Homework 2

Unsupervised Learning • STAT GR 5244 • Fall 2025

Assigned: October 15

Due: October 31

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) Mixture Models. For this problem, download the `authors` dataset from `https://github.com/DataSlingers/clustRviz/tree/master/data` or the course website. This data set consists of word counts from book chapters (observations) of four English language authors for stop words (features), or common words that are typically filtered out in natural language processing. (You should remove the Book ID for this analysis.) For this problem, you should remove the author labels and only use these to validate your approaches.

   (a) EM Algorithm Derivation. Since we are modeling count-valued data, suppose the data arises from a mixture of independent Poisson distributions. That is, $p(x_i; \pi, \lambda) = \sum_{k=1}^{K} \pi_k \text{Pois}(x_i; \lambda_k)$. Derive an EM algorithm to fit this distribution and estimate the cluster centroids and soft-cluster memberships.

   (b) Write a python function implementing your EM algorithm for a mixture of Poisson distributions.

   (c) Fit your mixture of Poisson distributions to the author data with $K = 4$ clusters. Also fit a Gaussian mixture model to this data. Compare and reflect upon your results. Interpret the words representing each cluster centroid and the book chapters that have low certainty in the soft cluster labels.

2. (50 points) Open-Ended Cluster Analysis - Breast Cancer gene expression data. For this problem, please use the BRCA data available from the course webpage and that you analyzed in homework 1. 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 clustering techniques to explore this data set. You should apply K-means, hierarchical clustering (using several linkages), spectral clustering, Gaussian mixture models, and DBSCAN. You may also want to try applying dimension reductions techniques first before applying clustering methods.

(b) Implement cluster validation techniques including the Silhouette score and stability and generalizability-based methods. Which values of K did the various methods select? Are the results consistent across methods and validation techniques? Select what you believe the be the single best clustering result and justify your choice; discuss your results.

(c) Interpret your cluster findings with respect to the metadata provided. Do the clusters you found coincide with any clinical categories? Provide at least one visualization of your cluster findings that illustrate these interpretations.