PUC-SP

Mineração de Dados

Prof. Dr. Daniel Rodrigues da Silva

Mineração de Dados

PUC-SP

MeanShift

Bibliografia básica:

Introdução à mineração de dados : conceitos básicos, algoritmos e aplicações. Leandro Nunes de Castro e Daniel Gomes Ferrari. – São Paulo : Saraiva, 2016.

Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina . André Carlos Ponce de Leon Ferreira et al. 2 Ed. LTC, 2024.

O método MeanShift (deslocamento médio)

O MeanShift é usado para identificar clusters em datasets onde o número de clusters não é conhecido a priori.

O método procura clusters deslocando iterativamente os pontos de dados para as regiões mais densas no espaço de recursos.

Por isso, o método é particularmente útil em aplicações como reconhecimento de objetos, onde segmenta imagens com base na intensidade e na cor do pixel, e também para rastrear objetos em sequências de vídeo.

PUC-SP

O método MeanShift (deslocamento médio)

Como um algoritmo de busca, ele rotula os clusters encontrando os modos, ou picos, na distribuição de dados.

Na prática, o método destaca as áreas mais densas.

Isso é feito deslocando iterativamente os centros de cluster para regiões de maior densidade de dados.

Inicialização: Inicie considerando cada ponto de dados como um potencial candidato para o centro do cluster.

Densidade Estimativa: Para cada ponto do seu dataset, defina uma janela ao redor dele (raio), e calcule a média dos pontos de dados dentro desse raio.

Mudança: Desloque cada ponto para a posição média. Essa etapa move o ponto em direção à região de maior densidade.

Convergência: Retorne às etapas 2 e 3 iterativamente até a convergência, ou seja, quando o deslocamento for menor que um limite predeterminado.

Uma mudança muito pequena no deslocamento indica que os pontos se estabilizaram em torno dos máximos locais da função de densidade.

Ao fim do processo, os pontos de dados estarão agrupados, formando os clusters.

A mudança de média é particularmente flexível porque se baseia na distribuição real dos dados em vez de assumir uma forma predefinida para os clusters, o que permite lidar com clusters de formas bem arbitrárias.

PUC-SP

Onde usar o MeanShift:

O **MeanShift** é uma ferramenta poderosa para descobrir a estrutura subjacente dos dados sem fazer nenhuma suposição sobre seus parâmetros, como o número de clusters ou sua forma. Considerando sua abordagem baseada em densidade que se concentra em regiões de alta densidade de dados, ela é robusta contra outliers.

Essa robustez o torna adequado para datasets do mundo real, que geralmente contêm irregularidades e ruídos, e também para aplicativos que exigem adaptabilidade e precisão.

Algumas aplicações do MeanShift:

Segmentação de imagens

O **MeanShift** pode segmentar imagens em regiões com base na intensidade ou na cor do pixel sem exigir conhecimento prévio do número de segmentos.

Essa flexibilidade o torna altamente eficaz para tarefas de segmentação de imagens, pois os clusters gerados com o **MeanShift** podem ter qualquer forma ou tamanho.

Rastreamento de objetos em análise de vídeo

PUC-SP

O MeanShifts é também útil e utilizado frequentemente no rastreamento de objetos em fluxos de vídeo.

Sua capacidade em identificar e seguir objetos dinamicamente à medida que eles se movem pelos quadros, o torna muito eficaz para aplicativos de rastreamento em tempo real.

Segmentação de clientes em marketing

O MeanShift é capaz de analisar o comportamento do cliente, os padrões de compra e os dados demográficos para descobrir agrupamentos, independentemente do tamanho do grupo.

Isso ajuda a criar estratégias de marketing direcionadas e experiências personalizadas, melhorando, em última análise, o envolvimento e a retenção do cliente.

PUC-SP

PUC-SP

Implementação:

```
import pandas as pd
from sklearn.cluster import MeanShift, estimate_bandwidth
from sklearn.datasets import make_blobs
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
 import warnings
warnings.filterwarnings("ignore") # Ignorar Avisos

# Cria um dataset aleatório
X , Y = make_blobs(n_samples = 500, centers = 5, cluster_std = 1, random_state = 27)
X1 = pd.DataFrame(X)
X1.columns = ['C1' , 'C2']
```

PUC-SP

```
# Estima a largura de banda usando a função estimate_bandwidth().
tam_banda = estimate_bandwidth(X1, quantile = 0.2, n_samples=500)
float(tam_banda)

# Desempenho do MeanShift
model = MeanShift(bandwidth = tam_banda, bin_seeding = True)
model.fit(X1)
labels = model.labels_
centers = model.cluster_centers_

# Plota os resultados
plt.scatter(X1['C1'], X1['C2'], c=labels, cmap='plasma', marker='p')
plt.scatter(centers[:, 0], centers[:, 1], s=250, c='blue', marker='X')
plt.title('Mean Shift')
plt.xlabel('Atributo 1')
plt.ylabel('Atributo 2')
plt.show()
```