Relatório 12 - Prática: Predição e a Base de Aprendizado de Máquina (II)

Lucas Augusto Nunes de Barros

Descrição das atividades

No card são abordados tópicos como regressão linear, regressão polinomial, além de abordar com novas práticas algoritmos já vistos como o Naive-Bayes, K-Means e RandomForest.

A regressão é uma técnica estatística usada para encontrar a equação que melhor correlaciona os dados existentes e fazendo uso desta equação é possível estimar resultados para novos dados de entrada. Pode ser dividida em regressão linear, que trabalha apenas com gráficos do primeiro grau, regressão polinomial, abrangendo os polinômios de grau 2 ou maior e ainda a regressão múltipla, onde é possível relacionar a variável de saída com mais de uma variável de entrada.

Existem diversos métodos que permitem encontrar a linha no gráfico que melhor relaciona os dados disponíveis, o método dos mínimos quadrados, que tem como objetivo minimizar a distância entre os pontos e a reta ideal, é bastante utilizado devido sua eficiência computacional.

O coeficiente de determinação, também chamado de R^2, é uma medida de ajuste para modelos de regressão, essa métrica mensura a variação dos dados, oscilando entre 0 e 1, onde 0 significa que o modelo não consegue regredir com eficiência, enquanto o valor 1 indica um ajuste excelente, onde o modelo consegue regredir bem com todas as variações de dados.

O algoritmo de Naive-Bayes, abordado anteriormente, é amplamente utilizado em classificadores. Opera analisando um conjunto de dados e considerando a probabilidade de cada uma das *features* pertencer a uma ou outra classe. Essa abordagem permite que o classificador avalie de forma eficiente as probabilidades. O exemplo prático para uso desse método foi um classificador de e-mail que avalia o conteúdo de cada e-mail e baseado na análise do texto classifica como spam ou não-spam.

Outro algoritmo já visto que foi abordado é o *K-Means*, uma técnica de aprendizado de máquina não supervisionada. Divide os dados em *K* grupos previamente definidos. De início são calculados os centróides, que são os k pontos calculados, onde cada um representa o centro de uma classe, feito isso o algoritmo relaciona cada ponto de dados ao centróide mais próximo. Para cada uma dessas

iterações, novos centróides são calculados como a média dos pontos dentro de cada classe.

A árvore de decisão é um algoritmo de aprendizado de máquina que funciona criando uma estrutura hierárquica semelhante a uma árvore, onde cada nó representa uma decisão baseada em características dos dados. Essas decisões são tomadas para dividir os dados em grupos cada vez mais específicos, até que se alcance uma previsão. Cada ramo da árvore representa uma escolha, enquanto as folhas representam os resultados.

Para classificar uma fruta como maçã ou laranja, a árvore de decisão deve buscar características relevantes, como a cor e o diâmetro. Por exemplo, a primeira questão poderia ser se a fruta é vermelha e com base na resposta o algoritmo pode seguir para outra pergunta, afunilando cada vez o número de possibilidades. A árvore de decisão é intuitiva e fácil de interpretar. No entanto, uma única árvore pode ser propensa a *overfitting*, ou seja, ajuste excessivo aos dados de treinamento. Por isso, em muitos casos, várias árvores são combinadas em métodos como o Random Forest, que melhora a precisão e a robustez do modelo.

Já o algoritmo *Random Forest*, como já foi deixado claro, combina várias árvores de decisão para criar um modelo mais robusto e menos propenso a *overfitting*. Ele funciona criando múltiplas árvores de decisão, cada uma treinada com um subconjunto aleatório dos dados e um conjunto aleatório de características. No final, as previsões de todas as árvores são combinadas, geralmente por meio de uma votação, em casos de classificação, ou da média, para casos de regressão, e então o resultado final é encontrado.

Um algoritmo especial abordado neste card foi o XGBoost (*Extreme Gradient Boosting*) é um algoritmo de aprendizado de máquina eficiente e poderoso, usado para problemas de classificação, regressão e ranking. Baseado no conceito de *Gradient Boosting*, ele constrói modelos sequenciais, onde cada novo modelo tenta corrigir os erros do anterior, utilizando técnicas de gradiente para minimizar a função de perda. O XGBoost se destaca por sua capacidade de evitar *overfitting* através de regularização, além de ser altamente otimizado. Sua flexibilidade e precisão o tornam uma escolha popular para aplicações reais.

Por fim, o SVM (*Support Vector Machine*) é um algoritmo usado para classificação e para regressão, funciona encontrando um hiperplano que melhor separa os dados em classes, maximizando a margem entre os pontos mais próximos de cada classe, chamados de vetores de suporte. Quando aplicado especificamente a problemas de classificação, como separar e-mails em "spam" ou "não spam", ele é chamado de SVC (*Support Vector Classification*). Já para problemas de regressão, como prever o preço de um carro, usa-se o SVR (*Support Vector Regression*). O SVM pode lidar com dados não linearmente separáveis, tornando-o abrangente em diversas aplicações.

Conclusão

O card abordou de maneira prática os algoritmos de aprendizado de máquina, destacando diversos métodos, permitindo compreender a importância e a funcionalidade de técnicas distintas para cada situação. É evidente a versatilidade dos algoritmos existentes hoje, por isso conhecer seu funcionamento e suas características é o primeiro passo para alcançar uma profunda compreensão sobre machine learning e suas aplicações práticas.

Referências

[1] card12