Homework 1

Unsupervised Learning • STAT GR 5244 • Fall 2025

Assigned: September 24

Due: October 10

Instructions:

- *You may work with others on this homework assignment but all solutions must be written up and submitted individually.*
- *All homework assignments must be submitted in pdf format and should be at most* **8 pages** *in length. Any material beyond 8 pages will not be graded.*
- *You must submit all code used to complete this homework assignment as an appendix (this can go beyond the 8 pages). Failure to submit code will result in 20% reduction.*
- *You are permitted 4 total late days on homework assignments throughout the semester. Any late days taken beyond these 4 will incur a 20% reduction per day.*

1. (50 points) Dimension Reduction on Digits Data. For this problem, please use the digits data available in `sklearn` (you should use all digits).

   (a) Apply linear dimension reduction techniques including PCA, NMF, and ICA. Visualize and interpret both the observation and feature patterns. Which linear dimension reduction technique is best for separating the digits? How did you select hyperparameters for these methods? Please employ a quantitative approach to answer the latter two questions.

   (b) Apply manifold learning approaches including Kernel PCA, Spectral Embedding, Classical MDS, Metric MDS, tSNE, and UMAP. Additionally, build and apply an autoencoder. Visually compare results from all of the above embeddings. Which approaches reveal the most separation between the digits? How did you select hyperparameters? Please employ a quantitative approach to answer the latter two questions.

   (c) Discussion. Reflect on your results. Why did some methods perform better than others? Which method is overall the best at separating the digits? Why? Present the best single plot visual summary of this data.

2. (50 points) Open-Ended Data Analysis - Breast Cancer gene expression data. For this problem, please use the BRCA data available from the course webpage. This data set consists of gene expression measurements for $n = 445$ breast cancer tumors and $p = 353$ genes taken from The Cancer Genome Atlas (TCGA). This subset of genes was selected based on whether they contain known somatic mutations in cancer. Additionally, this data contains clinical data on the (i) Subtype (denotes 5 PAM50 subtypes including Basal-like, Luminal A, Luminal B, HER2-enriched, and Normal-like), (ii) ER-Status (estrogen-receptor status), (iii) PR-Status (progesterone-receptor status), (iv) HER2-Status (human epidermal growth factor receptor 2 status), (v) Node (number of lymph nodes involved), and (vi) Metastasis (indicator for whether the cancer has metastasized).

   (a) Apply dimension reduction techniques to explore this data set. Use the clinical data to interpret patterns you find. Prepare visualizations of the data and the patterns you find. You should apply at least 5 different dimension reduction techniques and justify your choices. You should select hyperparameters in a principled manner. Compare results from your different dimension reduction methods. Which reveals the most interesting patterns in this data? Why? Present the best single plot visual summary of this data.