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

Yuri Vacelh Zamulhak Zdebski

Descrição da atividade

Assistir ao curso Machine Learning, Data Science and Deep Learning with Python, as seções previstas tinham como objetivo a introdução a modelos preditivos e machine learning com python.

Regressões

Linear

O modelo preditivo mais simples apresentado, consiste em encaixar uma reta nos dados de observação, e se utilizar dela para realizar predições (considerando que os dados sigam o mesmo padrão).

A implementação básica envolve minimizar a soma dos erros quadrados (diferença entre os valores reais e os valores previstos pelo modelo). O processo de treinamento ajusta os coeficientes (pesos) da equação da linha reta, $y = mx + b$, onde m é o coeficiente angular e b é o coeficiente linear.

Uma vez ajustado, o modelo pode ser utilizado para prever valores de y com base em novos valores de x. Algumas das formas de minimizar o erro são o gradiente descendente e iterar até encontrar uma linha que se encaixe no contorno dos dados.

Polinomial

Parecido com a regressão linear, porém ao invés de utilizarmos uma equação de reta, são utilizados polinômios, como $y = ax^2 + bx + c$. E modelos desse tipo é importante tomar cuidado para não utilizar mais graus do que o necessário, pois isso levaria a um overfitting, ou seja, o modelo se acostumaria demais com os dados de treino, podendo gerar confusão quando apresentados dados para validação.

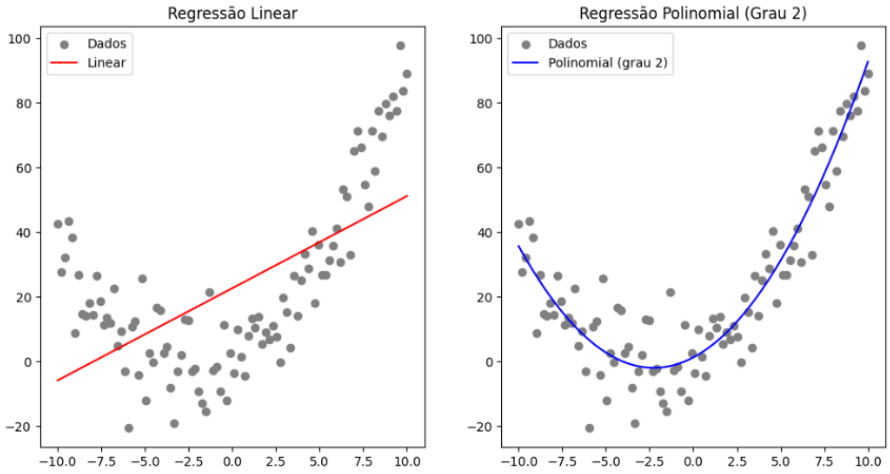
Múltipla e Multivariável

Uma regressão múltipla contém varias variáveis independentes, como por exemplo, para prever o valor de um carro, levamos em conta diversos fatores como a cor, a quilometragem, estado, ano etc.

Já uma regressão multivariável é uma regressão que tem diversas variáveis dependentes, ou seja, tenta prever varias coisas.

Quando se for utilizar uma técnica de regressão, é sempre importante escolher uma que se adéque aos dados obtidos.

Figura 1: Exemplo de regressão Linear e Polinomial



Fonte: Autoria própria

Modelos Multi-Nível

Foi apenas comentado no curso, mas consiste em analisar não apenas os dados, mas os diversos níveis que geraram eles, como por exemplo a nota de uma criança depende, além das respostas que ela marcou, de coisas como ambiente familiar, se ela descansou no dia anterior, a qualidade do ensino etc.

O que é Machine Learning

São algorítimos que aprendem processando dados de treino e podem fazer predições com base nisso. A parte do aprendizado pode ser dividida em supervisionado, não supervisionado e por reforço, mas nessa sessão vamos falar apenas dos 2 primeiros.

Não supervisionado

Consiste em dar os dados não tabelados para o treinamento do modelo, ou seja, o algoritmo vai encontrar por si só classes (no caso de um classificador) para os dados. É utilizado quando as nuncias dos dados não são claras, como agrupar pessoas por determinados interesses, classificar filmes e analisar texto.

Supervisionado

Os dados de treino são tabelados, ou seja, existem respostas certas e erradas durante o treinamento, isso possibilita a avaliação da performance do algoritmo. Quando se tem dados o suficiente, é interessante separar em conjuntos de treino e teste, criando assim métricas para validação.

Na pratica os conjuntos de treino e teste ajudam a evitar o overfitting, e quando necessário, podemos utilizar a validação cruzada para testar e encontrar parâmetros (como o grau de uma regressão) ótimos para o modelo.

Métodos Bayesianos

Se baseiam no teorema de Bayes, que é definido por:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

sendo que:

- P(A|B) é a probabilidade de A acontecer sendo que B aconteceu
- P(B|A) é a probabilidade de B acontecer sendo que A aconteceu
- P(A) é a probabilidade de A acontecer
- P(B) é a probabilidade de B acontecer

O exemplo mostrado no curso utilizava o Teorema de Bayes para calcular a probabilidade de uma mensagem ser spam com base em características específicas, como palavras ou expressões no texto. O sistema analisa a frequência dessas palavras em e-mails marcados como spam e não spam, ajustando suas previsões conforme mais dados são processados. Isso permite um modelo probabilístico que, ao encontrar um e-mail novo, estima a probabilidade de ser spam com base nas características observadas e as probabilidades anteriores, aprimorando a detecção de forma adaptativa.

K-Means Clustering

É um algoritmo de classificação de aprendizado não supervisionado, que consiste em separar o conjunto de dados em K grupos mais próximos a K centroides.

Funciona da seguinte forma, pega K centroides aleatórios, separa cada ponto do conjunto com base no centroide mais próximo, recomputa os centroides com base nas posições médias de cada ponto do cluster e repete até que os pontos parem de mudar de cluster.

Existem alguns truques para o funcionamento desse algoritmo. Para a escolha de um K ideal, o esquema é ir aumentando ele gradativamente, até que o erro quadrático pare de cair drasticamente. Rodar o algoritmo algumas vezes, para evitar mínimos locais do erro quadrático. Depois que o algoritmo estiver finalizado, analisar cada cluster para classificar cada um.

Entropia

Mede a desordem do dataset, ou seja, o quão diferentes os dados são entre si. Um dataset com entropia 0 teria todas as classes iguais, já um com diversas classes teria uma entropia alta.

Para computar a entropia, a seguinte formula é utilizada:

$$H(S) = -p_1 ln(p_1) - ... - p_n ln(p_n)$$

sendo p_i a i-ésima probabilidade para cada classe de dado tabelada.

Arvores de Decisão

É uma forma de aprendizado supervisionado, onde durante a fase de aprendizado, é produzida uma arvore, onde consultas podem ser feitas para gerar classificações.

Arvores são suscetíveis a overffiting dos dados, uma solução é a criação de florestas aleatórias, onde são construídas diversas arvores, e a resposta final depende da votação delas (daí o nome, floresta). A criação dessas florestas funciona da seguinte forma, cada árvore é treinada usando uma amostra aleatória dos dados e seleciona, de forma aleatória, subconjuntos das variáveis para tomar decisões.

Ensemble Learning

Significa, basicamente, criar vários modelos e deixar eles votarem na predição final. A floresta aleatória utiliza a técnica de bagging para implementar o ensemble learning, onde vários modelos são construídos a partir de subconjuntos aleatórios dos dados de treinamento. Em contraste, o boosting é uma técnica em que cada modelo subsequente no conjunto foca em melhorar a classificação dos erros cometidos pelo modelo anterior. O conceito de bucket of models envolve treinar diferentes modelos com os mesmos dados e escolher aquele que apresenta o melhor desempenho nos dados de teste. Já o stacking executa múltiplos modelos simultaneamente nos dados e combina seus resultados para uma previsão final mais robusta.

XGBoost

XGBoost (eXtreme Gradient Boosted Trees) é uma biblioteca muito robusta que se utiliza do ensemble learning. Nela são geradas diversas arvores que vão melhorando (boosting) a cada geração.

Support Vector Machines

Support Vector Machines, ou SVM é uma forma de aprendizado supervisionado, utilizado na classificação de dados de dimensões superiores. Funciona com o chamado truque do kernel, onde representa os dados em espaços com dimensões superiores para achar hiperplanos que não aparecem em dimensões menores. Na pratica utilizamos o chamado SVC para classificar dados utilizando SVM.

Conclusão

O curso trouxe uma ampla visão de algumas bibliotecas, técnicas e conceitos utilizados no contexto de aprendizado de maquinam, trazendo tanto exemplos práticos e teóricos que possibilitam uma maior compreensão dos assuntos abordados.