MSiA 421: Data Mining
Assignment -1

Individual Assignment (100 points)

Instructions:

- Submit the paper review as a word or pdf file.
- Submit code as a Python notebook (.ipynb) file along with the HTML version.
- Write elegant code with substantial comments. If you have referred to or reused code from a website add the links as reference.

1. Paper Review – Following the guidelines review any one of the technical papers from Group1 **(30)**

2. Generate random multidimensional (n=1000, D > 15) data using *sklearn*. **(20)**

   Use *numpy* and *matplotlib* to execute the following steps:

   - Calculate the Covariance Matrix of the data.
   - Calculate the Eigenvalues and Eigenvectors of the resulting Covariance Matrix.
   - Demonstrate that resulting Eigenvector that corresponds to the largest Eigenvalue can then be used to reconstruct a large fraction of the variance of the original dataset.
   - Compare the results above with Principal Components identified using *sklearn*.
   - Choose the most appropriate number of Principal Components and explain why that number was chosen.

3. Dimensionality reduction and visualization
   For the following dataset **(20)**

   http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

   http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data

   Perform a thorough Exploratory Data Analysis of the data including statistical summaries, correlation plots and pairwise plots.

   Apply the following dimensionality reduction techniques and visualize the data

   a) PCA
   b) Kernel-PCA
   c) t-SNE
   d) U-MAP

   Perform comparisons between these methods and summarize your observations.

4. We derived the maximum variance formulation for deriving the solution to PCA in class. Show that minimizing the reconstruction error in PCA would provide an equivalent solution. **(10)**

5. Entropy and KL Divergence **(20)**

   a) For the following events: {‘A’, 'B', 'C', ‘D’}

   P: 0.10, 0.40, 0.25, 0.25

   Q: 0.60, 0.15, 0.05, 0.20

   Calculate the entropy, cross entropy and KL divergence of above probability distributions P and Q:

   $$H(P), H(Q), H(P,Q), H(Q,P), \quad KL\ (P\ ||\ Q), KL\ (Q\ ||\ P)$$

   Summarize your observations.

   b) Compute the following:

   $$KL\ (P\ ||\ Q), KL\ (Q\ ||\ P), KL\ (P\ ||\ Q'), KL\ (Q'\ ||\ P)$$

   - P is a normal distribution with a mean of 0 and a standard deviation of 2
   - Q is a normal distribution with a mean of 2 and a standard deviation of 2.
   - Q’ is another distribution with a mean of 5 and a standard deviation of 3.

   Plot all three distributions to verify the output of the KL divergence measures.