

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

Willian Augusto Soder de Souza

O objetivo deste relatório é descrever os principais conhecimentos adquiridos ao assistir às duas seções indicadas do curso "Machine Learning, Data Science and Deep Learning with Python". Nessas seções, o professor apresenta diversos conceitos importantes sobre predições e sobre os fundamentos do Machine Learning, combinando teoria e prática de maneira didática utilizando Python e várias bibliotecas, como a scikit-learn.

Abaixo, segue um resumo dos principais conceitos teóricos ensinados nas aulas, que são extremamente importantes dentro do campo do Machine Learning:

- **Linear Regression:** é uma técnica estatística usada para modelar a relação entre uma variável dependente (ou variável resposta) e uma ou mais variáveis independentes (ou variáveis preditoras). A ideia central da regressão linear é encontrar a linha reta que melhor se ajusta aos dados observados. Esta linha reta é conhecida como a linha de regressão.
- **Polynomial Regression:** é uma extensão da regressão linear que permite modelar a relação entre a variável dependente e uma ou mais variáveis independentes quando essa relação não é linear. Em vez de ajustar uma linha reta aos dados, a regressão polinomial ajusta uma curva polinomial, que pode capturar melhor a complexidade dos dados.
- **Multiple Regression:** é uma extensão da regressão linear que permite modelar a relação entre uma variável dependente e várias variáveis independentes. Essa técnica é usada para entender como várias variáveis preditoras juntas afetam a variável resposta.
- **Multi-Level Models:** são uma extensão dos modelos de regressão que permitem a análise de dados com estruturas hierárquicas ou aninhadas.
- **Supervised Learning:** é uma abordagem de aprendizado de máquina em que um algoritmo é treinado com um conjunto de dados rotulados. Nesse contexto, "rotulados" significa que cada exemplo de treinamento é composto por uma entrada (input) e a saída esperada (output). O objetivo do algoritmo é aprender a mapear as entradas para as saídas corretas, de modo que possa fazer previsões precisas sobre novos dados não vistos.
- **Unsupervised Learning:** é uma abordagem de aprendizado de máquina em que um algoritmo é treinado com dados que não possuem rótulos ou saídas esperadas. O objetivo desse tipo de aprendizado é descobrir padrões ou estruturas ocultas nos dados. Em contraste com o aprendizado supervisionado, onde o modelo é treinado com pares de entrada-saída, no aprendizado não supervisionado o modelo trabalha apenas com entradas e tenta encontrar relações ou agrupamentos significativos dentro dos dados.
- **Train/Test:** o conjunto de treino é usado para ajustar o modelo, fornecendo exemplos de dados com suas respectivas respostas para que o modelo aprenda a fazer previsões. Já o conjunto de teste é um conjunto separado de dados que o modelo não viu durante o treinamento, utilizado para avaliar a performance do modelo de forma imparcial, garantindo que ele generalize bem para novos dados e não apenas memorize os exemplos de treino. Esses conjuntos devem ser bem ajustados para garantir um bom resultado do algoritmo.
- **K-fold Cross Validation:** é uma técnica para avaliar a performance de um modelo de aprendizado de máquina.
- **Bayes' Theorem:** é feito uma revisão sobre esse princípio fundamental da probabilidade que descreve como atualizar a probabilidade de um evento com base em novas evidências.

- **K-Mean Clustering:** é um algoritmo de agrupamento que visa dividir um conjunto de dados em K grupos (ou clusters) distintos, com base na similaridade entre os dados. O objetivo é minimizar a variância dentro de cada cluster, ou seja, a distância entre os pontos dentro de um mesmo cluster deve ser a menor possível.
- **Entropy:** é a medida da desordem de um conjunto de dados, indicando o quão homogêneo ou heterogêneo ele é.
- **Decision Trees:** são modelos de aprendizado de máquina que funcionam dividindo repetidamente um conjunto de dados em subconjuntos mais homogêneos, baseando-se em variáveis de entrada. O objetivo é criar uma estrutura de árvore onde cada nó interno representa uma decisão sobre um atributo, cada ramo representa o resultado dessa decisão, e cada nó folha representa uma previsão ou resultado final.
- **Ensemble Learning:** é uma técnica de aprendizado de máquina que combina múltiplos modelos para melhorar a performance preditiva em relação a um único modelo. A ideia é que, ao combinar as previsões de vários modelos, o ensemble possa corrigir os erros individuais dos modelos e proporcionar uma predição mais robusta e precisa.
- **Support Vector Machines (SVM):** são um conjunto de algoritmos de aprendizado de máquina supervisionado usados principalmente para tarefas de classificação, mas também podem ser aplicados para regressão e detecção de outliers. O principal objetivo das SVMs é encontrar um hiperplano em um espaço de alta dimensão que separa as classes de forma ótima.

CONCLUSÃO

Essas aulas ensinam conceitos extremamente importantes para o mundo do machine learning. Entender esses conceitos forma a base teórica e prática para a construção de modelos preditivos eficientes. Eles permitem entender como os dados podem ser manipulados, analisados e utilizados para fazer previsões precisas, identificar padrões ocultos, reduzir incertezas e tomar decisões informadas. Além disso, o conhecimento desses princípios é crucial para escolher e aplicar os algoritmos mais adequados a diferentes tipos de problemas, avaliar a performance dos modelos e garantir que eles se generalizem bem para novos dados.

Compreender esses conceitos teóricos também facilita a comunicação e colaboração com outros profissionais da área, contribuindo para o avanço de projetos e inovações tecnológicas. Portanto, um domínio sólido dessas técnicas é um diferencial significativo para aqueles que desejam se destacar e avançar em suas carreiras no campo de Data Science e Machine Learning.