TESTE 1 -

1) Exercício 1

Suponha que você possui uma base de dados rotulada com 10 classes não balanceadas, essa base é formada por 40 features de metadados e mais 3 de dados textuais abertos.
Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

Descreva como faria a modelagem dessas classes.
R: Utilizaria a biblioteca TensorFlow para construir modelos de aprendizado de máquina, para processamento de texto, talvez utilize NLTK ou KERAS para pré-processamento.
No pré-processamento usaria uma classe como oversampling ou outra para desbalanceamento.
Modelagem usaria diferentes experiencias com algoritmos de classificação, Random Forest ou redes neurais e ajustaria os parâmetros para entender qual melhor modelagem para solução.

Ao finalizar essa modelagem, como iria apresentar essa modelagem para a área contratante?
R: Apresentaria um relatório que inclua uma explicação da modelagem, métricas, desempenho e gráficos importantes. Destacaria a interpretabilidade do modelo e a justificativa da escolha realizada. Também apresentaria uma demonstração prática, mostrando como o modelo funciona na prática. Dando algumas pinceladas no funcionamento das predições trazendo uma experiência clara para o cliente.

Como faria a validação desse modelo?
R: Utilizaria alguns métodos de validação, como validação cruzada entendendo desempenho do modelo. Também consideraria métricas de precisão, recall, F1-score para medição. Faria validação específica para dados textuais, entendendo que o modelo esteja capturando padrões consideráveis.

Supondo que esses dados são recebidos diariamente, como iria trabalhar com esse desafio?
R: Num processamento diário eu automatizaria o pré-processamento como também na atualização do modelo para lidar com os dados talvez num D-1. Construiria uma pipeline que inclua a reavaliação e retreino rotineiro. Faria um dashboard para monitorar alguma possível degradação no desempenho do modelo no tempo.

Como levaria esse projeto para um ambiente produtivo?
R: Usaria a dockenização do modelo para facilitar a implantação para diferentes ambientes, faria uma API Rest expondo o modelo para facilitar a integração com outros sistemas. Prepararia para o escalonamento criando um mecanismo automático para lidar com picos de carga. Complementaria com um controle de versão de código do modelo e das pipelines, permitindo atualizações controladas.

Respondido por Alencar Porto

EXTRA - Existe mais algo que gostaria de relatar sobre esse caso?
R:Finalizaria com uma documentação do projeto, levaria em consideração questões de segurança, dados sensíveis para lidar com a LGPD e criar um canal de feedback para aprimorar o modelo com base nas observações dos usuários e do ambiente de produção.

2) Exercício 2:

Suponha que você tenha uma base de dados de vendas de uma loja de varejo que inclui informações sobre produtos, clientes, datas de compra e valores das vendas. A base de dados possui, em média, 10.000 registros diários.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

a) Como você iria explorar os dados para obter insights sobre o desempenho das vendas.
   R: Para explorar os dados utilizaria a biblioteca pandas para manipulação e análise de dados. Matplotlib para visualizações e o scikit-learn para análise estatística.
   Faria uma análise exploratória utilizando gráficos de dispersão, histogramas e estatística descritivas para entender a distribuição dos dados.
   Nos padrões identificados, exploraria correlações entre variáveis, como datas de compra, produtos e valores de vendas. Utilizaria visualizações temporais para entender padrões semestrais ou sazonais.

b) Como você responderia as seguintes questões:
   i. Qual é o desempenho de vendas ao longo do tempo?
      R: Criaria um dashboard com gráficos de linhas para visualizar tendências no tempo.
   ii. Quais são os produtos mais vendidos?
      R: Num dashboard faria uma relação dos produtos mais vendidos por contagem ou valor do produto total em vendas.
   iii. Como as vendas variam por categoria de produtos?
      R:Agruparia os dados dos produtos por categoria e analisaria as estatísticas descritivas.
   iv. Qual é a distribuição dos valores de venda?
      R: Usaria um histograma para visualizar uma participação dos valores nas vendas.
   v. Como os preços dos produtos afetam as vendas?
      R: Faria uma correlação entre preços e volume de vendas; criaria uma visualização de dispersão para explorar essa relação.
   vi. Qual é o perfil dos principais clientes em termos de compras?
      R: Primeiro agruparia os dados dos clientes, entenderia os padrões de compra, identificaria os clientes com maiores volumes de compra e sua sazonalidade.

Respondido por Alencar Porto

c) Como você faria para identificar grupos de clientes nessa base de dados?
R: Eu utilizaria a biblioteca scikit-learn para clusterização. Realizaria a normalização de dados, tratamento de valores ausentes e aplicar algoritmos de clusterização para identificar grupos de clientes com padrões de compra semelhantes.

d) Qual teste estatístico você usaria para provar uma hipótese referente aos segmentos de clientes? e como iria aplicá-lo?
R: Usando biblioteca Scipy para testes estatísticos. Aplicaria uma formula hipotética sobre como os segmentos de clientes podem ser comportar de maneira diferente. Analisaria um teste estatístico mais adequado e aplicaria o teste aos grupos identificados de clientes, comparando as métricas importantes.

Extra - Pensando nos dados acima, seria possível fazer mais algum tipo de análise?
R: Pensaria em aplicar uma análise de sentimento aos dados dos clientes, no caso de termos feedback textual para entender a experiência do cliente sobre o produto ou sobre o processo de compra.

3) Exercício 3

Suponha que você tenha uma base de dados contendo textos jurídicos, como decisões judiciais, petições e documentos legais. A base de dados inclui informações sobre o conteúdo do texto, data, jurisdição e outras informações relevantes. Seu objetivo é criar um sistema de recomendação que sugira textos jurídicos semelhantes a um texto de referência.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

Descreva como você desenvolveria o sistema de recomendação que recebe um texto de referência e sugere os textos mais semelhantes a ele na base de dados.
R: Penso em usar alguma biblioteca de processamento de linguagem natural, spaCy ou a Gensim para similaridade de texto. No pré-processamento de texto faria a tokenização, removendo os stop words, stemming, converteria os textos em representações numéricas usando TF-IDF ou outro talvez o embeddings. Aplicaria similaridade de texto com Jaccard ou cosseno. Com embeddings de palavras ou modelos pré-treinados em textos jurídicos. Então o sistema de recomendação dá um texto de referência, calcula a similaridade com outros textos na base de dados e apresenta as recomendações por score.

Respondido por Alencar Porto

a) *Como você avaliaria esse sistema de recomendação?*

R: Faria uma avaliação por métricas de precisão, recall, avaliação qualitativa, envolvendo especialistas do negócio para verificar a relevância das recomendações. Procuraria garantir que abranja também diversas jurisdições e tipos de documentos nessa avaliação.

c) *Suponha que novos textos jurídicos sejam adicionados diariamente. Como você manteria o sistema de recomendação atualizado e garantiria que ele continue a fornece recomendações relevantes?*

R: Faria uma rotina automatizada implantando um monitoramento contínuo, usando logs para identificar mudanças nas características dos textos e na distribuição de dados. Realizaria de forma sazonal o re-treino do modelo para incorporar novos textos e ajustaria a mudanças tendenciosas de dados. Talvez usaria um aprendizado incremental. Para novos textos, criaria uma pipeline automatizada para pré-processar e incorporar novos textos D-1. Manteria também histórico de versões para rastreamento de alterações no modelo. Aqui também criaria um canal de feedback dos usuários para melhorar a qualidade das recomendações.

Respondido por Alencar Porto

TESTE 2 –

1) Como funciona o teste de hipóteses e qual é a sua finalidade na análise estatística?
   R: O teste de hipóteses é como uma investigação estatística intrigante, onde buscamos respostas para questões sobre o universo dos dados. E como se fossemos detetives, e as hipóteses nossas teorias sobre o que está acontecendo. O teste nos guia a decidir se as evidências que coletamos apoiam ou contradizem nossas teorias. É como encontrar pistas em um caso complicado e ver as evidências se são fortes o bastante, podemos rejeitar ou aceitar as hipóteses e chegar em uma conclusão relevante.
2) O que são redes generativas adversárias (GANs) e quais são os possíveis usos dessas redes?
   R: As GANs são como dois artistas que competem para criar obras de artes. Sendo um gerador e outro um discriminador, que jogam um jogo para melhorar continuamente. O gerador cria as imagens e o discriminador avalia sua autenticidade. Como uma dança, onde o gerador tenta engar o discriminador e vice-versa. Uma GAN está na vanguarda da criação de imagens realistas, desde imagem de rostos humanos até obras de arte. Uma simulação de pintura em que cada um tenta aprimorar suas habilidades, resultando em imagens de obra de arte impressionantes.
3) O que são modelos de linguagem? Qual a diferença entre LLMs e modelos de linguagem tradicionais?
   R: Os modelos de linguagem é um tipo de escrita imaginativa que tenta prever qual será próxima palavra a ser dita numa história. São treinados em grandes volumes de textos para entender padrões e contextos linguísticos. A diferença básica entre o Large Language Models e os tradicionais está na escala e compreensão. LLMs exemplo GPT-3 são autores prolificamente experientes, enquanto os tradicionais são como novatos, limitados em suas capacidades.
4) Suponha que você tenha um conjunto de dados com três ou mais grupos para comparar e deseja determinar se há diferenças significativas entre eles. Descreva como você escolheria entre o teste ou outras técnicas estatísticas
   R: E um desafio realizar essa comparação de vários grupos, é como ser o maestro de sinfonia estatística. A escolha entre testes ou outra técnica é como decidir se precisamos de violinos suaves ou trompetes vibrantes. Testes não paramétricos são diferentes instrumentos na orquestra, cada um com sua melodia única. Essa decisão depende da distribuição dos dados e da natureza do experimento, escolher o método certo é como conduzir a harmonia da orquestra perfeita, eu diria tarefa bem difícil.

Respondido por Alencar Porto

5) Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?
R: Pré-processamento de texto tem como objetivo afinação de piano por exemplo, antes de uma apresentação musical. NLP precisa garantir que os textos estejam afinados e isso é crucial. Depois vem outras etapas como tokenização, remoção stop words, lematização são ajustes finos que garantem que o modelo escute a música do texto de forma clara. A base do pré-processamento, o palco onde a transformação do texto desorganizado se transforma em uma sinfonia de dados estruturados, preparando-o para análise significativa.

6) Descreva o processo de vetorização de texto e como modelos de linguagem como o Word2Vec ou o TF-IDF podem ser usados para representar palavras e documentos.
R: Na vetorização de texto é como traduzir palavras e documentos para um idioma que os modelos de máquina entendam. São palavras como notas musicais e documentos como composições inteiras. Modelos como TF-IDF são como dicionários de tradução, convertendo palavras e documentos em vetores, representando a essência da linguagem em coordenadas numéricas. É como compor uma partitura única para cada palavra, sendo legíveis e compatíveis para o modelo.

7) O que é a análise de sentimento em NLP e quais são os principais métodos para realizar essa tarefa? Como você avaliaria a eficácia de um modelo de análise de sentimento?
R: Análise de sentimento em (Natural Language Processing) é como decifrar a melodia emocional de um texto. É como um crítico literário que avalia se um livro é uma tragédia ou uma comédia. Usando métodos de análise léxica ou usando modelos pré-treinados, como BERT, VaderSentiment, entre outros são como a sintonia fina emocional. Avaliar a eficácia é como julgar a precisão de um pianista ao transmitir emoção.

8) Qual é a diferença entre a classificação de texto e o agrupamento (clustering) de texto em NLP? Em que situações cada um é mais apropriado?
R: Classificação de texto é como organizar livros no armário, cada um rotulado com o seu gênero. É sobre atribuir rótulos claros. Portanto, o agrupamento de texto é como criar prateleiras dinâmicas, agrupar por similaridades, ou por ausência de rótulos. A classificação é como organizar, enquanto o agrupamento é como descobrir afinidades entre os textos.

9) Explique o conceito de reconhecimento de entidades nomeadas (NER) em NLP e suas aplicações práticas.
R: Reconhecimento de entidades nomeadas é como identificar personagens principais em uma trama complexa. É como livros com enredos ricos, e o NER é com um roteiro que destaca o protagonista. Na prática é como descobrir o personagem principal em um romance jurídico, como identificar nomes de empresas, datas importantes e locais cruciais. Ferramenta essencial para desvendar os mistérios dentro dos textos.

Respondido por Alencar Porto

10) Como você lidaria com problemas de desequilíbrio de classe em tarefas de classificação de texto em NLP? Quais estratégias seriam eficazes?

R: Eu lidaria com uma estratégia de oversampling, undersampling no peso das classes para ajustar o volume de cada texto. Objetivo é garantir que nenhuma classe seja subestimada, mantendo a harmonia e a precisão na classificação dos textos.

Respondido por Alencar Porto

TESTE 3 - CASE

canada_amostra.csv

Amostra:

Contextualização:

O Base de dados canada_amostra em formato CSV representa um conjunto de empresas do Canadá com a respectiva descrição de seus produtos, dados econômicos e localização.

Assim, podemos caracterizar cada variável:

name: nome da empresa;

description: descrição do produto da empresa;

employees: número de empregados da empresa;

total_funding: Total de investimento já recebido pela empresa;

city: cidade;

subcountry: estado;

lat: latitude da cidade;

lng: Longitude da cidade.

1) Problema:

Deseja-se prospectar empresas que possuam soluções em **tratamento de água** , principalmente, elativas à : **solutions on waste and water, Improve water quality and water efficiency use, water contamination, water for human consumption, water resources** .

a) EXERCÍCIO 1 - Aplique um algoritmo de ML (ou um conjunto deles) capaz de selecionar as principais empresas indicadas para desenvolver a solução de acordo com seu alinhamento com o tema (Justifique a escolha do algoritmo).

   R: Utilizaria um algoritmo de processamento de linguagem natural para analisar a descrição do produto da empresa. Um modelo de machine learning baseado em TF-IDF pode ser eficaz na tarefa, permitindo a extração de características relevantes dos textos.

   Justifico o uso da técnica de NLP que permitirá a análise semântica das descrições das empresas, identificando aquelas que mencionam explicitamente ou estão semanticamente relacionadas com tratamento de água.

Respondido por Alencar Porto

b) EXERCÍCIO 2 - Faça uma análise exploratória dos resultados acrescentando as demais variáveis contidas no dataset. Quais insights você pode obter a partir desses dados? Quais são as principais cidades (pólos de desenvolvimento) para essa solução?
R: Faria análise estatística descritiva para compreender a distribuição de empregados e financiamento entr as empresas relacionadas. Quais insights possíveis, distribuição geográfica (identificando as cidades e estados com maior concentração de empresas em tratamento de água). Econômica e financeira (Análise a relação entre o número de empregados, financiamento total e o tipo de solução oferecida pela empresa). Possíveis polos de desenvolvimento (classificação das cidades e estados com base na concentração das empresas e variáveis econômicas para identificar potencial de crescimento no setor).

c) EXERCÍCIO 3 - EXTRA - Se você terminou o desafio de forma rápida, temos mais algumas perguntas para serem respondidas. Elas, como dito, não são obrigatórias, então sinta-se à vontade em não as responder ou até mesmo respondê-las parcialmente. Essa parte visa observar seu entendimento de um ambiente real de produção.

i) a) organize seus códigos em pacotes garantindo seu versionamento e documentação (bibliotecas auxiliares, etc.).
R:

Estrutura de Diretórios:
Organizei meu projeto em uma estrutura clara de diretórios, como src/, tests/, data/, etc.

Versionamento:
Utilizei uma ferramenta de versionamento como o Git para controlar as versões do código.

ii) b) construa testes automatizados para validação do seu pacote.
R: Testes Unitários:
Escreva testes unitários para cada função ou componente crítico do seu pacote.

iii) c) crie uma imagem Docker capaz de executar suas análises em um ambiente de produção.
R: Dockerfile:
Crie9 um Dockerfile na raiz do projeto.

Respondido por Alencar Porto

Especifiquei as dependências necessárias e os comandos para configurar o ambiente.

iv) d) crie um GitHub público e suba todo o código do Teste 3, e disponibilize para avaliação.

R:
Repositório Público:
Criei um repositório público no GitHub.

Respondido por Alencar Porto