Avanti Bootcamp - Machine Learning - 2025.1

Atividade I

Alan Delon Sousa Rocha , Março 2025.

1. **Explique, com suas palavras, o que é machine learning?**

Machine Learning é uma disciplina que faz parte do estudo de Inteligência Artificial que possui como objetivo criar modelos capazes de aprender a partir de dados e que possam, posteriormente, serem capazes de generalizar o aprendizado a novos dados. Esse aprendizado se dá pela junção de um conjunto de técnicas de algoritmos estatísticos, conhecimentos computacionais e otimização matemática. O machine mearning, é o que está por traz do funcionamento de chatbots, sistemas de reconhecimento de voz, visão computacional, reconhecimento de fraudes em sistemos financeiros e etc, podendo ser usado em diversos contextos, como a área da saúde, de finanças, de engenharia, de segurança e etc.

---

2. **Explique o conceito de conjunto de treinamento, conjunto de validação e conjunto de teste em machine learning.**

Para o treinamento de um modelo de Machine Learning, se faz necessário subdividir o dataset em 3 conjuntos de dados:
Treinamento: Esse conjunto é usado para ensinar o modelo durante a fase de modelagem, pode corresponder de 60% a 80% dos dados totais disponíveis.
Validação: Usado para avaliar o treinamento do modelo em tempo de execução, com isso pode se utilizar de técnicas de aperfeiçoamento do treinamento como o ajuste de hiperparâmetro, implementar early stopping, avaliação do modelo e ajuste dinâmico. Pode corresponder de 10% a 20% do total do datset.
Teste: O conjunto de teste é a porção dos dados que é utilizado na fase de avaliação, é por meio dele que será possível gerar métricas e gráficos que irão atestar a eficácia geral do modelo. Pode ser utilizado de 10% a 20% do dataset total.

---

3. **Explique como você lidaria com dados ausentes em um conjunto de dados de treinamento.**

Em um dataset em que uma coluna possui mais que 80% dos valores ausentes, talvez a melhor opção seja excluir a coluna do treinamento, mas em casos que os valores ausentes são a minoria mas ainda expressivos pode ser feito o processo de imputação dos valores.

A imputação ou preenchimento dos valores ausentes é feito com base no tipo de váriavel da coluna em questão.

Variáveis numéricas: Substituição os valores ausente por alguma métrica de tendência central como média, moda, mediana a depender do perfil de distribuição dos dados.

Variáveis categóricas: Substituição pela categoria mais comum ou adcionar uma categoria de desconhecido.
Variáveis temporais: Substuição por forward fill ou backward fill, que é quando se preenche os valores ausentes pela útima ou pela próxima data válida, respectivamente.

Existem técnicas mais avançadas, mas essas são as mais frequentemente necessárias.

---

4. **O que é uma matriz de confusão e como ela é usada para avaliar o desempenho de um modelo preditivo?**

A matriz de confusão é um métrica de avaliação que mostra o número de predições corretas e incorretas a depender do tipo de resposta. Ela é importante no processo de avaliação por fornecer um representação real e númerica do acertos e erros do modelo, e por gerar outras métricas derivadas da matriz como acurácia e precisão. Consiste uma tabela de N x N em que N é o número de classes no problema e cada linha são os resultados reais do conjunto de teste e as colunas são os resultados previstos pelo modelo. Os elementos na diagonais mostram as previsões corretas e os elementos de fora são as previsões erradas. Exemplo, N=2.

```
                    | Predito Positivo | Predito Negativo |
————————————————————|——————————————————|——————————————————|
Real Positivo       |       VP         |       FN         |
————————————————————|——————————————————|——————————————————|
Real Negativo       |       FP         |       VN         |
```

Onde:

VP (Verdadeiro Positivo): Predição positiva correta.
VN (Verdadeiro Negativo): Predição negativa correta.
FP (Falso Positivo): Erro Tipo I - predição positiva quando o real é negativo.
FN (Falso Negativo): Erro Tipo II - predição negativa quando o real é positivo.

Métricas diretamente derivadas:

Acurácia: (VP + VN) / (VP + VN + FP + FN)
Precisão: VP / (VP + FP)
Sensibilidade: VP / (VP + FN)
Especificidade: VN / (VN + FP)

Cada uma da métricas pode ser mais importante que outra a depender do contexto do problema.

---

5. **Em quais áreas (tais como construção civil, agricultura, saúde, manufatura, entre outras) você acha mais interessante aplicar algoritmos de machine learning?**

Para mim, as áreas mais interessantes para se aplicar algoritmos de machine learning são: Sáude, finanças, ciência/pesquisa e energia/sustentabiidade.

Sáude pela possibilidade de acelerar processos de diagnótico e cura de doenças.
Financeira pela possibilidadede ajudar uma instituição a diminuir fraudes ou melhorar a avaliação de crédito para quem precisa.
Ciência/pesquisa pela possibilidade de lidar com treinamento de modelos com dados biológicos ou astronômicos.
Energia/sustentabiidade pela possibilidade de ajudar a monitorar os uso de energias não renováveis, prever eventos climáticos extremos ou mapear áreas de conservação ambiental.

Mas acredito que praticamente todas ás áreas podem se beneficiar do uso do machine learning.