CSE472 (Machine Learning Sessional)

Assignment 4: Principal component analysis (PCA) and expectation-maximization (EM) algorithm

Introduction

Principal component analysis (PCA) and the expectation-maximization (EM) algorithm are two of the most widely used unsupervised methods in machine learning. In this assignment, you will implement PCA for dimensionality reduction and implement the EM algorithm for the Poisson mixture model.

Principal component analysis (PCA)

Dataset: You are given a space separated file titled "pca_data.txt" to be used as the dataset for this. The file contains 1000 rows and 500 columns. The 1000 rows correspond to 1000 sample points and each sample is represented by a 500 dimensional feature vector.

Tasks:

- Write code to perform PCA. You can call library functions to perform matrix operations such as eigendecomposition but do not call library functions to perform the entire PCA.
- Now project your data along the two eigenvectors corresponding to the two highest eigenvalues and create a 2D scatter plot showing the data.
- Use library functions to create UMAP and tSNE plots of the original data

Expectation-maximization (EM) algorithm

Dataset: You are given a space file titled "em_data.txt" to be used as the dataset for this. The file contains 1000 rows. The 1000 rows correspond to the number of children in 1000 (hypothetical) families.

Tasks:

- You are given numbers of children in 1000 families. Some families were given family planning advice and some were not. But you do not know which of the 1000 families were given family planning advice and which weren't. Implement the EM algorithm and run it on this dataset to estimate the mean number of children in families with

and without family planning. Also estimate the proportion of families with and without family planning. You can assume that the number of children in the two types of families are Poisson distributed.

Report Writing

1. You must provide clear instructions on how to run your codes in the report.
2. Include plots for PCA (your own code), UMAP and tSNE (library).
3. Report the mean number of children in families with and without family planning as well as the proportion of families with and without family planning.

Marking Rubrics

PCA Implementation and plot	40%
UMAP and tSNE plots	10%
EM implementation	45%
Code clarity and proper submission	5%

Submission Format

1905xyz
|-- codes
|-- report_1905xyz.pdf

Zip the folder and rename it to **[Student_ID].zip**

Deadline: 22-Nov-2024 (Friday), 10:00 PM.

Warning

1. Don't copy! We regularly use copy checkers.
2. First time wrongdoers (either copier or the provider) will receive **negative** marking because of dishonesty.
3. Repeated occurrences of copying will lead to severe departmental action and jeopardize your academic career. We expect fairness and honesty from you. Don't disappoint us!