Python for Computer Science and Data Science 2 (CSE 3652)

Major Assignment-3: Machine Learning- Classification, Regression and Clustering

Overview

Imagine you've been hired by *FashionX*, a fast-growing online fashion retailer struggling to organize its expanding product catalog. The company needs an automated system to classify fashion items into categories like T-shirts, Dresses, Shoes, etc., based on images.

As a data scientist, your task is to develop a machine learning model to classify 28x28 pixel images from the *Fashion MNIST* dataset into 10 fashion categories. You'll apply algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) for classification. Additionally, you'll use t-SNE to visualize the data, uncovering patterns and clusters in the high-dimensional space.

This project will help FashionX scale its operations by efficiently automating product categorization, making it easier for customers to find what they're looking for.

Tasks

Task 1: Data Preprocessing

- Download the Fashion MNIST dataset using TensorFlow. You can load it directly using the following code:

```
import tensorflow as tf
from tensorflow.keras.datasets import fashion_mnist

# Load the dataset
(train_images, train_labels), (test_images, test_labels) = \
fashion_mnist.load_data()
```

- The dataset consists of 60,000 training images and 10,000 test images of fashion products, with 10 distinct categories. Each image is 28x28 pixels.
- Perform the following preprocessing steps:
  1. Normalize the images (values should be between 0 and 1).
  2. Flatten the 28x28 pixel images into 1D arrays (784 pixels per image).
  3. Handle any missing or incorrect values (if any).
- Split the dataset into training and test sets.

Task 2: K-Nearest Neighbors (KNN) Classification

- Implement the K-Nearest Neighbors (KNN) algorithm using the preprocessed dataset.
- Experiment with different values of k (e.g., $k = 3$, $k = 5$, $k = 7$) to see how the model performance changes.
- Evaluate the model performance using accuracy on the test dataset.
- Provide a comparison of different k values and the impact on accuracy.

Task 3: Support Vector Machine (SVM) Classification

- Train a Support Vector Machine (SVM) classifier on the same preprocessed dataset.
- Experiment with different kernels (linear, polynomial, radial basis function) and hyperparameters such as C.
- Evaluate the model performance using accuracy on the test dataset.
- Compare the SVM performance with that of the KNN model.

Task 4: Data Visualization with t-SNE

- Use the t-SNE technique to reduce the dimensionality of the data from 784 features to 2 or 3 dimensions.
- Visualize the 2D or 3D representation of the Fashion MNIST dataset and observe the clustering of different fashion categories.
- Analyze the plot to identify how well the categories are separated, and discuss the results.

Task 5: Model Evaluation and Reporting

- Evaluate the performance of both the KNN and SVM models using accuracy, precision, recall, F1-score, and confusion matrix.
- Discuss which model performs better and why, based on the evaluation metrics.
- Write a report summarizing the approach used in each task, the results obtained, and insights derived from the visualizations.