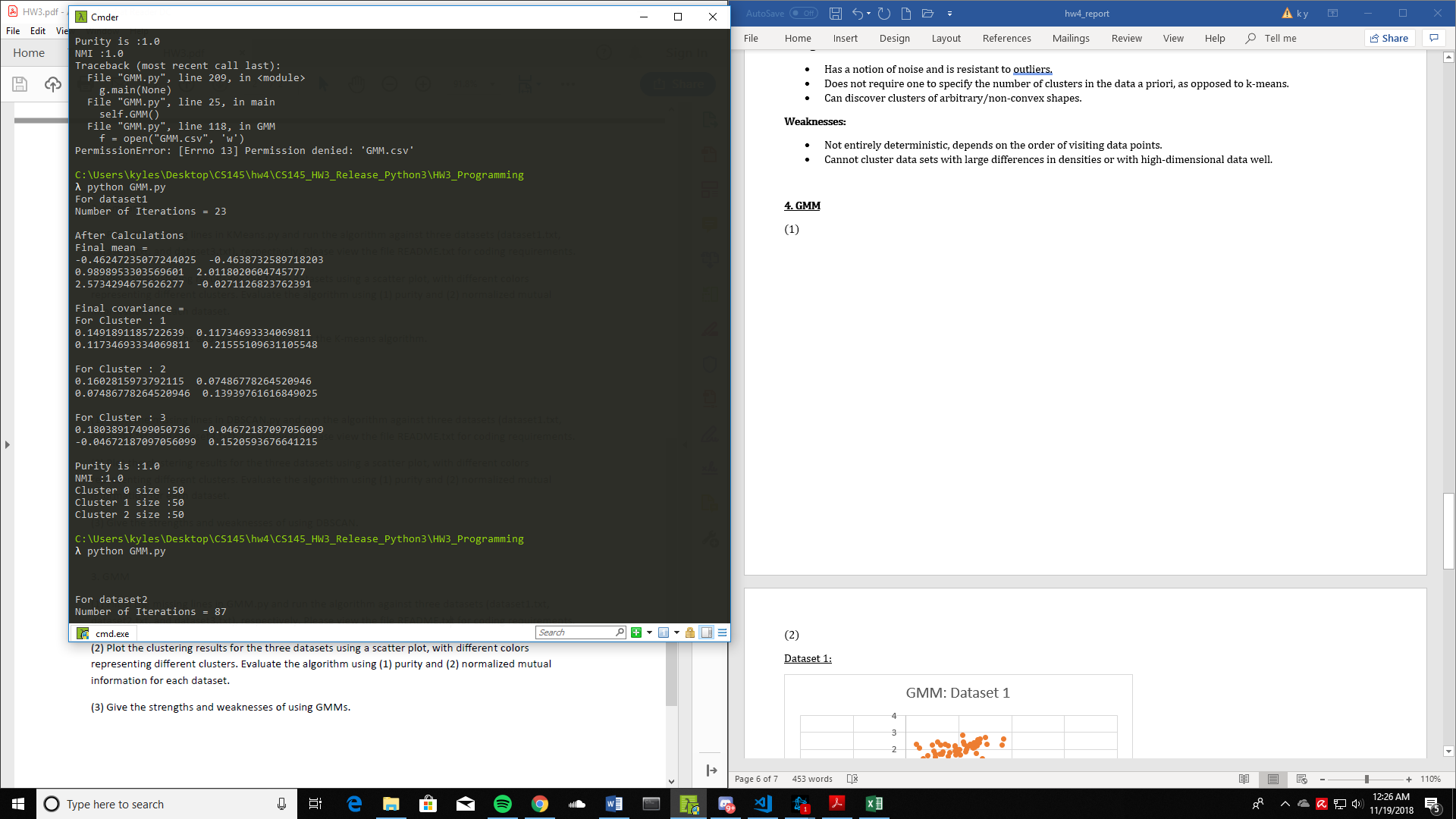
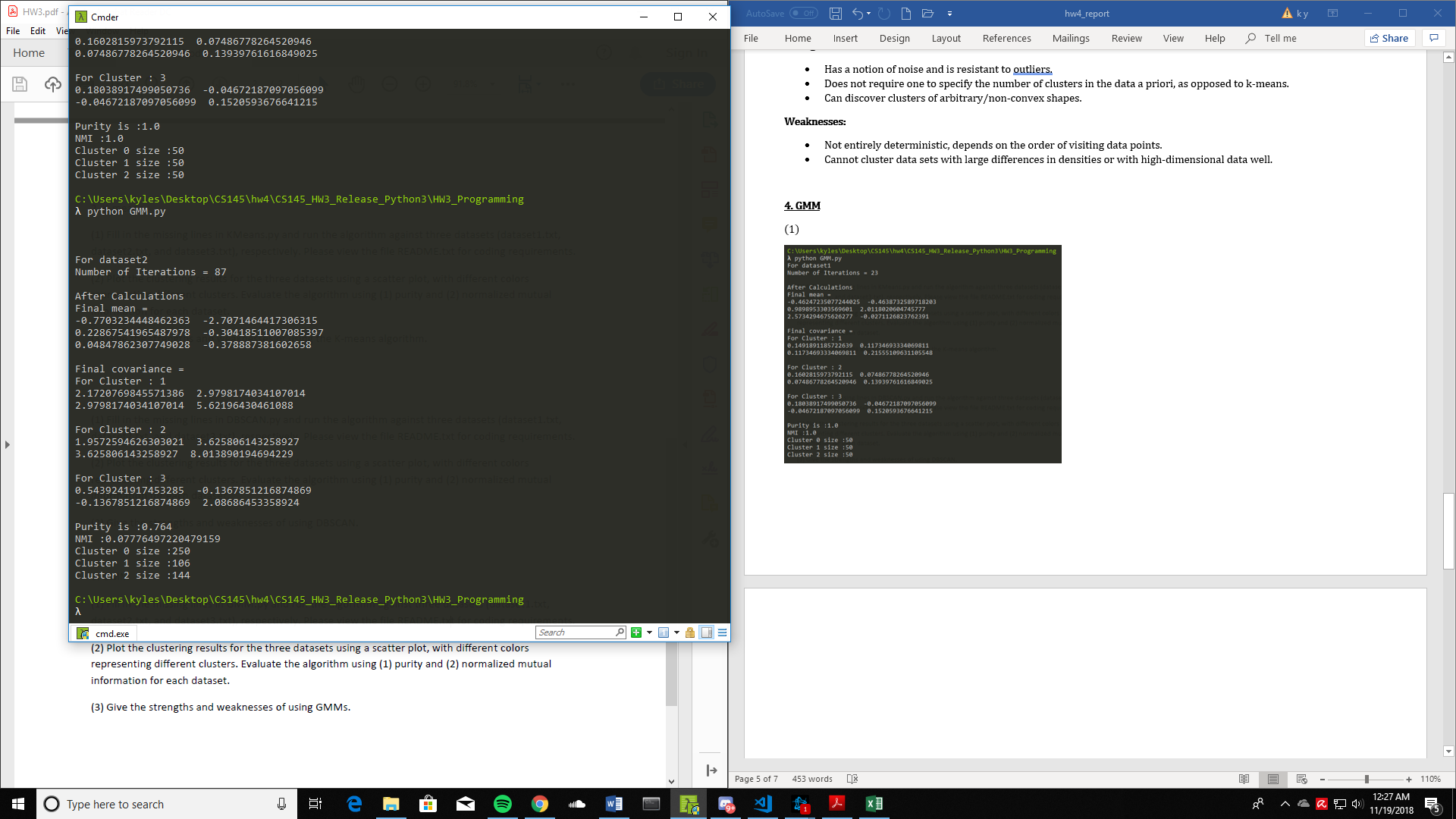
* Has a notion of noise and is resistant to outliers.
* Does not require one to specify the number of clusters in the data a priori, as opposed to k-means.
* Can discover clusters of arbitrary/non-convex shapes.

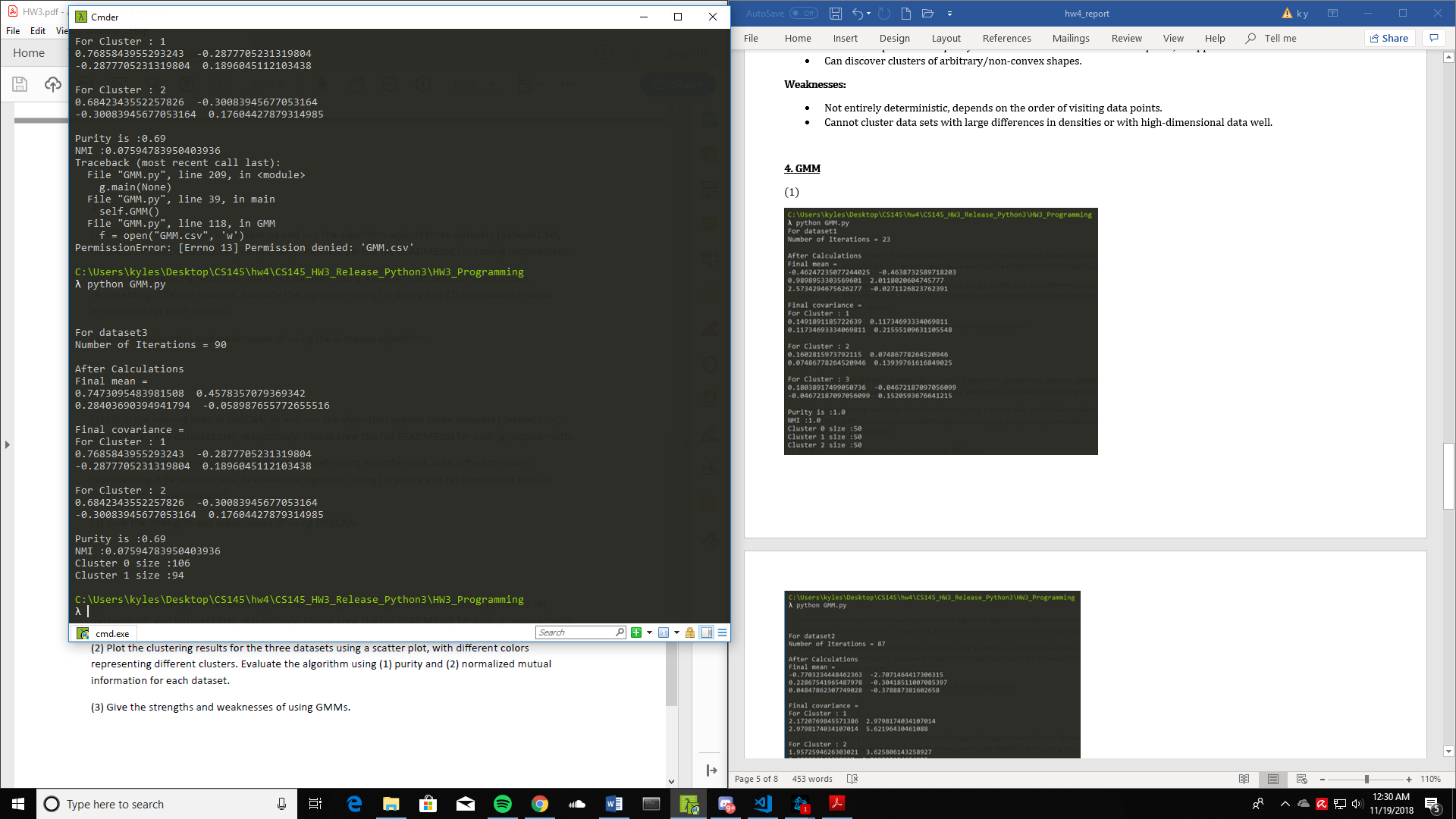
**Weaknesses:**

* Not entirely deterministic, depends on the order of visiting data points.
* Cannot cluster data sets with large differences in densities or with high-dimensional data well.

**4. GMM**

(1)



(2)

Dataset 1:

Purity: 1.0

NMI: 1.0

Dataset 2:

Purity: 0.764

NMI: 0.07776497220479159

Dataset 3:

Purity: 0.69

NMI: 0.07594783950403936

(3)

**Strengths:**

* Mixture models are more general than partitioning: can work well with clusters of different densities and sizes
* Clusters can be characterized by a small number of parameters
* The results may satisfy the statistical assumptions of the generative models

**Weaknesses:**

* Converges to local optima. (can overcome by running multiple times with random initializations)
* Computationally expensive if the number of distributions is large
* Hard to estimate the number of clusters
* Can only deal with spherical clusters