Assignment 2B

Introduction to Machine Learning: CS-3410-1 (Spring 2025)

April 23, 2025

NOTE: The dataset for Q-1 and Q-2 is uploaded separately on the classroom.

1. Programming – Healthcare Resource Distribution Using K-means Clustering

The World Health Organization is developing a strategic plan for distributing medical resources and establishing healthcare infrastructure. Using the provided dataset that contains healthcare indicators, demographic data, and infrastructure metrics, implement k-means clustering to categorize countries based on their healthcare needs and existing capabilities.

Dataset features

- **child mort**: Death of children under 5 years of age per 1000 live births
- **health**: Total health spending as % of GDP
- **income**: Net income per person
- **inflation**: Annual growth rate of the Total GDP
- **life expec**: Average life expectancy at birth
- **total fer**: Fertility rate (children per woman)
- **gdpp**: GDP per capita

Tasks

(a) **Data Preparation** [3 points]

(i) Load and examine the dataset structure.

(ii) Plot histograms for each individual feature.

(iii) Perform feature scaling/standardization (normalize mean to 0 and standard deviation to 1 for each feature).

(b) **K-means Implementation** [8 points]

(i) Implement the k-means algorithm from scratch with $k = 4$:

- Initialize centroids randomly.
- Assign countries to the nearest centroid using Euclidean distance.
- Update centroids by computing the mean of assigned countries.
- Repeat until convergence or a maximum of 100 iterations is reached.

(ii) Run the algorithm with five random initializations.

(iii) Implement and explain your convergence criteria.

(c) **Analysis** [5 points]

(i) Create 2D scatter plots for:

- *child mort* vs. *health*
- *income* vs. *life expec*
- *total fer* vs. *gdpp*

where each cluster is colour-coded.

(ii) For each cluster, analyze and report:

- Number of countries
- Average values of key indicators
- Characteristic features of the cluster

2. Programming - Development Aid Analysis Using PCA and Clustering

The United Nations Development Programme (UNDP) is analyzing global development indicators to identify patterns and groupings among countries for more targeted development assistance. The provided dataset contains key socio-economic and health indicators for different countries:

Dataset features

- **child mort**: Death of children under 5 years of age per 1000 live births
- **health**: Total health spending as % of GDP
- **income**: Net income per person
- **inflation**: Annual growth rate of the Total GDP
- **life expec**: Average life expectancy at birth
- **total fer**: Fertility rate (children per woman)
- **gdpp**: GDP per capita

Tasks

(a) **Principal Component Analysis** [6 points]

(i) Implement PCA on the standardized dataset:

- Calculate the covariance matrix.
- Compute all the eigenvalues and eigenvectors.
- Sort the principal components by eigenvalue in descending order.

(ii) Compute the explained variance ratio for each component (ratio of each eigenvalue to the sum of all eigenvalues).

(iii) Determine the minimum number of principal components needed to explain at least 80% of the variability.

(b) **2D Analysis and Visualization** [6 points]

(i) Create scatter plots of the first two principal components.

(ii) Apply k-means clustering ($k = 4$) on the 2D reduced data.

(iii) Show a 2D plot for the dimension-reduced data, colour-coded by cluster assignment.

(c) **3D Analysis and Comparison** [6 points]

(i) Create pairwise scatter plots of the first three principal components.

(ii) Apply k-means clustering ($k = 4$) on the 3D reduced data.

(iii) Show pairwise 2D plots for the dimension-reduced data, colour-coded by cluster assignment.

(iv) Compare clustering results across:

- Original high-dimensional clustering
- 2D PCA clustering

- 3D PCA clustering

(d) **Requirements**

- Use Python with `numpy`, `scikit-learn`, and appropriate visualization libraries.
- Include clear documentation and comments in your code.
- Provide visualizations that effectively communicate the results.
- Include error handling in your implementation.
- Discuss the practical implications of your findings for development aid allocation.

3. Programming – Expectation–Maximization for Mixture of Gaussians

Consider the following data points from a mixture of two univariate Gaussian distributions:

$$-0.39,\ 0.12,\ 0.94,\ 1.67,\ 1.76,\ 2.44,\ 3.72,\ 4.28,\ 4.92,\ 5.53,$$
$$0.06,\ 0.48,\ 1.01,\ 1.68,\ 1.80,\ 3.25,\ 4.12,\ 4.60,\ 5.28,\ 6.22$$

Implement the Expectation–Maximization Algorithm to find the maximum likelihood estimates of the means and variances of the two Gaussians.

4. VC Dimension

Let the input domain of a learning problem be $X = \mathbb{R}$. Give the VC dimension for each of the following classes of hypotheses. In each case, if you claim that the VC dimension is d, then you need to show that the hypothesis class can shatter d points, and explain why there are no $d + 1$ points it can shatter.

- $h(x) = 1\{a < x\}, \quad a \in \mathbb{R}.$
- $h(x) = 1\{a < x < b\}, \quad a, b \in \mathbb{R}.$
- $h(x) = 1\{a \sin x > 0\}, \quad a \in \mathbb{R}.$
- $h(x) = 1\{\sin(x + a) > 0\}, \quad a \in \mathbb{R}.$

5. VC dimension

Suppose that we are to use semicircles in the 2D plane to classify a given collection of 2-dimensional data points (diameters of these semicircles need not be parallel to either coordinate axis). Each semicircle classifier labels points in its interior as 0, and other points as 1. Let us call this collection of classifiers $\mathcal{H}$.

(i) Which of the following point sets can $\mathcal{H}$ shatter?

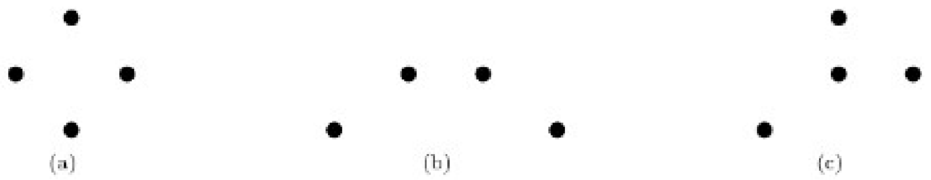

Figure 1: Candidate point sets for shattering by semicircle classifiers.

(ii) What does this tell you about the VC dimension of $\mathcal{H}$?