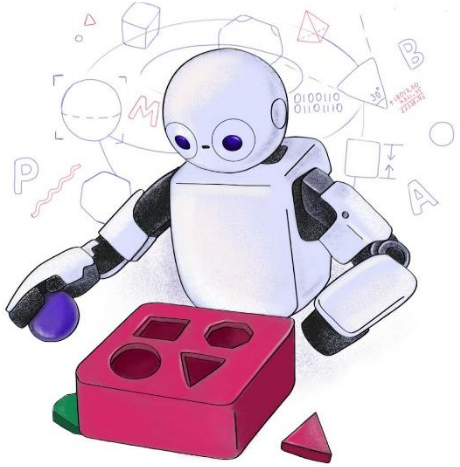
T319 - Introdução ao Aprendizado de Máquina: *Regressão Linear (Parte VI)*



Inatel

Felipe Augusto Pereira de Figueiredo
felipe.figueiredo@inatel.br

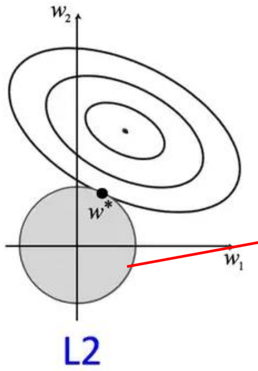
Recapitulando

- Anteriormente, vimos como escolher o melhor modelo de regressão utilizando as técnicas de ***validação cruzada***: holdout, k-Fold e leave-P-out.
- Escolhemos sempre o modelo menos complexo, mas que generaliza bem.
- Ou seja, escolhemos o modelo que apresenta valores baixos para ambos os erros, de treinamento e de validação.
- Uma abordagem alternativa é ***minimizar conjuntamente*** o erro e a complexidade da ***função hipótese***.
- Como veremos, esta abordagem combina erro e complexidade ***em uma única função de erro***, possibilitando que encontremos a melhor ***função hipótese*** de uma só vez.
- Portanto, hoje, veremos as seguintes abordagens para se encontrar o melhor modelo de regressão:
  - **Regularização**: penaliza ***funções hipótese*** muito complexas, ou seja, muito flexíveis.
  - **Early-stop**: encerra o treinamento de ***algoritmos iterativos*** quando o erro de validação for o menor possível.

Regularização: penalizando a complexidade dos modelos

- Grandes magnitudes dos pesos de um modelo são um claro sinal de um modelo mais complexo que sobreajustou os dados de treinamento.
- ***Regularização***: deixar o modelo menos flexível (ou seja, complexo).
- A ideia por trás da ***regularização*** é penalizar, explicitamente, ***funções hipótese*** complexas.
- Técnicas de ***regularização*** reduzem o risco de ***sobreajuste*** do modelo ao conjunto de treinamento, aumentando sua capacidade de ***generalização***.
  - Quanto menos graus de liberdade o modelo tiver, mais difícil será para ele se ***sobreajustar*** aos dados de treinamento.
- O risco de ***sobreajuste*** é reduzido incorporando-se ***penalizações*** proporcionais a alguma ***norma*** do ***vetor de pesos*** ao processo de treinamento.
- As principais técnicas de ***regularização*** são: *Rigde*, LASSO e *elastic-net*.
- A ***regularização*** força o algoritmo de aprendizado não apenas a se ajustar aos dados, mas também a manter os pesos do modelo os menores possíveis.

Ridge Regression



Região de factibilidade: possíveis valores que os pesos podem assumir. O raio do círculo é dado pelo fator de regularização, λ.

- Ao invés de minimizarmos apenas o ***erro quadrático médio***, como fizemos antes, introduzimos um ***termo de penalização*** proporcional à ***norma Euclidiana*** (ou seja, a ***norma L2***) do vetor de pesos:

Início em 1 e não em 0.

$$\min_{\boldsymbol{a} \in \mathbb{R}} \left(\|\boldsymbol{y} - \boldsymbol{\Phi a}\|^2 + \boxed{\lambda \|\boldsymbol{a}\|_2^2} \right) \text{ com } \|\boldsymbol{a}\|_2^2 = \sum_{i=1}^{K} {a_i}^2$$

onde $\lambda \geq 0$ é o ***fator de regularização***, $\boldsymbol{\Phi}$ é a matriz de atributos e $\boldsymbol{a}$ é o vetor de pesos.

- Podemos re-escrever o ***problema de regularização*** como um ***problema de otimização com restrições*** da seguinte forma

$$\min_{\boldsymbol{a} \in \mathbb{R}} \|\boldsymbol{y} - \boldsymbol{\Phi a}\|^2$$
$$\text{s.}\,a.\ \|\boldsymbol{a}\|_2^2 \leq c,$$

- Se c diminui, $\|\boldsymbol{a}\|_2^2$ também diminui até que se $c \rightarrow 0$, então $a_i \rightarrow 0$.
- Se c aumenta, $\|\boldsymbol{a}\|_2^2$ pode assumir valores maiores até que se $c \rightarrow \infty$, então a_i pode assumir qualquer valor.
- c define o tamanho da região de factibilidade.

onde c ***restringe a magnitude dos pesos*** e é inversamente proporcional à λ.

- Portanto, λ altera a complexidade (ou seja, a flexibilidade) da função hipótese.
- **OBS.**: o peso a_0 não é considerado no cálculo da ***norma L2***, pois a ***complexidade*** se deve à ordem do modelo e a_0 apenas dita o deslocamento em relação ao eixo y.

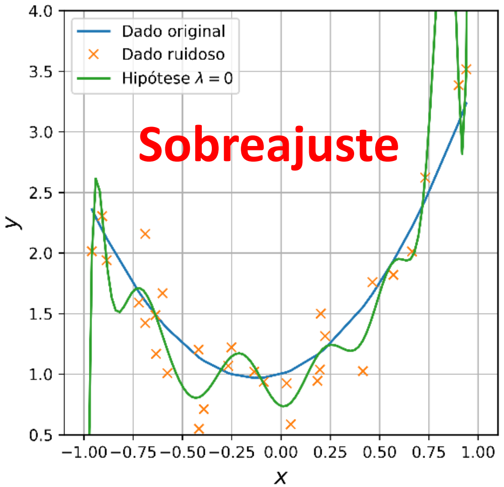
Ridge Regression

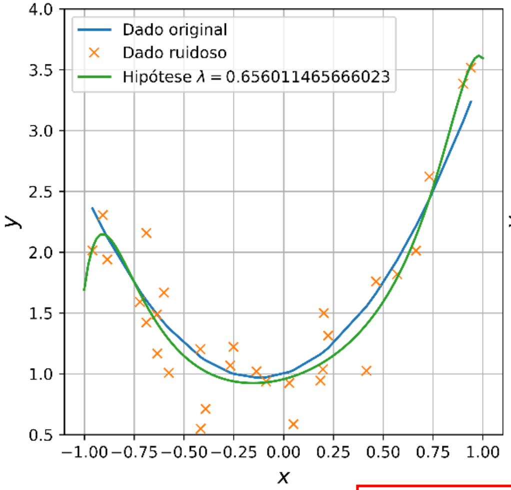
- A ***equação de erro regularizado***, $\|\boldsymbol{y} - \boldsymbol{\Phi a}\|^2 + \lambda\|\boldsymbol{a}\|_2^2$, continua sendo quadrática com relação aos pesos, e portanto, a superfície de erro continua sendo convexa.
- Desta forma, encontramos uma solução de forma fechada seguindo o mesmo procedimento que usamos para encontrar a ***equação normal***:

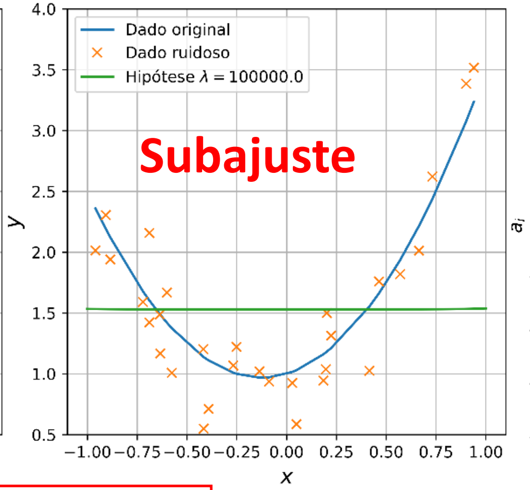
$$\boldsymbol{a} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} \boxed{+ \lambda\boldsymbol{I}'})^{-1}\,\boldsymbol{\Phi}^T\boldsymbol{y}, \text{ onde } \boldsymbol{I}' = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

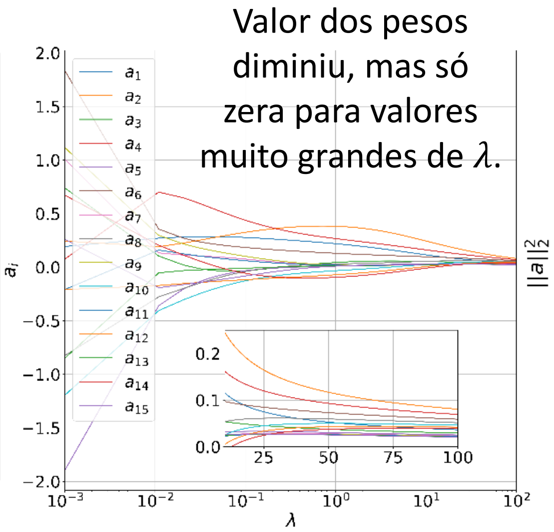
  - **OBS.1**: mesmo que a matriz $\boldsymbol{\Phi}$ não possua ***posto completo*** (i.e., matriz singular), a inversa na equação acima sempre existirá por conta da adição do ***termo de regularização*** à diagonal principal da matriz quadrada $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$.
  - **OBS.2**: como a ***norma L2*** é diferenciável, os problemas de aprendizagem usando a regularização de Ridge também podem ser resolvidos iterativamente através do ***algoritmo do gradiente descendente***.
  - **OBS.3**: o ***termo de regularização*** deve ser adicionado apenas à função de erro durante o treinamento. Depois que o modelo é treinado, a avaliação do seu desempenho não utiliza a regularização.

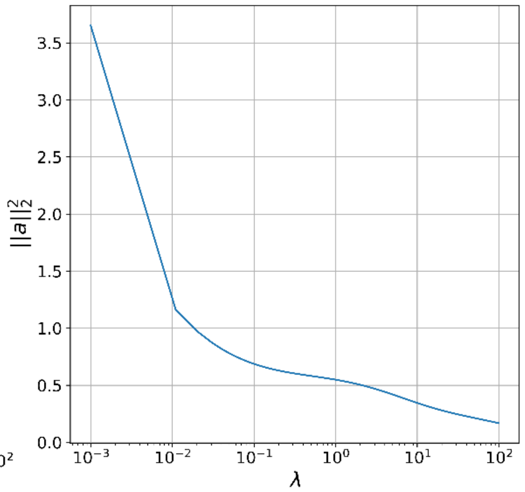
Ridge Regression: Exemplo







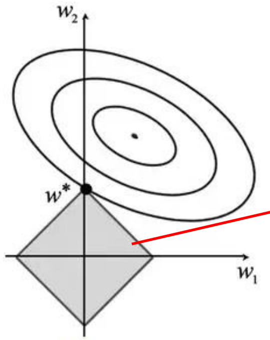




- Função observável: $y_{\text{noisy}} = \boxed{1 + 0.5x + 2x^2} + w$, onde $x \sim U(-1,1)$ e $w \sim N(0, 0.09)$.
- ***Função hipótese polinomial*** de ordem 15 treinada com 30 amostras geradas a partir de y_{noisy}.
- Com $\lambda = 0$, regressão Ridge se torna uma regressão polinomial sem regularização e sobreajusta.
- Conforme λ aumenta, o modelo se "*contorce*" menos e passa a se ajustar à função verdadeira.
- Se λ continuar aumentando, todos os pesos acabarão muito próximos de zero e o resultado será uma reta que passa pela ***média dos dados de treinamento***.
- **O aumento de λ leva a hipóteses menos complexas.** Isso reduz a variância do modelo, mas aumenta seu bias. Ou seja, ele tende a ***subajustar***.
- Conforme λ aumenta, os pesos e a norma L2 do vetor de pesos diminuem (figuras 4 e 5).
- Devemos utilizamos técnicas de ***validação cruzada*** para encontrar o valor ideal de λ.

[Exemplo: ridge_regression.ipynb](ridge_regression.ipynb)

LASSO Regression



L1

Região de factibilidade: possíveis valores que os pesos podem assumir. A área do quadrado é dada pelo fator de regularização.

- A ***regressão LASSO*** (*Least Absolute Shrinkage and Selection Operator*) adiciona à função de erro um ***termo de penalização*** proporcional à **norma L1** do vetor de pesos.

$$\min_{\boldsymbol{a}\in\mathbb{R}}(\|\boldsymbol{y} - \boldsymbol{\Phi a}\|^2 + \boxed{\lambda\|\boldsymbol{a}\|_1}),$$

onde $\|\boldsymbol{a}\|_1 = \sum_{i=1}^{K}|a_i|$ e $\lambda \geq 0$ é o ***fator de regularização***.

- Podemos re-escrever o ***problema de regularização*** acima como um ***problema de otimização*** com restrições da seguinte forma

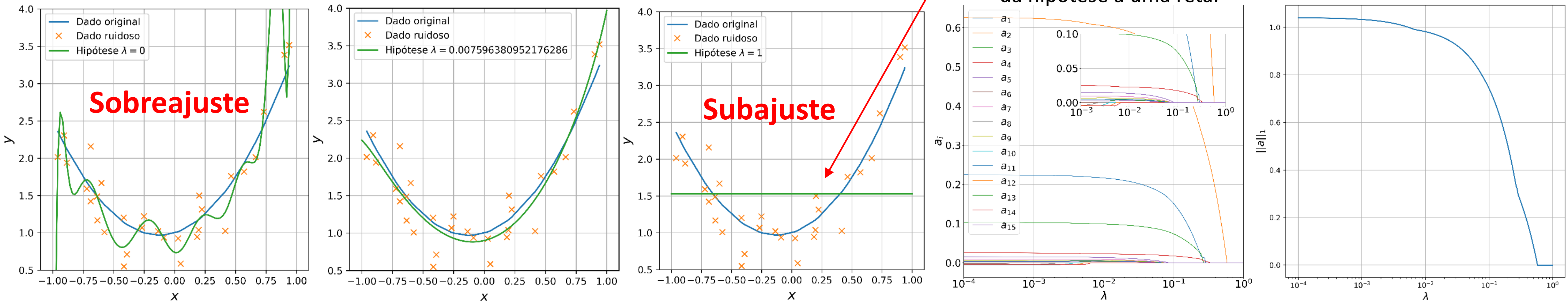
$$\min_{\boldsymbol{a}\in\mathbb{R}}\|\boldsymbol{y} - \boldsymbol{\Phi a}\|^2$$
$$s.a. \|\boldsymbol{a}\|_1 \leq c,$$

onde c restringe a magnitude dos pesos e é inversamente proporcional à λ.

OBS.: a_0 também não faz parte do cálculo da norma.

LASSO Regression

Valor dos pesos se torna igual a zero, restringindo a flexibilidade da hipótese a uma reta.

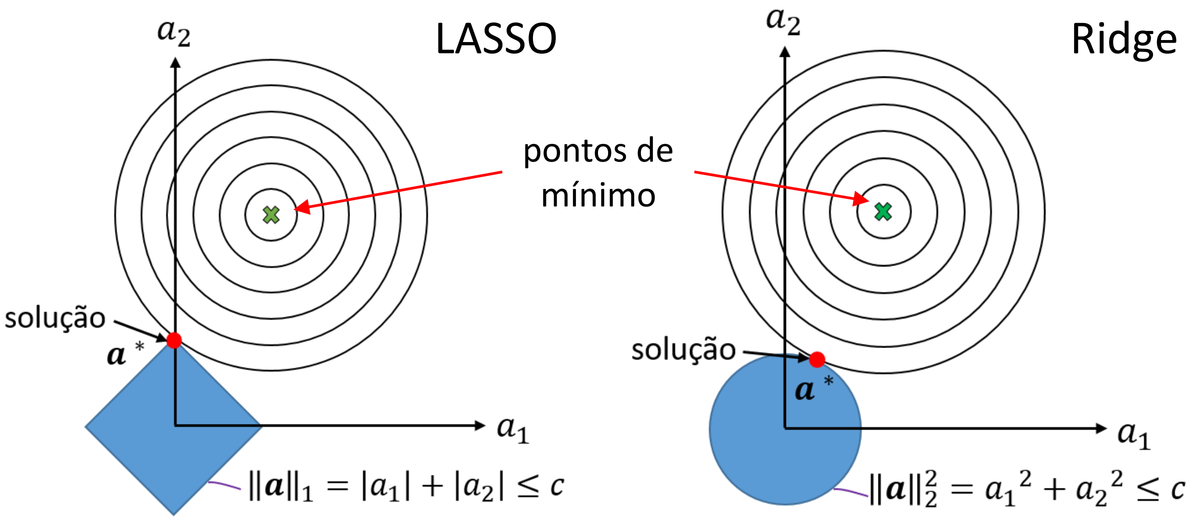


- Mesmas funções observável e hipótese do exemplo anterior.
- Valores pequenos de λ fazem o regressor LASSO se comportar com um regressor tradicional e valores muito grandes fazem os pesos serem anulados.
- A regularização com ***norma L1*** tem como vantagem a produção de ***modelos esparsos***.
  - Ou seja, vários elementos do vetor de pesos acabam sendo ***anulados***, indicando que os atributos correspondentes são irrelevantes para o processo de regressão.
- Isso sugere a ocorrência implícita de um processo de ***seleção automática de atributos*** e leva a ***modelos*** mais ***regulares***, ou seja, ***menos complexos***.
- **Desvantagem**: como a ***norma L1*** não possui derivada no ponto $a_i = 0, \forall i$, o problema da minimização não possui solução em forma fechada, mas pode ser implementada com o GD.
- Utiliza-se técnicas de validação cruzada para encontrar o valor ideal de λ.

Exemplo: lasso_regression.ipynb

Por que LASSO produz modelos esparsos?

- O quadrado azul representa o conjunto de pontos $\boldsymbol{a}$ no espaço de pesos bidimensional que tenham norma L1 menor do que c.
- A solução deve estar dentro do quadrado, o mais próximo do mínimo.



- O círculo azul representa o conjunto de pontos $\boldsymbol{a}$ no espaço de pesos bidimensional que tenham norma L2 menor do que c.
- A solução deve estar dentro do círculo, o mais próximo do mínimo.

- A figura mostra as ***curvas de nível*** da ***função de erro*** de um problema de regressão linear com dois pesos (a_1 e a_2) e as regiões do ***espaço de hipóteses*** onde as restrições L1 (esquerda) e L2 (direita) são válidas.
- A solução para ambos os métodos corresponde ao ponto, dentro da ***região de factibilidade*** (área em azul), mais próximo do ponto de mínimo da função de erro.
- É fácil ver que para uma posição arbitrária do mínimo, será comum que um ***canto*** (ou ponta) do quadrado seja o ponto mais próximo do ponto de mínimo da função de erro.
- Os ***cantos*** na ***região de factibilidade*** da restrição L1 aumentam as chances de alguns pesos assumirem o valor zero.
- E claro, os ***cantos*** são os pontos que possuem valor igual a 0 em alguma das dimensões (i.e., pesos).

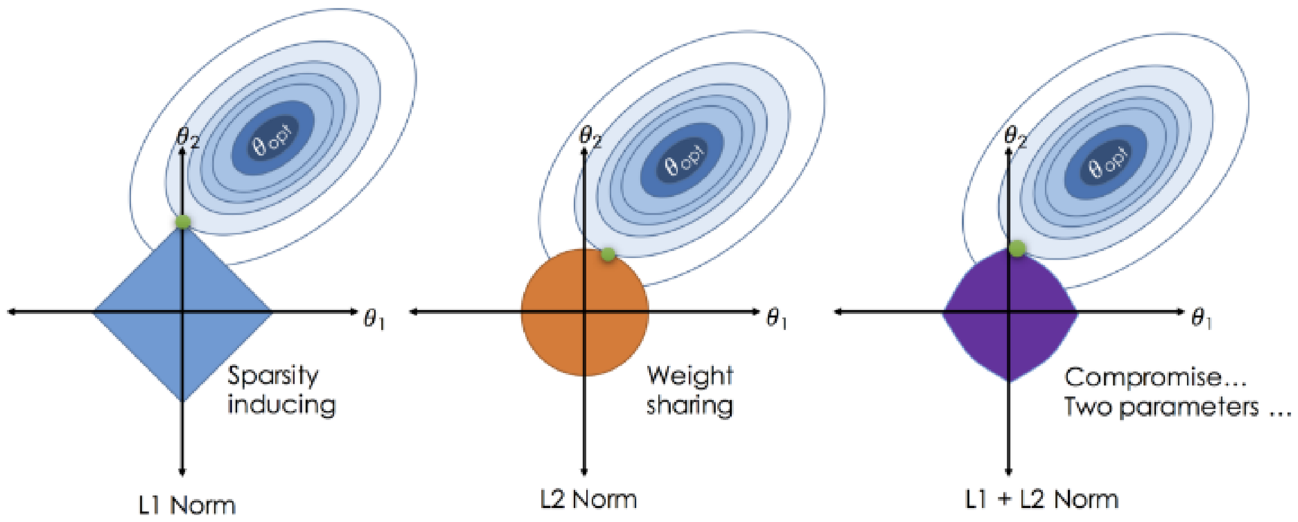
Elastic-net

- *Elastic-net* é uma solução intermediária entre as regressões Ridge e LASSO.
- É uma ***combinação linear*** entre as penalizações baseadas nas normas L1 e L2 do vetor de pesos.

$$\min_{\boldsymbol{a}\in\mathbb{R}}(||\boldsymbol{y}\ -\boldsymbol{\Phi a}||^2+\lambda\,[\kappa||\boldsymbol{a}||_1+(1-\kappa)||\boldsymbol{a}||_2^2]),$$

onde $\kappa\ \epsilon\ [0,1]$ é o termo de mistura ou parâmetro de elasticidade entre as duas normas.

- Quando κ = 0, a *Elastic-net* é equivalente a regressão Ridge e quando κ = 1, ela é equivalente a regressão LASSO.
- A seleção dos hiperparâmetros κ e λ pode ser feita por meio de ***validação cruzada***. Isso também se aplica ao dois outros métodos anteriores.



O hiperparâmetro κ dita a relação de compromisso entre as duas regularizações.

[Exemplo: elastic_net_regression.ipynb]

Quando utilizar regressão LASSO, Ridge ou Elastic-Net?

- **Regressão Ridge**: um bom começo. No entanto, se você suspeitar que apenas alguns atributos são realmente úteis, você deve preferir LASSO ou *Elastic-Net*.
- **Regressão LASSO**: boa para ***seleção automática de atributos***. No entanto, se o número de atributos, K, for maior que o número de exemplos de treinamento, N, ou quando houverem atributos fortemente correlacionados, deve-se usar a regressão *Elastic-Net*.
- **Elastic-Net**: é mais versátil que as anteriores, pois o parâmetro de elasticidade κ é ajustável. Uma proporção de 50% entre as penalizações L1 e L2 é uma boa escolha inicial para esse parâmetro.

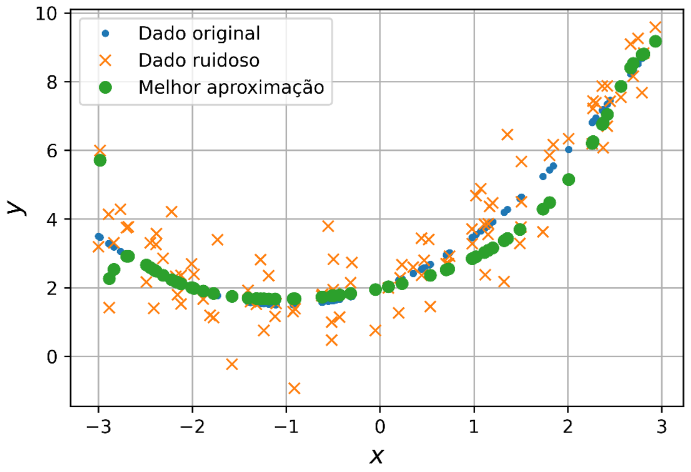
Early-stop: Parada antecipada

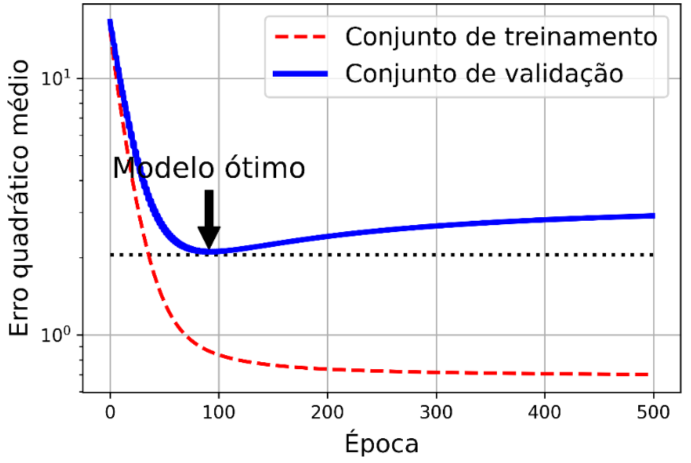
- O algoritmo do ***gradiente descendente*** tende a aprender modelos cada vez mais ***complexos*** à medida que o número de épocas aumenta.
- Ou seja, ele se ***sobreajusta*** ao conjunto de treinamento ao longo do tempo.
- Uma forma de se ***regularizar*** algoritmos de ***aprendizado iterativo***, como o ***gradiente descendente***, é interromper seu treinamento assim que o ***erro de validação*** comece a crescer sistematicamente.
- Essa abordagem é chamada de ***early-stop*** e pode ser vista como uma ***regularização temporal***.
- Assim como as outras abordagens, ela tem o objetivo de evitar o ***sobreajuste*** de um modelo.
- Ao se regularizar no ***tempo***, a complexidade do modelo pode ser controlada, melhorando sua ***generalização***.
- Mas como saber quando interromper o treinamento? Ou seja, qual é o critério de parada?

Exemplo: Early-stop

$$y_{\text{noisy}} = \boxed{2 + x + 0.5x^2} + w,$$
onde $x \sim U(-3{,}3)$ e $w \sim N(0{,}1)$

- Existem duas estratégias para se definir o critério de parada:
  - Interromper o treinamento quando o erro de validação aumenta por ***P*** (paciência) épocas sucessivas.
    - **Problema**: como o erro de validação pode oscilar bastante (e.g., SGD), nem sempre é fácil desenvolver detectores automáticos de mínimos e encerrar o treinamento.
  - Permitir que o treinamento prossiga por um determinado número de épocas, mas sempre armazenando os pesos associados ao ***menor erro de validação***.
- A figura mostra um modelo de regressão polinomial com grau igual a 90 sendo treinado usando o ***gradiente descendente estocástico*** e apenas 100 amostras de treinamento.
- À medida que as épocas passam, o algoritmo aprende e seu erro quadrático médio no conjunto de treinamento diminui, juntamente com o erro no conjunto de validação.
- No entanto, após algumas épocas, o erro de validação para de diminuir e começa a crescer.
- Isso indica que o modelo começou a ***sobreajustar*** aos dados de treinamento.





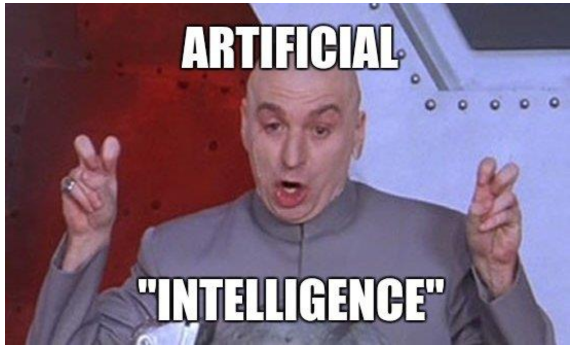
[Exemplo: early_stopv2.ipynb]()

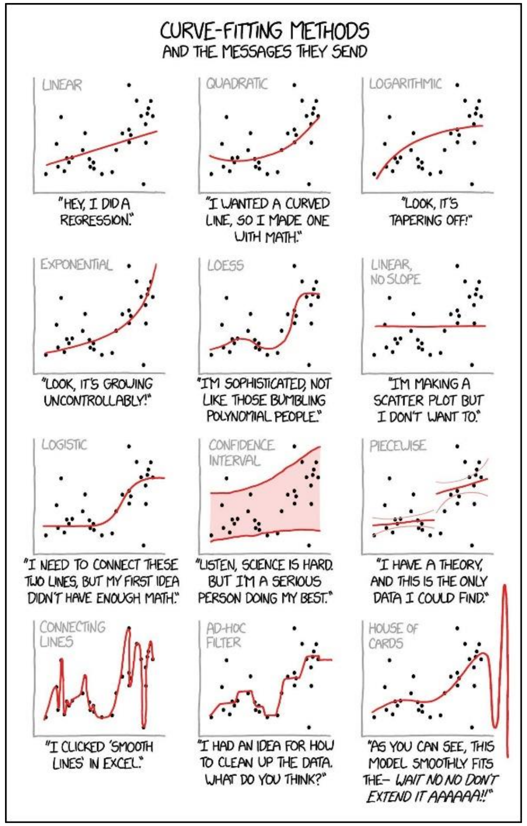
Tarefas

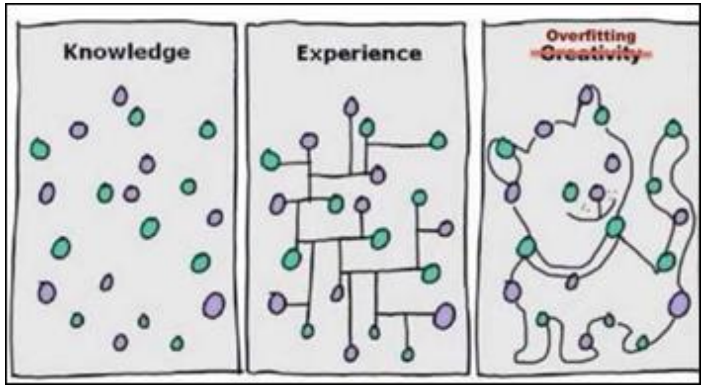
- **Quiz**: “*T319 - Quiz - Regressão: Parte VI*” que se encontra no MS Teams.
- **Exercício Prático**: Projeto Prático.
  - **Projeto pode ser feito em grupo de no máximo 3 alunos.**
  - **Atentem-se ao prazo de entrega definido na tarefa do MS Teams (11/12/2022).**
  - **Entregas fora do prazo não serão aceitas.**
  - **Leiam os enunciados atentamente.**

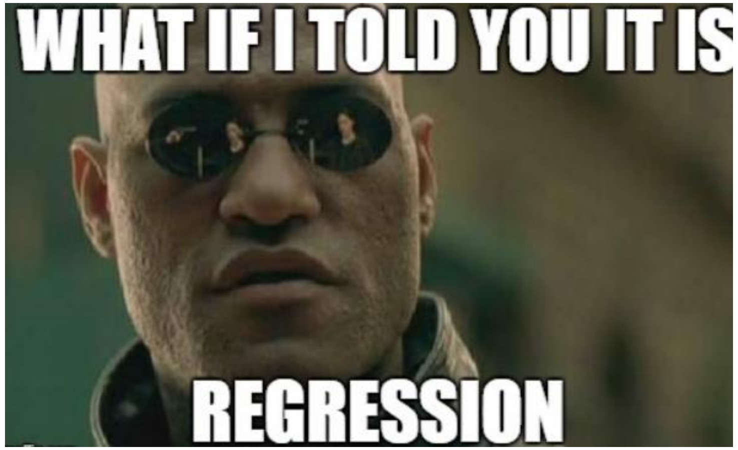
Obrigado!


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


ARTIFICIAL
"INTELLIGENCE"


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"


Knowledge
Experience
Overfitting
~~Creativity~~


WHAT IF I TOLD YOU IT IS
REGRESSION


WHOA
I KNOW LINEAR REGRESSION


LETS LEARN
REGRESSION
memegenerator.net