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

Machine learning ou aprendizado de máquina é um modelo no qual a máquina é capaz de auferir coisas com base nos dados a ela oferecidos. Ou seja, em outras palavras, enquanto um programa ou código comum funciona parecido como uma receita de bolo, com passo a passo para cada situação. Já os modelos de machine learning são instruções para como separar, identificar ou fazer previsões a partir de um conjunto de dados.

Portanto, em resumo, os modelos de machine learning são capazes de aprender a realizar alguma auferição baseados em técnicas computacionais e estatísticas com um conjunto de dados a ela oferecidos.

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

Quando se está criando algum modelo de machine learning para auferir alguma coisa com base em um conjunto de dados é necessário antes treinar o modelo. Ou seja, é necessário que o modelo antes seja treinado com uma quantidade de dados, posteriormente é necessário testar o modelo com outra quantidade de dados para verificar se a auferição treinada está dando o resultado desejado. Porém, entre o treinamento e o teste é necessário um passo intermediário que seja capaz de verificar os parâmetros e outras coisas.

Vamos supor que após treinar o modelo se verifique que este aufere perfeitamente com o conjunto de dados do treinamento porém quando se usa o conjunto de dados de teste os resultados são diferentes. Pode ter ocorrido que o modelo se ajuste muito bem ao conjunto de dados de teste mas seja ineficiente para outros conjuntos de dados, e o resultado seria portanto um modelo ruim. Portanto, tanto na situação de ajuste de parâmetros, comparação com outros modelos e evitar esta situação de um sobreajuste( situação do modelo se ajustar a um conjunto e ser ineficaz com outros conjuntos de dados) se tem o conjunto de validação.

Enfim podemos resumir dizendo que o conjunto de treinamento é uma parte dos dados separados para o treinamento propriamente dito do modelo. O conjunto de validação é outra parte dos dados usados para ajustar o modelo, e o conjunto de testes para aferir o resultado do modelo.

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

Não tenho prática em tratar dados ausentes em modelos de Machine Learning, portanto nunca foi algo que eu particularmente fiz. Porém já tratei muitos dados ausentes para análises exploratórias. Acredito que de certa forma o processo deva ser semelhante, com ressalvas é claro. Ou seja, é necessário fazer uma exploração estatística dos dados para verificar a distribuição, verificar a quantidade de dados faltantes e que dados são estes. Após analisar estes dados faltantes se tem algumas opções.

Caso a quantidade de dados faltantes seja uma porcentagem ínfima, uma possibilidade é remover os dados ausentes. Outra opção é tentar completar os dados ausentes pesquisando as informações, porém em muitos casos estas informações não estão disponíveis. Uma possibilidade que acredito que seja muito usada é a substituição dos valores ausentes por medidas de tendência centrais, tais quais mediana, moda ou média, além disso é possível também segmentar os dados faltantes e usar informações do próprio conjunto de dados para imputar alguma medida de tendência central segundo alguma outra informação mais completa no próprio conjunto de dados. Ressaltando que para tal é necessário analisar e ver qual medida seria melhor para aquele conjunto de dados e porque. Por último podemos utilizar técnicas de imputação avançada, se valendo de algoritmos de machine learning como KNN, regressão linear e outras. Pessoalmente nestes casos eu não tenho muita experiência, apenas um pouco de estudo sobre, já chegando a fazer alguns exemplos, mas sem usos reais.

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

A matriz de confusão permite avaliar o desempenho de um modelo preditivo criando uma matriz com os resultados do modelo. Ou seja em uma matriz você tera os valores positivos reais e negativos reais e os resultados os valores de positivos verdadeiros, positivos falsos, negativos verdadeiros e negativos falsos e assim você consegue com os respectivos valores calcular algumas métricas que nos auxiliam a avaliar se o resultado foi o esperado ou não, além de nos ajudar a verificar o que deve ser melhorado caso o resultado seja diferente do desejado.

As principais métricas que nos permitem avaliar verificar o desempenho do modelo com a matriz de confusão são: A acurácia, que mede a proporção de previsões corretas; A precisão, que nos trará a proporção de verdadeiros positivos em relação ao total de previsões positivas; O recall, que nos traz o total de verdadeiros positivos em relação ao total de positivos reais (O recall é muito util em situações nas quais o falso negativo gera problemas muito grandes, como no caso de doenças.); o F1-Score que nos trará um balanço entre precisão e recall, sendo um resultado de uma média harmônica das duas métricas mencionadas( muito util em modelos onde o falso positivo e o falso negativo tem uma importância grande); a Especificidade, que nos trará a proporção de negativos verdadeiros em relação ao total de negativos(muito util em situações onde identificar o negativo é muito importante).

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

Pensando de maneira insípida e rasa, acredito que em todo lugar onde haja necessidade de processos de mensuração, de classificação, distinção e previsão os algoritmos de machine learning são interessantes de aplicar. De maneira geral, eles nos permitem automatizar processos que são custosos em tempo, mão de obra e material. Portanto acredito que na verdade eles devam ser largamente utilizados de maneira a melhorar todas as áreas.

Agora quais áreas seriam mais interessantes sob meu ponto de vista específico. Bom, acredito que de maneira geral a iniciativa privada já se vale do Machine Learning em tudo que pode ou está correndo atrás de melhorar seus processos para aumentar o lucro. Embora eu pessoalmente tenha muito interesse na área construção, financeira, agricultura e saúde, acredito que devemos também nos voltar para processos que beneficiem além do interesse privado o interesse público. Então criar projetos que auxiliem a combater queimadas, grilagem de terra, conflito por terra e contra populações desfavorecidas, que alivia e automatiza processos do SUS e assim por diante.

Ou seja em resumo, acredito que devemos olhar e nos debruçar sobre a aplicação de machine learning sempre que for uma opção viável para a resolução do problema, afinal tenho a firme crença de que os problemas devem sempre ser o foco, a paixão, e as ferramentas, bom as ferramentas são feitas para ajudar a resolver os problemas e não para existirem por si só. Então vejo na saúde e em projetos que beneficiem o social um local talvez mais urgente pois provavelmente tem menos poder aquisitivo por si só para poder se utilizar de tudo que a humanidade produziu de melhor para poder resolver os problemas de maneira duradoura e eficiente.