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.

Para realizar o processo de ajuste e previsão eu utilizaria a biblioteca scikit-multilearn que é apropriada para realizar classificações multi classes e tem diferentes modelos disponíveis, como o RakelD. As 10 classes seriam modeladas utilizando o BinaryRelevance, e testaria diferentes classificadores, incluindo o RandomForest, XGBoost, LightBoost, CatBoost, Multilable KNN.

Escolheria métricas que levam em conta o desbalanceamento de classes, como precisão, recall e F1 por classe. Caso todas as classes sejam importantes, calcularia a média de todas as métricas para ter uma medida de erro global.

Em relação as covariaveis concatenaria os 3 dados textuais abertos e utilizaria o TfidfVectorizer do sklearn, utilizando de técnicas de pré-processamento de texto previamente. Além da matriz com os pesos do TfidfVectorizer utilizaria as 40 features de metadados aplicando one hot enconder para as variáveis categórias, e normalização para as variáveis contínuas para ficar mais próximo do range de valores do TfidfVectorizer.

Essa seria a abordagem mais clássica. Com o objetivo de melhorar a performance testaria os modelos transformers como o BertModel ao invés de utilizar o TfidfVectorizer. Caso isso não seja o suficiente, testaria modelos de Deep Learning para melhorar a acertividade.

Ainda faria otimização de hiperparâmetros utilizando busca bayesiana.

1. Ao finalizar essa modelagem, como iria apresentar essa modelagem para a área contratante?

* Faria um diagrama para apresentar os passos da metodologia utilizada.
* Deixaria um slide que mostre todas as abordagens testadas.
* Focaria no benefício e as vantagens da implementação utilizada.
* Apresentaria a métrica de erro do conjunto de teste que foi acordada junto com a área contratante.
* Colocaria exemplos de como essa modelagem funcionaria com novos textos e metadados.

1. Como faria a validação desse modelo?

A validação dos modelos seria feita utilizando validação cruzada maximizando a métrica escolhida. Se possível, deixaria 10% dos dados de forma intocável, como se fossem novos dados, e avaliaria o modelo nesse conjunto de dados.

1. Supondo que esses dados são recebidos diariamente, como iria trabalhar com esse desafio?

As minhas escolhas seriam pautadas no tempo que o modelo precisa para fazer o treino e a predição. Avaliaria o tempo que leva para treinar, a necessidade de retreino do modelo, e o tempo para fazer a predição. Buscaria fazer um modelo em que a predição fosse boa o suficiente, e que não precisasse rodar todo dia. No entanto, faria validações na entrada de dados para entender se o comportamento de data drifting ocorre; e só então faria um novo retreino. Também acompanharia métricas do negócio que podem estar relacionadas a uma perca de performance do modelo. Não utilizaria um modelo baseado em KNN pois a predição é lenta.

1. Como levaria esse projeto para um ambiente produtivo?

Criaria uma API para fazer a consulta do modelo utilizando Flask. Deixaria a API rodando no serviço AWS Lambda que é serverless, sendo possível escalar facilmente caso ocorra um aumento no número de requisições. Nesse caso também é necessário configurar a API Gateway para deixar que a rede consiga acessar o serviço da AWS. Utilizaria o Docker se necessário.

1. EXTRA - Existe mais algo que gostaria de relatar sobre esse caso?

Faria análise descritiva para entender o desbalanceamento da classe resposta, e entenderia a necessidade de predizer as atuais 10 classes separadamente. Avaliaria se seria possível agrupar algumas classes a depender do contexto. Faria uma análise descritiva dos 40 metadados também para entender se existe variabilidade suficiente nas covariáveis. Pensaria em utilizar computação em paralelo ou ambiente spark caso o tempo computacional com a solução atual fosse um limitante.

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.

Poderia utilizar tanto o TfidfVectorizer do sklearn para converter o texto de entrada na matriz de pesos do Tfidf; ou um transformer para obter as word embeedings. Ao obter o novo texto fornecido do usuário calcularia a similaridade do cosseno para obter os textos mais similares em relação ao texto de entrada e retornaria os 5 mais similares para o usuário. Caso os textos sejam longos, faria um resumo dos textos previamente, ou obteria apenas a seção mais importante do texto para ser o input do modelo, resultando em cálculos mais rápidos. Assim, a informação principal não é perdida.

1. Como você avaliaria esse sistema de recomendação?

Suponha que sejam feitas 5 recomendações por texto de entrada, eu utilizaria a métrica MAP pois leva em conta a ordenação e o número de recomendações escolhidas. Essa métrica avaliaria qual das diferentes abordagens (modelos) de recomendação seria a mais eficaz. Também pediria o feedback do usuário (positivo e negativo) e avaliaria qual ferramenta obteve maior percentual de feedbacks positivos.

1. 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?

Novos textos que entrassem diariamente já entrariam para compor a base de textos a serem comparados. Assim, cada nova informação já seria considerada para o cálculo de similaridade garantindo recomendações relevantes. Caso a métrica do modelo perdesse performance, uma nova recalibração do modelo seria feita.

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 se a média representa bem 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, e assim, melhor a média representao conjunto de dados.

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 normalmente 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 diferem 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 decide-se 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. Isso é 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?

Eu conduziria o experimento na página até atingir uma amostra que possa ser considerada como representativa da população. A cada novo usuário a alocação da página ‘A’ ou ‘B’ poderia ser feita de forma aleatoriamente. Ou ainda; aplicar a amostragem aleatória estratificada proporcional para melhor representar um determinado estrato, a partir, por exemplo, da idade, sexo, perfil, entre outros.

As métricas utilizadas para esse experimento são Número de links clicados, Tempo de permanência na página, Número de conversões (compras) e número de compartilhamentos realizados.

Avaliaria no final se houve diferença em cada uma das métricas em relação ao grupo (‘A’ ou ‘B’) utilizando teste t por exemplo, ou o teste de mann-whitney, a depender de ter distribuição normal.

Outra opção ainda, é utilizar análise de regressão (estatística) e considerar outras variáveis que podem afetar as métricas do experimento como controle.

1. Após realizar um teste ANOVA e obter um valor de F significativo, como você determinaria quais grupos são estatisticamente diferentes entre si?

Eu usaria o teste de Tukey como a análise de post hoc para definir se os grupos são estatisticamente diferente entre si. Como o número de comparações duas a duas pode ser grande em virtude do número de categorias das variáveis a serem testadas, aumenta a chace de ocorrer uma diferença estatística ao acaso. Dessa forma aplico a correção de valor-p utilizando a False Discovery Rate (FDR) que leva em consideração a proporção esperada de erros Tipo I entre as hipóteses que foram rejeitadas (ou seja, daqueles contrastes que foram significativos, faz-se a correção de valor-p).

1. 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

Para escolher o teste é necessário definir o tipo da variável que se deseja estudar, bem como entender o tipo do experimento realizado.

Experimentos que não envolvem medidas repetidas precisam ser testados diferente daqueles que envolvem. Bem como experimentos realizado ao longo do tempo demandam de técnicas específicas.

Se restrigirmos o escopo para um experimento não pareado, em que os grupos foram dividos de forma aleatoriamente, indivíduos foram mensurados apenas uma vez, e se a variável de interesse for contínua podemos utilizar testes como anova ou kruskal-wallis e fazer a comparação múltiplica utilizando testes de tukey e de Dunn respectivamente com correção de FDR.

Casos os grupos fossem pareados, poderia ser utilizado o teste da ANOVA para medidas repetidas caso a variável em cada grupo tivesse distribuição normal (sendo isso verificado através da análise de resíduo), ou o teste de Friedman em caso da falha da distribução Normal.

Ainda, os problemas que aplica-se testes bivariados podem ser resolvidos através de modelos de regressão lineares generalizados escolhendo a distribuição apropriada.

1. Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?

A etapa de pré-processamento é importante para realizar a limpeza de um texto para que ele possa ser utilizando em um modelo de ML. Ele diminui a variabilidade do input, que por se tratar de dado textual já é grande e aumenta a padronização dos dados. Essa limpeza de texto envolve padronizar para minúsculo, remoção de stopwords, caracteres especiais, algoritmos de stemming ou lematização. As stopwords são palavras que não agregam significado para a frase, como por exemplo: então, pois, dele, ?, !; usa-se o stemming obtém a raiz da palavra; enquanto que a lematização obtém o sentido da palavra.

1. 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.

Após o pré-processamento de texto é necessário utilizar a vetorização de texto para converter um texto livre para uma forma que os algoritmos de ML consigam ler. O TF-IDF é uma matriz de pesos que para cada documento e palavra ele dá maior importância para aquelas palavras que aparecem muito em um dado texto, mas pouca frequência em todos os documentos. Já o Word2Vec atribui um vetor de pesos para cada palavra conforme a similaridade semântica entre elas. Palavras mais semelhantes estão mais próximas no campo vetorial.

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

Para lidar com esse problema podemos utlizar métricas que não sofrem com o problema da classe majoritária, como por exemplo, F1, precision, recall. Garantir que o treino e o teste tenham a mesma proporção das classes. Utilizar modelos em que seja possível penalizar mais o erro quando ele ocorre na classe menos frequente, como por exemplo a Regressão Logística e Support Vector Machine no sklearn através do argumento class\_weight. Técnicas como downsampling (diminuir o número de dados da classe majoritária) e upsampling (aumentar o número de dados da classe minoritária) também são amplamente testados, mas a eficácia delas estão sendo questionadas recentementes. Nas vezes que eu testei, também não obtive sucesso com o SMOTE.

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