Métricas de Regressão

1. Introdução ao Problema
2. MSE
3. RMSE
4. MAE
5. Coeficiente de Determinação
6. R^2 Ajustado
7. Implementações

INTRODUÇÃO

A modelagem preditiva de regressão é a tarefa de aproximar uma função de mapeamento (f) de variáveis de entrada (X) para uma variável de saída contínua (y). Uma variável de saída contínua é um valor real, como um número inteiro ou valor de ponto flutuante. Freqüentemente, são quantidades, como quantidades e tamanhos. Um problema de regressão requer a previsão de uma quantidade. Uma regressão pode ter variáveis de entrada de valor real ou discretas. Um problema com múltiplas variáveis de entrada é freqüentemente chamado de problema de regressão multi-variada.

Como um modelo preditivo de regressão prevê uma quantidade, a habilidade do modelo deve ser relatada como um erro nessas previsões. Há muitas maneiras de estimar a habilidade de um modelo preditivo de regressão, mas talvez a mais comum seja calcular a raiz quadrada média do erro, abreviado pela sigla RMSE.

Existe uma sobreposição entre os algoritmos de classificação e regressão; por exemplo:

Um algoritmo de classificação pode prever um valor contínuo, mas o valor contínuo está na forma de uma probabilidade para um rótulo de classe.

Um algoritmo de regressão pode prever um valor discreto, mas o valor discreto na forma de uma quantidade inteira.

Alguns algoritmos podem ser usados para classificação e regressão com pequenas modificações, como árvores de decisão e redes neurais artificiais. Alguns algoritmos não podem ou não podem ser facilmente usados para ambos os tipos de problemas, como regressão linear para modelagem preditiva de regressão e regressão logística para modelagem preditiva de classificação.

É importante ressaltar que a maneira como avaliamos as previsões de classificação e regressão varia e não se sobrepõe, por exemplo:

As previsões de classificação podem ser avaliadas usando precisão, enquanto as previsões de regressão não.

As previsões de regressão podem ser avaliadas usando o RMSE, enquanto as previsões de classificação não podem.

MSE

MSE ou erro quadrático médio é uma das métricas preferidas para tarefas de regressão. É simplesmente a média da diferença quadrática entre o valor alvo e o valor previsto pelo modelo de regressão. À medida que eleva ao quadrado as diferenças, ele penaliza até mesmo um pequeno erro que leva a uma superestimação de quão ruim é o modelo. É mais preferido do que outras métricas porque é diferenciável e, portanto, pode ser melhor otimizado
RMSE

A média dos termos de erro individuais é calculada para que possamos relatar o desempenho de um modelo com relação a quanto erro o modelo comete geralmente ao fazer previsões, em vez de especificamente para um determinado exemplo. As unidades do MSE são unidades quadradas. Por exemplo, se seu valor alvo representa "dólares", então o MSE será "dólares ao quadrado". Isso pode ser confuso para as partes interessadas; portanto, ao relatar os resultados, geralmente o erro quadrático médio da raiz é usado (discutido na próxima seção).

O erro quadrático médio entre os valores esperados e previstos pode ser calculado usando a função mean_squared_error () da biblioteca scikit-learn. A função pega uma matriz unidimensional ou lista de valores esperados e valores previstos e retorna o valor de erro quadrático médio.

É uma boa ideia estabelecer primeiro um MSE de linha de base para o seu conjunto de dados usando um modelo preditivo ingênuo, como prever o valor alvo médio do conjunto de dados de treinamento. Um modelo que atinge um MSE melhor do que o MSE para o modelo ingênuo tem habilidade.

RMSE

RMSE é a métrica mais usada para tarefas de regressão e é a raiz quadrada da diferença quadrada média entre o valor de destino e o valor previsto pelo modelo. É mais preferido em alguns casos porque os erros são primeiro ao quadrado antes da média, o que representa uma grande penalidade para erros grandes. Isso implica que o RMSE é útil quando grandes erros são indesejados

Você deve se lembrar que a raiz quadrada é o inverso da operação quadrada. MSE usa a operação quadrada para remover o sinal de cada valor de erro e punir erros grandes. A raiz quadrada inverte essa operação, embora garanta que o resultado permaneça positivo.

A raiz quadrada média do erro entre seus valores esperados e previstos pode ser calculada usando a função mean_squared_error () da biblioteca scikit-learn.

Por padrão, a função calcula o MSE, mas podemos configurá-la para calcular a raiz quadrada do MSE definindo o argumento "quadrado" como False.

MAE

MAE é a diferença absoluta entre o valor alvo e o valor previsto pelo modelo. O MAE é mais robusto para outliers e não penaliza os erros tanto quanto o mse. MAE é uma pontuação linear, o que significa que todas as diferenças individuais são ponderadas igualmente. Não é adequado para aplicações em que você deseja prestar mais atenção aos valores discrepantes.

O erro médio absoluto entre seus valores esperados e previstos pode ser calculado usando a função mean_absolute_error () da biblioteca scikit-learn.

A função pega uma matriz unidimensional ou lista de valores esperados e valores previstos e retorna o valor de erro absoluto médio

Um valor de erro absoluto médio perfeito é 0,0, o que significa que todas as previsões corresponderam exatamente aos valores esperados.

Isso quase nunca é o caso e, se acontecer, sugere que seu problema de modelagem preditiva é trivial.

Um bom MAE é relativo ao seu conjunto de dados específico.

É uma boa ideia estabelecer primeiro uma MAE de linha de base para o seu conjunto de dados usando um modelo preditivo ingênuo, como prever o valor alvo médio do conjunto de dados de treinamento. Um modelo que alcança um MAE melhor do que o MAE, pois o modelo ingênuo tem habilidade.

R2

Coeficiente de determinação ou R^2 é outra métrica usada para avaliar o desempenho de um modelo de regressão. A métrica nos ajuda a comparar nosso modelo atual com uma linha de base constante e nos diz o quanto nosso modelo é melhor. A linha de base constante é escolhida tomando a média dos dados e desenhando uma linha na média. R^2 é uma pontuação sem escala que implica que não importa se os valores são muito grandes ou muito pequenos, o R^2 será sempre menor ou igual a 1.

R2 AJUSTADO

R^2 ajustado representa o mesmo significado que R^2, mas é uma melhoria dele. O R^2 sofre do problema de que as pontuações melhoram em termos crescentes, embora o modelo não esteja melhorando, o que pode confundir o pesquisador. O R^2 ajustado é sempre menor do que o R^2, pois se ajusta para os preditores crescentes e só mostra melhora se houver uma melhora real

IMPLEMENTACAO

1. Criando modelo

É importante ter em mente que, neste exemplo, estamos usando dados arbitrários. Podemos fazer este exercício com qualquer conjunto de dados. X é uma matriz de inteiros de 0–9. Y é uma matriz dos primeiros 10 dígitos da sequência de Fibonacci. Depois de traçar os pontos de dados, ajustaremos nossa linha de regressão de mínimos quadrados comum.

Esta linha parece se ajustar relativamente bem. Os pontos azuis não estão muito longe da linha de regressão. Quanto mais próximos os pontos estão da linha, menor é a nossa variância. Quanto menor for a variância, melhor será o nosso modelo!

2. Linha média

A seguir, colocaremos outra linha em nossos dados. Se teoricamente tivéssemos apenas os dados Y (e nenhum X), o melhor modelo preditivo que seríamos capazes de fazer seria adivinhar a média de Y todas as vezes. Esta é uma etapa fundamental no cálculo de nosso r-quadrado, como você verá em um minuto.

3. Diferença entre os dados e o modelo linear

Se medirmos a diferença entre cada ponto de dados e a linha de regressão linear, elevar ao quadrado cada diferença e totalizá-los, obteremos a variância que existe dentro do modelo de regressão.

4. Diferença entre os dados e a média

Se medirmos a diferença entre cada ponto de dados e a linha horizontal, elevarmos ao quadrado cada diferença e totalizá-los, obteremos a variância total que existe apenas no conjunto de dados Y. Lembre-se, este é o modelo que usaríamos independentemente da existência de nossos dados X