TASK

Capstone Project – Unsupervised Machine Learning

Visit our website

Introduction

WELCOME TO THE UNSUPERVISED MACHINE LEARNING CAPSTONE PROJECT!

Well done on making it this far! For this Capstone project, we will perform principal component analysis (PCA) and clustering on the dataset provided.

Developer portfolio

Developers who have the edge are those who find ways to apply their newfound skills from the get-go. As you may know, a **developer portfolio** (a collection of online creations that you have made) allows you to demonstrate your skills rather than just telling people about them. It's a way of bringing your CV to life and introducing yourself to the world. As you learn more skills and put these into practice, each project that you complete will become more efficient and eye-catching.

These capstone projects give you the means to create projects for your very own developer portfolio, allowing you to walk away from this course not only with a certificate but, more importantly, with a head start to your tech career!

The task at hand

In this task, you will explore the underlying relationships between different crime rates across the United States and urban density in 1973. You will use PCA to reduce the dimensionality of the **US Arrests dataset**, identifying the principal components that capture the most significant variation in the crime rates. The US Arrests dataset contains information about arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states. The percentage of people living in urban areas is also provided.

Instructions

A key focus of this project will be ensuring that your code is correct, well-formatted, readable, and that it adheres to the **PEP 8 style guide**. In this regard, **make sure that you do the following** (and double check before submitting your work to avoid losing marks unnecessarily!):

1. Identify and remove all syntax, runtime, and logical errors from your code.
2. Construct readable code. To ensure this, add comments to your code, use descriptive variable names, and make good use of whitespace and indentation.
3. Create modular code. Create functions to perform specific units of work.
4. Write efficient code. How you choose to write code to create the solution to the specified problem is up to you. However, make sure that you write your code as efficiently as possible.
5. Make sure that all outputs that your program provides to a user are easy to read and understand. Label all data that you output.

Practical task

1. Create a Jupyter notebook called **unsupervised_task.ipynb**.
2. Use the dataset **UsArrests.csv** included in this folder to generate an in-depth PCA of the data.
3. Explore the data and motivate any preprocessing steps you take.
4. Perform a correlation analysis and interpret the output.
5. Plot and interpret a biplot of the first two principal components.
6. Select the number of principal components to use in further analyses. Justify your choice.
7. Apply two clustering techniques and an analysis of the clusters they generate. Comment on similarities between data points within each cluster.

Important: Be sure to upload all files required for the task submission inside your task folder and then click "Request review" on your dashboard.

Share your thoughts

Please take some time to complete this short feedback **form** to help us ensure we provide you with the best possible learning experience.

References

(n.d.). *Violent Crime Rates by US States*. GitHub. Retrieved 24 November 2024. **https://vincentarelbundock.github.io/Rdatasets/doc/datasets/USArrests.html**