

RELATÓRIO

LISTA DE EXERCÍCIOS 1

INTRODUÇÃO AO APRENDIZADO DE MÁQUINA (COE782) | COPPE | UFRJ

## Luis Paulo Albuquerque Guedes

Programa de Engenharia Elétrica

Universidade Federal do Rio de Janeiro

Av. Athos da Silveira Ramos, 149 - Bloco H, 3º andar  
[luis.albuquerque@marinha.mil.br](mailto:luis.albuquerque@marinha.mil.br)

**1 INTRODUÇÃO**

O conceito de Aprendizagem de Máquina (*Machine Learning -* ML) aborda duas questões estreitamente relacionadas: a criação de sistemas de computador que podem melhorar automaticamente com base na experiência e a identificação das leis estatísticas, computacionais e teóricas que regem todos os processos de aprendizado, sejam eles computacionais, humanos ou organizacionais. Essa disciplina desempenha um papel central não apenas na resolução dessas questões fundamentais do ponto de vista científico e de engenharia, mas também na criação de software altamente prático que é amplamente utilizado, principalmente para auxiliar na tomada de decisões em uma variedade de setores, como saúde, manufatura, educação, modelagem financeira, segurança pública e marketing [1].

O avanço na aprendizagem de máquina tem sido impulsionado por novos algoritmos, teorias e pela crescente disponibilidade de dados online e recursos computacionais acessíveis. Isso resultou na adoção generalizada de métodos intensivos em dados em campos diversos, promovendo uma abordagem mais orientada por evidências na tomada de decisões, abrangendo várias esferas da vida, desde a assistência médica até a indústria, educação e áreas como segurança pública e marketing [1].

Neste contexto, ainda que possam ser confundidos, cabe ressaltar que Inteligência Artificial (IA) e Aprendizagem de Máquina guardam particularidades distintas, cujos conceitos, não necessariamente, se confundem. Em [2], apesar de projetos de aprendizagem de máquina serem genericamente chamados de “tecnologia de IA”, os sistemas subjacentes não envolvem raciocínio ou pensamento de alto nível, como pressupõe as atividades cognitivas do pensamento humano.

A doutrina, nesta mesma linha, também adota a posição de que aprendizagem de máquina se relaciona a atividades que envolvem, principalmente, tomada de decisão, ao passo que IA, propriamente dita, se situa em um nível mais elevado de abstração, sendo, até o presente momento um objetivo inatingível: “No futuro próximo, os computadores não serão capazes de se equiparar aos humanos em sua capacidade de raciocinar abstratamente sobre situações do mundo real” [2].

Neste contexto, é a apresentada a disciplina Introdução ao Aprendizado de Máquina (COE782), que tem como objetivo fornecer uma perspectiva abrangente acerca dos conceitos básicos de Aprendizado de Máquina, com ênfase no aprendizado supervisionado e nos desafios associados à regressão e classificação, bem como apresentar a abordagem metodológica necessária para lidar com conjuntos de dados, bem como algumas técnicas clássicas para resolver problemas relacionados a esta temática.

Lançado em 2006, o livro *Pattern Recognition and Machine Learning* (PRML), de Christopher Bishop, destaca-se como uma obra altamente popular na área da aprendizagem de máquina, além de ser a principal referência utilizada na disciplina COE782 do Programa de Engenharia Elétrica (PEE) da COPPE/UFRJ.

No ensejo destas breves considerações, o presente relatório busca atender aos questionamentos propostos pela primeira lista de exercícios da disciplina COE782. Para consecução desse objetivo, este relatório encontra-se estruturado em três capítulos, sendo o primeiro deles, a presente introdução. O capítulo 2 apresenta o conhecimento prévio relevante e notas de aula que servem de base teórica para a resolução das questões propostas, que são apresentadas no decorrer no capítulo 3.

**2 CONHECIMENTO PRÉVIO RELEVANTE E NOTAS DE AULA**

**2.1 CONCEITOS BÁSICOS**

**2.1.1 PROBABILIDADE**

A ideia geral da probabilidade é frequentemente dividida em dois conceitos relacionados [3]:

1. **Probabilidade clássica, também chamada de "a priori" ou "teórica":** descreve uma série de eventos futuros cuja ocorrência é determinada por fenômenos físicos aleatórios. Por exemplo, em um lançamento de dado justo, no qual são identificáveis ​​seis números (faces) como possíveis resultados, tem-se que cada face possui a mesma probabilidade de se concretizar, sendo atribuída a cada uma delas uma probabilidade de 1/6. Portanto, se o evento desejado for "número ímpar", a probabilidade associada a esse resultado específico é de 3/6, ou seja, 1/2. Esta perspectiva tem a vantagem de ser conceitualmente simples para muitas situações. No entanto, é limitado, uma vez que muitas situações não têm um número finito de resultados igualmente prováveis. Lançar um dado ponderado é um exemplo em que temos um número finito de resultados, mas eles não são igualmente prováveis.
2. **Probabilidade bayesiana, também chamada de "a posteriori" ou "frequentista"**: reflete as incertezas humanas sobre afirmações quando não se possui conhecimento completo das circunstâncias causadoras. Trata-se de uma perspectiva que define a probabilidade por meio de um experimento mental, na qual as probabilidades podem fornecer uma quantificação da incerteza. A visão frequentista é o que dá credibilidade às estimativas padrão baseadas em amostragem. Por exemplo, se uma amostra aleatória suficientemente grande de uma população (escolha aleatória de uma amostra de 1.000 estudantes da população de todos os 50.000 estudantes matriculados em uma determinada universidade), então a média de alguma medida (por exemplo, despesas universitárias) para a amostra é uma estimativa razoável da média da população. Vale ressaltar que, ao longo do livro PRML, o autor destaca em várias instâncias uma abordagem fundamentada na probabilidade Bayesiana.

O **Teorema de Bayes** apresenta a seguinte forma [1]:

Em que:

representa a probabilidade "a priori";

representa a probabilidade "a priori"; e

representa a*likelihood.*

Isso nos permite avaliar a incerteza depois dos termos observados na forma da probabilidade posterior *.*

**2.1.1 TEOREMA CENTRAL DO LIMITE (TCL)**

O Teorema Central do Limite (TCL) é um resultado notável na teoria da probabilidade, pois, a partir dele, é possível realizar procedimentos de inferência estatística sem necessidade de conhecer a distribuição da população. Em resumo, ele descreve como a soma de variáveis aleatórias independentes converge para um modelo normal quando convenientemente padronizada. O TCL afirma que à medida que o tamanho da amostra aumenta, a distribuição da média da amostra se aproxima cada vez mais de uma distribuição normal. Em outras palavras, quando a amostra é suficientemente grande, a distribuição da média é uma distribuição aproximadamente normal [4].

**2.1.2 TEOREMA DE FUBINI**

Sejam A e B espaços de medida completos. Suponha uma função mensurável. Se

em que a integral é tomada com relação à medida produto associada ao espaço , então:

em que as duas primeiras integrais são integrais iteradas com relação a duas medidas, respectivamente, e a terceira é uma integral com relação ao produto dessas duas medidas [5].

**2.1.3 MUDANÇA DE VARIÁVEIS EM INTEGRAIS DUPLAS: CASO PARTICULAR**

Para o caso particular da mudança de variáveis do sistema de coordenadas cartesianas para o sistema de coordenadas polares em que e , o jacobiano é dado por:

Portanto, o jacobiano da transformação a mudança de variáveis na integral dupla toma a forma [6]:

**2.2 NOTAÇÃO**

Os vetores são representados por letras romanas minúsculas em negrito, como, e todos os vetores são assumidos como vetores coluna. Um sobrescrito pode denotar uma matriz transposta ou um vetor, de modo que será um vetor linha. Letras romanas maiúsculas em negrito, como **M**, representam matrizes. A notação denota um vetor linha com M elementos, enquanto o vetor coluna correspondente é escrito como .

O valor esperado de uma função em relação a uma variável aleatória é definida por . Em situações em que não houver ambiguidade quanto à variável sobre a qual está sendo realizada a média, isso será simplificado omitindo o sufixo, por exemplo, . Se a distribuição de estiver condicionada a outra variável , então o valor esperado correspondente será expresso por . Da mesma forma, a variância é indicada por , e para variáveis vetoriais, a covariância é escrita como Também usaremos como uma notação abreviada para

Se tivermos valores de um vetor x de D-dimensões podemos combinar as observações em uma matriz de dados , na qual a n-ésima linha decorresponde ao vetor linha . Assim, o elemento de corresponde ao i-ésimo elemento da n-ésima observação

**2.3 FORMULÁRIO**

(1.152)

**3 RESOLUÇÃO DOS EXERCÍCIOS PROPOSTOS**

* 1. Considere a função de erro da soma dos quadrados dada por (1.2), na qual a função é dada pelo polinômio em (1.1). Mostre que os coeficientes que minimizam essa função de erro são dados pela solução do seguinte conjunto de equações lineares:

Em que:

Aqui, tem-se que ou denota o índice de um componente, enquanto ( denota elevado à potência de .

**Solução:**

Para minimizar a função de erro E, é necessário calcular a sua derivada em relação ao vetor **w** e, em seguida, igualar o resultado a zero, ou seja, . Como consequência, isso resultará na solução para **w** = {} que minimiza a função E.

Considerando a Eq. (1.2), temos:

Derivado em relação a , obtemos:

Fazendo a distributiva dentro do somatório, temos:

Em aula foi mencionado que a derivada de um ponto em relação a um vetor resulta em uma matriz. A derivada de uma matriz em relação a um vetor resulta em um tensor etc. Percebemos, portanto, que a derivada em relação a um vetor acarreta um aumento da dimensionalidade. Portanto, abaixo, temos um caso em que a derivada de um vetor em relação a um vetor resulta em uma matriz. Vejamos:

Substituindo , temos:

Aplicando as propriedades de operações com somatórios:

Sabemos que e concluímos, então, que a equação:

pode ser escrita como:

Como queríamos demonstrar.

* 1. Escreva o conjunto de equações lineares acopladas, da forma da Eq. (1.122), satisfeitas pelos coeficientes que minimizam a função erro regularizada da soma dos quadrados dada pela Eq. (1.4).

**Solução:**

Para minimizar a função de erro E, é necessário calcular a sua derivada em relação ao vetor **w** e, em seguida, igualar o resultado a zero, ou seja, . Como consequência, isso resultará na solução para **w** = {} que minimiza a função E.

Derivado em relação a , obtemos:

Fazendo a distributiva dentro do somatório, temos:

Substituindo , temos:

Manipulando o somatório duplo, em atenção às propriedades da linearidade, temos:

Temos, portanto, a relação da Eq. .

Para trazer o termo para dentro do somatório, temos que trabalhar com o delta de Kronecker Pela definição:

Portanto:

Em que:

* 1. Utilizando a definição em (1.38), mostre que a variância de satisfaz a equação (1.39).

**Solução:**

Expandindo o quadrado do termo, temos:

Sabe-se que . Assim:

Como queríamos demonstrar.

* 1. Mostre que se duas variáveis e são independentes, então a covariância entre elas é igual a zero.

**Solução:**

Para que e sejam independentes, devemos mostrar que , uma vez que a independência entre duas variáveis pressupõe .

Portanto, devemos partir da seguinte relação:

Sabendo que:

Quando e são independentes, temos que . Assim, podemos fazer:

Se:

Então, concluímos que:

Ou seja, as variáveis em análise têm covariância igual a zero, portanto, são independentes, como queríamos demonstrar.

* 1. Neste exercício, provamos a condição de normalização (1.48) para a Gaussiana univariada. Para fazer isso, considere a integral:

Que podemos avaliar escrevendo seu quadrado na forma:

Agora, faça a transformação das coordenadas cartesianas para coordenadas polares e, em seguida, substitua Mostre que, realizando as integrações em relação a e e, em seguida, tomando a raiz quadrada de ambos os lados, obtemos:

Por fim, utilize este resultado para demonstrar que a distribuição Gaussiana está normalizada.

**Solução:**

Fazendo a mudança das variáveis do sistema de coordenadas cartesianas para o sistema de coordenadas polares de , temos:

Resolvendo a integral, temos que:

Retornando à variável :

Portanto:

Como queríamos demonstrar.

Por fim, demonstraremos que a distribuição Gaussiana está normalizada, ou seja:

Sabemos que a distribuição normal ou gaussiana ( tem a seguinte forma geral para a sua função de densidade de probabilidade:

Então, podemos fazer:

Usando a transformação , podemos reescrever:

Como queríamos demonstrar.

* 1. Usando uma mudança de variáveis, verifique que a distribuição Gaussiana univariada dada por (1.46) satisfaz (1.49). Em seguida, ao diferenciar ambos os lados da condição de normalização:

em relação a , verifique que a Gaussiana satisfaz (1.50). Por fim, mostre que (1.51) é verdadeira.

**Solução:**

Fazendo , podemos escrever:

Em :

Em :

Já sabemos que pelo exercício 1.7 Portanto:

Assim:

Concluímos, então, que a distribuição Gaussiana univariada dada por (1.46) satisfaz (1.49), sendo esta expressa por:

Em seguida, ao diferenciar ambos os lados da condição de normalização em relação a , podemos verificar que a Eq. (1.50) é satisfeita. Para tanto, devemos demonstrar a segunda parte deste problema partindo da seguinte relação:

Diferenciando ambos os lados de em relação a

Temos que a relação acima se trata da definição de para . Vejamos:

Portanto:

Assim, finalmente, temos:

Por fim, mostraremos que (1.51) é verdadeira. Vejamos:

Sabemos que:

Como queríamos demonstrar.

* 1. Mostre que a moda (ou seja, o valor máximo) da distribuição Gaussiana (1.46) é dado por . Da mesma forma, mostre que a moda da distribuição Gaussiana multivariada (1.52) é dada por μ.

**Solução:**

Sabe-se que a função de densidade de probabilidade (PDF) da distribuição normal univariada é dada por:

Sabe-se, também, que o máximo de uma distribuição é conhecido como sua moda. Portanto, temos que, para encontrar a moda da distribuição Gaussiana, fazer o seguinte:

Chegamos à conclusão de que , ou seja:

Similarmente, faremos o mesmo processo para demonstrar que a moda da distribuição Gaussiana multivariada (1.52) também é dada por μ.

A Eq. é dada por:

Fazendo:

Temos:

A partir da Eq. 81 de [7], podemos afirmar que:

E, também, que:

Portanto:

Chegamos à conclusão de que , ou seja: como queríamos demonstrar.

* 1. Suponha que as duas variáveis x e z sejam estatisticamente independentes. Mostre que a média e a variância da sua soma satisfazem:

**Solução:**

Como x e z são variáveis independentes, podemos fazer:

Pela Teorema de Fubini, temos:

Para o cálculo da , temos que:

Podemos representar da seguinte forma:

Desmembrando a integral:

Expandindo o quadrado, temos:

Desmembrando a integral, temos:

Aplicando o Teorema de Fubini, encontramos:

Sabe-se que , então:

Expandindo termo , chegamos:

* 1. Ao igualar as derivadas da função de log-verossimilhança (1.54) em relação a e a zero, verifique os resultados (1.55) e (1.56).

Derivando em relação a , temos:

Fazendo:

Obtemos:

Obtemos, assim, a relação da Eq. :

Como queríamos demonstrar.

Analogamente, derivando em relação a , temos:

Fazendo

Obtemos, assim, a relação da Eq. para ML:

Como queríamos demonstrar.

* 1. Suponha que a variância de uma Gaussiana seja estimada usando o resultado (1.56), mas com a estimativa de máxima verossimilhança substituída pelo valor real da média. Mostre que este estimador tem a propriedade de que sua expectativa é dada pela verdadeira variância .

Substituindo por conforme o enunciado, temos:

Expandindo o quadrado, temos:

Aplicando as propriedades do somatório, chegamos no seguinte:

Como queríamos demonstrar.

* 1. Considere dois números não negativos e, mostre que, se , então . Use este resultado para mostrar que, se as regiões de decisão de um problema de classificação de duas classes forem escolhidas para minimizar a probabilidade de classificação incorreta, esta probabilidade irá satisfazer

**Solução:**

Extraindo a raiz quadrada de ambos os lados, temos:

Multiplicando ambos os lados pelo termo :

Como queríamos demonstrar.

Para a segunda parte do problema, devemos saber que:

Usando a relação de desigualdade provada acima, temos:

De acordo com [1], para minimizar, devemos fazer com que cada x seja atribuído à classe que tiver o menor valor do integrando em (1.78). Então, podemos fazer:

Como queríamos comprovar.

* 1. Dada uma matriz de perdas com elementos, o risco esperado é minimizado se, para cada , escolhermos a classe que minimiza (1.81). Verifique que, quando a matriz de perdas é dada por = 1 − , em que são os elementos da matriz identidade, isso se reduz ao critério de escolha da classe com maior probabilidade posterior. Qual é a interpretação desta forma de matriz de perdas?

**Solução:**

Para = 1 − temos:

Dessa forma, se para qualquer, temos que o critério para minimizar o risco esperado é escolher a classe que maximiza . Isso significa escolher a classe com maior probabilidade a posteriori.

A interpretação dessa forma da matriz de perdas é que penaliza as classificações incorretas atribuindo uma perda de 1 quando a classe verdadeira é classificada incorretamente (elementos fora da diagonal da matriz de perdas), e uma perda de 0 quando a quando a classe verdadeira é classificada corretamente (elementos diagonais da matriz de perdas), já que, de acordo com [1]: dada uma matriz de perdas cujas entradas denotam a perda de atribuir uma observação à classe quando na realidade está na classe 𝑘. Essa função de perda está alinhada com o objetivo de escolher a classe que possui a máxima probabilidade posterior, pois minimiza o risco esperado.

Tomando com a matriz de perdas, temos que a seria expressa, por exemplo, da seguinte forma:

* 1. Derive o critério de minimização da perda esperada quando existe uma matriz geral de perdas e probabilidades anteriores gerais para as classes.

**Solução:**

A partir de [1], podemos afirmar que dada uma matriz de perda cujas entradas denotam a perda de atribuir uma observação à classe quando na realidade está na classe ) precisamos atribuir uma observação à classe que minimiza a quantidade

Assim, temos que

podemos expressar a quantidade que queremos minimizar em termos de

1.25 Considere a generalização da função de perda quadrada (1.87) para uma única variável alvo para o caso de múltiplas variáveis ​​alvo descritas pelo vetor t dado por

Utilizando o cálculo de variações, mostre que a função para a qual essa perda esperada é minimizada é dada por Mostre que este resultado se reduz a (1.89) para o caso de uma única variável alvo .

**Solução:**

Para minimizar , devemos fazer . Assim, temos:

Desmembrando a integral, podemos escrever:

Como queríamos demonstrar.

1.31 Considere duas variáveis e com distribuição conjunta Mostre que a entropia diferencial desse par de variáveis satisfaz

(1.152)

com igualdade se, e somente se, *x* e *y* são estatisticamente independentes.

**Solução:**

Para demonstrar o que se pede, devemos calcular . Assim, pode ser calculado da seguinte forma:

Em (1.93), é definido como:

Fazendo para , temos que:

Desmembrando a integral, temos:

Assim, temos:

Segundo [1]: é chamada de informação mútua entre as variáveis e ***.*** A partir das propriedades da divergência de Kullback-Leibler, observamos que com igualdade se, e somente se, e são independentes.

1.33 Suponha que a entropia condicional entre duas variáveis aleatórias discretas e seja zero. Mostre que, para todos os valores de x, nos quais , a variável deve ser uma função de . Em outras palavras, para cada x, há apenas um valor de y tal que 0.

**Solução:**

Tratando-se de variáveis aleatórias discretas, podemos expressar a entropia condicional da seguinte forma:

Novamente, podemos calcular a entropia condicional a partir da entropia de uma variável. A equação (1.98) nos fornece o seguinte:

Se é uma função de , podemos obter o valor de assim que observarmos um. Portanto, não obteremos informações adicionais ao observar um dado um já observado.

1.37 Usando a definição (1.111) juntamente com a regra do produto da probabilidade, prove o resultado (1.112).

A equação nos fornece o seguinte:

Usando a definição acima, juntamente com a regra do produto, devemos provar a relação expressa pela Eq. (1.112):

Primeiramente, devemos partir do cálculo de **:**

Desmembrando a integral, temos o seguinte:

1.39 Considere duas variáveis binárias e com a distribuição conjunta dada na Tabela 1.3. Avalie as seguintes quantidades.

Desenhe um diagrama para mostrar a relação entre essas diversas quantidades.

**Solução:**

1. I[x, y]

Por fim, a partir dos resultados acima expostos, podemos chegar nas seguintes relações entre as variáveis:

Graficamente, temos:

Diagrama, Diagrama de Venn

Descrição gerada automaticamente

1.41 Usando as regras de soma e de produto da probabilidade, mostre que a informação mútua satisfaz a relação (1.121).

**Solução:**

Obtemos, assim, a relação da Eq. :

Como queríamos demonstrar.

**Exercícios Extra:**

E1) Considere a função custo/objetivo dada na equação (1.2) em que está definido na equação (1.1). Mostre que (1.2) pode ser escrito como:

Em que:

Determine as dimensões e os elementos (também chamados de entradas) da matriz A. Mostre que o vetor de coeficientes que minimiza esta função objetivo pode ser escrito como

Compare com o exercício 1.1 e note a vantagem de se usar Álgebra Linear para trabalhar com uma notação mais compacta.

**Solução:**

Seja a equação (1.2):

E a equação (1.1):

Podemos mostrar que (1.2) pode ser escrito como:

Para tanto, temos que:

Então:

Portanto, A tem colunas e linhas.

Para minimizar a função , temos:

Então:

Dessa forma:

Multiplicando ambos os lados da igualdade por , temos:

Portanto, temos que o vetor que minimiza a função custo pode ser expresso da seguinte forma:

Como queríamos demonstrar.

E2) Mesma ideia de E1, porém agora considerando a função objetivo dada na equação (1.4) do livro. Escrevê-la de forma matricial. Encontre o vetor de coeficientes ótimos (em fórmula fechada).

**Solução:**

Dada a seguinte função:

Temos que para minimizá-la, devemos derivá-la em relação a Vejamos:

Portanto, temos que o vetor que minimiza a função custo pode ser expresso da seguinte forma:

E3) (Exercício Computacional) Replique o experimento computacional denominado “*Polynomial Curve Fitting*” usado diversas vezes no livro texto (veja páginas 4 e 5 do livro, bem como Apêndice A)

Faça:

1. Replique os resultados da Figura 1.4 e da Figura 1.6 para validar seu código (i.e., ter certeza de que ele está funcionando adequadamente);

**Solução:**

**Figura 1.4**

Gráfico, Gráfico de linhas, Histograma

Descrição gerada automaticamente

**Figura 1.6**

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

(b) Simule uma base de dados que não tenha relevância estatística, isto é, que seja uma amostra que NÃO representa bem o todo (a população). Verifique alguns resultados experimentais para compreender a importância de ter uma amostra relevante. Explique qual a relação entre o caso simulado e casos práticos envolvendo vetores de dimensão elevada.

**Dica**: Para a simulação, ao invés de pegar dados igualmente espaçados no intervalo [0,1], você̂ pode forçar com que seus dados sejam amostrados apenas do semiciclo positivo (ou apenas do negativo) do modelo gerador.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

**Análise:**

Conforme o grau do polinômio aumenta (de 0 para 9), o modelo se torna mais flexível e tenta ajustar-se aos dados de treinamento de forma mais precisa. Para o grau 0, o modelo é uma linha reta, que não consegue capturar a diversidade e a complexidade da função seno. O ajuste é muito simplista. Conforme o grau se eleva para 1 e 3, o modelo começa a se aproximar do padrão senoidal, mas ainda não é capaz de capturar a complexidade completa da função. Percebe-se que, analogamente ao caso do item a do E3 para M=3, a curva vermelha (*fitting*), que é a estimativa da curva verde, tende a se ajustar a senoide. Entretanto, neste caso, não é possível extrair conclusões significativas para as áreas em que não há dados disponíveis (semiciclo positivo). Isso representa um desafio inerente quando se considera um conjunto de dados que é incapaz de representar fielmente a totalidade da população real.

Para o grau 9, o modelo se ajusta perfeitamente aos pontos de treinamento, o que já era esperado pela análise do tem a do E3 para M=9, no entanto, além do problema inerente ao caso de superajuste (*overfitting*)(um modelo superajustado tende a não generalizar bem para novos dados), tem-se a incapacidade de se obter informações da totalidade da população real, em virtude da distribuição desigual dos dados da amostra, que, forçosamente, se concentraram no semiciclo negativo da função seno (x = np.linspace(0.6, 1, sample\_size)).

Conclui-se, portanto, que se a amostra de dados não representa adequadamente a população como um todo, os modelos construídos podem não ser representativos e terão dificuldade em fazer previsões precisas. O exemplo simulado demonstra como uma amostra não representativa (que inclui dados apenas no semiciclo negativo da função seno) pode levar a modelos de ajuste inadequado.

(c) Simule uma base de dados em que 1 dos dados seja *outlier*. O que ocorre com a curva vermelha, estimativa da curva verde (modelo gerador), neste caso?  
**Dica**: Para a simulação, você̂ pode gerar seus dados de treinamento normalmente, igual feito no item (a), e ao final do processo escolher 1 desses dados pra atribuir um novo valor de *target* que seja completamente “maluco” (por exemplo, target = 10).

Interface gráfica do usuário, Gráfico

Descrição gerada automaticamente

**Análise:**

Quando um outlier é introduzido nos dados o que ocorre é que a curva vermelha (*Fitting*) é influenciada significativamente pelo *outlier*, uma vez que o modelo de regressão polinomial tentará ajustar a curva vermelha de forma a minimizar o erro quadrático médio.

No caso concreto, em que um *outlier* com um valor muito maior do que os outros dados é adicionado (y\_train[outlier\_index] = 10), a curva vermelha é "puxada" em direção a esse outlier.

Percebe-se que a curva vermelha não mais se amolda suavemente à curva verde (modelo gerador), em vez disso, a curva vermelha é fortemente distorcida para acomodar o *outlier*. Como consequência, isso resulta em uma estimativa inadequada da função geradora real dos dados, já que o *outlier* exerce uma influência desproporcional no ajuste do modelo.

Em resumo, a introdução de um outlier pode distorcer a estimativa do modelo, fazendo com que a curva vermelha se ajuste inadequadamente aos dados, especialmente se o *outlier* for significativamente diferente dos outros pontos de dados, como foi proposto pelo exercício.

**Observação:**

Os códigos e os gráficos se encontram no seguinte endereço do **Google Colab**:

<https://colab.research.google.com/drive/1ZGF_RQcenMDSbzUJ5kPfOZFnRH6L_rem?authuser=1#scrollTo=QppSgtZvHE7N>

**REFERÊNCIAS:**

[1] C. Bishop, "Reconhecimento de padrões e aprendizado de máquina", 1ª ed. Springer, 2011.

[2] K. Pretz, "Stop Calling Everything AI, Machine-Learning Pioneer Says", IEEE Spectrum, disponível em: https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says# toggle-gdpr , Acesso em: 01/11/2023.

[3] "O que é probabilidade?". Disponível em: https://web.ma.utexas.edu/users/mks/statmistakes/probability.html. Acesso em: 03/11/2023.

[4] DJM Soares, TEA Soares, PC Emiliano, "Uma aplicação do teorema do limite central", Revista Brasileira de Desenvolvimento. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/ article/view/5620/5083. Acesso em: 03/11/2023.

[5] J. Lebl, "Análise Básica II Introdução à Análise Real, Volume II por Jiří Lebl. Royden, HL (1968). Real Analysis 2. ed. [Sl]: Macmillan". Disponível em: https:// www .jirka.org/ra/realanal2.pdf . Acesso em: 04/11/2023.

[6] "Mudança de Variáveis ​​em Integrais Duplas". Disponível em: https://cesad.ufs.br/ORBI/public/uploadCatalago/11351916022012C%C3%A1lculo\_III\_aula\_2.pdf . Acesso em: 04/11/2023.

[7] K. B. Petersen and M. S. Pedersen, "The Matrix Cookbook," Nov. 15, 2012. [Online]. Disponível em: http://matrixcookbook.com. Acesso em: 03/11/2023.