Relatório 12 - Predição e a Base de Aprendizado de Máquina

Lucas Scheffer Hundsdorfer

Descrição da atividade

Seção 3 - Predictive Models

Aula 1 - Linear Regression

Regressão linear é basicamente ajustar uma linha reta a um conjunto de dados,a regressão funciona utilizando os mínimos quadrados, minimizando a soma dos erros quadrados entre cada ponto e a linha. Uma outra maneira de pensar em regressão linear é uma linha que representa a probabilidade máxima de uma observação. R-quadrado mede o quão bem a linha está ajustada aos dados, que é baseada na fração da variação total em y que é capturada pelo seu modelo, vai de 0 a 1 sendo 0 terrível e 1 perfeito. Após a explicação foi feito os testes práticos dentro do Jupyter Notebook.

Aula 2 - Polynomial Regression

A regressão polinomial é aplicada quando as relações não são em linhas retas, a fórmula da regressão linear era ($y = mx + b$), porém essa é uma fórmula de primeira ordem, uma de segunda ordem é assim ($y = ax^2 + bx + c$), quantos mais ordens mais produz curvas. Porém se você utilizar mais ordens do que o necessário pode ocorrer overfitting.

Aula 3 - Multiple Regression

Acontece quando você tenta prever algo porém com mais de variável influente no que você deseja, exemplo prever o preço de um carro baseado em quilometragem, estilo, marca, modelo, ou pode acontecer de quando se está tentando prever mais de alguma variável, usando o mesmo exemplo do carro, o quanto ele vale ou quanto ele demora para ser vendido.

Aula 4 - Multilevel Models

O conceito é que alguns efeitos acontecem em vários níveis, exemplo a saúde a sua saúde depende de uma hierarquia da saúde das suas células, órgãos, sua família, sua cidade, o seu mundo. O seu dinheiro pode ser também um exemplo, além de depender do seu próprio trabalho, pode depender dos seus parentes.

Seção 4 - Machine Learning with Python

Aula 1 - Supervised vs Unsupervised Learning and TrainTest

Primeiramente o machine learning consiste em algoritmos que conseguem ler dados e fazer observações e fazer previsões baseadas neles, e existem tipos de machine learning, como o não supervisionado, que consiste em não te dar

respostas ele te dá observações baseadas na métrica passada. Ele ajuda quando você não sabe o que está olhando.

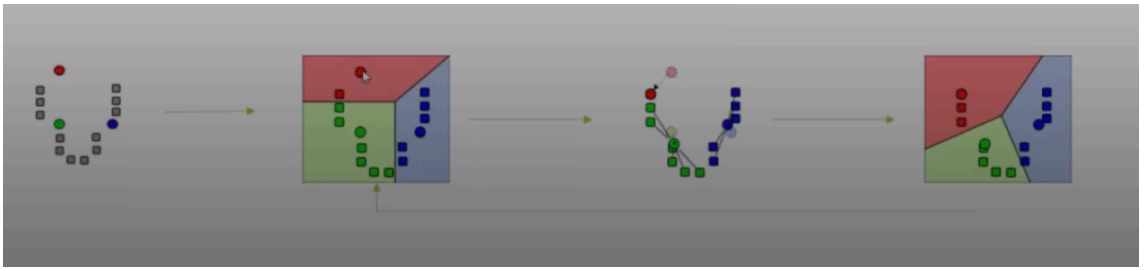
Já o aprendizado supervisionado aprende com os dados ‘corretos’, exemplo você dá dados para ele sobre precificação de carro e ele saberia prever o quanto custaria um carro X. E existe uma maneira de avaliar o aprendizado supervisionado, é basicamente pegar todos os dados e separar em dados de testes e dados de treinamento, os dados para treinamento são usados para treinar o modelo selecionado e para medir o quão bom está sendo é usar os dados de teste onde já é sabido qual é o resultado. Porém esse conjunto de dados tem de ser grande o suficiente para ter variações e outliers.

Aula 3 - Bayesian Methods Concepts

O teorema de Bayes é uma fórmula de probabilidade que calcula a possibilidade de um evento acontecer, com base em um conhecimento que pode estar relacionado ao evento, existe o naive Bayes que é um classificador probabilístico que utiliza esse teorema é uma ferramenta muito útil e está contida dentro da biblioteca de python scikit-learn.

Aula 5 - K-Means Clustering

É a tentativa de juntar os dados em K grupos onde eles estão mais próximos do K centroide, é um tipo de aprendizado não supervisionado que usa apenas as posições dos dados, exemplo:



Algumas das limitações desse método é que tem que escolher o valor certo para K e outro que não tem como colocar um nome na relação dos dados. A aula seguinte é o teste disso no Jupyter.

Aula 7 - Measuring Entropy

É basicamente a medida que mostra o quão um conjunto de dados está desordenado , ou quão diferente ou iguais dois conjuntos são.

Aula 11 - Decision Trees Concepts

Baseia-se em construir um fluxograma que te ajuda a tomar decisões tomadas a partir de uma variável através do machine learning, exemplo, criar um sistema que filtra baseado em elaborar currículos com base em dados históricos de contratações, tem uma banco de dados com alguns importantes atributos dos candidatos e você sabe quais serão contratados e quais não serão, é possível treinar uma árvore de decisão com esses dados para no futuro conseguir prever quais candidatos vão ser contratados ou não. Ela toma

funciona a partir de cada passo a árvore acha um atributo que podemos usar para particionar o conjunto de dados para minimizar a entropia dos dados na próxima etapa. Um problema com a árvore de decisão é que elas são muito propensas a overfitting, a resolução é usar uma técnica chamada floresta aleatória baseia-se em que podemos construir várias árvores de decisão alternativas e deixá-las votar na classificação final.

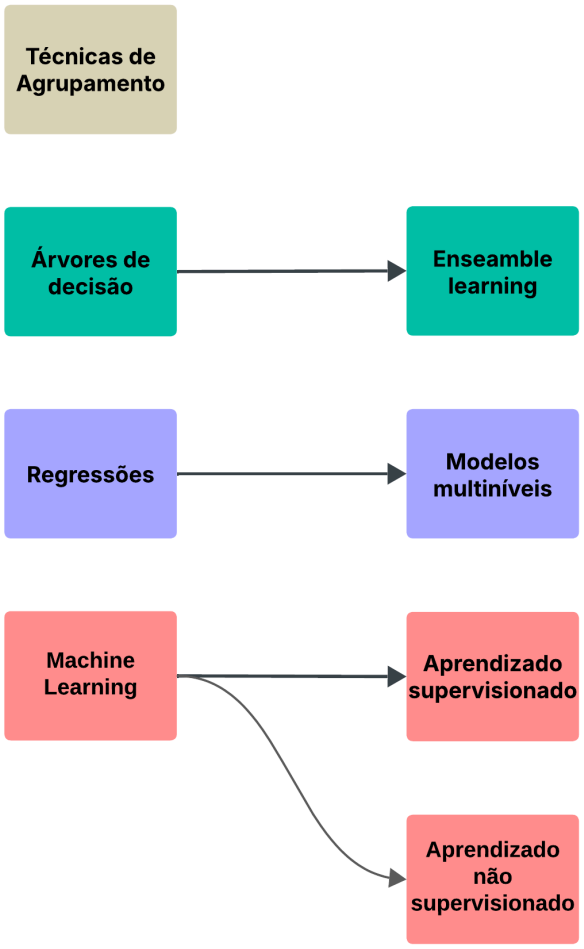
Aula 13 - Ensemble Learning

Um exemplo de aprendizado em conjunto é a técnica usada na floresta aleatória ou Random Forests, basicamente o aprendizado em conjunto é quando é escolhido vários modelos para tentar e solucionar o mesmo problema. Boosting ou impulsionamento é uma técnica que ele impulsiona o atributo errado para o modelo subsequente de uma forma que a intenção seja que o próximo modelo consiga identificar o erro. A bucket of models é quando é pegado múltiplos modelos para um único conjunto de dados e é escolhido o modelo que funciona melhor, Stacking é quando é feito o mesmo processo porém os resultados finais são combinados.

Aula 14 - XGBoost

Quase que uma junção do aprendizado em conjunto e as árvores de decisão, e é o melhor algoritmo nos dias de hoje para o machine learning, é utilizado o boosting já explicado no aprendizado em conjunto nas árvores de decisão, o atributo errado é impulsionado para que cada árvore subsequente não erre. XGBoost tem vários recursos como o ‘regularized boosting’ que previne o overfitting, outro recurso é que ele descobre a melhor maneira de lidar com os valores ausentes, seu processo pode ser realizado em paralelo, isso o que o torna tão eficiente, logo pode ser usado para big data, pode ser realizado validação cruzada em cada etapa, conseguindo avaliar cada passo com o objetivo de ver se cada etapa está sendo realmente útil, é possível também o treinamento incremental que é parar o treinamento e continuar depois, pode ser separados em períodos de tempo.

De insight visual original eu montei um mapa conceitual com algum dos conceitos aprendidos em curso, e vou dar um breve resumo sobre eles:



1. **Técnicas de agrupamento**: É uma técnica de aprendizado não supervisionado usada para agrupar dados semelhantes em conjuntos.
2. **Árvores de decisão**: São modelos de aprendizado supervisionado (com rótulos), usados tanto para classificação quanto regressão.
3. **Ensemble Learning**: É uma técnica que junta vários modelos simples para criar um modelo mais preciso e confiável.
4. **Regressões**: É uma técnica de aprendizado supervisionado usada para prever um valor contínuo com base em variáveis de entrada.
5. **Modelos Multiníveis**: São modelos estatísticos que consideram a hierarquia natural dos dados, permitindo que você modele variações em diferentes níveis.
6. **Machine Learning**: É um subcampo da inteligência artificial que cria sistemas capazes de "aprender com dados" e fazer previsões ou tomar decisões sem serem explicitamente programados para isso.
7. **Aprendizado Supervisionado**:É um tipo de machine learning onde você treina o modelo com dados rotulados, ou seja, cada exemplo do conjunto de dados tem entradas e uma saída conhecida.
8. **Aprendizado não supervisionado**: É um tipo de machine learning em que o modelo recebe apenas dados de entrada, sem rótulos ou respostas.

Conclusões

Foi possível entender a importância de escolher o modelo certo para cada tipo de problema e os desafios enfrentados, como o overfitting e escolha de hiperparâmetros adequados. Além disso, o curso destacou a importância da validação e do uso de métricas para avaliar a eficácia dos modelos. No geral, o aprendizado proporcionou uma base sólida para aplicar machine learning de forma prática e eficiente.

Referências