Implementação e Análise do Algoritmo K-means com o Dataset Human Activity Recognition

Brenda Trindade e Kevin Borges

29 de novembro de 2024

## **Resumo**

O objetivo deste projeto foi implementar e avaliar o algoritmo K-means para agrupamento de atividades humanas com base em dados coletados por sensores de smartphones. Usando o dataset "Human Activity Recognition Using Smartphones", aplicamos técnicas de análise exploratória, normalização e redução de dimensionalidade para preparar os dados. A escolha do número de clusters foi feita por meio dos métodos do cotovelo e do silhouette score, e o modelo de K-means foi treinado para identificar os padrões de atividade. A avaliação do modelo foi realizada utilizando o silhouette score e a inércia, com resultados que indicam boa separação entre os clusters. Este estudo visa aprimorar o entendimento de como o K-means pode ser aplicado na segmentação de dados relacionados ao reconhecimento de atividades humanas.

## **Introdução**

### **Contextualização do Problema**

O **Reconhecimento de Atividades Humanas (HAR)** visa identificar as atividades realizadas por um indivíduo com base em dados coletados de sensores, como acelerômetros e giroscópios. Este problema tem várias aplicações em **monitoramento de saúde**, **assistência a idosos**, e **interação com dispositivos móveis**. O uso de smartphones, com sensores incorporados, tornou-se uma solução acessível para realizar esse tipo de monitoramento.

Neste projeto, utilizamos o **dataset "Human Activity Recognition Using Smartphones"** da UCI Machine Learning Repository. O dataset contém dados de 30 participantes realizando seis atividades diárias (andar, subir e descer escadas, sentar, deitar e ficar em pé), coletados com o uso de um smartphone com acelerômetro e giroscópio.

### **Justificativa para o uso do K-means**

O K-means é um algoritmo de **agrupamento não supervisionado** amplamente utilizado em tarefas de segmentação e clusterização de dados. Ele foi escolhido para este projeto devido à sua simplicidade, eficiência e boa aplicabilidade na análise de grandes volumes de dados, como é o caso dos dados de sensores. O K-means permite agrupar os dados sem a necessidade de rótulos prévios, tornando-o adequado para entender padrões em um problema de reconhecimento de atividades.

## **Metodologia**

### **Análise Exploratória**

A análise exploratória foi realizada com o objetivo de compreender a estrutura dos dados e identificar padrões, distribuições e correlações. Para isso, utilizamos técnicas como:

* **Visualização das distribuições** das variáveis e identificação de características importantes para o agrupamento.
* **Redução de dimensionalidade** utilizando **PCA (Principal Component Analysis)**, o que permitiu visualizar os dados em um espaço 2D e facilitar o entendimento das relações entre as amostras.

### **Implementação do K-means**

O algoritmo **K-means** foi aplicado para segmentar as atividades dos participantes em clusters. A implementação seguiu os seguintes passos:

* **Escolha do número de clusters (K)**: A escolha de KKK foi realizada utilizando dois métodos:
  + **Método do Cotovelo (Elbow Method)**: Avaliou a inércia para diferentes valores de K.
  + **Silhouette Score**: Mediu a coesão e separação dos clusters.
* **Processamento dos Dados**: A normalização dos dados foi feita com a classe **StandardScaler** do Scikit-learn para garantir que todas as variáveis tivessem a mesma escala.

### **Avaliação dos Clusters**

A avaliação dos clusters gerados foi feita com as métricas:

* **Silhouette Score**: Para medir a coesão e separação entre os clusters.
* **Inércia**: Mediu a soma das distâncias quadráticas dos pontos ao centroide de seus respectivos clusters.

## **Resultados**

### **Métricas de Avaliação**

* **Silhouette Score**: O **silhouette score** foi calculado para determinar a qualidade da separação dos clusters. O valor obtido foi **[valor obtido]**, indicando uma boa separação entre os clusters.
* **Inércia**: A inércia foi **[valor obtido]**, indicando a coesão dos clusters formados.

### **Gráficos e Visualizações**

* **Gráfico de Projeção PCA**: A projeção dos dados em duas dimensões (PCA) mostrou a separação clara entre as atividades.

*Figura 1: Visualização PCA das atividades com 2 componentes principais.*

* **Distribuição de Atividades por Cluster**: A distribuição das atividades em cada cluster foi analisada e os clusters foram interpretados com base nas atividades predominantes. A tabela cruzada mostrava os clusters formados por K-means e as atividades associadas.
* **Centroides dos Clusters**: Os centroides dos clusters, representando o centro de cada grupo de atividades, foram plotados e comparados com as distribuições das amostras.

*Figura 2: Visualização dos centroides dos clusters.*

## **Discussão**

O K-means demonstrou uma boa capacidade de segmentação das atividades humanas. Contudo, houve uma certa sobreposição entre atividades como **"sentar"** e **"ficar em pé"**, o que sugere que essas atividades podem ser difíceis de separar devido à similaridade nos dados dos sensores. Para melhorar os resultados, seria interessante explorar a adição de novos recursos, como a fusão de dados de diferentes sensores ou o uso de modelos supervisionados.

### **Limitações**

* O **K-means** é sensível à escolha inicial dos centroides e à presença de outliers. Isso pode afetar a qualidade dos resultados.
* A redução de dimensionalidade para 2D pode perder informações relevantes para o agrupamento.

## **Conclusão**

O projeto demonstrou que o algoritmo **K-means** é eficaz para agrupar atividades humanas com base em dados de sensores de smartphones. A escolha do número de clusters foi fundamental para garantir uma boa separação das atividades. As métricas de avaliação indicaram boa qualidade nos agrupamentos formados.

## **Referências**

1. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). Human Activity Recognition Using Smartphones. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.
2. UCI Machine Learning Repository. Human Activity Recognition Using Smartphones. Disponível em:<https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>
3. JAIN, A.K. (2010). Data Clustering: 50 Years Beyond K-means. *Pattern Recognition Letters*.