STA221—HW1

Zhikuan Quan (917800911) Bohao Zou (917796070)

April 30, 2020

1 True or False?

(a) A rectangular matrix of size n $\times$ m is a linear transformation.
True. It is a linear transformation to a vector by definition.

(b) Only square matrices have Eigenvalue decompositions.
True. Since the definition of an eigenvalue is only for square matrices. If A is not a square matrix, for example, the dimension of A is $m \times n$, then Ax is the dimension m x 1. but the dimension of λx shoule be $n \times 1$.

(c) Power Method can be used to find only eigenvectors (and not singular vectors).
False. Power methods can also be used to compute singular value decomposition.

(d) Singular vectors are orthogonal to each other.
False. Only left singular vetors are orthogonal to lefe sigular vetors, and the right ones are orthogonal to the right.

(e) Kernel PCA is a linear dimension reduction technique.
False. It is non-linear with diverse kernel function.

(f) Spectral Clustering is a non-linear dimension reduction technique.
True. It is by definition with non-convex curve.

2 Python Practice

(a) It shows that the deviation $\bar{X}_n - \mu$ converges to 0 as n increases. (See Figure 1)

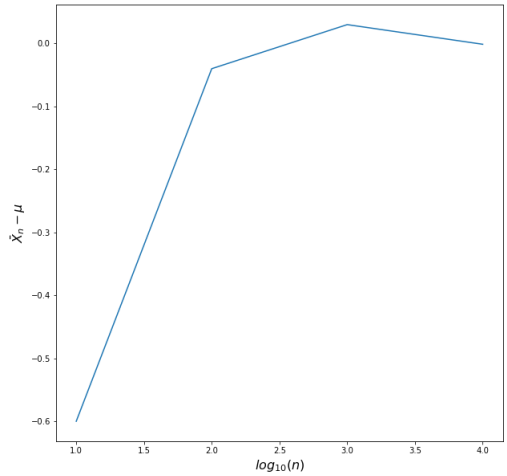

Figure 1: Test Result

(b) This plot illustrates the convergence of empirical averages to true expectation. (See Figure 2)

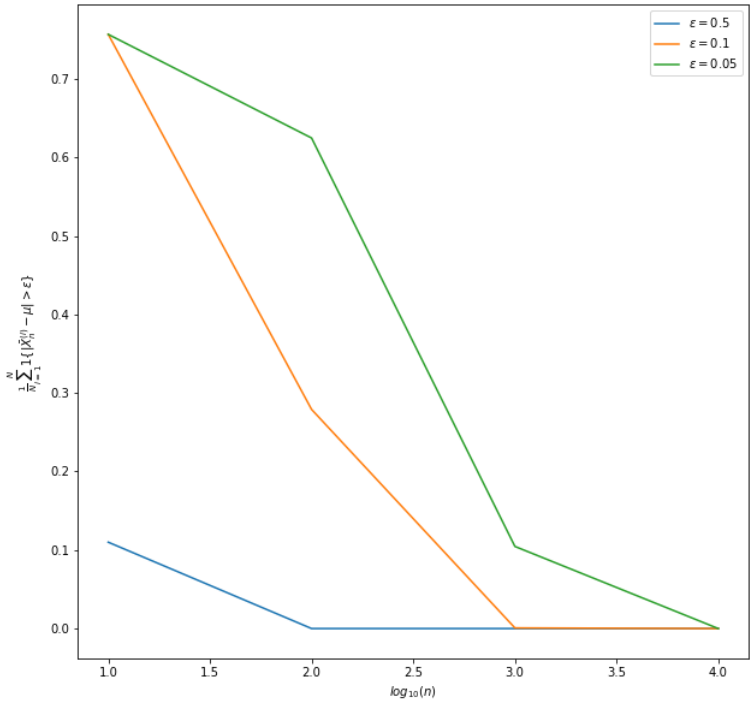

Figure 2: Test Result

(c) This plot illustrates the Central Limit Theorem: as n becomes bigger, the distribution tends to be normal. (See Figure 3)

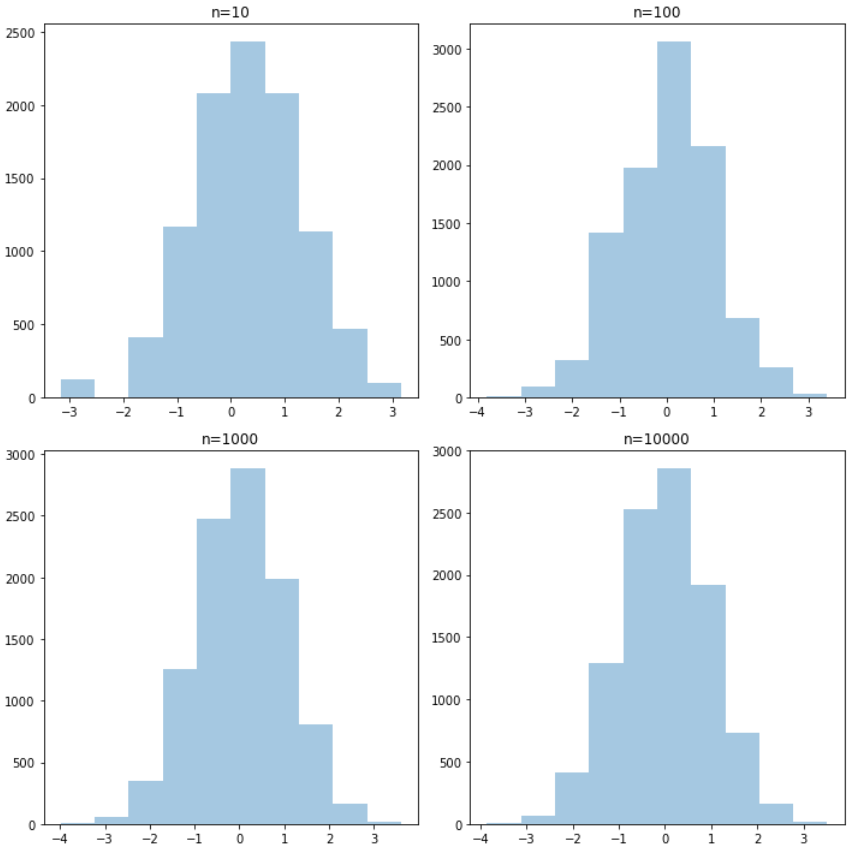

Figure 3: Test Result

3 Amazon Review Analysis

(a)-(d) All the codes are attached in the .ipynb file. The Document-Term matrix (1312×5186) and TF-IDF matrix (1312×5806) are constructed.
(e) There are 2 rating value in this data set (1 and 5). The number reviews of 1 rating is 656; The number reviews of 5 rating is 656.
(f) The plots above show the result of PCA and Kernel PCA(cosine kernel). For the Document-Term Matrix, it seems that they are overlapped for two clusters. However, both PCA and Kernel PCA can seperate the clusters in TF-IDF Matrix. (See Figure 4)

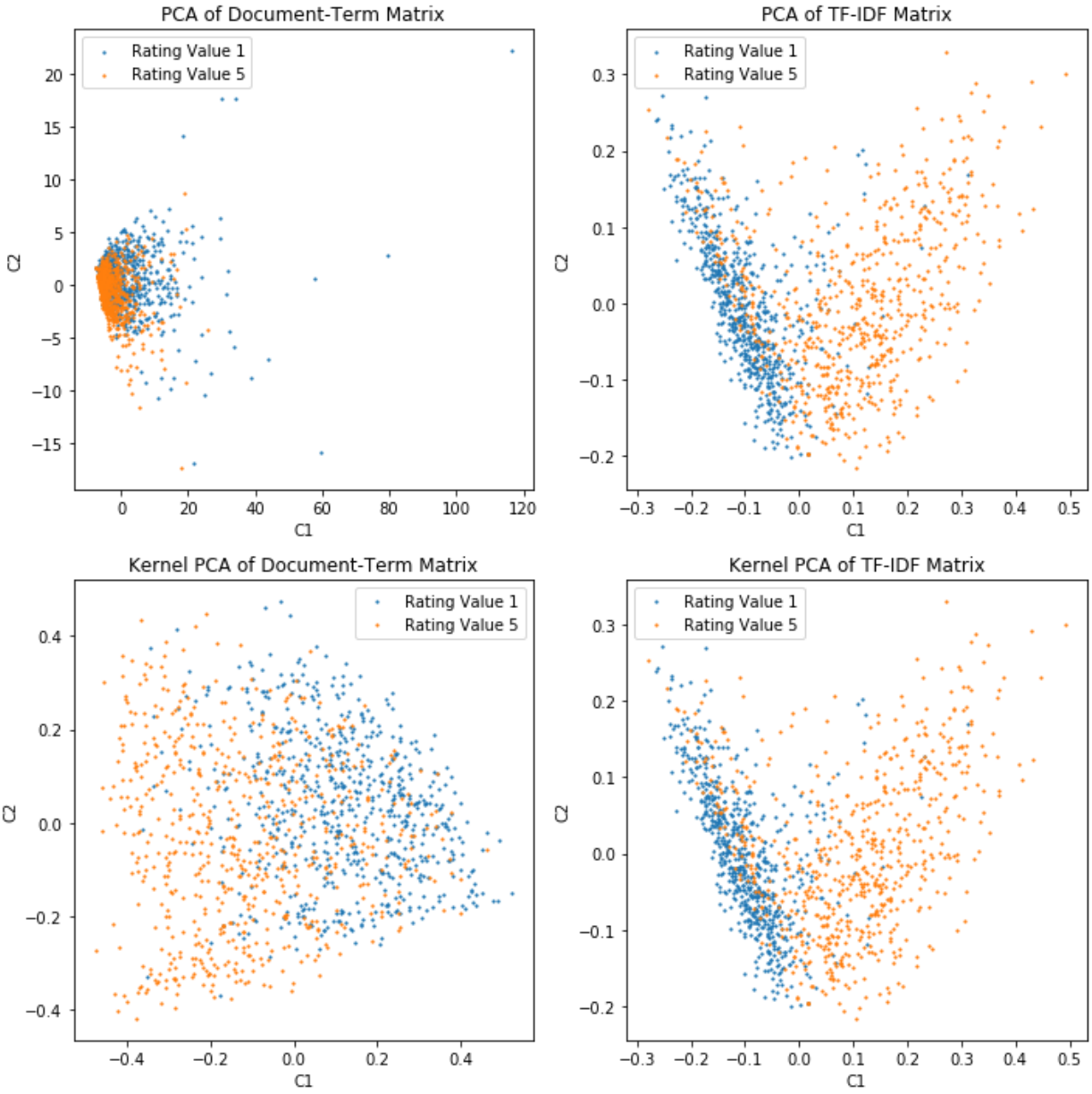

Figure 4: Test Result

(g) The confusion matrix are shown in the following. And then we can also acquire the accuracy of each algorithm.

	Pred 5	Pred 1
True 5	624	32
True 1	528	128
Accuracy	0.573	/

Table 1: K-means: Document-Term Matrix

	Pred 5	Pred 1
True 5	472	184
True 1	27	629
Accuracy	0.839	/

Table 2: K-means: TF-IDF Matrix

	Pred 5	Pred 1
True 5	538	118
True 1	452	204
Accuracy	0.566	/

Table 3: Spectral: Document-Term Matrix

	Pred 5	Pred 1
True 5	475	181
True 1	33	622
Accuracy	0.837	/

Table 4: Spectral: TF-IDF Matrix

In this case, the cluster assignment we obtained is almost consistent with the true rating of each document.

Pledge:

Please sign below (print full name) after checking (✓) the following. If you can not honestly check each of these responses, please email me at kbala@ucdavis.edu to explain your situation.

- ✓ We pledge that we are honest students with academic integrity and we have not cheated on this homework.
- ✓ These answers are our own work.
- ✓ We did not give any other students assistance on this homework.
- ✓ We understand that to submit work that is not our own and pretend that it is our is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs.
- ✓ We understand that suspected misconduct on this homework will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through Dismissal from the University and a grade penalty up to a grade of "F" for the course.

Team Member 1 Zhikuan Quan

Zhikuan Quan

Team Member 2 Bohao Zou

Bohao Zou