Relatório LAMIA 11
Prática: Predição e a Base de Aprendizado de Máquina (II)

Kaique Medeiros Lima

1 Introdução

Esse card aborda conceitos e técnicas de aprendizado de máquina para aprimorar a compreensão sobre o tema. Serão discutidos métodos de aprendizado supervisionado e não supervisionado, validação cruzada, modelos preditivos, entropia, árvores de decisão, *ensemble learning*, *XGBoost*, Support Vector Machine e outros tópicos relevantes.

2 Descrição da atividade

2.1 Unsupervised Learning

No *Unsupervised Learning*, o modelo não recebe respostas ou rótulos para aprender. Ele deve encontrar padrões e fazer descobertas por conta própria, apenas com base em observações dos dados. É uma abordagem útil quando não sabemos previamente quais classificações ou agrupamentos estamos procurando nos dados.

2.2 Supervised Learning

No *Supervised Learning*, o modelo é treinado com dados rotulados para aprender a prever saídas desconhecidas. Ele utiliza exemplos com respostas corretas para ajustar suas previsões e, uma vez treinado, é capaz de fazer previsões sobre novos dados.

2.3 Train/Test Split

A técnica de *Train/Test Split* divide os dados em dois subconjuntos: um para treinamento e outro para teste. Essa abordagem avalia a capacidade do modelo de generalizar. A divisão deve ser aleatória, com conjuntos grandes o suficiente para capturar variações importantes, ajudando a evitar *overfitting*.

2.3.1 K-Fold Cross Validation

Na validação cruzada *K-Fold*, os dados são divididos em múltiplos segmentos (*k-folds*). Um desses segmentos é reservado como conjunto de testes, enquanto os outros são usados para o treinamento do modelo. O processo é repetido até que todos os segmentos tenham sido usados como conjunto de testes pelo menos uma vez. Isso garante uma avaliação mais robusta do modelo.

2.4 Métodos Bayesianos

Os métodos bayesianos dividem os dados em partes e utilizam o *Teorema de Bayes* para calcular a probabilidade de um dado pertencer a uma determinada classe. É uma técnica baseada em probabilidades condicionais.

2.5 K-Means Clustering

O *K-Means Clustering* é um algoritmo de *Unsupervised Learning* usado para agrupar dados em K clusters, onde cada ponto de dado pertence ao grupo cujo centróide está mais próximo. O processo envolve:

- Escolher K centróides aleatórios.
- Atribuir cada ponto de dado ao centróide mais próximo.
- Recalcular os centróides com base nos novos agrupamentos e repetir o processo até que não haja mais mudanças.

2.6 Modelos Preditivos

2.6.1 Regressão Linear

A regressão linear ajusta uma linha reta a um conjunto de observações, com o objetivo de prever valores para dados novos ou não observados. Pode ser usada para prever pontos tanto no passado quanto no futuro, apesar do nome *regressão*. A eficiência do modelo é medida pelo coeficiente R^2, onde 0 indica uma correlação fraca e 1 indica uma boa correlação.

2.6.2 Regressão Polinomial

A *regressão polinomial* é usada quando uma linha reta não se ajusta adequadamente aos dados. Nesse caso, uma curva polinomial de maior grau é ajustada ao gráfico. Quanto maior o grau do polinômio, mais complexa será a curva.

2.6.3 Regressão Múltipla

A *regressão múltipla* permite analisar a relação entre uma variável dependente e múltiplas variáveis independentes. Esse modelo ajuda a prever o comportamento de uma variável com base em várias outras.

2.6.4 Modelos Multiníveis

Os modelos multiníveis são usados para a análise de dados com estrutura hierárquica ou *nested*, onde diferentes níveis de dados podem influenciar uns aos outros. Esses modelos são úteis para situações em que há dependência entre as observações.

2.7 Entropia

A *entropia* é uma medida de quantificação da desordem nos dados, ou seja, o quão iguais ou diferentes eles são. Uma entropia de 0 indica que os dados são totalmente iguais, enquanto uma entropia de 1 sugere que são completamente diferentes.

2.8 Árvores de Decisão

As *árvores de decisão* são um modelo de aprendizado supervisionado que segmenta os dados repetidamente subconjuntos com base em certas condições. Essa segmentação é representada em uma estrutura de árvore, onde cada folha corresponde a uma decisão e cada galho representa a consequência dessa decisão.

2.9 Ensemble Learning

O *ensemble learning* é uma técnica de aprendizado de máquina onde vários modelos são treinados para resolver o mesmo problema, permitindo a combinação de seus resultados para melhorar o desempenho geral. Essa abordagem oferece resultados superiores em comparação ao uso de modelos isolados. Entre as técnicas de *ensemble learning*, destaca-se o *boosting*, onde os modelos são treinados em sequência, ajustando os erros cometidos pelo modelo anterior, tornando cada etapa mais precisa. Outra abordagem é o *bucket of models*, na qual diversos modelos são treinados, mas apenas o que obtém o melhor desempenho nos dados de treinamento é escolhido. Já na técnica *stacking*, múltiplos modelos são treinados simultaneamente e seus resultados são combinados para criar uma solução final mais robusta.

2.10 XGBoost

O *XGBoost* é uma implementação do *gradient boosting* que combina diversos modelos de aprendizado para formar um modelo mais forte. Ele lida automaticamente com valores ausentes e aproveita o processamento paralelo para acelerar o treinamento.

2.11 Support Vector Machine (SVM)

A *Support Vector Machine* (SVM) busca encontrar o hiperplano que separa os dados em diferentes classes com máxima eficiência. Esse método é especialmente útil quando os dados possuem múltiplas dimensões, oferecendo uma separação eficaz em contextos complexos.

Conclusão

Todos esses conceitos e técnicas são fundamentais para a compreensão e aplicação de aprendizado de máquina em diversos contextos. A escolha do método mais adequado depende do problema a ser resolvido e dos dados disponíveis. A prática contínua e a exploração de diferentes abordagens são essenciais para aprimorar as habilidades em aprendizado de máquina e obter resultados mais precisos e confiáveis.