Big Data Analytics and Security

CSEC 5311/CETE 4392

Spring 2025

TEXAS A&M UNIVERSITY SAN ANTONIO

Assignment-Machine Learning & MapReduce

Part-A: Machine Learning

Objective

Deep learning has revolutionized data generation and decision-making using models like Autoencoders, VAEs, GANs, and Reinforcement Learning. In this part, you will implement one generative model to showcase your understanding.

Choose **ONE** of the following generative models to implement and analyze.

Option 1: Implement a Variational Autoencoder (VAE) [50 Marks possible]

- Train a **VAE on the MNIST dataset (MNIST Dataset).**
- Show an **original image, reconstructed image, and generated samples.**
- Visualize the **latent space interpolations.**
- Explain challenges faced while training.

OR

Option 2: Implement a GAN for Image Generation [50 Marks possible]

- Train a **GAN on MNIST (MNIST Dataset) or CIFAR-10 (CIFAR-10 Dataset).**
- Show **generated images** after different epochs of training.
- Explain **mode collapse and how you handled it.**

Part-B: MapReduce

Objective:

You are given a web server log file (web_log_large.txt) containing 5000 records in Common Log Format (CLF). Your task is to implement two MapReduce programs in Python using **mrjob** to analyze this log data.

Q-1: Identify the Top 10 Most Frequent HTTP Status Codes and Their Counts (25 marks possible)

Task:

- Extract all **HTTP status codes** from the log file.
- Count the occurrences of each status code.
- **Sort the results** in descending order based on frequency.
- **Output only the top 10** most frequent HTTP status codes and their counts.

Q-2: Identify the Top 5 IP Addresses Generating the Most Errors (4xx and 5xx Status Codes) (25 marks possible)

Task:

1. Identify **IP addresses** that made requests resulting in **client (4xx) or server (5xx) errors**.
2. Count the **total number of error requests** made by each IP.
3. **Sort the results** in descending order based on the number of errors.
4. **Output only the top 5 IPs** generating the most errors.