Documentation Report: Human Activity Recognition using K-Nearest Neighbors Classifier

Introduction

This comprehensive documentation report provides an extensive account of a noteworthy project centered on activity recognition using the k-Nearest Neighbors Classifier (k-NN) in conjunction with GridSearch. The focal point of this endeavor was to leverage the power of machine learning to classify and understand activities undertaken by participants, based on an extensive data set brimming with accelerometer data. The dataset in question serves as a treasure trove of insights, derived from participants wearing two 3-axial accelerometers over a span of approximately 2 hours in a free-living setting.

Activity recognition, a burgeoning field with myriad applications across industries, has never been more relevant. It finds its utility in diverse domains, from health monitoring to sports analytics, where understanding and categorizing human activities can unravel invaluable insights and propel innovation. In this context, the project at hand sought to harness the potential of machine learning to demystify the nuances of human movement.

The dataset itself incapsulates the essence of this project, with each participant's recordings meticulously stored in separate CSV files. These files lay bare a wealth of information, including timestamps, 3-axial accelerometer readings from both back and thigh sensors, and annotated activity codes. The activities, meticulously labeled, range from commonplace actions like walking and sitting to more intricate endeavors such as cycling and stair climbing. With this data set serving as

our foundation, our objective was to construct a robust model capable of accurately classifying these activities.

The k-Nearest Neighbors Classifier (k-NN) emerged as the chosen vehicle for this journey. This supervised machine learning algorithm operates on the principle of proximity, classifying data points based on the majority class among their k-nearest neighbors in the feature space. The project includes an additional layer of sophistication with the incorporation of GridSearch, a hyperparameter timing technique that optimizes the k-NN algorithm for superior accuracy.

The results achieved in the endeavor are indeed noteworthy. The k-NN classifier, supercharged by GridSearch attained an impressive accuracy of 96.6%. This accuracy metric underscores the project's success in building the model that excels in recognizing and categorizing activities based on the nuanced readings from the accelerometers.

This documentation report will traverse the intricacies of the project, diving deeper into data preprocessing, model selection, hyperparameter tuning, and the significance of the results. It will provide a comprehensive account of the technical intricacies behind the scenes, offering a blueprint for others seeking to embark on similar journeys in the realm of activity recognition. Ultimately, this project is a testament to the power of machine learning in unravelling the secrets hidden within the accelerometer data, paving the way for future innovations in this burgeoning field.

Dataset Description

The foundation of this project lies in a meticulously curated dataset, a rich repository of insights obtained from 22 participants who voluntarily wore two 3-axial accelerometers for an extensive period of approximately 2 hours. This data collection occurred in an unencumbered, free-living setting, mirroring the real-world scenarios where activity recognition holds immense significance.

Each participant's recorded data is thoughtfully segregated into separate CSV files, offering a granular and participant-specific perspective on their activities. The data set showcases a remarkable level of detail, featuring the following columns:

- **'Timestamp':** This column provides a comprehensive record of the date and time for each recorded data sample, offering a temporal dimension to the analysis.
- **'Back_x', 'Back_y', 'Back_z':** These columns represent the triaxial accelerometer readings from the back sensor. 'Back_x' captures the acceleration in x-direction(downward), 'Back_y' captures the acceleration in y-direction(leftward), and 'Back_z' captures the acceleration in z-direction(forward). These readings are expressed in units of gravity(g), offering a precise measure of motion.
- **'Thigh_x', 'Thigh_y', 'Thigh_z':** These columns mirror the structure of the back sensor readings but pertain to the triaxial accelerometer data from the thigh sensor. 'Thigh_x' represents the x-direction(downward), 'Thigh_y' the y-direction(rightward) and 'Thigh_z' the z-direction(backward).
- **'Label':** Annotated Activity Code: This pivotal column encapsulates the essence of the data set, assigning a numeric code to each recorded activity. These codes correspond to the diverse range of activities performed by the participants during the data collection phase.

The data set is characterized by its meticulous completeness; there are no missing values or gaps, ensuring the integrity of the data. These annotated activities span a gamut of human movements and actions, each associated with a unique code:

- **‘1’:** Walking
- **‘2’:** Running
- **‘3’:** Shuffling
- **‘4’:** Ascending the stairs
- **‘5’:** Descending the stairs
- **‘6’:** Standing
- **‘7’:** Sitting
- **‘8’:** Lying
- **‘13’:** Cycling (Sitting)
- **‘14’:** Cycling (Standing)
- **‘130’:** Cycling (Sitting, inactive)
- **‘140’:** Cycling (Standing, inactive)

Each activity code provides a precise label for the respective action, empowering the project to discern and categorize activities with remarkable accuracy.

This dataset is the cornerstone of our activity recognition project, and its intricate structure serves as a testament to the comprehensive sections, we all delve into the preprocessing of this data and the application of the k-Nearest Neighbors Classifier with GridSearch to unlock the valuable insights hidden within this rich dataset.

Data Preprocessing

The journey towards accurate activity recognition from the accelerometer data set embarked with meticulous data preprocessing. This critical phase was the gateway to ensuring that the data was appropriately formatted and redefined for analysis and classification. Here, we elaborate on the steps taken to prepare the data for the subsequent stages of the project.

One of the initial tasks during data preprocessing was the assessment of the dataset for any superfluous or redundant columns. It was discovered that some of the datasets contained an 'index' column that did not carry any meaningful information for the classification task at hand. Therefore, as part of the data cleaning process, this column was systematically removed to streamline the data set and reduce unnecessary complexity. This pruning step was undertaken to ensure that the subsequent analysis focused solely on the pertinent attributes, enhancing the efficiency and interpretability of the model.

With the removal of the 'index' column, the attention moved towards the selection of relevant features from the data-set. Feature selection is a pivotal step in machine learning, as it determines the attributes that will be used for training and classification. In this context, the relevant features encompassed the accelerometer readings from both the back and thigh sensors along the x, y, z axes.

The accelerometers readings, denoted as 'back_x', 'back_y', 'back_z', 'thigh_x', 'thigh_y', 'thigh_z', and the timestamp columns were retained for further analysis. These features were selected due to their intrinsic value in characterizing human movement across various activities. These features, expressed in units of gravity(g), provide precise measurements of motion in different directions, offering a rich source of information for the classification model.

To assess the performance of the classification model accurately, the data set was divided into two distinct subsets: a training set and a testing set. The conventional split was employed, with 80% of the data allocated to the training set and the remaining 20% reserved for the testing set. This partitioning ensured that the model could be trained on one portion of the data and then evaluated for its predictive accuracy on a separate, unseen portion, thus avoiding potential issues of overfitting.

The final preprocessing step involved the encoding of activity labels. While the dataset provided human-readable activity labels in the form of numerical codes, the machine learning model required these labels to be converted into a numeric format. This transformation is known as label encoding and was implemented to facilitate the model's understanding of the target variable. For instance, the activity 'walking' with a label of '1' was encoded as '1', making it compatible with the algorithm's requirements.

The data preprocessing phase was pivotal in shaping the dataset into a format suitable for classification. By carefully selecting relevant features, eliminating extraneous columns, and preparing the target variable through label encoding, we laid the foundation for subsequent stages. With the dataset primed and redefined, we proceeded to apply the k-Nearest Neighbors Classifier with GridSearch, as detailed in the subsequent sections, unlocking the potential for accurate activity recognition.

Hyperparameter Tuning with GridSearch

The success of our activity recognition model hinged not only on the choice of the k-Nearest Neighbors Classifier(k-NN) algorithm but also on the selection of appropriate hyperparameters.

Hyperparameters are settings that are not learned from the data but must be specified prior to training the model. To identify the optimal combination of hyperparameters for our k-NN model, we employed the powerful technique of GridSearch.

Parameter Grid Definition: GridSearch is a systematic approach to hyperparameter tuning that involves exploring predefined combinations of hyperparameter values to identify the set that yields the best model performance. For our project, we meticulously constructed a parameter grid, considering the specific hyperparameters that could significantly influence the k-NN model's accuracy. The parameter grid consisted of the following hyperparameter options.

- **N neighbors:** We considered a range of values for the number of neighbors(k) to be considered when making predictions. The choices in our parameter grid were '[3, 5, 7, 9]'. This exploration allowed us to assess the impact of different neighborhood sizes on the model's ability to classify activities accurately.
- **Weights:** The choice of weighting function used in prediction is a crucial hyperparameter for k-NN. We included two weight options in the grid:
  '**uniform'**: This weight assigns equal importance to all neighbors in the prediction.
  'distance': This weight assigns greater importance to closer neighbors, which can be particularly relevant in scenarios where closer neighbors are more likely to have similar activity patterns.

- **P (Power Parameter for Distance Metric)**: The Minkowski distance metric used in k-NN classification is parameterized by the power parameter 'p'. We explored two options for 'p':
  '1': Corresponding to the Manhattan distance metric, where the distance is calculated as the sum of absolute differences between feature values.
  '2': Corresponding to the Euclidian distance metric, which calculates distance as the square root of the sum of squared differences between feature values.

GridSearch meticulously evaluated the performance of our k-NN model across all possible combinations of the hyperparameter options in the parameter grid. It systematically trained and tested the model with each combination, employing cross-validation techniques to assess how well the model generalizes to unseen data.

Upon completion of the GridSearch process, the results revealed that the optimal combination of hyperparameters for our activity recognition model was as follows:

- **N Neighbors**: 3
- **Weights**: 'distance'
- **P**: 1

In essence, the best-performing model selected three nearest neighbors, applied to the 'distance' weighing function, and employed the Manhattan distance metric (with p set to 1) to classify activities. This combination demonstrated the highest accuracy on our testing data, indicating that it was the most effective configuration for our specific activity recognition task.

The hyperparameter tuning phase with GridSearch was pivotal in enhancing our model's accuracy, fine-tuning it to perform optimally for our data set and classification objectives. By systematically exploring a range of hyperparameter combinations and identifying the most suitable set, we ensured that our k-NN model was primed to excel in recognizing and categorizing activities based on accelerometer data.

Results

The culmination of our project saw the application of the k-Nearest Neighbors Classifier (k-NN) with GridSearch for activity recognition based on an extensive dataset containing accelerometer data. This section provides an in-depth analysis of the results obtained through our rigorous approach.

Model Approach: The k-NN model, finely tuned through GridSearch, exhibited exceptional performance in the recognition of various human activities. The primary evaluation metric used to assess the model's accuracy was the classification accuracy, representing the proportion of correctly classified instances in the testing data.

The results were highly promising, with the k-NN model achieving an impressive classification accuracy of 96.6% on the testing data. This level of accuracy underscores the model's efficacy in accurately categorizing activities based on accelerometer readings. It is a testament to the robustness of our approach, reflecting the meticulous data preprocessing, feature selection, and hyperparameter tuning that culminated in a high-performance model.

Key Findings

Through the comprehensive analysis of the model's performance, several key findings emerged:

Optimal Hyperparameters: The GridSearch process identified the optimal hyperparameters for the k-NN model as follows:

- **Number of Neighbors:** 3
- **Power Parameter for Distance Metric:** 1(Manhattan distance)
- **Weights:** 'distance'

These hyperparameters collectively contributed to the model's accuracy by enabling it to consider the nearest three neighbors, assign greater importance to closer neighbors using the 'distance' weighting, and employ the Manhattan distance metric.

Impact of Data Preprocessing: The meticulous data preprocessing steps, including feature selection, column removal, and label encoding, played a pivotal role in enhancing the model's performance. They ensured that the model received relevant and well-structured data, minimizing noise and facilitating accurate classification.

Generalization: The model's impressive accuracy on the testing data highlighted its ability to generalize effectively to previously unseen data. This is a crucial characteristic for any machine learning model, indicating that it can be relied upon to perform well in real-world scenarios beyond the training dataset.

Conclusion

In conclusion, our project represents a significant achievement in the realm of activity recognition using accelerometer data. The application of the k-Nearest Neighbors Classifier with GridSearch yielded compelling results, attaining a remarkable accuracy rate of 96.6% in classifying diverse human activities. This achievement is a testament to the effectiveness of our approach and the potential of machine learning in understanding and categorizing physical activities.

Practical Implications

The implications of our work extend across various domains:

- **Health monitoring:** Accurate activity recognition can be instrumental in health monitoring applications, helping healthcare professionals track patient activity levels and identify deviations from normal patterns.
- **Sports Analytics:** Our model's ability to recognize activities like running and cycling has direct relevance in sports analytics, enabling coaches and athletes to gain insights into training and performance.
- **Human-Computer Interaction:** In the realm of human-computer interaction, our model can be integrated into systems that respond to human activities, fostering intuitive and responsive technology.

Future Directions:

While our project has achieved significant success, there remain avenues for further exploration and improvement:

- Multi-Sensor Integration: Incorporating data from multiple sensors could enhance the accuracy and richness of activity recognition.
- Real-Time Recognition: Developing a real-time recognition system could enable immediate feedback and intervention in scenarios like fall detection for the elderly.

Summary

In summary, our project serves as a foundational stepping-stone in the domain of activity recognition. Its robustness, accuracy, and versatility position it is a valuable tool with diverse applications, poised to drive innovation and insights across various fields. The journey of understanding and categorizing human activities through machine learning continues, offering limitless potential for future advancements.