



Microsoft Power BI Para Business Intelligence e Data Science

Microsoft Power BI Para Business Intelligence e Data Science

O Que é Detecção de Anomalias?

A detecção de anomalias, também conhecida como detecção de outliers, é uma técnica em Machine Learning e Estatística que visa identificar padrões incomuns, inesperados ou anômalos nos dados. Esses padrões podem ser diferentes das observações normais de várias maneiras, como magnitude, frequência ou comportamento. A detecção de anomalias é importante porque as anomalias podem indicar problemas, erros, falhas, fraudes ou atividades maliciosas e, em muitos casos, é crucial identificar e analisar esses eventos anômalos para tomar decisões informadas e apropriadas.

Existem várias abordagens para detectar anomalias em Machine Learning, algumas das quais incluem:

Métodos Estatísticos: Esses métodos baseiam-se na análise estatística dos dados, como testes de hipóteses, distribuições de probabilidade e medidas de dispersão (por exemplo, desvio padrão e intervalos interquartis). Observações que estão significativamente distantes da média ou fora dos intervalos esperados são consideradas anômalas. Esses métodos são estudados na Formação Análise Estatística aqui na DSA.

Aprendizado Supervisionado: Nesta abordagem, um modelo de Machine Learning é treinado usando um conjunto de dados rotulado, que inclui exemplos de observações normais e anômalas. O modelo aprende a distinguir entre as duas classes e, em seguida, pode ser usado para classificar novas observações como normais ou anômalas. Esse método é estudado na Formação Cientista de Dados aqui na DSA.

Aprendizado Não Supervisionado: Neste caso, os algoritmos de Machine Learning são usados para analisar dados não rotulados e identificar padrões ou agrupamentos naturais neles. As anomalias são identificadas como pontos de dados que não se encaixam bem em nenhum desses agrupamentos ou que estão significativamente distantes de outros pontos de dados. Alguns exemplos de algoritmos de aprendizado não supervisionado usados para detecção de anomalias incluem clustering (por exemplo, K-means) e técnicas de redução de dimensionalidade (por exemplo, PCA). Esse método é estudado na Formação Análise Estatística aqui na DSA e será usado agora neste capítulo.

Aprendizado Semi-Supervisionado: Esta abordagem combina elementos de aprendizado supervisionado e não supervisionado. Os algoritmos são treinados em um conjunto de dados parcialmente rotulado, que contém exemplos de observações normais e um pequeno número de exemplos anômalos. O modelo aprende a distinguir entre as classes e identificar novas anomalias com base nos padrões aprendidos.

Métodos Baseados em Densidade: Esses métodos identificam anomalias como pontos de dados que estão localizados em áreas de baixa densidade do espaço de recursos (atributos). Um exemplo popular de algoritmo de detecção de anomalias baseado em densidade é o DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

Métodos Baseados em Vizinhança: Esses métodos comparam a distância ou similaridade entre pontos de dados e seus vizinhos para identificar anomalias. Os pontos de dados que têm vizinhos significativamente diferentes de si mesmos são considerados anômalos. Exemplos de algoritmos que empregam essa abordagem incluem o k-NN (k-Nearest Neighbors) e o LOF (Local Outlier Factor). O k-NN é estudado na Formação Cientista de Dados e o LOF no curso de Cyber Security Data Science da Formação Engenheiro de IA.

A escolha do método mais adequado para detecção de anomalias depende do contexto, da natureza dos dados, do conhecimento técnico do profissional de análise e do tipo de problema que desejamos resolver.