Universidade de Brasília
Departamento de Estatística

Interpretação de modelos de Aprendizado de Máquina

Aluno: Lucas de Moraes Pinto Pereira

Pesquisador Responsável: Guilherme Souza Rodrigues

Resumo

Neste estudo, realizamos uma extensa comparação entre os resultados obtidos por modelos probabilísticos e de aprendizagem de máquina no ajuste de um conjunto de dados reais. Tal exercício ilustrou o uso de alguns dos principais métodos de interpretação e de inferência estatística em modelos preditivos. Assim, verificamos que além de melhorar a precisão das predições, os modelos de aprendizagem de máquina conseguiram identificar as covariáveis mais importantes.

Abstract

In this study, we performed an extensive comparison between the results obtained by probabilistic and machine learning models when fitting a real dataset. This exercise illustrated the use of some of the main methods of interpretation and statistical inference in predictive models. Thus, we found that in addition to improving the accuracy of predictions, machine learning models were able to identify the most important covariates.

Introdução

Contextualização

Em virtude do crescimento da complexidade e do volume dos dados atualmente disponíveis, métodos algorítmicos de aprendizagem de máquina passaram a oferecer soluções

muito bem sucedidas, sobretudo para as atividades de previsão, de classificação e de agrupamento. Entretanto, tais benefícios trouxeram consigo perdas substanciais na capacidade de abstração, de interpretação e de generalização. Apesar dessas problemáticas, existem métodos já disponíveis que ajudam a emular o que se faziam com os modelos probabilísticos de uma forma indireta, via simulação e que ajudam na interpretação e na inferência desses modelos.

Aprendizado de Máquina

Como Arthur Samuel definiu [1], aprendizado de máquina é o "campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados". Ou seja, utilizando algoritmos para coletar e aprender com os dados, podemos fazer previsões e entender comportamentos.

Objetivos do Projeto

Dentro destas perspectivas apresentadas, o objetivo principal deste trabalho é analisar e ilustrar alguns dos principais métodos de interpretação e visualização de algorítmos de predição e também realizar comparações entre a abordagem probabilística e a de aprendizado de máquina.

Fundamentos Teóricos

Regressão Linear Múltipla

O modelo de regressão linear múltipla estimado por mínimos quadrados ordinários (MQO) permite modelar a relação entre uma variável dependente e um conjunto de variáveis explicativas. Este modelo assume uma relação linear entre uma variável dependente Y e um conjunto de variáveis explicativas $x_{i0}, x_{i1}, ..., x_{iK}$. x_{ik} estas também são chamadas de variáveis independentes ou covariáveis. Para uma revisão detalhada sobre Regressão Linear, veja [2].

Redes Neurais Artificiais

Uma rede neural artificial é um modelo de aprendizado de máquina composto por camadas, que é uma coleção de neurônios, com conexões entre elas. Essas camadas transformam os dados, primeiro calculando a soma ponderada das entradas e, depois, normalizando-a, ao usar as funções de ativação atribuídas aos neurônios. A camada mais à esquerda em uma Rede Neural é chamada de camada de entrada e a camada mais à direita é chamada de camada de saída. As camadas entre a entrada e a saída são chamadas de camadas ocultas. Qualquer rede neural tem uma camada de entrada e uma camada de saída. No entanto, o número de camadas ocultas difere entre diferentes redes, dependendo da complexidade do problema. Além disso, cada camada oculta pode ter sua própria função de ativação.

Qualquer rede neural com pelo menos duas camadas ocultas é chamada de Rede Neural Profunda. Uma rede neural faz previsões aprendendo os pesos de cada um dos neurônios em cada camada. O algoritmo através do qual eles aprendem é chamado de "Back Propagation". Para um maior entendimento de Redes Neurais Artificiais, veja [3].

XGBoost (eXtreme Gradient Boosting)

XGBoost significa *Extreme Gradient Boosting*, que foi proposto pelos pesquisadores da Universidade de Washington [4]. É uma biblioteca escrita em C++ que otimiza o treinamento para *Gradient Boosting*. Neste algoritmo, as árvores de decisão são criadas de forma sequencial. O peso das variáveis preditas incorretamente pela árvore é aumentado e essas variáveis são, portanto, alimentadas na segunda árvore de decisão. Esses classificadores/preditores individuais se agrupam para fornecerem um modelo forte e mais preciso, que pode funcionar em problemas de regressão, classificação e previsão definida pelo usuário.

Metodologia

Introdução

Foi realizada uma comparação de capacidade de predição, por meio dos coeficientes R^2 e o EQM (Erro quadrático médio), entre um modelo probabilístico de Regressão Linear Múltipla, modelos de Redes Neurais Artificiais e um modelo XGBoost. Com o intuito de interpretar os modelos de aprendizagem de máquina, foram utilizadas simulações de Monte

Carlo para obter coeficientes que representam a importância de cada variável na predição em cada modelo os valores de SHAP (*SHapley Additive exPlanations*) , introduzido em [5]. As análises foram feitas com a linguagem de programação Python.

Conjunto de Dados

As técnicas utilizadas foram aplicadas e ilustradas em um conjunto real de dados, o "California Housing", que contém dados do censo norte americano de 1990.

Neste conjunto de dados, temos informações sobre a demografia (renda, população e ocupação das casas) nos bairros, a localização (latitude e longitude) e informações gerais sobre as casas (número de quartos, número de salas e idade das casas). A variável de interesse é o valor médio das casas para os distritos da Califórnia, expresso em centenas de milhares de dólares ($100.000). O conjunto de dados possui 20443 observações e foi dividido em treino e teste para a realização do estudo, deixando 80% para o treinamento dos modelos e 20% para a validação.

Na figura a seguir, é possível visualizar os dados, por meio de um mapa referente à California, em que os pontos em vermelho representam as casas com um valor médio mais alto e os pontos em azul representam as casas com os valores mais baixos.

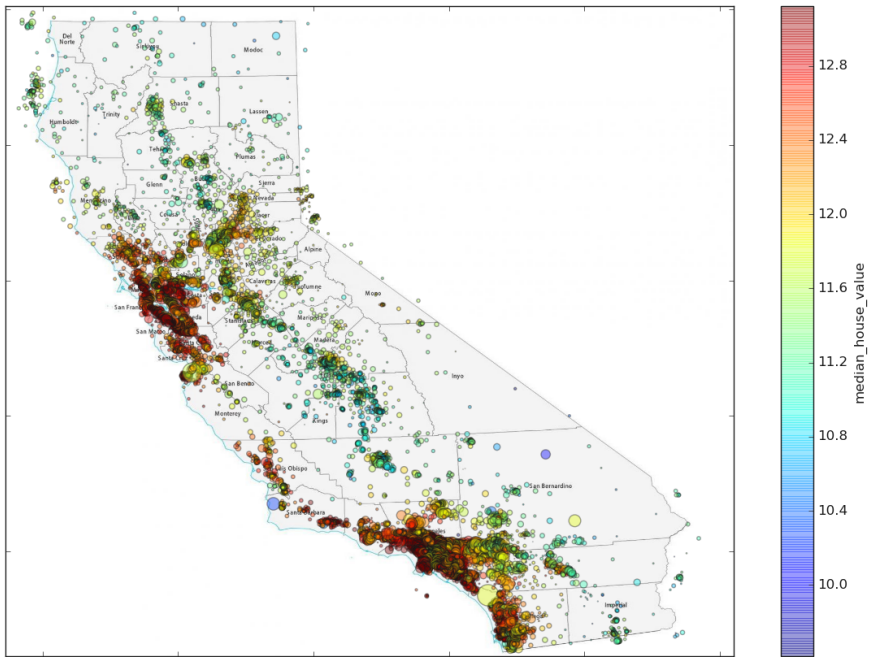


Figura 1: Mapa da Califórnia

Pré-processamento

O passo seguinte foi o pré-processamento dos dados. Com o auxílio das bíbliotecas Pandas, Numpy e SKlearn, na linguagem Python, foram realizadas as seguintes tarefas para a padronização dos dados e adequação ao estudo:

- Aplicação da transformação log na variável resposta Y, com o intuito de obter uma maior linearidade e, ainda, manter a interpretabilidade da variável.
- Adição de duas variáveis com distribuição Normal Padrão ao conjunto de dados ("Norm1" e "Norm2"), sendo elas totalmente independentes ao valor das casas. Essa adição foi feita com o objetivo de enteder a capacidade desses modelos de identificar variáveis irrelevantes ao valor da variável resposta.
- Normalização das variáveis explicativas para deixá-las em uma mesma escala de valores.

Modelagem

Para realizar tais comparações entre os modelos, foi utilizado o modelo probabilístico de Regressão Linear Múltipla e, para os de aprendizado de máquina, foram utilizados as Redes Neurais Artificiais e o algoritmo XGBoost. Nesse sentido, para ajustar os modelos, foram utilizadas as bíbliotecas SKlearn, Keras e a XGBoost, respectivamente.

Durante a análise exploratória dos dados, notou-se que os valores das casas tinham um teto máximo de 500.001$ e, por isso, durante a predição dos modelos, as previsões que ultrapassaram este teto máximo foram reajustadas para 500.001$.

Interpretação dos Modelos

A fim de interpretar os modelos, foram utilizadas três metodologias. Exclusivamente para a Regressão Linear, foram utilizados os coeficientes estimados de cada variável e seus respectivos p-valores. Entretanto, não é possível interpretar os demais modelos da mesma forma e, nesse sentido, foram trabalhadas outras formas que podem ser utilizadas em qualquer modelo preditivo.

Utilizando os métodos de simulação de Monte Carlo, foram gerados coeficientes que representam a relevância da variável para a predição dos modelos. Tais coeficientes foram gerados, ao embarallhar aleatoriamente os valores de uma variável e calcular o Erro Quadrático Médio (EQM) das previsões para o banco sintético de dados. Esse procedimento foi realizado mil vezes para cada váriavel e para cada modelo e, finalmente, calculou-se uma média aritmética desses valores. Com os mil EQMs gerados, por meio de simulação,

foi possível calcular o p-valor desses coeficientes. Assim, quanto maior o valor do coeficiente gerado, em comparação ao EQM real do modelo, maior relevância a variável possui na predição do modelo e, quanto menor o valor, menos a variável é relevante.

Outro método utilizado para quantificar a importância da variável para a predição dos modelos, foi o dos valores de SHAP, no qual, quanto maior o valor, mais importante a variável.

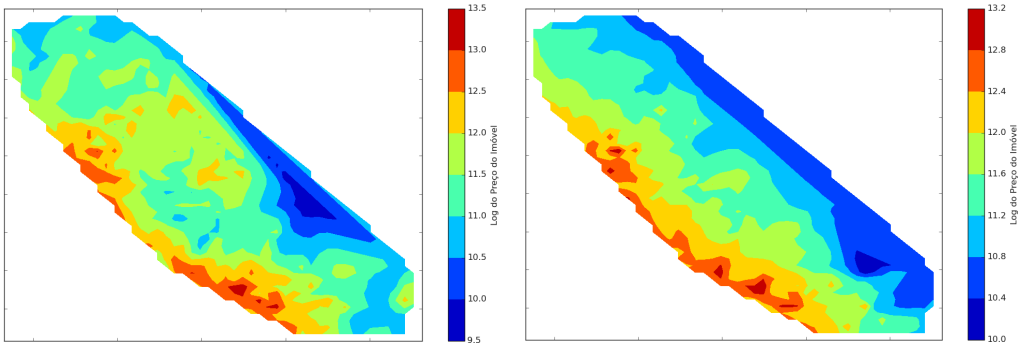
Resultados

Com os coeficientes obtidos pelo método de simulação de Monte Carlo, foi possível observar qual importância os modelos atribuíram a cada variável, uma vez que todos os obtiveram as mesmas variáveis como mais relevantes para a predição. Contudo, graus de relevância diferentes foram apresentados, fato evidenciado quando foram comparados aos valores de SHAP e coeficientes da Regressão Linear, pois todos estavam de acordo. Outrossim, observando os coeficientes e os p-valores obtidos pela simulação de Monte Carlo e pelo modelo probabilístico, todos os modelos foram capazes de identificar a irrelevância das duas variáveis que foram artificialmente adicionadas aos dados.

Tabela 1: EQM e R^2

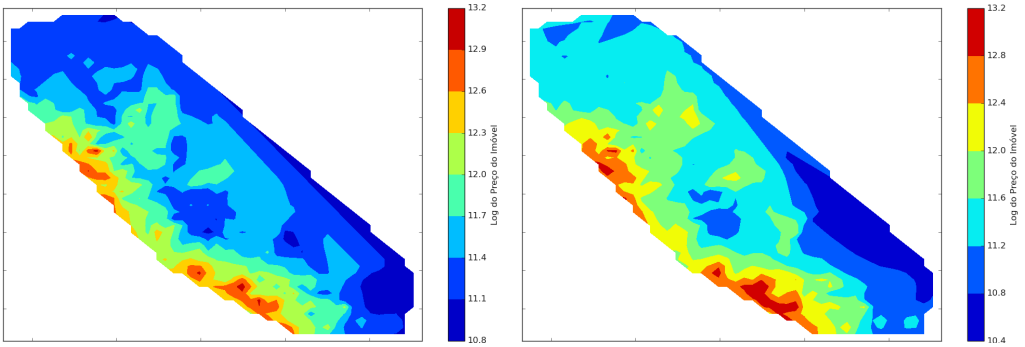
Modelo	***EQM***	R^2
Regressão Linear	0.112	0.662
Rede Neural	0.073	0.779
XGBoost	0.060	0.818

Pôde-se observar, por meio dos coeficientes R^2 e o EQM, uma melhor capacidade preditiva dos modelos de aprendizado de máquina em relação ao probabilístico. Observando a Tabela 1, percebemos que a regressão linear obteve um erro quadrático médio superior, comparado aos demais, e, pelo R^2, vemos que a regressão linear explica cerca de 66% da variabilidade dos dados, quando os outros modelos conseguem explicar por volta de 80% da variabilidade.



(a) Valores das casas reais

(b) Regressão Linear



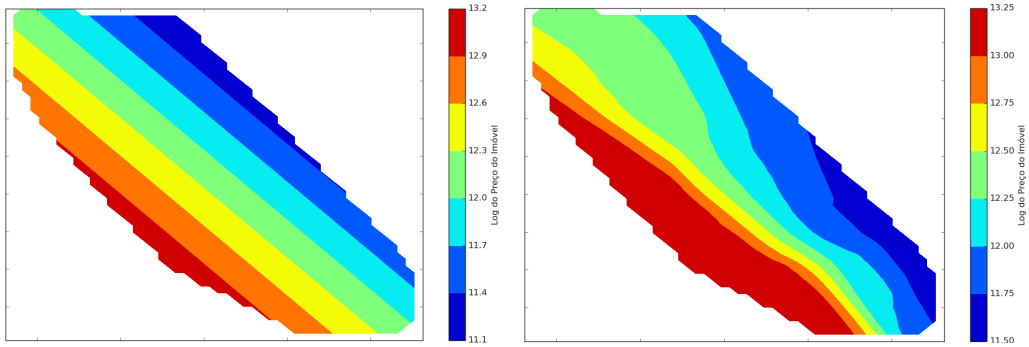
(c) Rede Neural Artificial

(d) XGBoost

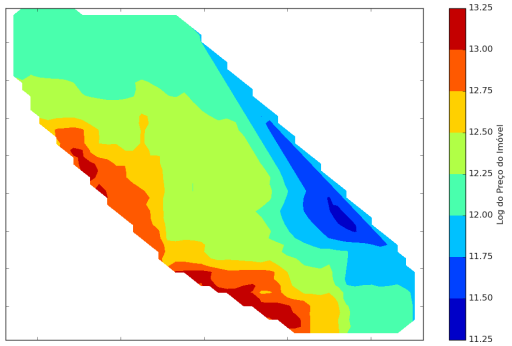
Figura 2: Mapa de curva de nível com os valores reais das casas e com as previsões dos valores das casas para cada modelo.

Os gráficos da Figura 2 são mapas de curva de nível referentes ao território da Califórnia, em que as regiões mais vermelhas representam os locais com os valores de casa mais elevados.

Um dos motivos para essa diferença de desempenho foi a falta de capacidade do modelo probabilístico para captar as não-linearidades de algumas das variáveis. Para perceber a não linearidade dos dados, analisa-se o gráfico da Figura 2(a) e conclui-se que a latitude e longitude interferem no valor das casa de forma não linear. Nesse sentido, para visualizar e confirmar a relação não linear entre estas variáveis, foram feitas previsões com os modelos ajustados, variando apenas a longitude e latitude e fixando as demais variáveis em suas respectivas médias. Essas previsões foram representadas pelos mapas de curva de nível a seguir:



(a) Regressão Linear (b) Rede Neural Artificial



(c) XGBoost

Figura 3: Mapa de curva de nível da predição dos preços das casas variando apenas a latitude e longitude e fixando as demais variáveis em suas respectivas médias.

Por meio da análise da Figura 3(a), percebe-se que a Regressão Linear não conseguiu capturar a complexidade das relações. Entretanto, analisando as Figuras 3(b) e 3(c), a Rede Neural e o XGBoost conseguiram captar essa não lienaridade de forma mais eficiente.

Ocasionalmente, pode-se suspeitar que uma variável explicativa particular X_i não é muito útil, isto é, que a sua influência sobre a variável dependente não é significativa. Para o modelo de regressão, basta analisar os coeficientes das variáveis (β_i), e, para saber se é significativo, testamos a hipótese nula de que o coeficiente para esta variável é nulo, com nível de significância de 5%:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

Tabela 2: Coeficientes estimados pela Regressão Linear e os respectivos p-valores

Variáveis	**Coef. Regressão Linear**	**P-valor**
longitude	-0.551788	0.000
latitude	-0.595822	0.000
housing_median_age	0.040401	0.000
total_rooms	-0.070470	0.000
total_bedrooms	0.200682	0.000
population	-0.186556	0.000
households	0.086362	0.000
median_income	0.337624	0.000
Norm1	-0.000383	0.886
Norm2	0.007481	0.005

Observando a Tabela 2, vale ressaltar que as variáveis latitude e longitude foram as que tiveram maior peso na predição do preço das casas, e as variáveis adicionadas artificialmente Norm1 e Norm2 tiveram as menores contribuições. Ao nível de significância de 0.05, há evidências para concluir que, exceto a variável Norm1, todas foram úteis na predição do preço das casas.

Mas para os modelos de aprendizado de máquina essa interpretação foi realizada por meio dos coeficientes gerados por simulacao e pelos valores de SHAP. Com os coeficientes obtidos pelo método de simulação de Monte Carlo, foi possível observar qual importância os modelos atribuíram a cada variável,

Tabela 3: Coeficientes gerados por simulações baseadas no EQM para cada modelo

Variáveis	**Regressão Linear**	**Rede Neural**	**XGBoost**
longitude	0.634	0.335	0.318
latitude	0.747	0.403	0.388
housing_median_age	0.116	0.090	0.066
total_rooms	0.120	0.121	0.074
total_bedrooms	0.189	0.106	0.079
population	0.173	0.161	0.097
households	0.129	0.148	0.064
median_income	0.3357	0.238	0.210
Norm1	0.113	0.074	0.061
Norm2	0.113	0.074	0.061
EQM real	0.112	0.073	0.060

Analisando os valores obtidos na Tabela 3, percebe-se que em todos os modelos as variáveis longitude e latitude foram as que apresentaram o maior peso na predição do preço

das casas, pois, ao serem embaralhadas foram as que mais elevaram o EQM comparado ao EQM real. Entretanto, as variáveis Norm1 e Norm2 elevaram muito pouco o valor do EQM quando comparado ao EQM real dos modelos, indicando que essas variáveis tem pouco peso na predição do preço das casas .

para saber se é significativo, foi necessário formular hipóteses diferentes das apresentadas na Regressão Linear, então, testamos a hipótese nula de que o EQM simulado pelo embaralhamento da variável é menor ou igual ao EQM real do modelo, com nível de significância de 5%:

$$\begin{cases} H_0 : EQMsimulado \leq EQMreal \\ H_1 : EQMsimulado > EQMreal \end{cases}$$

Tabela 4: P-valores referentes aos coeficientes baseados no MSE para cada modelo

Variáveis	**Regressão Linear**	**Rede Neural**	**XGBoost**
longitude	0.000	0.000	0.000
latitude	0.000	0.000	0.000
housing_median_age	0.000	0.000	0.000
total_rooms	0.000	0.000	0.000
total_bedrooms	0.000	0.000	0.000
population	0.000	0.000	0.000
households	0.000	0.000	0.000
median_income	0.000	0.000	0.000
Norm1	0.150	0.067	0.060
Norm2	0.368	0.509	0.178

Observando a Tabela 4, percebe-se que para os 3 modelos a hipótese nula não foi rejeitada para as variáveis adicionadas artificialmente Norm1 e Norm2, ou seja, como o esperado, ao nível de significância de 5%, há evidências para concluir que essas variáveis não influenciam no valor da predição dos preços das casas.

Outra maneira de quantificar a relevância das variáveis para os modelos é por meio dos valores de SHAP.

Tabela 5: Valores de SHAP para cada modelo

Variáveis	**Regressão Linear**	**Rede Neural**	**XGBoost**
longitude	0.500	0.376	0.207
latitude	0.548	0.434	0.251
housing_median_age	0.033	0.031	0.017
total_rooms	0.042	0.062	0.033
total_bedrooms	0.131	0.056	0.047
population	0.119	0.161	0.082
households	0.057	0.091	0.012
median_income	0.252	0.160	0.225
Norm1	0.000	0.005	0.005
Norm2	0.006	0.005	0.006

Observando os valores de SHAP presentes na Tabela 5, chegamos às mesmas conclusões tiradas pelos métodos de interpretação anteriores, ou seja, as variáveis latitude e longitude são as mais relevantes para os três modelos preditivos, pois possuem os valores mais altos e todos os modelos foram capazes de identificar a irrelevância das variáveis Norm1 e Norm2 para as predições, atribuindo-as valores muito baixos.

Conclusão

O presente trabalho comparou modelos probabilísticos com os de aprendizagem de máquina, buscando, assim, explorar potenciais técnicas para a interpretação dos modelos conhecidos como "caixa-preta". Nesse sentido, observa-se que, para dados mais complexos, a Rede Neural e o XGBoost tiveram maior capacidade de predição comparados à Regressão Linear Múltipla e, por isso, interpretar estes modelos é de suma importância. Assim, com as técnicas de simulação de Monte Carlo e os valores SHAP, foi possível esclarecer as previsões dos algoritmos “caixa-preta” e ter um ganho significativo na interpretabilidade desses modelos, tornando possível identificar as covariáveis mais importantes para os modelos de aprendizado de máquina.

Bibliografia

[1] SAMUEL, A. L. Some studies in machine learning using the game of checkers. IBM Journal ofresearch and development, IBM, v. 3, n. 3, p. 210–229, 1959.

[2] Neter, John, et al. "Applied linear statistical models."(1996): 318.

[3] Haykin, S., Network, N. (2004). A comprehensive foundation. Neural networks, 2(2004), 41.

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[5] Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.