TESTE 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.

1. Descreva como faria a modelagem dessas classes.
2. Ao finalizar essa modelagem, como iria apresentar essa modelagem para a área contratante?
3. Como faria a validação desse modelo?
4. Supondo que esses dados são recebidos diariamente, como iria trabalhar com esse desafio?
5. Como levaria esse projeto para um ambiente produtivo?
6. EXTRA - Existe mais algo que gostaria de relatar sobre esse caso?

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.

1. 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.
2. Como você avaliaria esse sistema de recomendação?
3. 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?

TESTE 2 -

1. O que é um desvio padrão e qual é o seu papel na medição da dispersão dos dados?

O desvio padrão é uma medida de variabilidade, utilizada para quantificar quão dispersos são os dados com domínio de 0 até infinito. Ele é calculado através da raiz quadrada da variância. Com ele, podemos entender como a média representa um conjunto de dados. Quanto menor o valor do desvio padrão, menor é a variabilidade dos dados. Uma forma de avaliar o quanto uma variável é dispersa em relação as outras é calcular o coeficiente de variação. O coeficiente de variação é obtido através da divisão do desvio padrão pela média. Quanto menor o coeficiente, menos disperso é o dado. Isso é muito utilizado nas séries temporais para avaliar a previsibilidade de uma série temporial. Ainda, outro exemplo é uma releitura dessa estatística através da raiz quadrada do erro quadrático médio, largamente utilizada em modelos de previsão de uma variável contínua para avaliar o erro de previsão.

1. Como funciona o teste de hipóteses e qual é a sua finalidade na análise estatística?

O teste de hipóteses é ponto central do pensamento estatístico. Ele é definido a partir de duas hipóteses, a hipótese nula e a alternativa. A hipótese nula usualmente é a mais conservadora; e a alternativa é o resultado que você gostaria de obter com os dados através do novo método aplicado. Ao testarmos uma hipótese estatísticamente tem-se dois possíveis resultados: rejeitar a hipótese nula, ou seja, o novo método de fato é mais eficaz; ou não rejeitar a hipótese nula, nesse caso, o novo método não é mais eficaz que o usual. Para exemplificar podemos definir o seguinte cenário:

Um farmacêutico está interessado em testar se o novo medicamento que ele está desenvolvendo contra a AIDS diminui o número de sintomas comparado ao tratamento atual. Logo ele define as seguintes hipóteses:

Hipotese Nula: O número médio de sintomas não é menor nos pacientes que foram submetidos ao novo medicamente.

Hipotese Alternativa: O número médio de sintomas é menor nos pacientes que foram submetidos ao novo medicamente.

Uma vez definida a hipótese o farmacêutico conduz um experimento a cegas definindo uma cohort e separando o grupo de pacientes em dois grupos (um que irá receber o novo medicamento; e outro que irá receber o medicamento usual). Ao final do experimento registra-se o número de sintomas por paciente e avalia-se, por exemplo, através de um teste t de Student (caso o pressuposto de normalidade em cada grupo seja atendido) se as médias diferentes entre os grupos. As hipóteses mencionadas acima estão diretamente relacionadas à estatística de teste do teste de t de Student. Por fim, a estatística do teste resulta em um valor-p, que a depender do nível de significância podemos decidir pela rejeição ou não da hipótese nula. Ou seja, a aceitação ou não do novo medicamento.

1. O que é aprendizado supervisionado e como ele difere do aprendizado não supervisionado?

Esses termos são utilizados para categorizar dois tipos de problemas em Machine Learning (ML), ou ainda, categorizam dois diferentes tipos de modelos. Na aprendizagem supervisionada o problema a ser estudando dispõem de uma base rotulada, na qual sabemos o output desejado, e queremos obter o mecanismo que gera esse output. Já na aprendizagem não supervisionada não temos uma base rotulada, e estamos interessados em obter, por exemplo, novas categorias através dos dados, gerando assim, novos rótulos. Existem algoritmos próprios para cada tipo de problema. Por exemplo, modelos como gradiente boosting e regressão linear são exemplos de modelos utilizados para resolver problemas através de um aprendizado supervisionado; enquanto que algoritmos de cluster são utilizados na aprendizagem não supervisionada.

1. O que é transfer learning e como ele é usado em deep learning?

O transfer learning é uma forma de retrainar um modelo de deep learning com a inservação de uma nova quantidade pequena de dados, específica para um caso de uso. Isso permite calibrar os parâmetros do modelo com os dados específicos da sua aplicação. Utilizado no campo da visão computacional e NLP para LLM.

1. Você está conduzindo um experimento A/B em um site de comércio eletrônico para determinar a eficácia de uma nova página de destino na conversão de visitantes em clientes. Como você projetaria o experimento, escolheria as métricas apropriadas para avaliação e realizaria a análise estatística para tirar conclusões significativas?
2. Após realizar um teste ANOVA e obter um valor de F significativo, como você determinaria quais grupos são estatisticamente diferentes entre si?
3. 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
4. Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?
5. 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.
6. Como você lidaria com problemas de desequilíbrio de classe em tarefas de classificação de texto em NLP? Quais estratégias seriam eficazes?

**OBS: Se for necessário, crie um repositório público no GitHub e compartilhe com a gente para a avaliação.**