Pré Processamentos

Os pré-processamentos de dados em aprendizado de máquina (ML) são técnicas utilizadas para preparar os dados de treinamento antes de alimentá-los para o modelo. Elas são importantes para garantir que os dados estejam no formato correto e sejam relevantes para o problema específico que está sendo resolvido, aumentando a precisão e eficiência do modelo.

Adequação dos tipos de dados

A adequação dos tipos de dados refere-se ao processo de certificar-se de que os dados estejam no formato esperado e sejam compatíveis com o algoritmo de aprendizado de máquina que será utilizado. Isso é importante porque muitos algoritmos de aprendizado de máquina esperam que os dados estejam em um determinado formato ou tipo, como valores numéricos para variáveis contínuas ou valores categóricos para variáveis categóricas.

Por exemplo, alguns algoritmos de aprendizado de máquina só podem lidar com dados numéricos e, portanto, é necessário converter variáveis categóricas para valores numéricos antes de utilizá-los. Isso pode ser feito através de técnicas como codificação de variáveis categóricas.

Além disso, é importante verificar se os dados estão completos, ou seja, que não contenham valores faltantes ou valores inválidos. Se necessário, esses valores devem ser tratados ou removidos antes de utilizar os dados para treinar o modelo.

Concluindo, adequar os tipos de dados é um passo importante no pré-processamento de dados, pois garante que os dados estejam no formato esperado e compatíveis com o algoritmo de aprendizado de máquina que será utilizado, e assim aumenta a qualidade do modelo de aprendizado.

Dados Missing

Dados faltantes são valores que não foram fornecidos ou não podem ser encontrados em um conjunto de dados. Eles podem ocorrer devido a várias razões, como erros de coleta de dados, falhas na entrada de dados ou simplesmente a falta de informação. Dados faltantes podem ser representados de várias maneiras, como valores nulos, valores vazios ou valores marcados como "não disponível". Esses dados faltantes podem afetar negativamente a precisão e a eficiência de um modelo de aprendizado de máquina, por isso é importante tratá-los adequadamente antes de utilizá-los para treinar um modelo.

Existem várias abordagens para lidar com dados faltantes, e a escolha depende do conjunto de dados e do objetivo do modelo de aprendizado de máquina. Algumas das opções comuns incluem:

- Excluir as linhas ou colunas que contêm dados faltantes: essa é uma solução simples, mas pode resultar em perda de dados significativa se houver muitos dados faltantes.
- Preencher os dados faltantes com valores específicos, como a média, mediana ou moda dos dados disponíveis: essa é uma opção rápida e fácil, mas pode introduzir distorções nos dados se os valores faltantes forem diferentes dos valores preenchidos.
- Utilizar técnicas de aprendizado de máquina para prever os valores faltantes a partir dos valores disponíveis: essa abordagem pode ser mais precisa, mas requer mais tempo e esforço para ser implementada.
- Utilizar técnicas de processamento de linguagem natural para inferir os valores faltantes a partir de contexto e outros dados disponíveis.

Além disso, é importante avaliar se a presença de dados faltantes pode estar relacionada com a variável alvo, caso sim é importante tomar cuidado com a sua exclusão dos dados.

Feature Selection: Correlação

A seleção de características (feature selection) é um processo no qual se busca selecionar um subconjunto de características (atributos ou colunas) do conjunto de dados para utilizar no modelo de aprendizado de máquina. Isso pode ser feito com o objetivo de melhorar a precisão do modelo, reduzir a complexidade do modelo ou aumentar a interpretabilidade dos resultados.

Uma das maneiras de realizar a feature selection é através de medidas de correlação, onde se busca identificar quais características estão mais relacionadas com a variável alvo. A correlação é uma medida estatística que expressa a relação linear entre duas variáveis. Ela pode ser medida por meio de diferentes coeficientes, como o coeficiente de Pearson, que varia entre -1 e 1. Valores próximos a 1 indicam uma forte correlação positiva, enquanto valores próximos a -1 indicam uma forte correlação negativa. Valores próximos a zero indicam ausência de correlação. Usando essas medidas de correlação, é possível selecionar as características que apresentam maior correlação com a variável alvo, e assim incluir somente essas características no modelo de ML.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

onde x1, x2, … xn e y1, y2, … yn são os valores medidos de ambas as variáveis

Entretanto, manter mais de uma variável altamente correlacionada no modelo pode causar problemas conhecidos como multicolinearidade. A multicolinearidade ocorre quando duas ou mais variáveis independentes em um modelo estatístico são altamente correlacionadas entre si. Isso pode causar problemas porque as estimativas dos coeficientes do modelo podem ser imprecisas e instáveis, tornando difícil a interpretação dos resultados. Além

disso, a multicolinearidade pode afetar negativamente a capacidade do modelo de prever novos dados, aumentando o erro do modelo.

Para evitar esses problemas, é recomendado remover as variáveis altamente correlacionadas antes de treinar o modelo. Isso pode ser feito através de técnicas de seleção de características, como a correlação mencionada anteriormente, onde somente uma das variáveis altamente correlacionadas é selecionada e incluída no modelo. Isso pode ajudar a melhorar a precisão e a estabilidade do modelo, além de tornar a interpretação dos resultados mais fácil.