Applied Machine Learning [COMS-4995_033]

Project Proposal - Group 17

Title : Physical Activity Prediction using Wearable Device Data

Background :

Wearable devices like smartwatches and fitness trackers generate vast amounts of physiological and motion-related data through sensors such as heart rate monitors and gyroscopes. The ability to gather useful high-level information like the physical activity the user is currently performing (e.g., walking, running, jogging, exercising), will open up to various real-world applications. This data can be used by healthcare professionals to monitor the patient's physical activities, providing insights to daily lifestyle activities (recommending standing breaks, encouraging exercising, etc) or enabling app developers to create innovative applications and systems to understand and cater to individual user needs based on their current physical activities.

Problem Statement :

Despite the availability of data from wearable devices, translating these raw signals into meaningful insights about the physical activities of users is challenging. Accurately predicting the specific physical activity remains a complex task due to the variability in individual movement patterns, environmental conditions, and sensor data accuracy.

The primary objective of this project is to evaluate and develop machine learning models that accurately predict physical activities based on data from wearable devices, such as heart rate, gyroscope, and accelerometer data. The project will focus on:

1. Data preprocessing and feature extraction from sensor readings.
2. Building and training a classification model to recognize various activities like walking, running, jogging, and exercising.
3. Evaluating the model's performance and optimizing it for accurate predictions.

Dataset to be used :

PAMAP2 - Physical Activity Monitoring -
https://archive.ics.uci.edu/dataset/231/pamap2+physical+activity+monitoring

The PAMAP2 Physical Activity Monitoring dataset contains data of 18 different physical activities (such as walking, cycling, playing soccer, etc.), performed by 9 subjects wearing 3 inertial measurement units and a heart rate monitor.

- Raw sensory is available in the space-separated text-files (.dat)
- Missing values are indicated with NaN.
- Each line corresponds to one time stamped and labeled instance of sensory data.
- The data files contain 54 columns
- Each line consists of a timestamp, an activity label (the ground truth) and 52 attributes of raw sensory data.

We'll be performing data pre-processing techniques with data cleaning, feature extraction and performing exploratory data analysis (EDA) on the above data to pass on to ML models to predict and evaluate the predictions.

Proposed ML Techniques :

Since this is a multi-class classification problem with labeled data (e.g., activity types like walking, running, etc.), we'll be incorporating a set of supervised learning techniques to predict the different classes based on the different features and evaluating the different models with the test sets using different evaluation metrics.

Few of the models and techniques that we'll be exploring would be :

1) Logistic Regression
2) Decision Trees
3) Random Forests
4) Support Vector Machines (SVM)
5) Deep Neural Networks

Techniques :

1) Gradient Boosting
2) Bagging
3) Model Ensembling
4) Hyperparameter tuning with Cross Validation (Grid Search/ Random Search)
5) AutoML