INSTITUTO FEDERAL
Ceará

Programa de Pós-Graduação
em Ciência da Computação

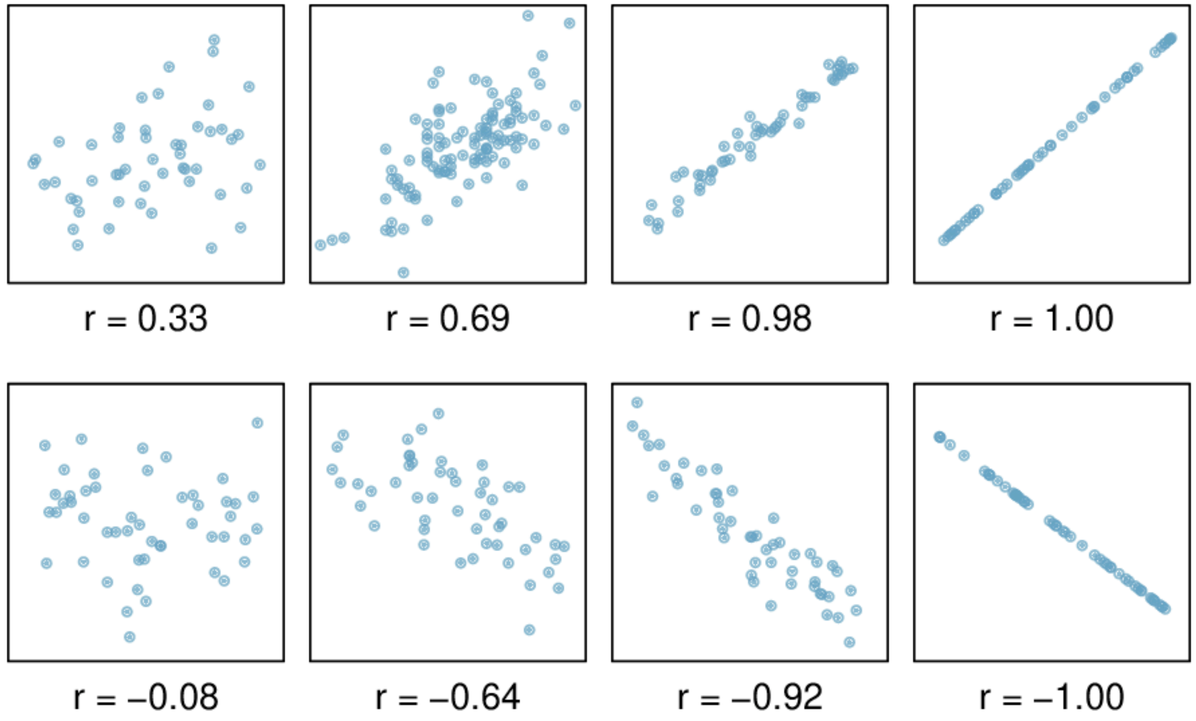
06 PROBLEMINHAS COMUNS

APRENDIZAGEM PROFUNDA

PPGCC – 2023.1

Prof. Saulo Oliveira <saulo.oliveira@ifce.edu.br>

CORRELAÇÃO LINEAR



+1

- Correlação linear positiva perfeita;
- Correlação linear positiva forte;
- Correlação linear positiva fraca;

0 - Ausência de correlação linear;

- Correlação linear negativa fraca;
- Correlação linear negativa forte;
- Correlação linear negativa perfeita;

-1

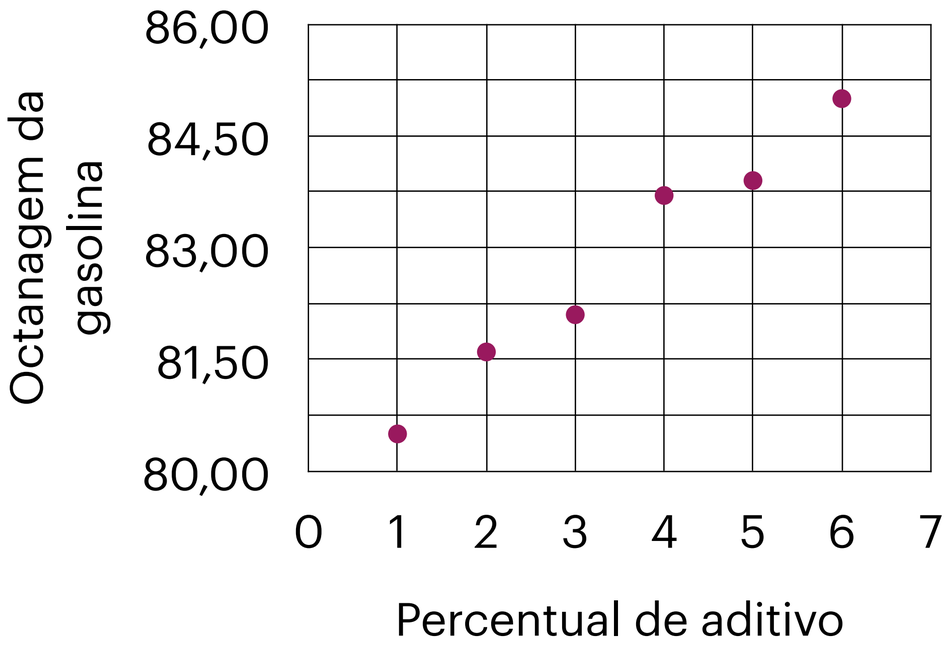
https://spot.pcc.edu/math/ahss/ed2/lineFittingResidualsCorrelation.html

PROBLEMINHAS

EXPERIMENTO 01

Considere um experimento em que se analisa a octanagem da gasolina em função da adição de um aditivo. Para isto, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados dos ensaios seguem na Tabela abaixo. Há como descrever, bem, uma relação linear entre estas duas variáveis?

Percentual do aditivo	Octanagem da gasolina
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0



EXPERIMENTO 01

	Percentual do aditivo	Octanagem da gasolina	x^2	x * y
	1	80,5	1	80,5
	2	81,6	4	163,2
	3	82,1	9	246,3
	4	83,7	16	334,8
	5	83,9	25	419,5
	6	85,0	36	510
Σ	**21**	**496,8**	**91**	**1754,3**

$$\alpha = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}.$$

$$\alpha = \frac{6(1754{,}3) - (\sum x_i 21)(496{,}8)}{6(91) - (21)^2}.$$

$$\alpha = 0{,}886.$$

$$\beta = \frac{\sum y_i - \alpha \sum x_i}{n}.$$

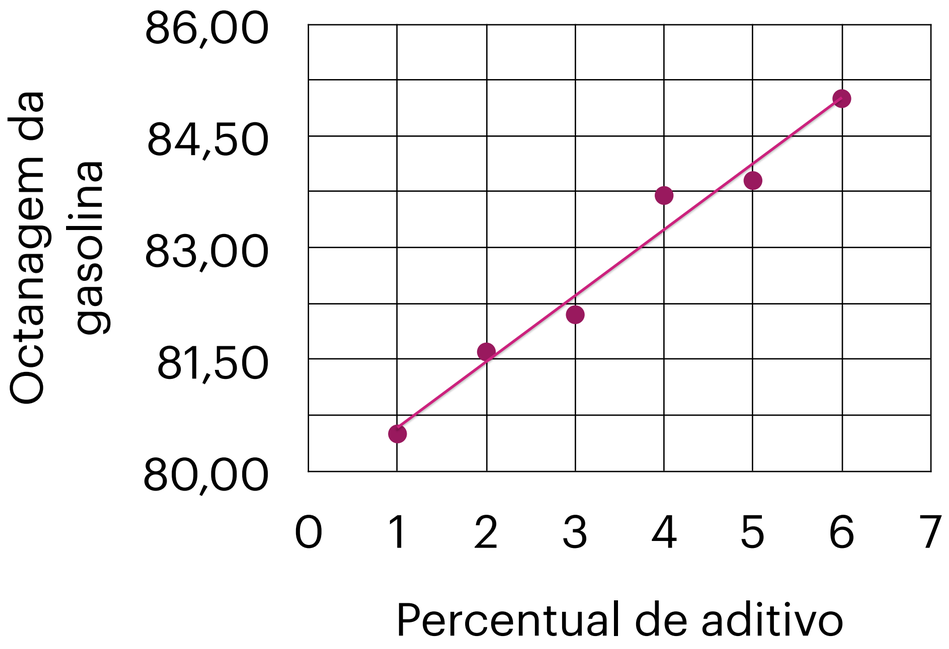
$$\beta = \frac{496{,}8 - (0{,}886)(21)}{6}.$$

$$\beta = 79{,}70.$$

EXPERIMENTO 01

Considere um experimento em que se analisa a octanagem da gasolina em função da adição de um aditivo. Para isto, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados dos ensaios seguem na Tabela abaixo. Há como descrever, bem, uma relação linear entre estas duas variáveis?

Percentual do aditivo	Octanagem da gasolina	Predição
1	80,5	80,59
2	81,6	81,47
3	82,1	82,36
4	83,7	83,24
5	83,9	84,13
6	85,0	85,02



PROPRIEDADES E SUPOSIÇÕES

PROPRIEDADES & SUPOSIÇÕES

Ao se adotar essa abordagem de estimação dos parâmetros são feitas as seguintes suposições:

- Que há de verdade um relacionamento linear entre as variáveis;
- A variável independente é oriunda de um processo não-estocástico e é obtida sem nenhum tipo de erro;
- Os erros oriundos do modelo são estatisticamente independentes;
- **Os erros seguem uma distribuição normal com média zero e desvio padrão constante.**

Erros comuns:

- Não verificar a linearidade entre as variáveis;
- **Confundir correlação com causalidade;**
- Confundir o R com R^2;

CORRELAÇÕES ESPÚRIAS

The NEW ENGLAND JOURNAL of MEDICINE

COFFEE AND CANCER OF THE PANCREAS

BRIAN MACMAHON, M.D., STELLA YEN, M.D., DIMITRIOS TRICHOPOULOS, M.D., KENNETH WARREN, M.D., AND GEORGE NARDI, M.D.

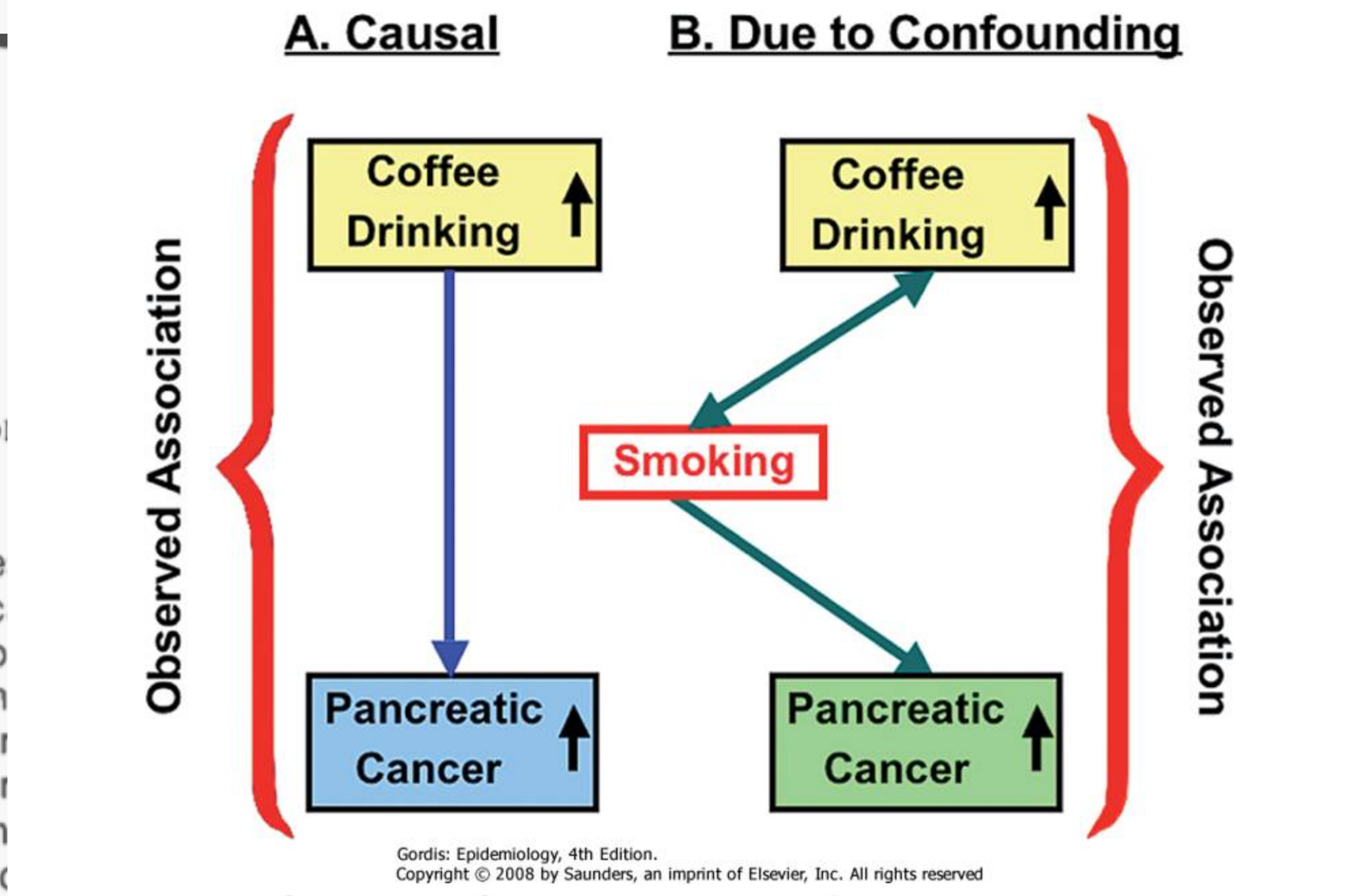
Abstract We questioned 369 patients with histologically proved cancer of the pancreas and 644 control patients about their use of tobacco, alcohol, tea, and coffee. There was a weak positive association between pancreatic cancer and cigarette smoking, but we found no association with use of cigars, pipe tobacco, alcoholic beverages, or tea. A strong association between coffee consumption and pancreatic cancer was evident in both sexes. The association was not affected by controlling for cigarette use. For the sexes combined, there was a significant dose-response relation ($P \sim 0.001$); after adjustment for cigarette smoking, the relative risk associated with drinking up to two cups of coffee per day was 1.8 (95 per cent confidence limits, 1.0 to 3.0), and that with three or more cups per day was 2.7 (1.6 to 4.7). This association should be evaluated with other data; if it reflects a causal relation between coffee drinking and pancreatic cancer, coffee use might account for a substantial proportion of the cases of this disease in the United States. (N Engl J Med. 1981; 304:630-3.)

The NEW ENGLAND JOURNAL of MEDICINE

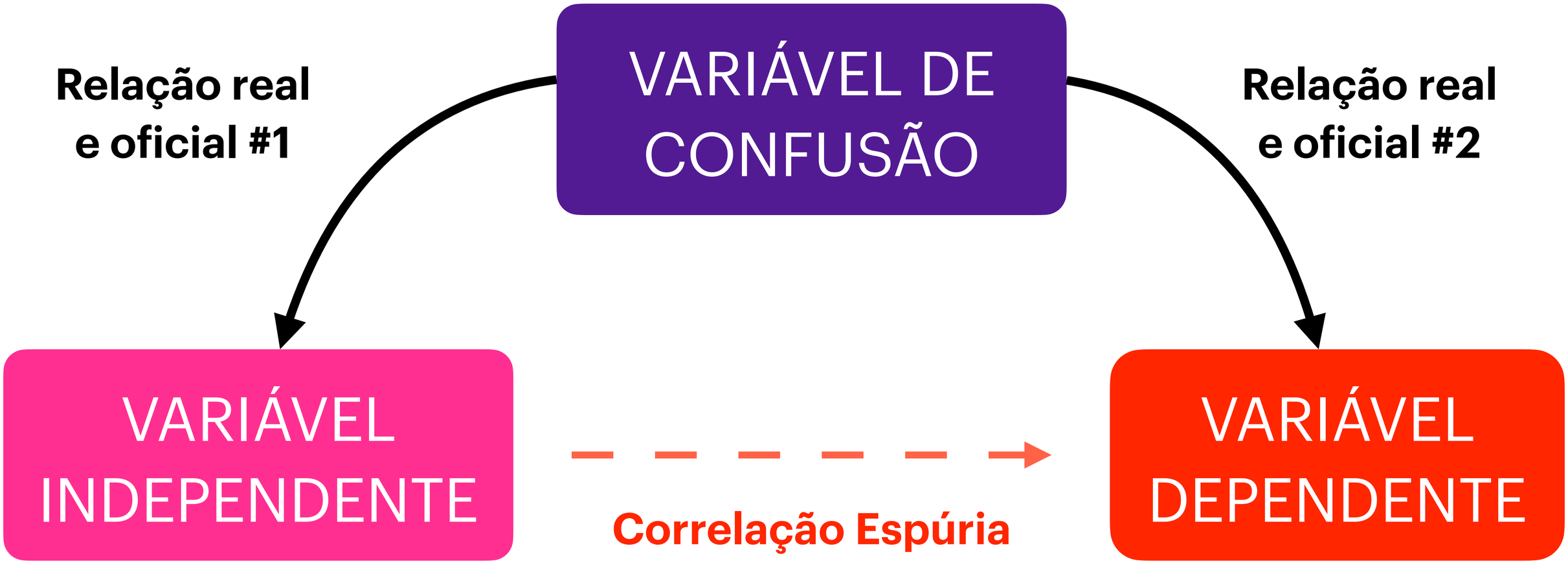
March 12, 1981

BRIAN MACMAHO[illegible] TH WARREN, M.D.,

Abstract We que[illegible] adjustment for cig-
logically proved c[illegible] isk associated with
control patients ab[illegible] fee per day was 1.8
tea, and coffee. Th[illegible] 1.0 to 3.0), and that
tion between pancr[illegible] was 2.7 (1.6 to 4.7).
ing, but we found [illegible] ated with other data;
pipe tobacco, alcoh[illegible] ween coffee drinking
sociation between [illegible] se might account for
ic cancer was evident in both sexes. The association was not affected by controlling for cigarette use. For the sexes combined, there was a significant dose-re- a substantial proportion of the cases of this disease in the United States. (N Engl J Med. 1981; 304:630-3.)



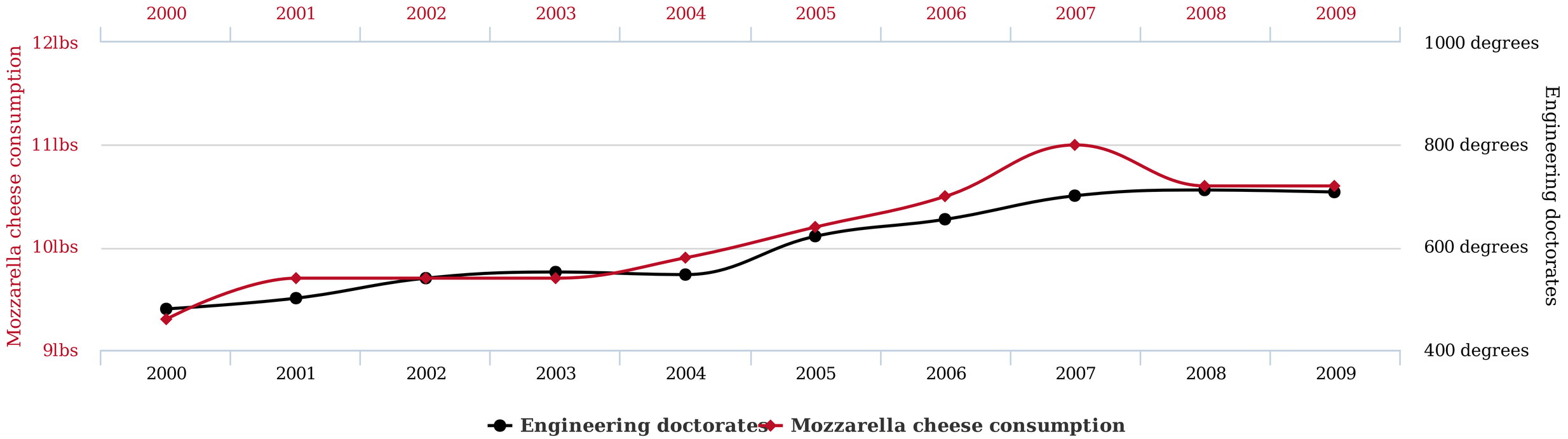
CAINDO EM GOLPES



Per capita consumption of mozzarella cheese

correlates with

Civil engineering doctorates awarded



tylervigen.com

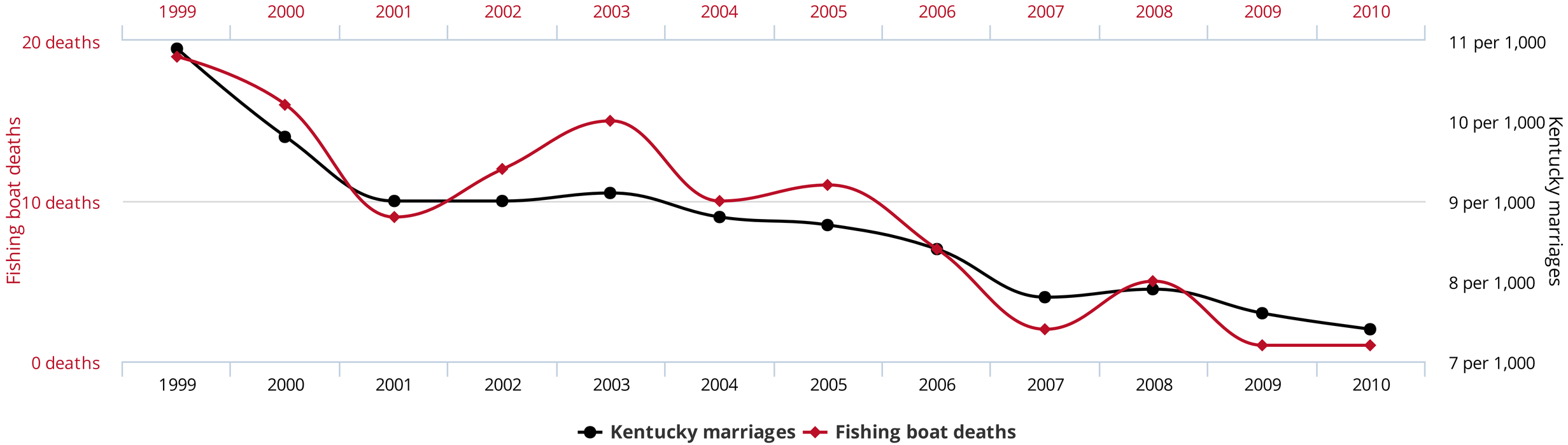
$R^2 = 99{,}79\,\%$

http://tylervigen.com/spurious-correlations

People who drowned after falling out of a fishing boat

correlates with

Marriage rate in Kentucky



tylervigen.com

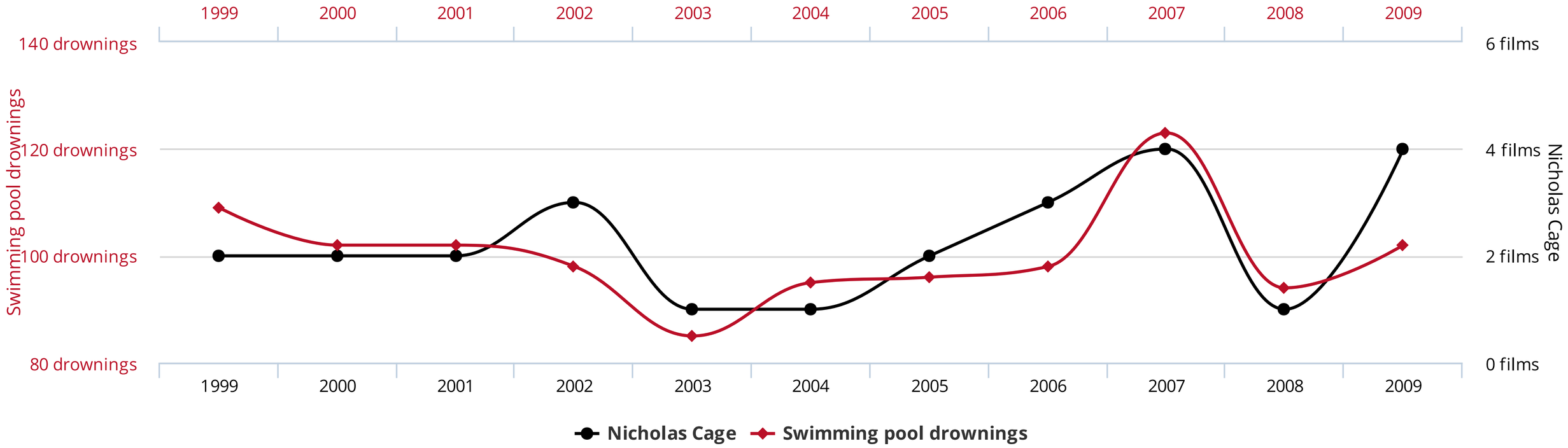
$R^2 = 95{,}24\,\%$

http://tylervigen.com/spurious-correlations

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



tylervigen.com

$R^2 = 66{,}60\,\%$

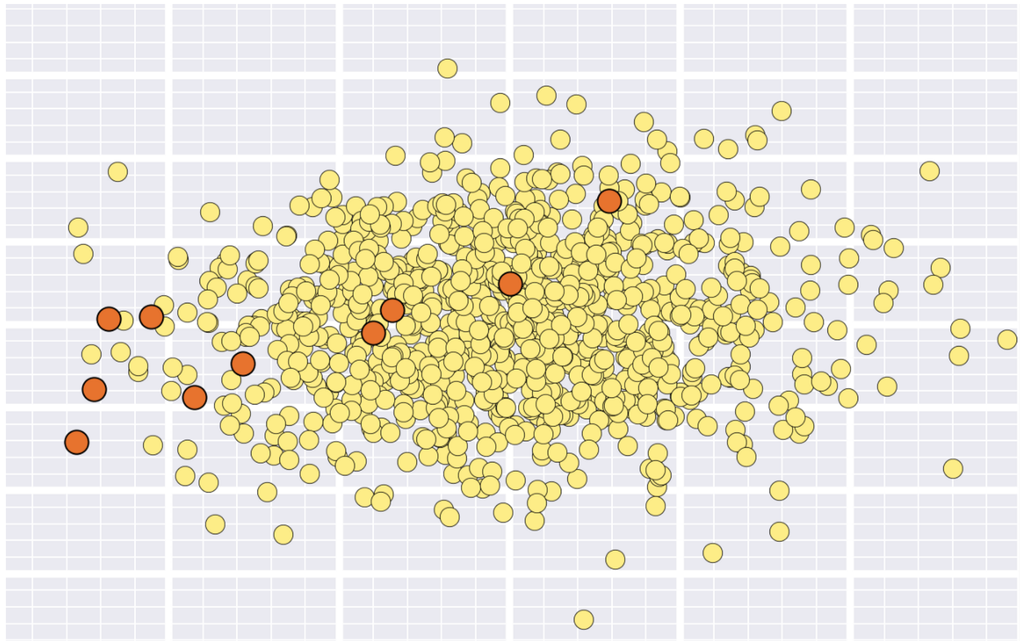
http://tylervigen.com/spurious-correlations

É FÁCIL ENCONTRAR PADRÕES
INTERESSANTES ONDE NÃO EXISTEM!

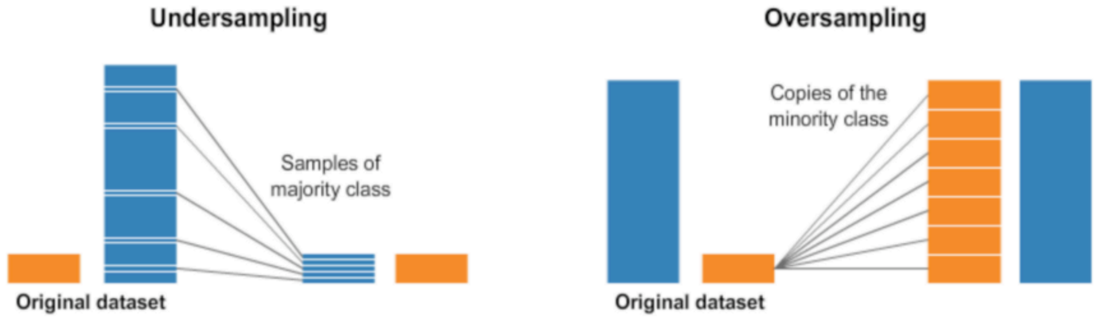
NEM TUDO SÃO ROSAS 🌷,
NEM EXISTE BALA DE PRATA 🔫🎯

QUANDO O PROBLEMA VEM DOS DADOS

DADOS DESBALANCEADOS

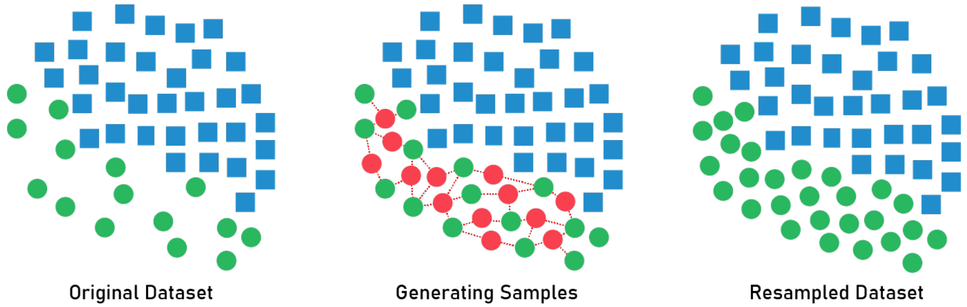


https://towardsdatascience.com/the-main-issue-with-identifying-financial-fraud-using-machine-learning-and-how-to-address-it-3b1bf8fa1e0c



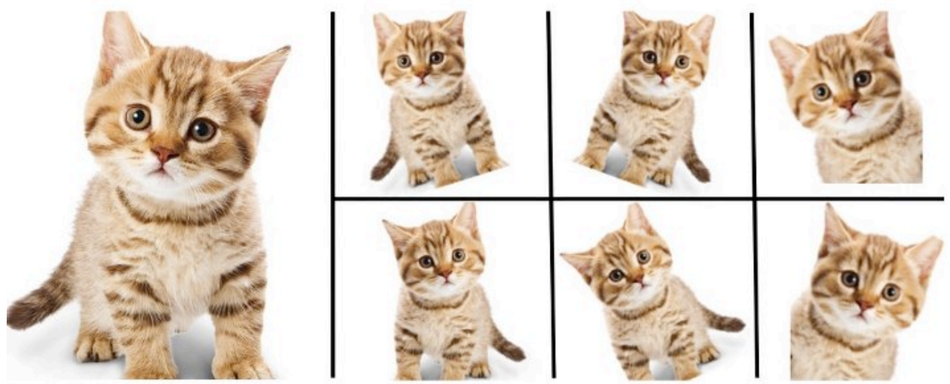
https://www.kdnuggets.com/2019/05/fix-unbalanced-dataset.html

Synthetic Minority Oversampling Technique



https://medium.com/analytics-vidhya/bank-data-smote-b5cb01a5e0a2

EXPANDINDO O CONJUNTO DE DADOS



https://www.kdnuggets.com/2020/02/easy-image-dataset-augmentation-tensorflow.html

Many customers initiated a return process of the product as it was not suitable for use

Many customers launched a return process of the product as it was not appropriate for use

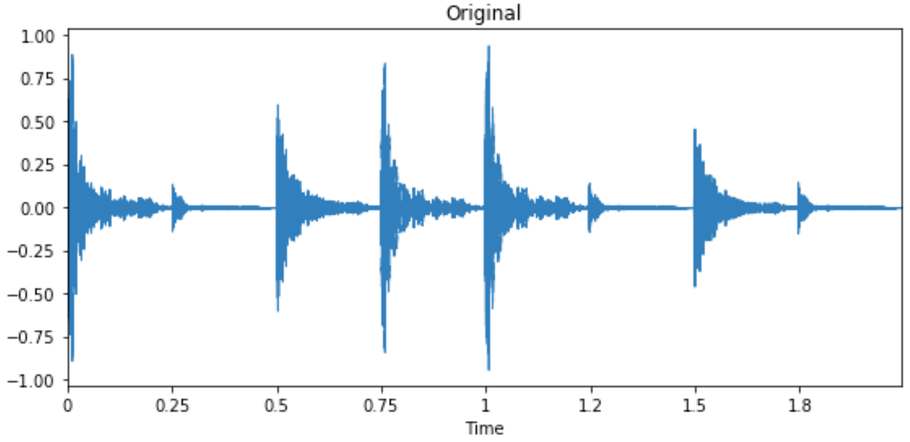
It was conditioned in very thin box which caused scratches on the main screen

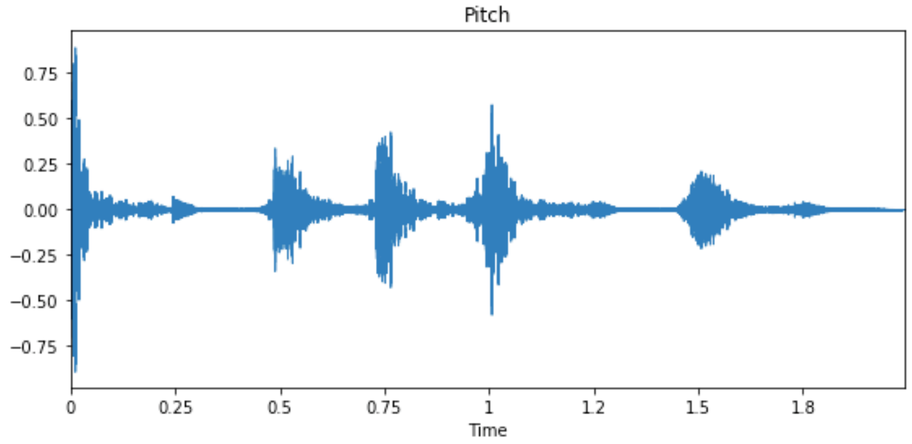
It was packaged in very thin table which provoked scratches on the main screen

The involved firms positively answered their clients who were fully refunded

The involved firms favourably answered their clients who were fully reimbursed

https://medium.com/opla/text-augmentation-for-machine-learning-tasks-how-to-grow-your-text-dataset-for-classification-38a9a207f88d

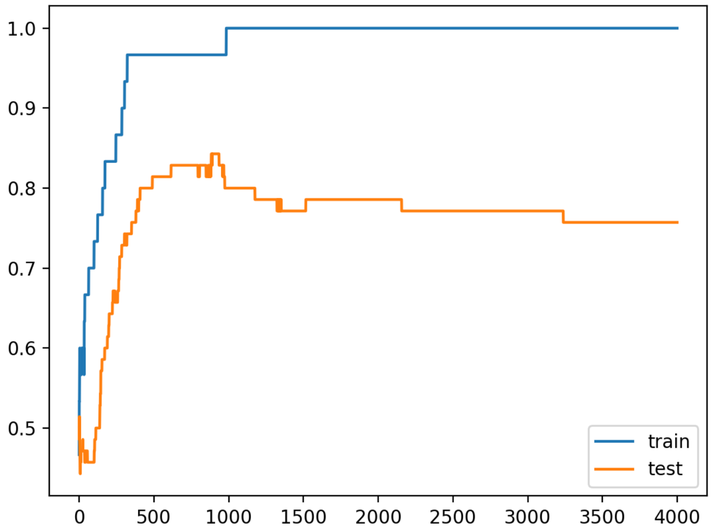




https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6

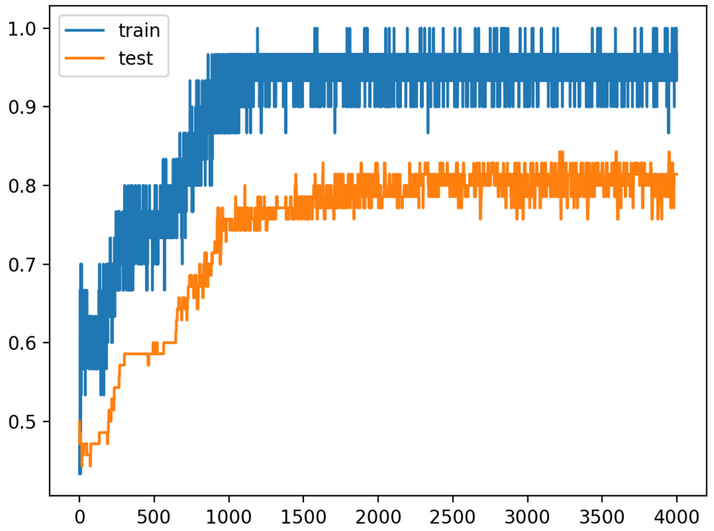
QUANDO VEM DO MODELO

ADICIONANDO RUÍDO NAS CAMADAS OCULTAS



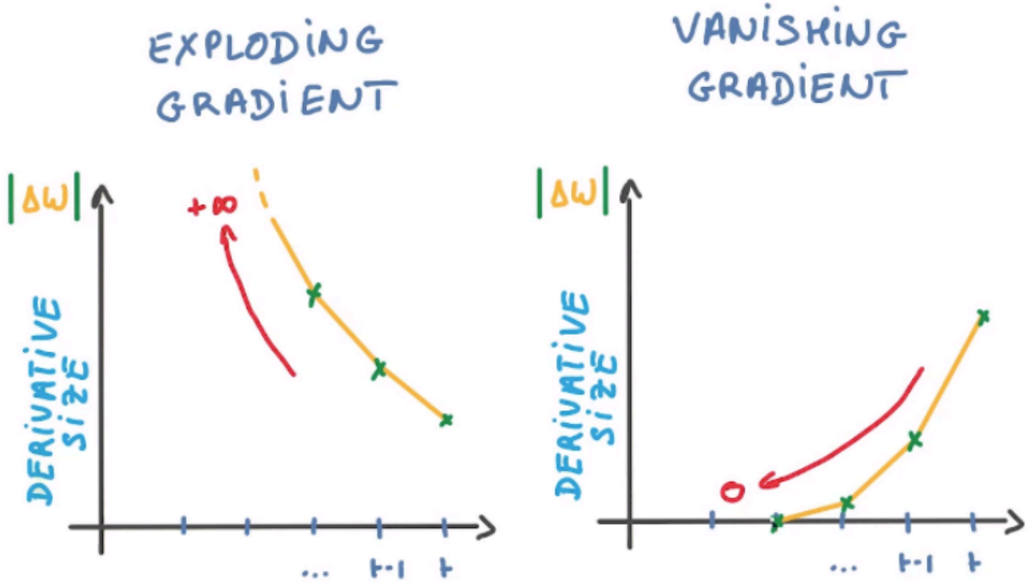
PYTHON: MLP WITH HIDDEN LAYER NOISE

```
model = Sequential( )
model.add(Dense(500, input_dim=2))
model.add(GaussianNoise(0.1))
model.add(Activation('relu'))
model.add(Activation(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy',
                    optimizer='adam',
                    metrics=['accuracy'])
```



https://machinelearningmastery.com/how-to-improve-deep-learning-model-robustness-by-adding-noise/

CÁLCULO DO GRADIENTE (1)



$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta\mathbf{w}^{(t)}$$

Calcular o $\Delta\mathbf{w}^{(t)}$, geralmente, envolve funções de ativação.

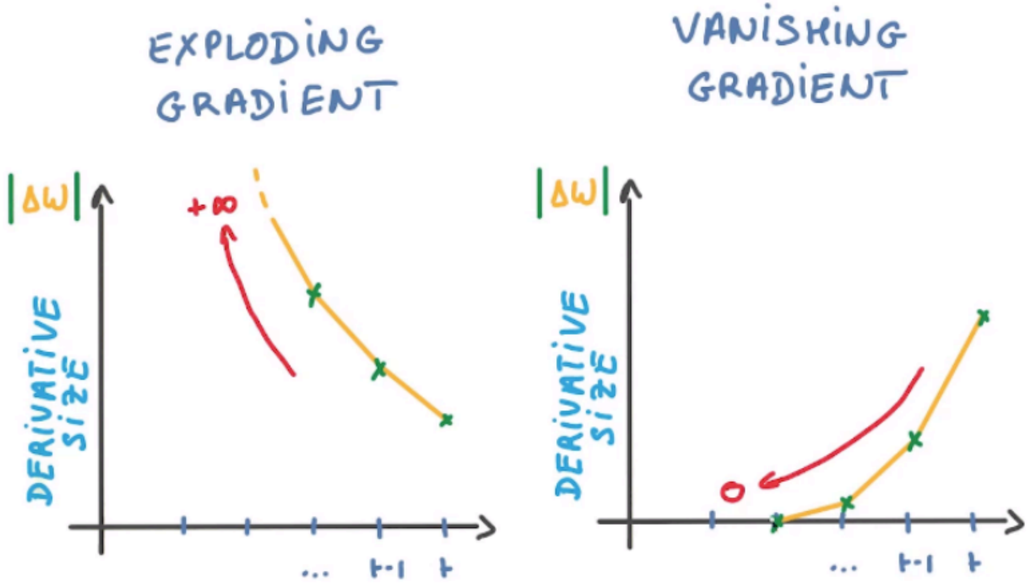
- **Explosão do gradiente**: acúmulo de grandes derivadas faz com que o modelo fique muito instável e incapaz de aprendizado efetivo. Pior cenário $\Delta\mathbf{w}^{(t)} \rightarrow +\infty$, resultando em NaNs;

- **Desaparecimento do gradiente**: acúmulo de pequenos gradientes que resulta em um modelo incapaz de aprender de forma significativa, haja vista que os pesos não serão atualizados efetivamente. Pior cenário com $\Delta\mathbf{w}^{(t)} \rightarrow 0$.

http://jashish.com.np/blog/posts/practical-aspects-of-deep-learning-part-2/

https://towardsdatascience.com/the-vanishing-exploding-gradient-problem-in-deep-neural-networks-191358470c11

CÁLCULO DO GRADIENTE (2)



$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta\mathbf{w}^{(t)}$$

Calcular o $\Delta\mathbf{w}^{(t)}$, geralmente, envolve funções de ativação.

COMPORTAMENTOS OBSERVADOS QUANDO OCORRE A EXPLOSÃO DO GRADIENTE:

- Incapacidade de aprender;
- Mudanças abruptas nos valores da função objetivo;
- Alguns NaNs podem aparecer durante treino.

COMPORTAMENTO OBSERVADO QUANDO OCORRE O DESAPARECIMENTO DO GRADIENTE:

- Convergência lenta;
- Pesos próximos à camada de saída atualizados cp, mais intensidade, enquanto os das camada mais próximas à entrada sofrem mudanças pífias;
- Pesos com valores bem pequenos, chegando a zero.

https://towardsdatascience.com/the-vanishing-exploding-gradient-problem-in-deep-neural-networks-191358470c11

FUNÇÕES DE ATIVAÇÃO

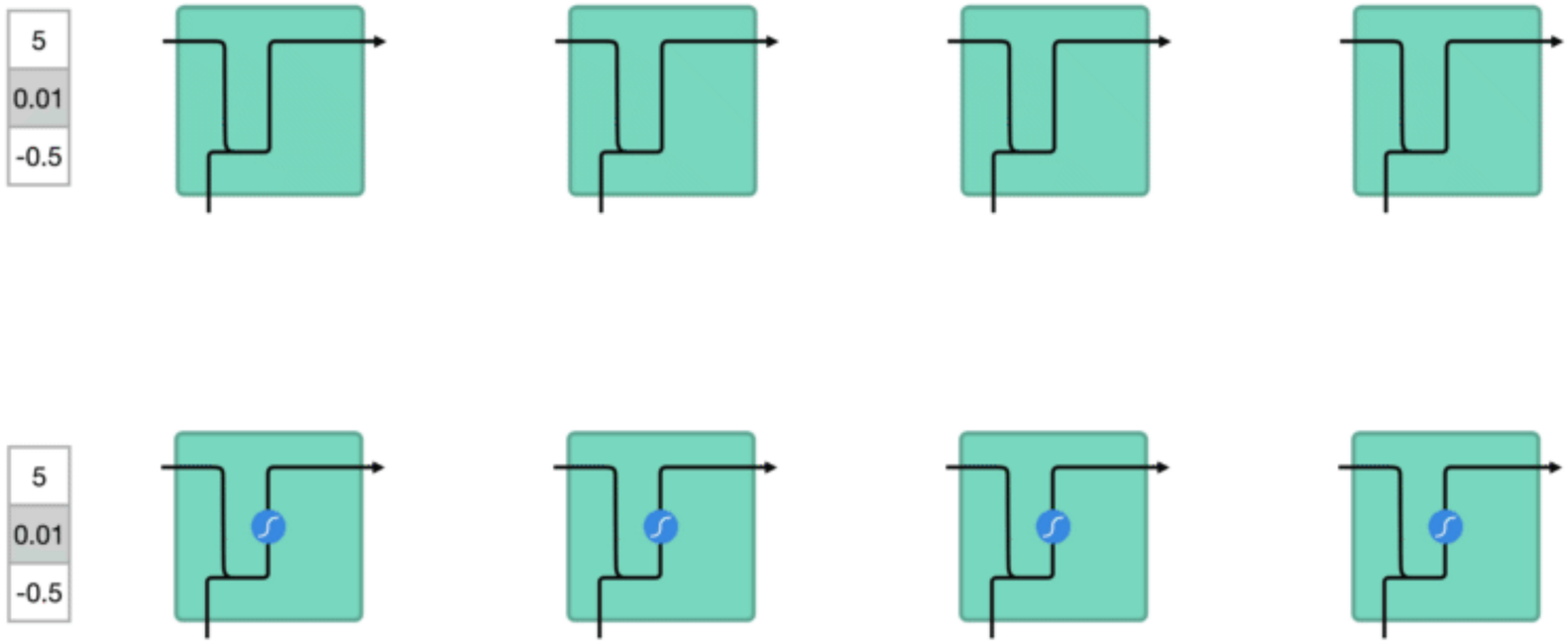
IMPORTÂNCIA

- As funções de ativação permitem que pequenas mudanças nos pesos e viés causem apenas uma pequena alteração na saída do neurônio (e da rede);

- Eles permitem a introdução de recursos não lineares na rede. Assim, as funções de ativação são cruciais para o modelo de rede neural aprender e entender funções não lineares complexas;

- Sem funções de ativação, as saída são apenas funções lineares simples (Alô, Adaline!). A complexidade das funções lineares é limitada. Logo, a capacidade de aprender mapeamentos de funções complexas a partir de dados é baixa;

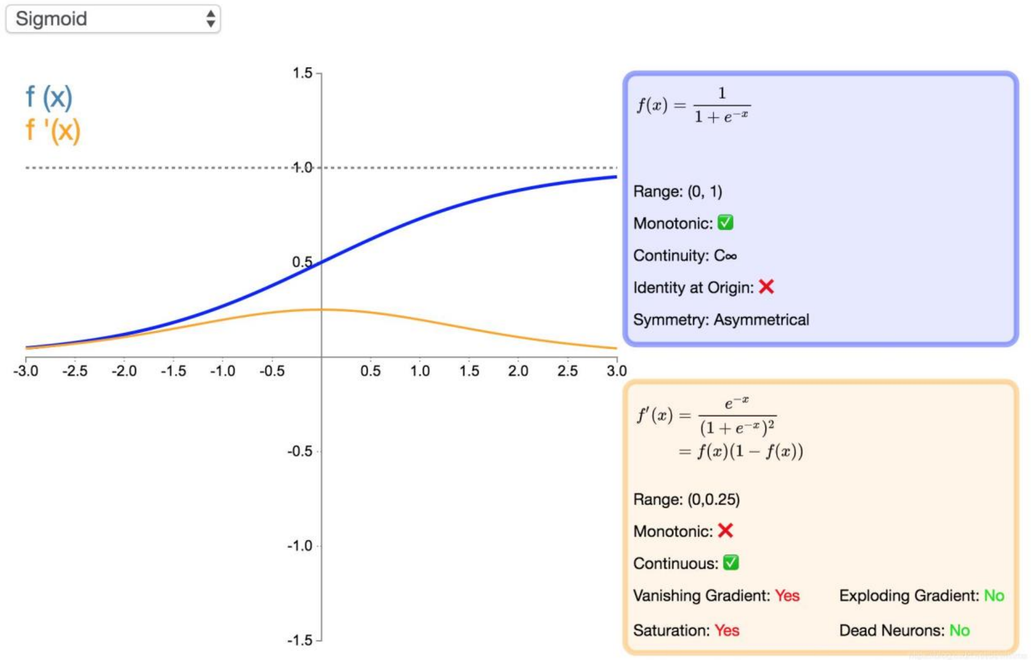
IMPORTÂNCIA



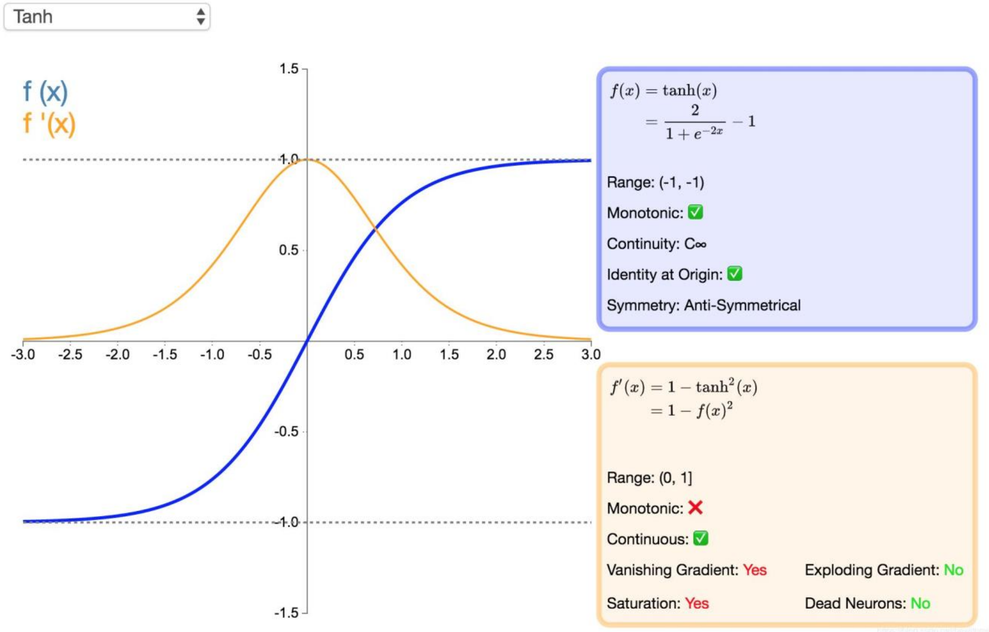
https://laptrinhx.com/introduction-to-neural-networks-and-their-key-elements-part-c-activation-functions-layers-1195957448/

AS FAMOSINHAS

Sigmoid



Tangente hiperbólica



https://laptrinhx.com/introduction-to-neural-networks-and-their-key-elements-part-c-activation-functions-layers-1195957448/

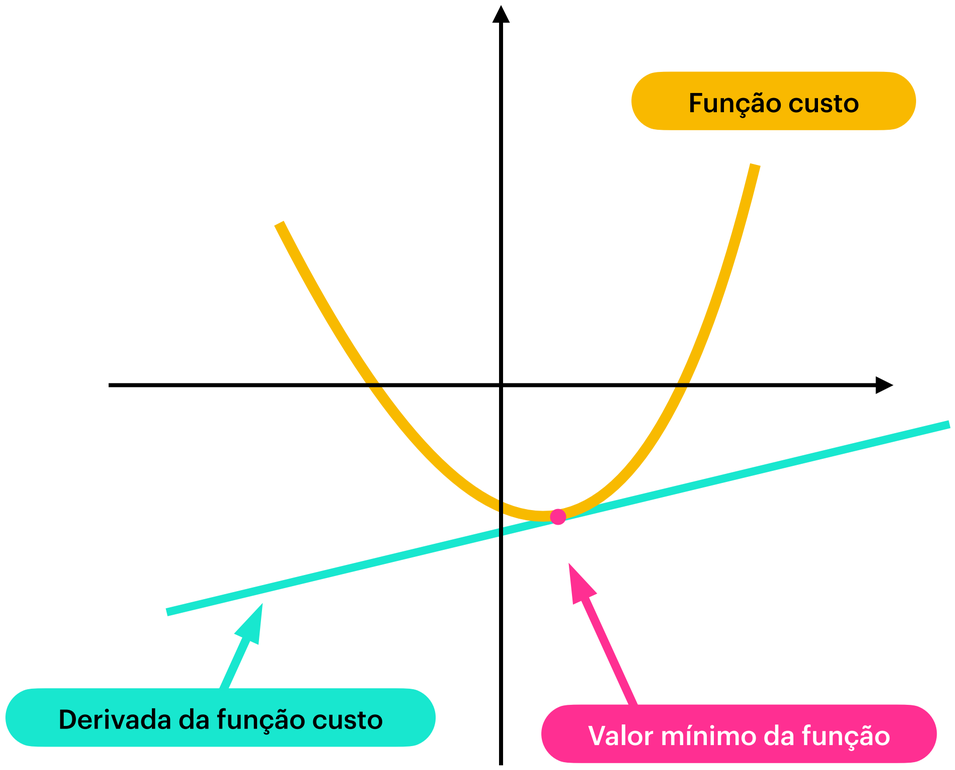
TREINANDO UM ADALINE (Recap)

- Premissa básica: achar o vetor de pesos que minimize o TODOS os erros;
- Forma de calcular o erro continua a mesma, i.e., $e = d - y = d - \varphi\left(\mathbf{x}^\intercal \mathbf{w}\right)$.
- Agregar TODOS os erros em uma função só:

$$J(\mathbf{e}) = \sum_i e_i^2.$$

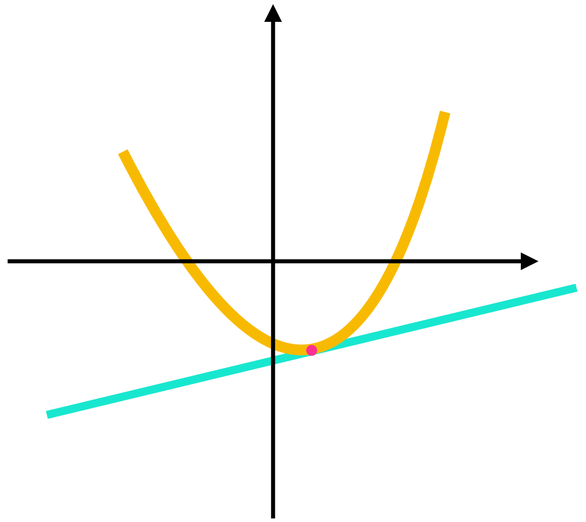
$$J(\mathbf{w}) = \sum_i \left(d_i - \varphi(\mathbf{x}_i^\intercal \mathbf{w})\right)^2.$$

$$\mathbf{w}^\star = \min_{\mathbf{w}} J(\mathbf{w})$$



TREINANDO UM ADALINE (Recap)

- Com o intuito de minimizar $J(\mathbf{w})$, é necessário que seja obtido a direção do ajuste a ser aplicado no vetor de pesos para aproximar a solução do mínimo de $J(\mathbf{w})$;
- De modo iterativo, iremos realizar $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta\mathbf{w}$;
- Para tal, usa-se o **gradiente** da função $J(\mathbf{w})$ no ponto $\mathbf{w}^{(t)}$;
- Sabe-se que o gradiente possui a mesma direção da maior variação do erro (aponta na direção do crescimento da função). Portanto, o ajuste deve ocorrer na direção contrária do gradiente. Logo a variação dos pesos pode ser descrita como:



$$\Delta\mathbf{w} \propto -\nabla\mathbf{w}.$$

TREINANDO UM ADALINE (Pulo do 🐈)

- Sabe-se que a função J depende de e, bem como sabe-se que e depende de y e, ainda, sabe-se que y depende de φ que depende de $\mathbf{w}$. Logo, regra da cadeia:

$$\frac{\partial J}{\partial w_i} = \frac{\partial J}{\partial e_i}\frac{\partial e_i}{\partial y_i}\frac{\partial y_i}{\partial w_i}, \qquad \text{em que} \qquad J(\mathbf{w}) = \frac{1}{2}\sum_i \Big(d_i - \varphi(\mathbf{x}_i^\intercal \mathbf{w})\Big)^2.$$

- Quanto valem os termos abaixo?

a) $\dfrac{\partial J}{\partial e_i} = 2e_i$ b) $\dfrac{\partial e_i}{\partial y_i} = -\varphi'(\mathbf{x}_i^\intercal \mathbf{w})$ c) $\dfrac{\partial y_i}{\partial w_i} = \mathbf{x}_i$ d) $\dfrac{\partial J}{\partial w_i} = -e_i\,\varphi'(\mathbf{x}_i^\intercal \mathbf{w})\,x_i$

- Adicionalmente:

a) $\text{sigmoid}'(u) = \text{sigmoid}(u)(1 - \text{sigmoid}(u))$ b) $\tanh'(u) = 1 - \tanh^2(u)$

TREINANDO UM ADALINE (Pulo do 🐈)

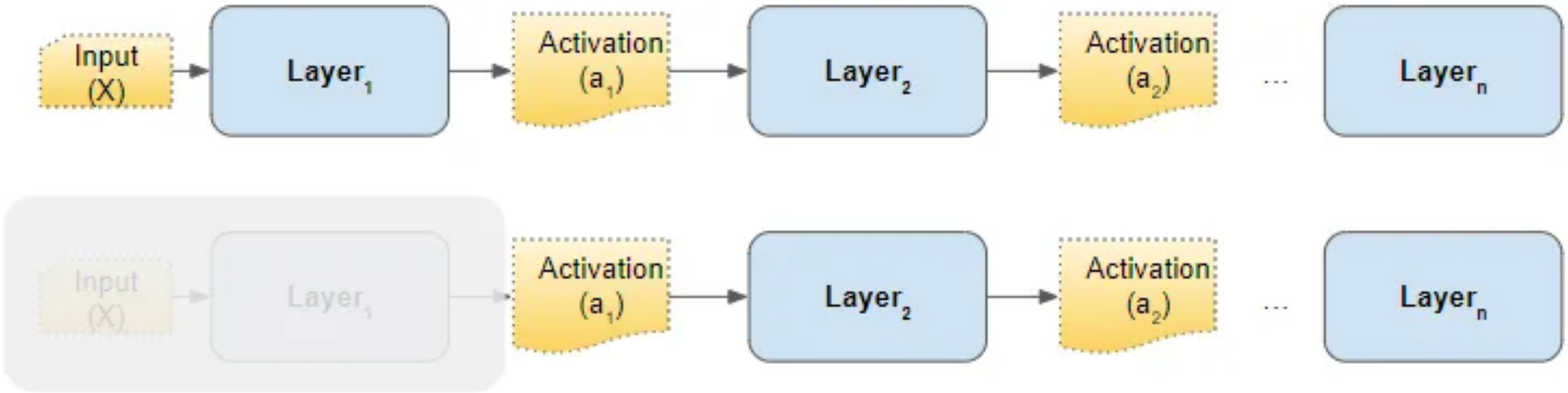
- Considerando que

a) $\frac{\partial J}{\partial e_i} = 2e_i$ b) $\frac{\partial e_i}{\partial y_i} = -\varphi'(\mathbf{x}_i^\intercal \mathbf{w})$ c) $\frac{\partial y_i}{\partial w_i} = \mathbf{x}_i$ d) $\frac{\partial J}{\partial w_i} = -e_i\,\varphi'(\mathbf{x}_i^\intercal \mathbf{w})\,x_i$

- Considerando também a regra de aprendizagem $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta\mathbf{w}$;
- Considerando que $\Delta\mathbf{w} \propto \nabla J$ e que $\frac{\partial J}{\partial \mathbf{w}} = -e\,\varphi'(\mathbf{x}^\intercal \mathbf{w})\,\mathbf{x}$, podemos reescrever a regra de atualização da seguinte forma:

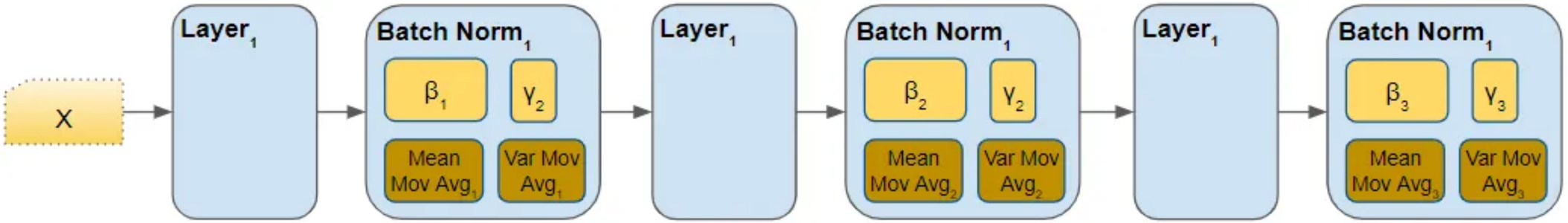
$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta e^{(t)} \varphi'\left(\mathbf{x}^\intercal \mathbf{w}^{(t)}\right) \mathbf{x}.$$

NORMALIZANDO MAPAS DE ATRIBUTOS



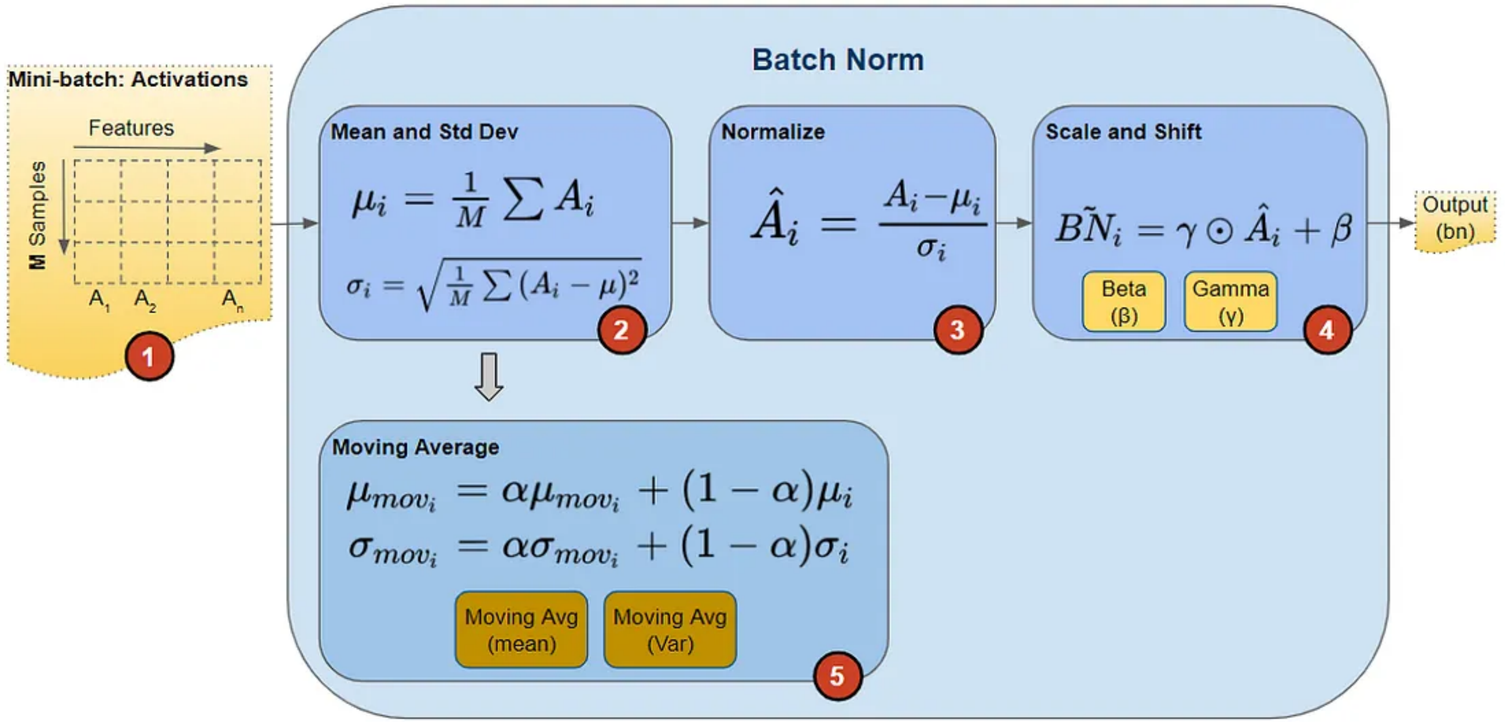
https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks-need-it-b18919692739

NORMALIZANDO MAPAS DE ATRIBUTOS



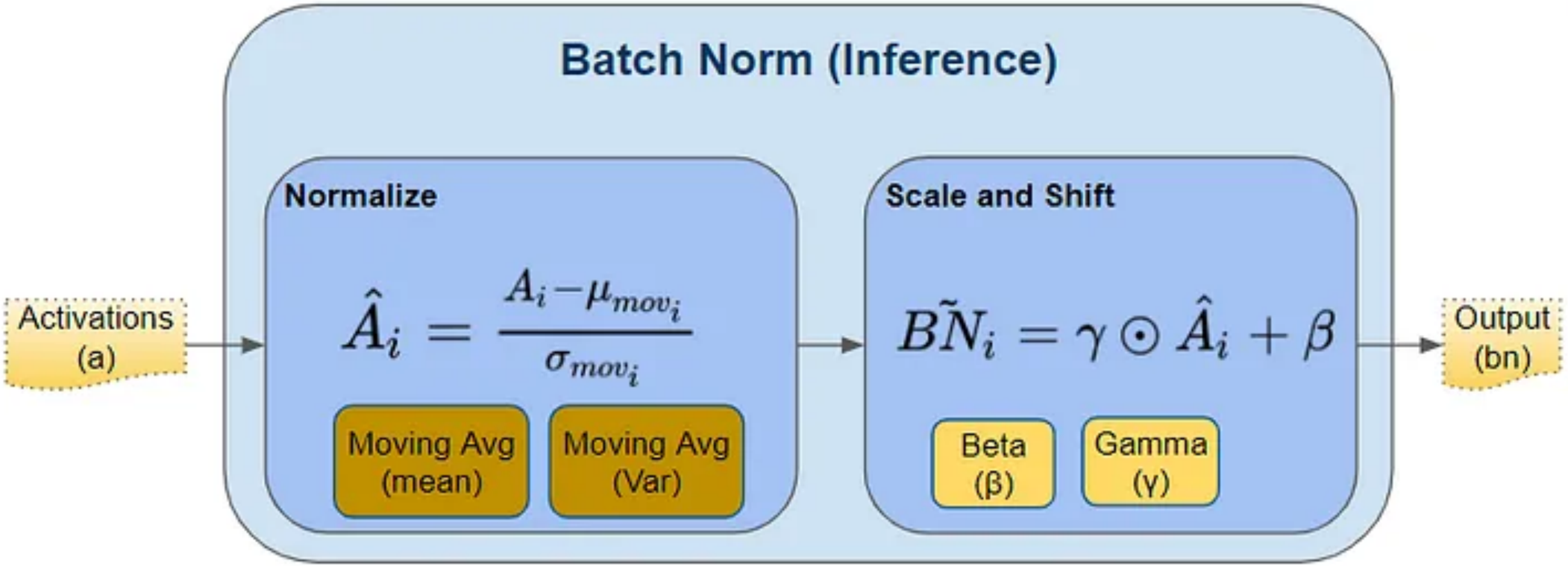
https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks-need-it-b18919692739

NORMALIZANDO MAPAS DE ATRIBUTOS



https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks-need-it-b18919692739

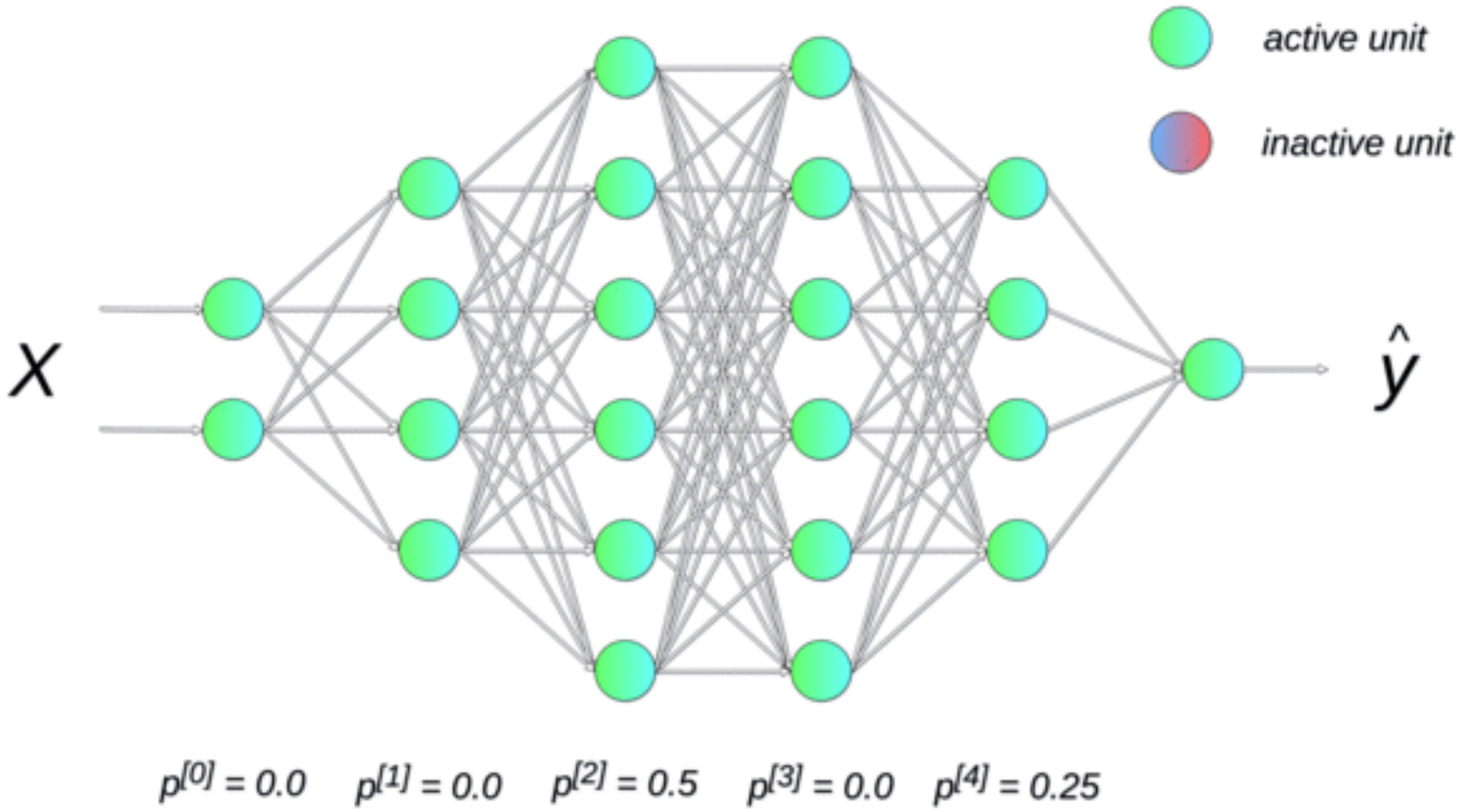
NORMALIZANDO MAPAS DE ATRIBUTOS



https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks-need-it-b18919692739

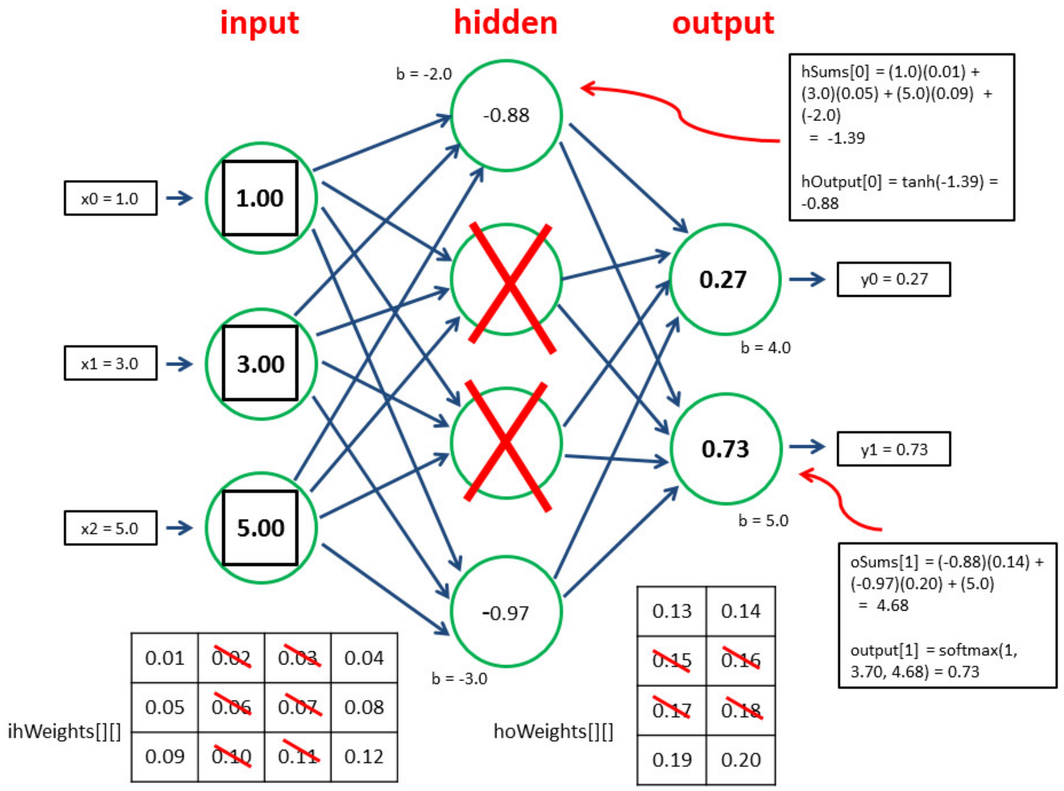
VOLTANDO AO MODELO

DROPOUT (1)



https://towardsdatascience.com/preventing-deep-neural-network-from-overfitting-953458db800a

DROPOUT (2)



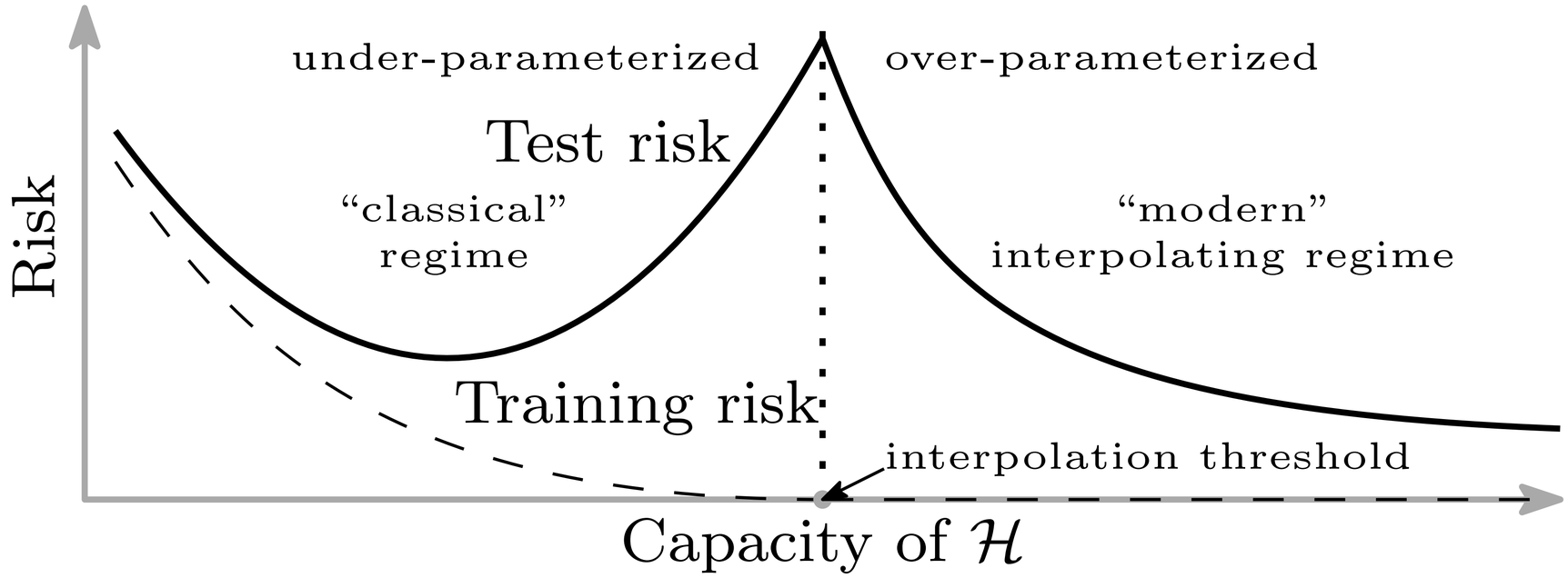
- Agora, imagine que terminamos o treinamento e removemos o *Dropout*. Agora as ativações dos neurônios de saída serão computadas com base em quatro valores da camada oculta. Isso provavelmente colocará os neurônios de saída em regime incomum, de modo que eles produzirão valores absolutos muito grandes, sendo superexcitados;

- Para evitar isso, o truque é multiplicar os pesos das conexões de entrada da última camada por 1-p (portanto, por 0,5);

- Alternativamente, pode-se multiplicar as saídas da camada oculta por 1-p, que é basicamente a mesma coisa.

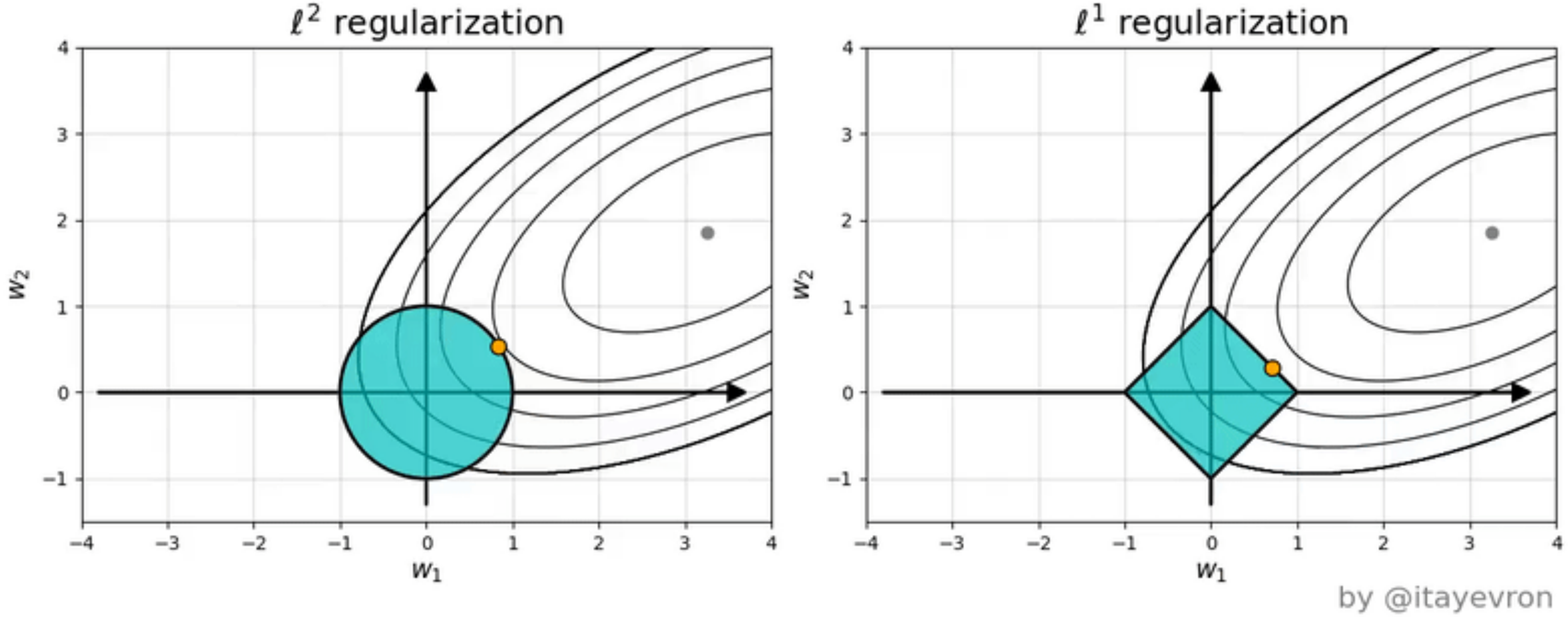
https://datascience.stackexchange.com/questions/44293/how-does-dropout-work-during-testing-in-neural-network

REGULARIZAÇÃO (1)



REGULARIZAÇÃO (1)

ℓ^1 induces sparse solutions for least squares

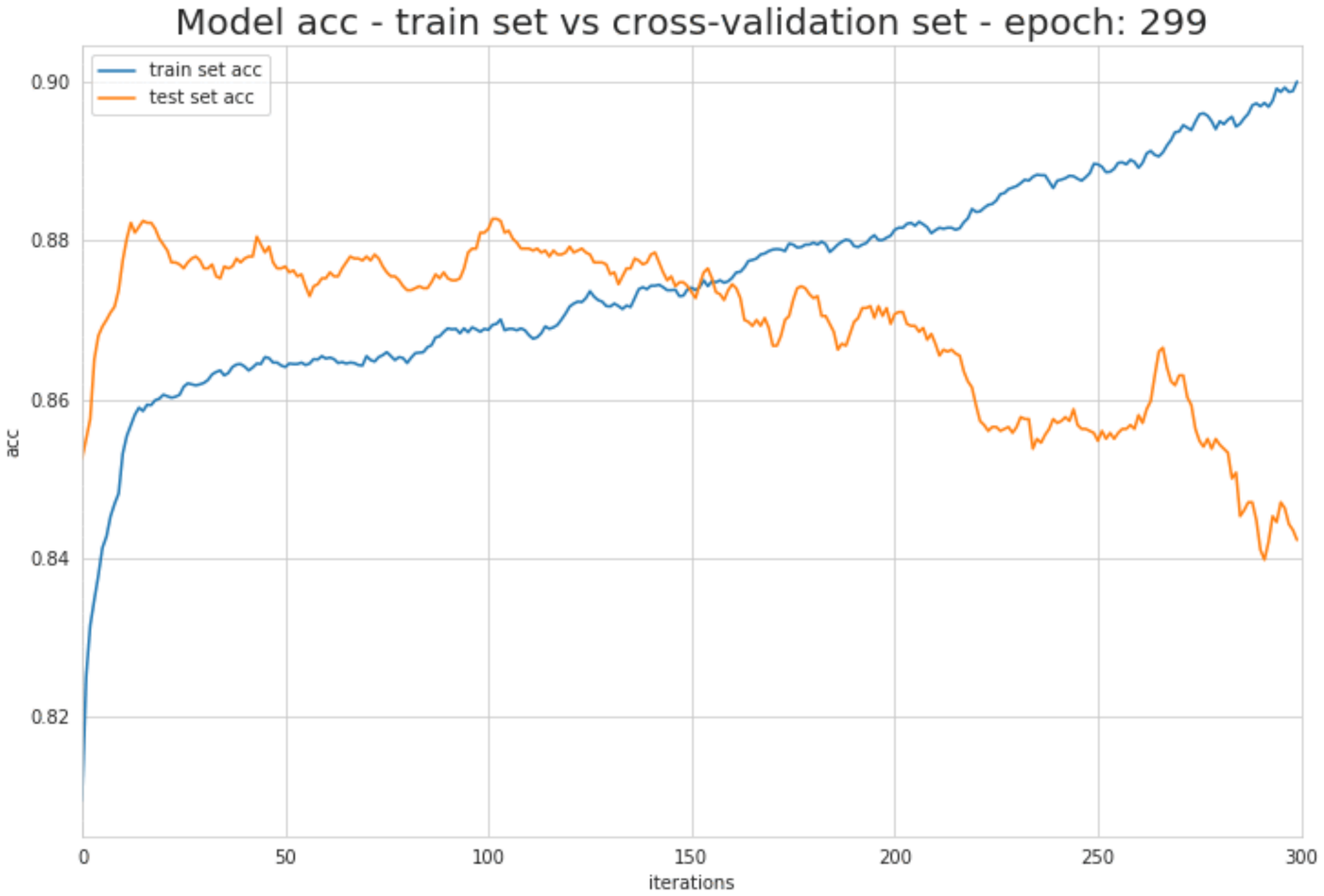


$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \underbrace{\eta \mathbf{w}^{(t)} + (1 - \alpha\eta) \nabla J\left(\mathbf{w}^{(t)}\right)}_{\Delta \mathbf{w}^{(t)}}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \underbrace{\eta \nabla J\left(\mathbf{w}^{(t)}\right) + \alpha \operatorname{sign}\left(\mathbf{w}^{(t)}\right)}_{\Delta \mathbf{w}^{(t)}}$$

https://www.quora.com/What-is-the-intuition-that-the-l-1-norm-leads-to-sparse-solutions/answer/Itay-Evron-1

PARADA PRECOCE



https://towardsdatascience.com/preventing-deep-neural-network-from-overfitting-953458db800a

Referências (1)

- AUTOMATION & **IIOT. What's Wrong with Deep Learning?** https://www.machinedesign.com/automation-iiot/article/21837994/whats-wrong-with-deep-learning. 2019, Accessed on Mar 2021.
- Gustavo Chávez. **The main issue with identifying Financial Fraud using Machine Learning (and how to address it)**. https://towardsdatascience.com/the-main-issue-with-identifying-financial-fraud-using-machine-learning-and-how-to-address-it-3b1bf8fa1e0c. 2019, Accessed on Mar 2021.
- Dipanjan (DJ) Sarkar. **A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning**. https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a. 2018, Accessed on Mar 2021.
- Dmytro Prylipko. **How does dropout work during testing in neural network?** https://datascience.stackexchange.com/questions/44293/how-does-dropout-work-during-testing-in-neural-network. 2019, Accessed on Jun 2021.
- Richard O. Duda, Peter E. Hart, David G. Stork. **Pattern Classification**. John Wiley & Sons, 2012.

Referências (2)

- Itay Evron. **What is the intuition that the l-1 norm leads to sparse solutions**? https://towardsdatascience.com/intuitions-on-l1-and-l2-regularisation-235f2db4c261. 2017, Accessed on Feb 2021.
- Kurtis Pykes. **The Vanishing/Exploding Gradient Problem in Deep Neural Networks.** https://towardsdatascience.com/the-vanishing-exploding-gradient-problem-in-deep-neural-networks-191358470c11. 2021, Accessed on Mar 2021.
- Matthew Mayo. **Easy Image Dataset Augmentation with TensorFlow.** https://www.kdnuggets.com/2020/02/easy-image-dataset-augmentation-tensorflow.html. 2020, Accessed on Feb 2021.
- Wanshun Wong. **What is Gradient Clipping?** https://towardsdatascience.com/what-is-gradient-clipping-b8e815cdfb48. 2020, Accessed on Mar 2021.
- Zaki Jefferson. **Bank Data**: SMOTE. https://medium.com/analytics-vidhya/bank-data-smote-b5cb01a5e0a2. 2020, Accessed on Mar 2021.