COMS4060A/7056A: Assignment # 3

University of the Witwatersrand
Muhammad Nasir
muhammad.nasir@wits.ac.za

September 2024

Introduction

This assignment is based on content covering dimensionality reduction. You will be required to perform some basic data cleaning and exploration techniques on a prescribed dataset. The aim is to explore the dataset and make observations. You are required to provide a Jupyter notebook and a PDF for all your reasoning. You must discuss all your answers on the PDF with the plots attached. You must provide one Jupyter notebook that clearly mentions the question number you are solving. The provided Jupyter notebook should be saved after a complete run, which will save your answers in the notebook as well. You are free to use any library/framework in Python.

NBA Dataset

The original dataset is available of Kaggle but you will be using the one attached. Some key features can be looked up on `https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular/data` but you are given some extra features as well.

1 Data Cleaning [5 marks]

1. Do a preliminary analysis of the dataset. Are there any points which seem odd? Decide weather or not to remove them and justify your decisions. Discuss methods of handling the missing values, if there are any.

2 Dimensionality Reduction [65 marks]

1. After cleaning your dataset, reduce the dimensionality to two dimensions through the following techniques:

(a) Autoencoders [10]

(b) Autoencoders + self-organising maps (SOMs) [10]

(c) Autoencoders + t-SNE [10]

(d) Autoencoders + UMAP [10]

(e) Variational Autoencoder [10]

2. Plot the reduced dataset in two dimensions, for each reduced dataset apply k-Means to cluster the reduced dataset. Visualise, compare and explore your results. Which of the methods gives the most insights after clustering? [15]

Submission

Work by yourself or in groups of up to four people. Submit your work to Moodle. Your submission should be a PDF and a Jupyter notebook. You should use NeurIPS latex styles as a format of your PDF

Total Marks Available: 70 Marks. (10 bonus marks)

Deadline: 28 October 2024