**MST 690 – Data Science Mathematics** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Session 6 – Linear Algebra**

**In-Class Exercise**

Instructions: Solve each problem. Show all work. Attached additional sheets as necessary. Unless otherwise stated, you may use Python or other tools to assist.

1. Show that the following is true (remember that variance is the square of the standard deviation):
2. Explain the benefits of dimensionality reduction on large data sets. In what ways might dimensionality reduction be detrimental? Recall from our Clustering lecture the “Curse of Dimensionality.” How might the utility of dimensionality reduction be explained in this context?
3. Calculate (by hand) the eigenvalues and the associated eigenvectors of matrix A:
4. Compute (by hand) the determinant of matrix A.
5. You are a data scientist at a three-letter agency. You have been following a group of suspected ISIL members on social media, and have derived 4 features from various profiles. You are developing a supervised learning algorithm for identifying ISIL members based on these features, and need to project your data onto two dimensions for clustering analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature 1 | Feature 2 | Feature 3 | Feature 4 |
| 5.1 | 3.5 | 1.4 | 0.2 |
| 4.9 | 3 | 1.4 | 0.2 |
| 4.7 | 3.2 | 1.3 | 0.2 |
| 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 3.6 | 1.4 | 0.2 |
| 5.4 | 3.9 | 1.7 | 0.4 |
| 4.6 | 3.4 | 1.4 | 0.3 |
| 5 | 3.4 | 1.5 | 0.2 |
| 4.4 | 2.9 | 1.4 | 0.2 |
| 4.9 | 3.1 | 1.5 | 0.1 |
| 5.4 | 3.7 | 1.5 | 0.2 |
| 4.8 | 3.4 | 1.6 | 0.2 |
| 4.8 | 3 | 1.4 | 0.1 |
| 4.3 | 3 | 1.1 | 0.1 |
| 5.8 | 4 | 1.2 | 0.2 |
| 5.7 | 4.4 | 1.5 | 0.4 |
| 5.4 | 3.9 | 1.3 | 0.4 |
| 5.1 | 3.5 | 1.4 | 0.3 |
| 5.7 | 3.8 | 1.7 | 0.3 |
| 5.1 | 3.8 | 1.5 | 0.3 |

Do the following:

1. Derive a covariance matrix from this data set.
2. Calculate the feature vector of eigenvalues from the covariance matrix.
3. Project the dataset into the appropriate principle component space.
4. Assuming the class of each record is known, explain how this reduced data set could be used to derive a supervised learning algorithm based on clustering.
5. BONUS: Graph the 2D principle components.