Projeto Final

Ciência dos Dados

Prevendo o Tempo com Regressão Linear

Quando estávamos em uma das primeiras fases de execução do projeto, a escolha do dataset, tínhamos em mente trabalhar com algo relativo a crimes ou acidentes, se utilizando de uma técnica diferente da classificação, haja vista que já tínhamos trabalhado com esta no segundo projeto. Sempre ouvíamos um milhão de coisas acerca da clusterização, a ponto de já acharmos que esta salvaria o mundo, ou algo assim, mas o que realmente nos encantou foi a técnica de regressão. Oras, por quê? Eu é que lhe pergunto. Não há qualquer coisa de mágico em prever o futuro? Para nós havia, de modo que fizemos o caminho inverso: não escolhemos o dataset para, posteriormente, escolher a técnica; escolhemos um dataset que se adaptasse à técnica que já havíamos fixado.

Trocamos de dataset diversas vezes, visto que as informações contidas nestes não nos renderiam boas regressões, dado que quase não haviam variáveis quantitativas, mas sim qualitativas. Queríamos fazer algo grande, que despertasse a atenção de quem nos ouvisse falar sobre o projeto. Até que, então, o professor citou algo que nos despertou a atenção. Por que não prever o tempo com regressão? Prever o tempo é uma tarefa muito complexa, há diversas variáveis envolvidas e, embora o modelo matemático desenvolvido tenha alta precisão, é quase impossível acertar sempre. Estava lançado o desafio. Não precisávamos de algo de outro mundo, apenas precisávamos desenvolver algo que resolvesse problemas reais e palpáveis do nosso dia a dia, do nosso mundo. Prever o tempo faz parte desse escopo, sem dúvidas.

O Dataset

O nosso dataset possui informações acerca de precipitação, temperatura, mínima e máxima, insolação, umidade e velocidade do vento, medidas de 01/01/2010 a 05/10/2017 na estação Mirante de Santana, a principal estação meteorológica do Instituo Nacional de Meteorologia (INMET) no município de São Paulo.

Proposta

Desenvolver uma regressão linear precisa para prever a precipitação diária (se vai ou não chover no dia seguinte).

Desenvolvimento

O primeiro passo foi separar o dataset em duas partes para fazer a cross-validation da técnica após concluir o seu desenvolvimento. Fazer a cross-validation de um modelo/técnica nada mais é do que treinar a técnica em uma base e então testá-la em outra para verificar o quão precisa ela consegue ser.

Em seguida, resolvemos gerar a tabela com os resultados da regressão OLS com uma ou outra variável para verificar quão relacionada a “precipitação” estava com essas variáveis. Na regressão por mínimos quadrados (OLS – Ordinary

Least Squares), a equação estimada que gera a melhor reta para acompanhar os dados é calculada determinando-se a equação que minimiza a soma dos quadrados das distâncias entre os pontos de dados amostrais e os valores preditos pela equação. Em suma, esse método atua minimizando a soma dos quadrados dos erros entre o que foi medido e o que foi predito.

Geramos a tabela OLS da "Precipitação" com todas as demais variáveis do dataset, mas não obtivemos resultados tão satisfatórios. A predição não era boa, não bastava levar em consideração apenas uma variável para prever se iria ou não chover no dia seguinte, aparentemente prever o tempo envolvia um esforço bem maior que esse.

Então, resolvemos partir da regressão linear simples para uma regressão linear múltipla. Por vezes, uma variável depende / é influenciada não somente por uma outra variável, mas sim por várias. Geramos a tabela com todas as demais variáveis e percebemos, a partir de "$P > |t|$", que nos conta qual é a probabilidade do coeficiente que acompanha cada variável ser 0, que a "Temperatura Máxima" e a "Velocidade do Vento Média" não eram tão relevantes para nossa análise (a probabilidade dos coeficientes que acompanhavam essas variáveis ser 0, em comparação com as demais, era muito alto – da ordem de 50%). Tendo retirado essas duas variáveis, nos sobraram apenas a "Temperatura Mínima", a "Insolação" e a "Umidade Relativa Média". Rodando o OLS somente com essas três variáveis, constatamos que a "Insolação" também havia deixado de ser relevante para nossa análise. Por fim, havíamos enxugado nossa regressão e haviam apenas duas variáveis sendo consideradas para prever a "Precipitação". Nessa última análise, conseguimos um R-Squared, ou coeficiente de determinação múltipla, propriedade que indica quanto da variação em y é correspondida pela variação em x de 0,101.

R-Squared ou Coeficiente de Determinação Múltipla:

Se traçarmos uma reta que vai de cada um dos pontos amostrais até o ponto correspondente na reta gerada pelo modelo, teremos o erro nesse ponto. Se somarmos os quadrados dos erros de todos os pontos, teremos o erro quadrado da reta (o OLS gera uma reta que torna esse erro total o menor possível, isto é, que melhor se adequa aos dados amostrais). E então, se traçarmos uma reta que represente o valor médio de y e somarmos os quadrados das distâncias de cada ponto amostral até essa reta teremos a variação total de y. Sabemos que se dividirmos o erro quadrado da reta pela variação total de y, obteremos a porcentagem da variação total que não é correspondida pela reta. Para saber quanto da variação é correspondida pela reta, basta subtrair esse valor de 1. Esse resultado é chamado de R-Squared ou Coeficiente de Determinação (Múltipla, nesse caso, porque estamos tratando de uma regressão linear múltipla). Quando a variação total que não é correspondida pela reta é muito próxima de 1, o R-Squared é muito próximo de 0, o que, na maior parte das vezes, indica que a reta gerada não é uma boa aproximação para os dados amostrais. Se, por outro lado, a variação total que não é correspondida pela reta

é muito baixa, o R-Squared se aproxima de 1, indicando, na maior parte das vezes, que a reta é uma boa aproximação para os dados amostrais.

Dada a explicação do que significa o R-Squared, é perceptível que, aparentemente, a reta gerada não explicava muito bem os dados. Plotamos, então, um gráfico tridimensional das variáveis envolvidas e da "Precipitação" e acabamos por concluir que, no nosso caso, ter um Coeficiente de Determinação relativamente baixo não era tão ruim. Observando o gráfico, vimos que a escala da "Precipitação", a variável predita, vai de 0 a 120mm. Percebemos também que a grande concentração dos dados está na parte inferior do gráfico, bem próxima à reta. No entanto, há alguns pontos esparsos bem acima, a altos níveis de "Precipitação". Se pararmos para pensar, esses altos níveis de precipitação se devem a eventos externos à ordem natural das coisas, a eventos passíveis de serem chamados de caóticos, como frentes frias que vêm do oceano, por exemplo. Em um primeiro momento, não estamos tentando prever esse tipo de evento. Estamos, unicamente, tentando prever o tempo em situações realmente passíveis de previsão.

O teste F-Statistic, em suma, testa a hipótese de ser possível modelar os dados com boa acurácia "setando" os coeficientes regressores em 0. Isso significa que, se a hipótese nula não for rejeitada, a nossa regressão "não serve para nada", não tem utilidade. Pelo contrário, se a hipótese nula for rejeitada, nós aceitamos a hipótese alternativa, de que pelo menos um dos coeficientes regressores não deve ser 0 para que consigamos modelar os dados com boa acurácia. Na tabela com os resultados do OLS, abaixo de "F-Statistic", nós temos "Prob(F-Statistic). Quanto mais baixo for esse valor, melhor, pois assim podemos rejeitar a hipótese nula e afirmar, com convicção, que a nossa regressão tem algum propósito. Em nosso caso, o valor da Prob(F-Statistic) foi de $2{,}21 \cdot 10^{-35}$ e, portanto, pudemos rejeitar a hipótese nula.

Ainda analisando a tabela de resultados do OLS, percebemos, observando os coeficientes de cada variável analisada que, para cada aumento de 1°C na temperatura mínima, a precipitação sofre um acréscimo de 0,4949mm, bem como para cada aumento de 1 na umidade relativa média, a precipitação sofre um acrescimento de 0,2943. Ao que parece, faz sentido.

Na parte inferior da tabela de resultados, são apresentados os resultados de diversos testes estatísticos. O primeiro deles é o Omnibus, que utiliza a "skewness" (assimetria) e a "kurtosis" (achatamento) para testar a hipótese nula de que a distribuição residual, isto é, do erro, é uma distribuição normal. Em nosso caso, dado que a "Prob(Omnibus)" retornada foi 0, podemos rejeitar a hipótese nula. Isso implica que a distribuição residual de nossa regressão não segue uma distribuição normal.

Em seguida, temos o Durbin-Watson, que checa a autocorrelação do resíduo/erro gerado. O intervalo estatístico em que o teste é calculado vai de 0 a 4. Se o valor retornado está próximo de 2, isso sugere que não há

autocorrelação. Valores maiores que 2 sugerem correlação negativa e valores menores que 1 sugerem correlação positiva. Olhando para o resultado gerado em nossa tabela, tivemos um retorno de 1,776, indicando que podemos considerar, praticamente, ausência de correlação no resíduo gerado.

Por fim, o "Conditional Number" mensura a sensibilidade da saída de uma função à sua entrada. Se este for maior que trinta, então a regressão pode ter multicolinearidade. Em nosso caso, o OLS nos retornou um "Conditional Number" de 617, indicando que, provavelmente, nossa regressão possui multicolinearidade. Essa propriedade, caso assuma valores muito altos, pode atrapalhar a acurácia do modelo. Nesse caso, é preciso se utilizar de técnicas de regularização para minimizá-la e impedi-la de atrapalhar a qualidade da regressão. Dois métodos de regularização passíveis de serem aplicados são as técnicas de regressão Ridge ou LASSO, que serão melhor explicadas posteriormente.

Mas, então, paramos para pensar: não faz sentido prevermos a precipitação do dia com variáveis medidas no próprio dia. Afinal, é mais interessante que possamos saber, com antecedência de, no máximo, um dia, se vai chover ou não no dia seguinte. Começamos, então, a criar variáveis no dataset referentes aos dias anteriores, bem como criar média de temperaturas ($\frac{temperatura\ mínima + temperatura\ máxima}{2}$) e outras variáveis que pudessem tornar nossa regressão mais precisa. Até que criamos uma variável da média de precipitações dos dias anteriores. E foi aí que surgiu o primeiro problema.

Caindo em Séries Temporais

Quando fizemos menção de utilizar a média da precipitação de dias anteriores para calcular a precipitação, o professor nos alertou sobre o fato de estarmos caindo no que são chamadas séries temporais. Em suma, uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo. Uma característica muito importante das séries temporais é que as observações vizinhas são dependentes e o interesse é analisar e modelar essa dependência. No nosso caso, para que pudéssemos utilizar a média das precipitações em dias anteriores para prever a própria precipitação, teria de ocorrer exatamente o contrário: não poderia haver correlação entre essas duas variáveis (autocorrelação da precipitação com ela mesma em instantes de tempo anteriores).

Para ter uma resposta definitiva, resolvemos investir em um teste de autocorrelação denominado Ljung-Box. Esse teste averigua a relação entre uma variável com ela mesma em instantes de tempo anteriores e retorna os coeficientes de autocorrelação. Se estes se distanciarem muito de 0, isso significa que os valores não são aleatórios e independentes ao longo do tempo. Em nosso caso, obtivemos valores muito negativos, o que nos sinalizou que não

poderíamos utilizar a média das precipitações de dias anteriores para prever a precipitação atual, já que se fizéssemos isso, cairíamos em um modelo de séries temporais.

Testando outras técnicas, que não OLS

Polinomial e Gaussiana

Na regressão polinomial, tentamos ajustar um polinômio de algum grau à distribuição dos dados. Bem como na regressão obtida a partir de um processo gaussiano, não obtivemos resultados tão satisfatórios.

Ridge

Essa é uma das técnicas derivadas da regressão que, junto à técnica LASSO e algumas outras menos conhecidas, constitui os métodos regressores de Shrinkage, ou de encolhimento.

Métodos Regressores de Shrinkage

Métodos geralmente empregados em regressões constituídas de diversos coeficientes regressores. Estes tornam o modelo como um todo muito mais complexo e podem tirar características de interpretabilidade. Esses métodos objetivam eliminar esse problema, que pode absorver o ruído dos dados e causar o overfitting (quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados), ao atuar retendo um subconjunto de coeficientes regressores, o que não somente reduz a complexidade do modelo e a forma que o mesmo é calculado e construído, bem como reduz o erro e minimiza qualquer possibilidade do modelo ter overfitting.

Em suma, a Ridge Regression é um método de regularização que objetiva suavizar atributos que sejam relacionados uns aos outros e que aumentam o ruído no modelo (A.K.A multicolinearidade - condição que ocorre quando algumas variáveis preditoras no modelo estão correlacionadas a outras variáveis preditoras; a multicolinearidade forte é problemática porque pode aumentar a variância dos coeficientes de regressão, tornando-os instáveis). Assim, com a retirada de determinados atributos do modelo, este converge para um resultado muito mais estável em que, com a redução desses atributos, a redução em termos de acurácia do modelo se mantém inalterada. O mecanismo algorítmico que faz isso é um mecanismo de penalização que coloca um viés e que vai reduzindo os valores dos betas até valores muito próximos de zero. Usando esse

mecanismo de penalização do viés, os atributos que contribuem menos para o poder preditivo do modelo são levados para a irrelevância.

Quando iniciamos o projeto, testamos diversas técnicas de regressão para verificar aquela que melhor se adaptava ao modelo e que gerava o menor erro. Embora não tivéssemos muitos coeficientes em nossa regressão, resolvemos testar a Ridge Regression para verificar os resultados que obteríamos.

LASSO (Least Absolute Shrinkage and Selection Operator)

O LASSO, assim como a Ridge Regression, tem um mecanismo de penalização dos coeficientes com um alto grau de correlação entre si. A diferença é que este usa o mecanismo de penalizar os coeficientes de acordo com o seu valor absoluto (soma dos valores dos estimadores), minimizando o erro quadrático. Isso é feito por meio da penalização do coeficiente até que este convirja para zero, o que, naturalmente, vai eliminar o atributo e reduzir a dimensionalidade do modelo.

Também testamos a regressão LASSO em nossos dados, mas, como é possível observar nos resultados gerados no código, esta não produziu resultados satisfatórios.

Finalizando o Projeto

Por fim, após testar a regressão pelo método OLS com diversas variáveis, constatamos que aquelas que geravam melhor resultado eram a “Temperatura Média do Dia Anterior” (média obtida a partir das temperaturas máxima e mínima do dia anterior), juntamente à “Média de Umidade de Dois Dias Anteriores”. Para tornar o resultado ainda melhor, previmos o logaritmo da “Precipitação” e não a “Precipitação” em si. As variáveis mencionadas, como é possível observar no código do projeto, geraram um R-Squared de 0,143, bem como uma prob(F-Statistic) de $1{,}04 \cdot 10^{-50}$, uma assimetria e um achatamento muito inferiores aos mencionados na regressão gerada em momento anterior e ainda uma certa probabilidade, embora muito baixa, de que a distribuição residual seja uma normal (Prob(Jarque-Bera)). Se observarmos o gráfico 3D gerado com as variáveis, é possível perceber que o plano se adapta consideravelmente bem aos dados.

Calculando o erro dessa regressão, obtivemos um valor de aproximadamente 2719, o que é um erro consideravelmente alto, sem dúvidas. Ao aplicar o método de regressão Ridge, obtivemos o mesmo valor residual, e o LASSO, um valor ainda maior (é possível observar, no código, dado o R^2 gerado (0), que a regressão LASSO não foi uma boa escolha para esses dados – apenas deixamos no código para que pudessem visualizar a aplicação da técnica em questão).

Por fim, concluímos que prever o tempo é uma tarefa realmente complexa, de modo que é muito difícil obter uma previsão de alta precisão a partir de um número tão restrito de variáveis. Para tornar nosso modelo cada vez melhor, seria necessário partir para o uso de séries temporais, bem como modelar fenômenos extraordinários, como os efeitos da aproximação de um El Niño, por exemplo. Prever o tempo é uma tarefa que, até o presente momento, ninguém conseguiu executar com alta precisão. Seria necessário um estudo mais aprofundado e o uso de técnicas mais sofisticadas que a Regressão Linear para fazê-lo.

Interface para Acesso aos Resultados

Montamos uma interface mais amigável para que o usuário possa inserir os dados necessários e obter uma resposta do algoritmo à pergunta: “Vai chover amanhã?”. Ainda não transformamos o algoritmo de regressão em uma função para gerar a resposta correta ao usuário, mas a interface imaginada já está pronta para acesso no GIT. No lugar da função, colocamos um algoritmo aleatório.

Trabalho em equipe:

Marco Moliterno Pena Piacentini: Criou a interface que facilita a utilização do programa pelo usuário.

Samara Barreto de Oliveira Gadelha: Pesquisa de *Dataset's,* ajuda em problemas pontuais com o código, pesquisa das técnicas utilizadas, elaboração da parte final do relatório e organização do GitHub.

Tainara Soares Mendes: Pesquisa de *Dataset's*, pesquisa das técnicas utilizadas, estruturação do código, elaboração do relatório e organização do código.