# Modelagem Preditiva Avançada

Rafael Lychowski

FGV IDE
In Company Management Online

FGV IDE In Company Management Online

# MAPA CONCEITUAL DA DISCIPLINA

| | PRIMEIRO DIA | SEGUNDO DIA |
|---|---|---|
| MANHÃ | • Revisão Modelagem Preditiva<br>• Estudo de Caso | • SVM<br>• Estudo de Caso<br>• Redes Neurais<br>• Estudo de Caso |
| TARDE | • Combinação<br>• Estudo de Caso | • Algoritmos Genéticos<br>• Estudo de Caso<br>• Trabalho Final |

# MAPA CONCEITUAL DA DISCIPLINA

## PRIMEIRO DIA

- Revisão Modelagem Preditiva
- Estudo de Caso

## TERCEIRO DIA

- SVM
- Estudo de Caso
- Redes Neurais
- Estudo de Caso

## SEGUNDO DIA

- Combinação
- Estudo de Caso

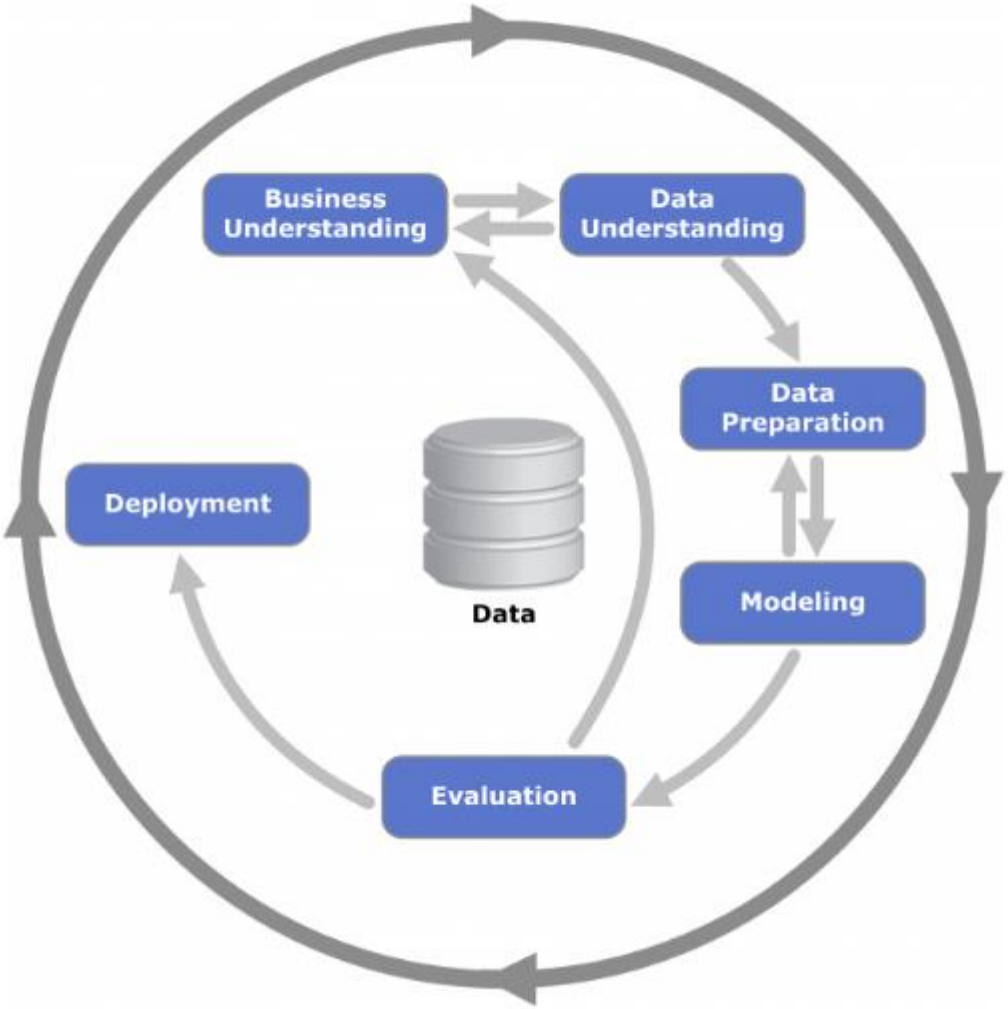
## QUARTO DIA

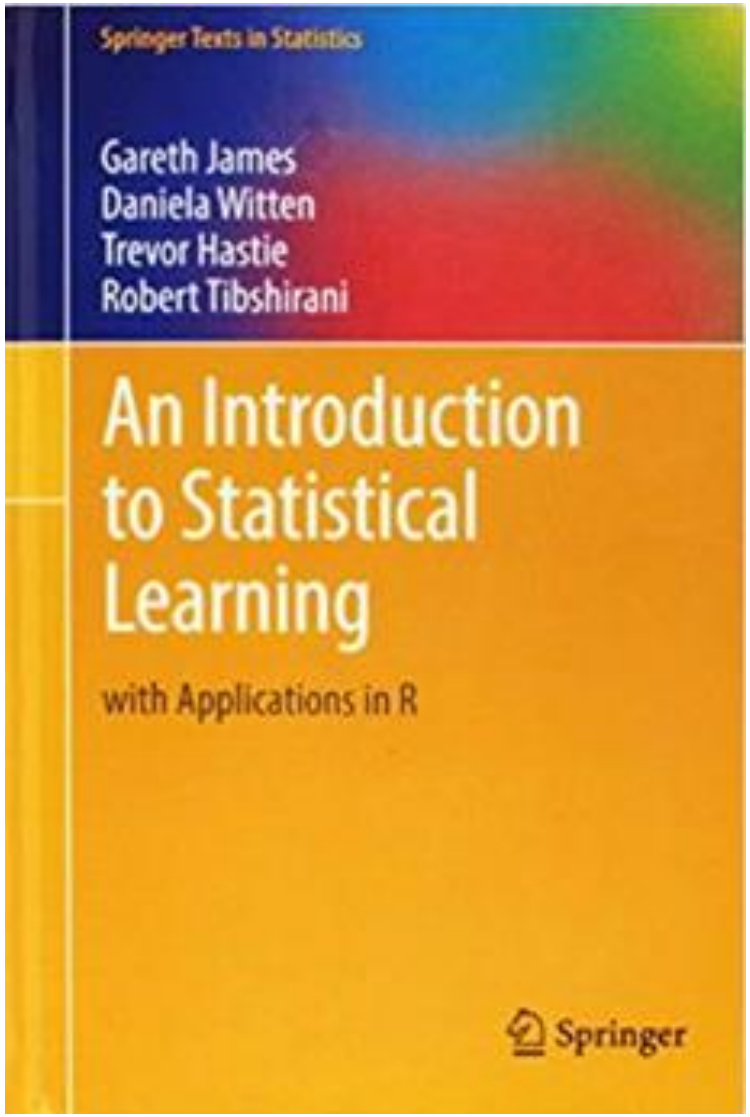
- Algoritmos Genéticos
- Estudo de Caso

## QUINTO DIA

- Trabalho Final

# CRISP – DM (Cross Industry Standard Process for Data Mining)




Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer

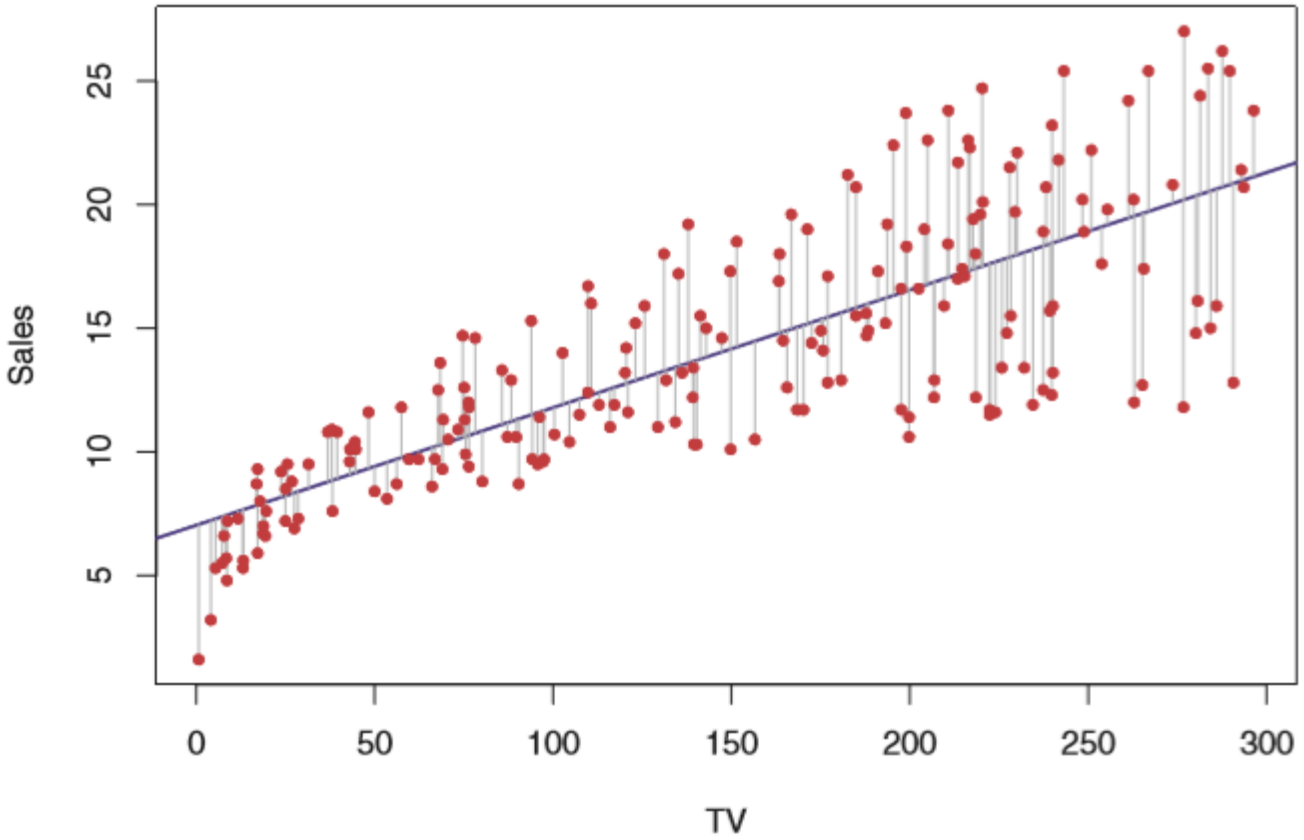
# Modelagem Preditiva Avançada

FGV IDE In Company Management Online

| Método | Sub Método | Objetivo | Caso de Uso | Algoritmos |
|---|---|---|---|---|
| **Supervisionados**<br><br>Para cada conjunto de entrada existe um valor alvo correspondente | Regressão | Estimar uma variável continua. Forecast, Time series | • Forecast da demanda de compras<br>• Predição da quantidade de chuva | • Linear Regression<br>• Neural networks<br>• Decision trees |
| | Classificação | Estimar uma variável discreta | • Prever a quebra de equipamentos<br>• Risco de crédito | • Logistic Regression<br>• SVMs<br>• Neural Networks<br>• Decision Trees |
| **Não Supervisionados**<br><br>Encontrar as relações entre diferentes entradas sem uma variável alvo definada | Clustering | Identificar objetos similares | • Segmentação de clientes (marketing)<br>• Análise de Redes Sociais | • K-Means<br>• DBSCAN<br>• HDBSCAN<br>• Hierarchical Clustering |
| | Redução de Dimensionalidade | Reduzir a complexidade dos dados | • Sistemas de Recomendação (Netflix, Amazon)<br>• Processamento de Linguagens Naturais | • PCA, SVD, ALS<br>• Latent dirichlet allocation<br>• t-SNE, MDS |

# Regressão Linear

$$Y = \beta_0 + \beta_1 X + \epsilon.$$



# Regressão Linear

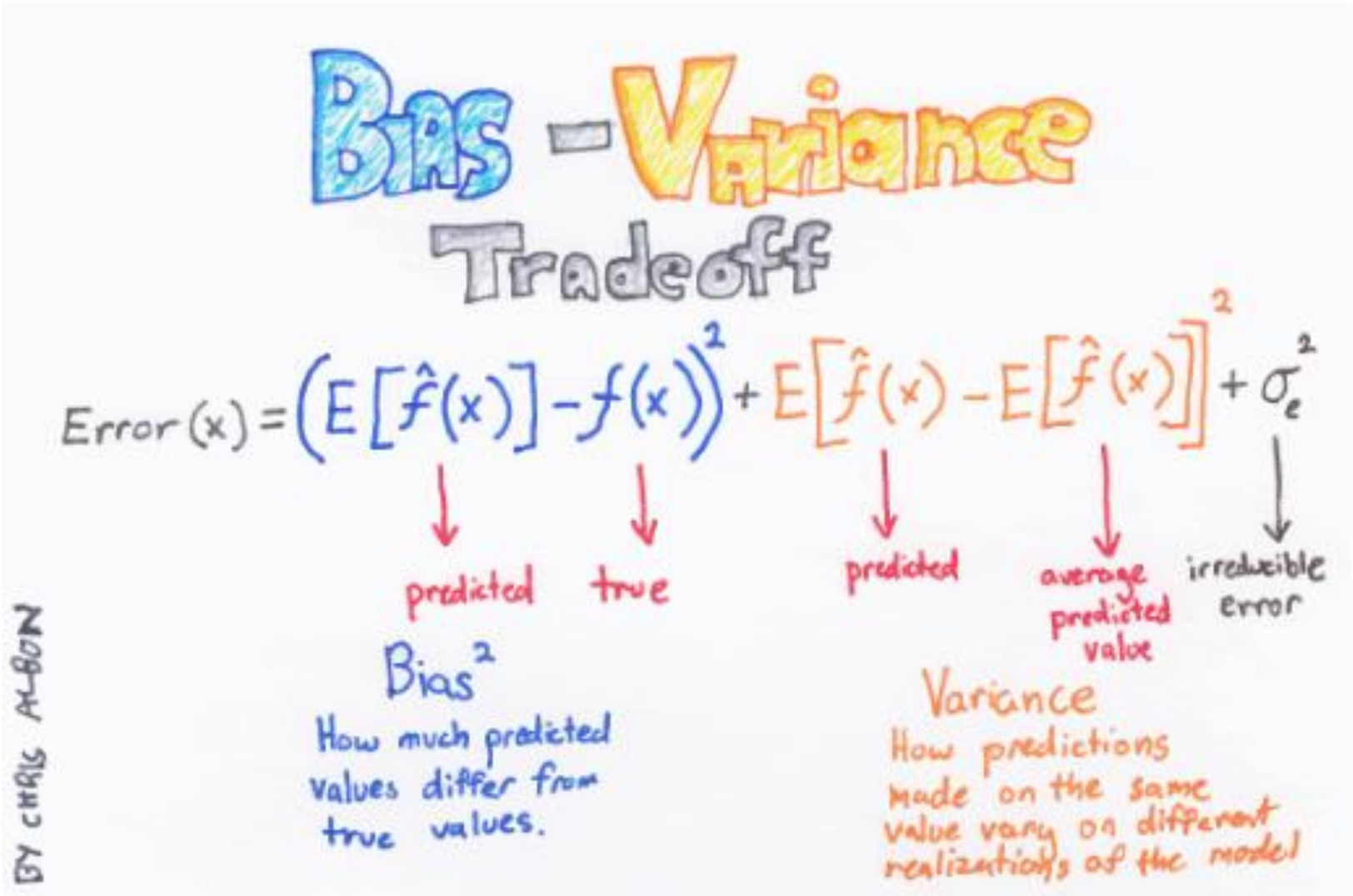
1. Existe alguma relação entre a variável de input e de output ?
2. O quão essa relação é forte ?
3. Qual variável contribui mais ? (importância)
4. Com qual acurácia podemos estimar o efeito de cada variável de input na variável de output ?
5. Com qual acurácia podemos estimar a variável de output ?
6. A relação das variáveis é linear ?
7. Existe sinergia entre as variáveis de input ?

# Regressão Linear

$$h_\theta(x) = \theta_0 + \theta_1 x$$

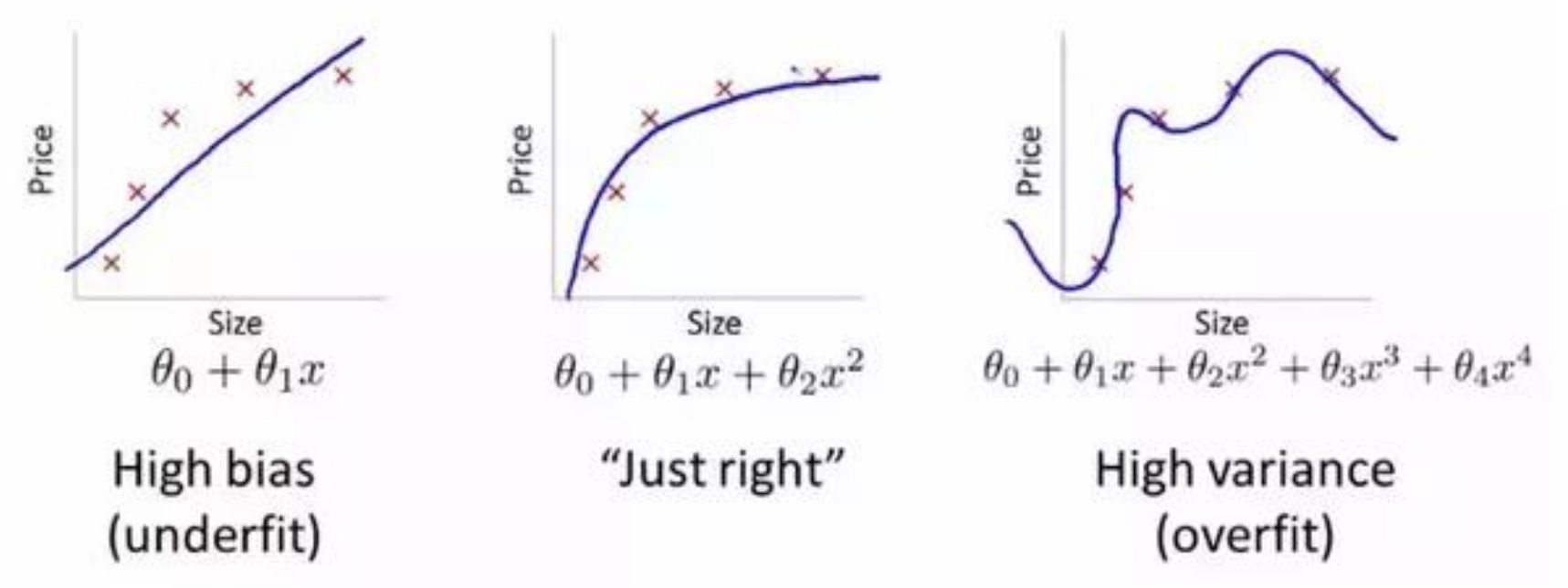
E se tivermos várias features ?

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$


Bias - Variance Tradeoff
$Error(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\hat{f}(x) - E[\hat{f}(x)]\right]^2 + \sigma_e^2$
predicted
true
predicted
average predicted value
irreducible error
Bias²
How much predicted values differ from true values.
Variance
How predictions made on the same value vary on different realizations of the model
BY CHRIS ALBON

# Revisão

## Bias x Variance

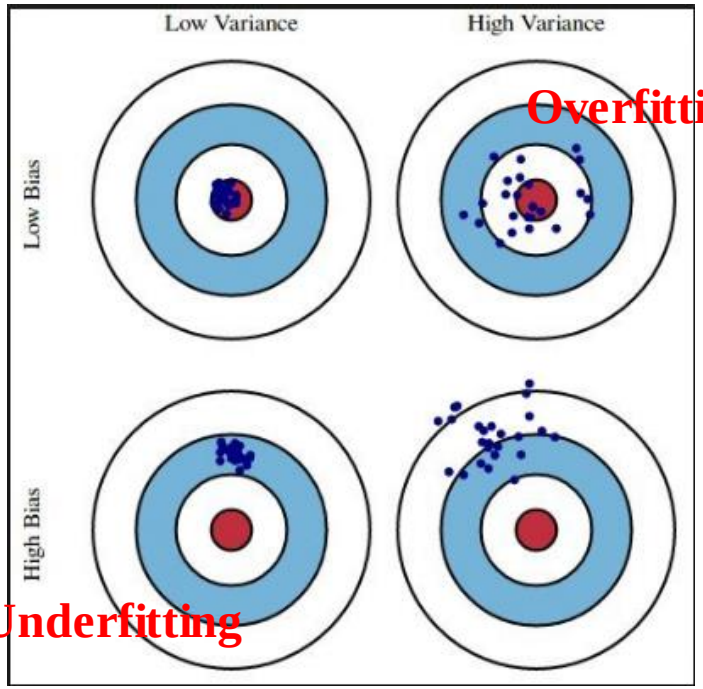


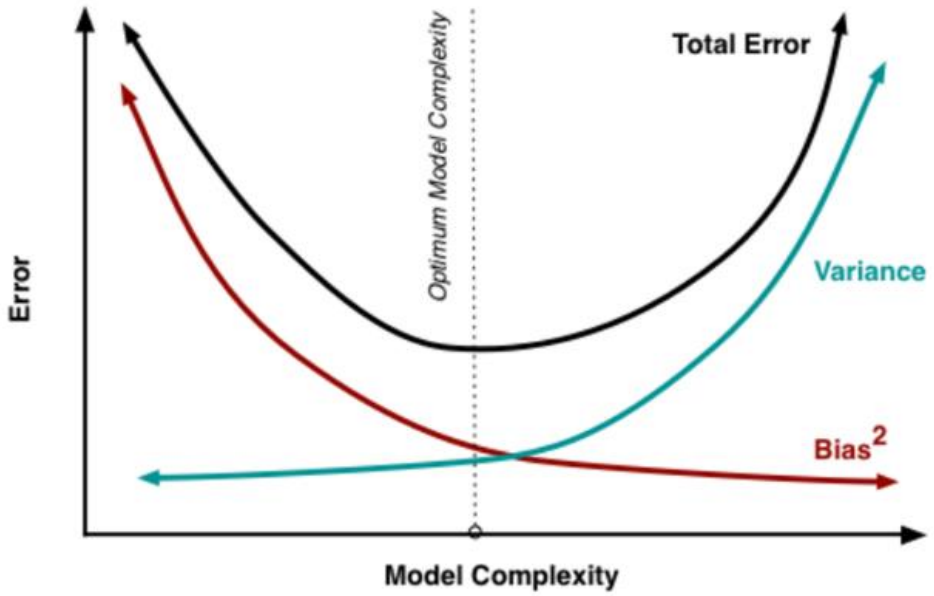

- Adicionar variáveis (features)
- Aumentar complexidade do modelo

- Adicionar mais dados (base de treino)
- Diminuir variáveis (features)

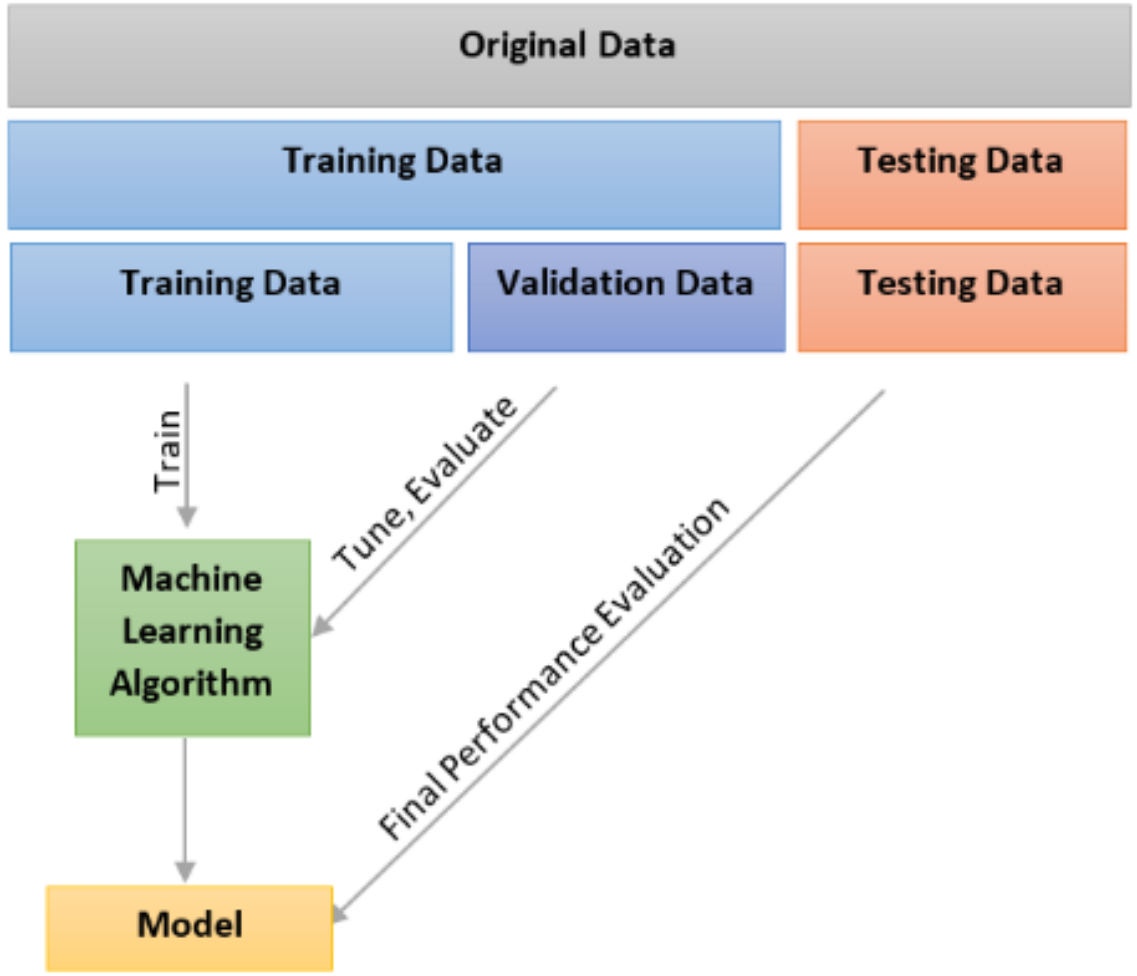
# Revisão

Bias x Variance



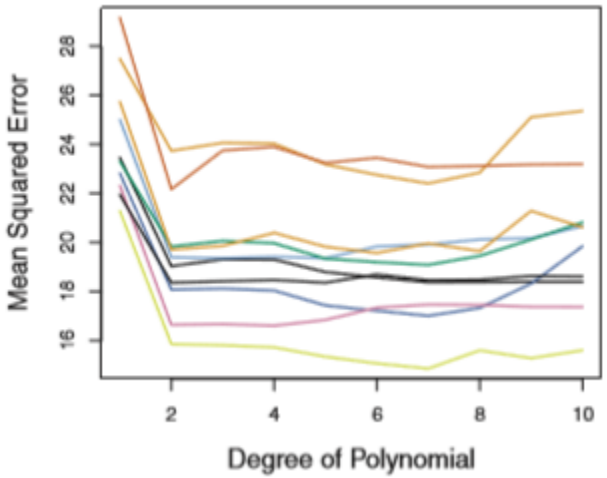



# Partições


Original Data
Training Data
Testing Data
Training Data
Validation Data
Testing Data
Train
Tune, Evaluate
Machine Learning Algorithm
Final Performance Evaluation
Model

# Resampling

- Aumenta confiabilidade do modelo
- Exige maior poder computacional
- Cross validation




Amostras diferentes geram resultados bem diferentes !!

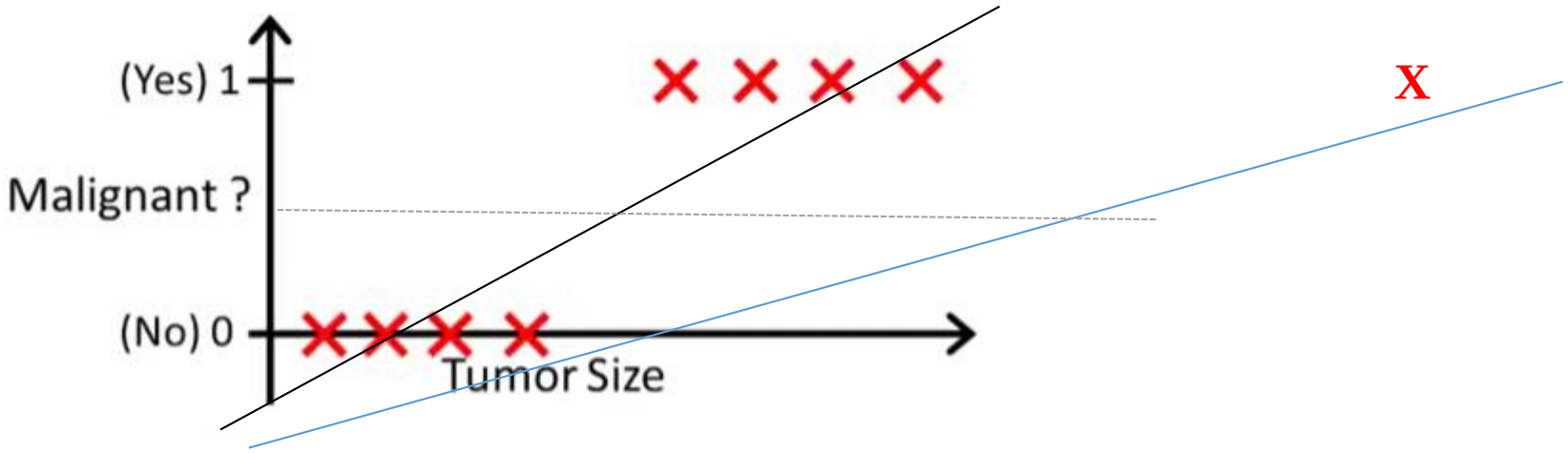
# Cross Validation

# Regressão Linear

- Parabéns! Você acaba de ser contratado como o mais novo Data Scientist de uma empresa global de Real Estate. Com o crescimento acelerado da cidade de Boston, devido sua proximidade a centros de excelência como Harvard e MIT, o mercado imobiliário da região apresenta uma oportunidade única. Para comprar os melhores imóveis aos melhores preços, você deve desenvolver um modelo capaz de receber os dados de um imóvel qualquer e dizer qual deve ser seu preço aproximado. Assim ao buscar por oportunidades na região poderá filtrar o que está caro demais e o que está barato.

| Feature | Descrição | Tipo |
|---|---|---|
| CRIM | Taxa de Crimes | Real |
| NROOMS | Número de Quartos | Real |
| AGE | Idade do Imóvel | Real |
| DISC | Distância do Centro | Real |
| BATH | Número de Banheiros | Real |
| TAX | Taxa de IPTU | Real |
| MEDV | Preço do Imóvel | Real |

# Classificação: por que não regressão ?




**Um único valor pode distorcer completamente o resultado**

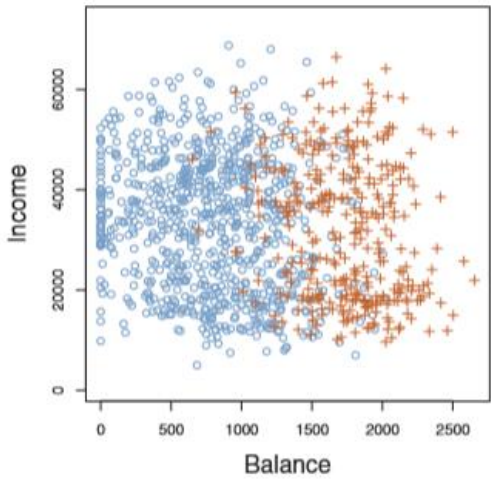
# Classificação: por que não regressão ?

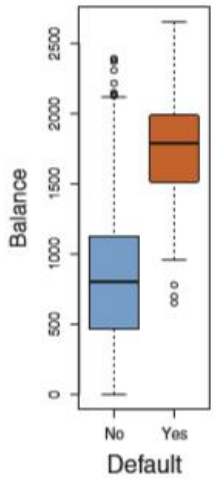
$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

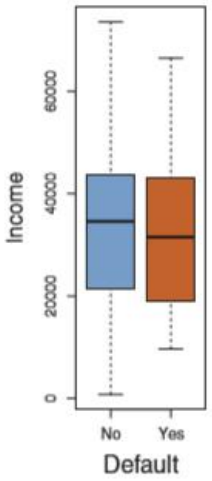
$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

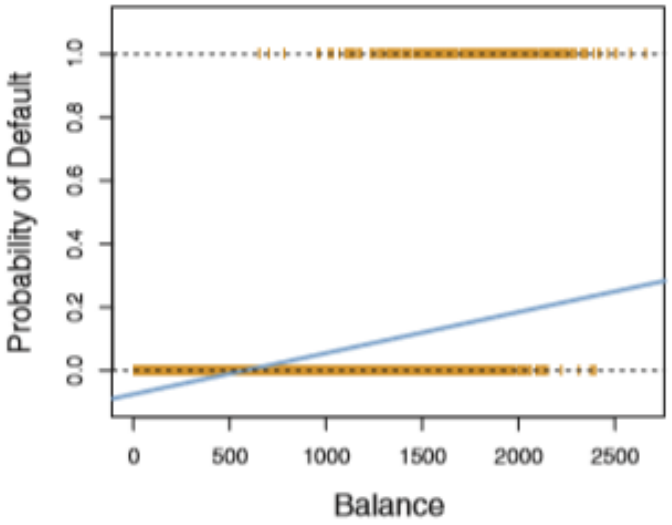
**Ordem e valores relativos não correspondem a realidade**

# "Regressão" Logística



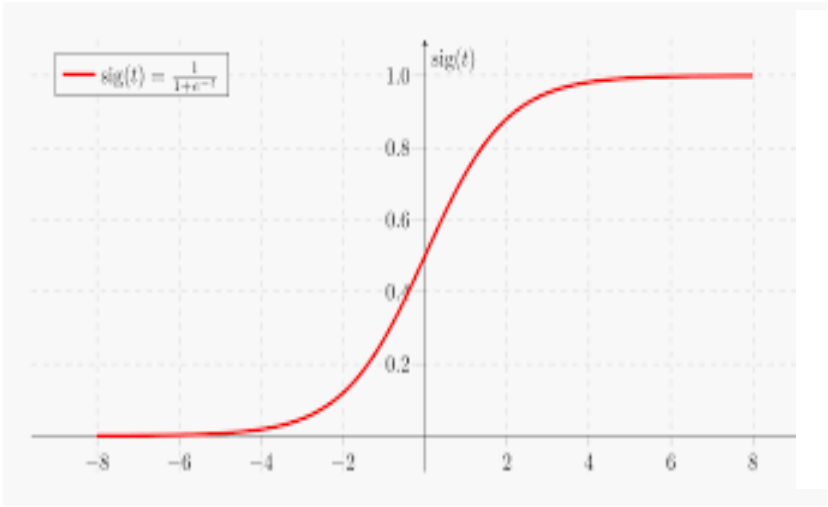



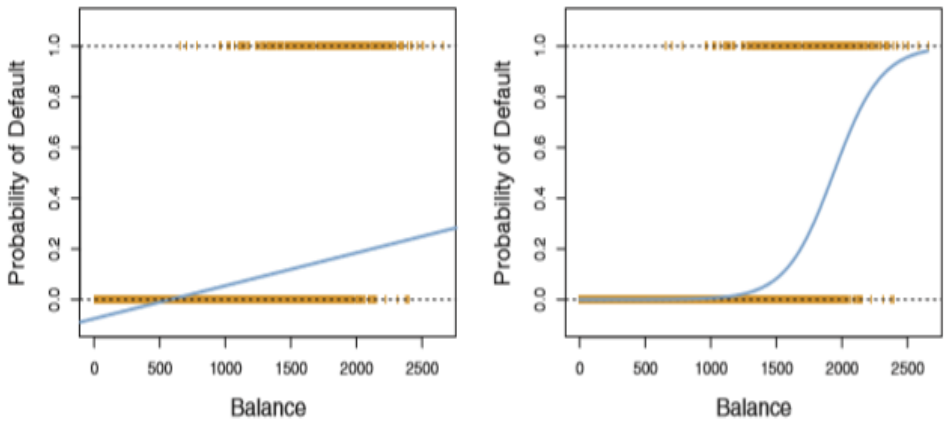

**Resultado da regressão pode exceder o intervalo (0 a 1)**

# "Regressão" Logística

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$




Sigmoid Function ou
Logistics Function




Logistic Regression: $0 \leq h_\theta(x) \leq 1$

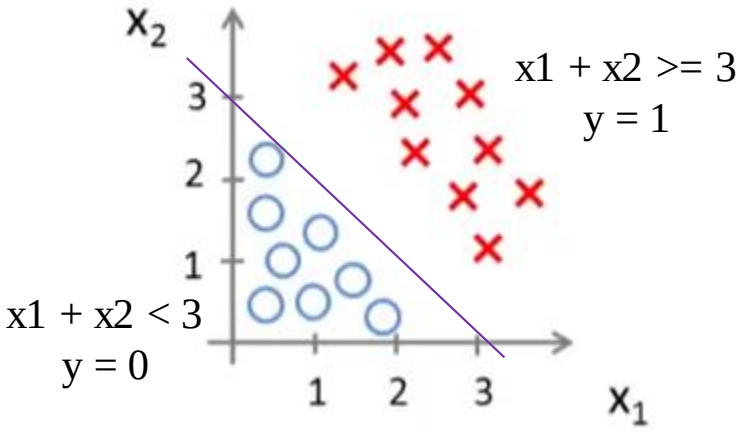
$$h_\theta(x) \geq 0.5 \rightarrow y = 1$$
$$h_\theta(x) < 0.5 \rightarrow y = 0$$

$$g(z) \geq 0.5$$
$$when\ z \geq 0$$

# "Regressão" Logística

$$h_\theta(x) \geq 0.5 \rightarrow y = 1$$
$$h_\theta(x) < 0.5 \rightarrow y = 0$$

$$g(z) \geq 0.5$$
$$when\ z \geq 0$$




x1 + x2 >= 3
y = 1

x1 + x2 < 3
y = 0

x1 + x2 = 3
h(x) = 0.5

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

O0 = -3
O1 = 1
O2 = 1

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

x1 + x2 >= 3

# "Regressão" Logística

- Simples x Múltipla: uma mesma variável pode ter um efeito sozinha e outra quando combinada (sessão 4.3.3 página 148)

| | Coefficient |
|---|---|
| Intercept | −3.5041 |
| student[Yes] | 0.4049 |

| | Coefficient |
|---|---|
| Intercept | −10.8690 |
| balance | 0.0057 |
| income | 0.0030 |
| student[Yes] | −0.6468 |

This is an important distinction for a credit card company that is trying to determine to whom they should offer credit. A student is riskier than a non-student if no information about the student's credit card balance is available. However, that student is less risky than a non-student with the same credit card balance!

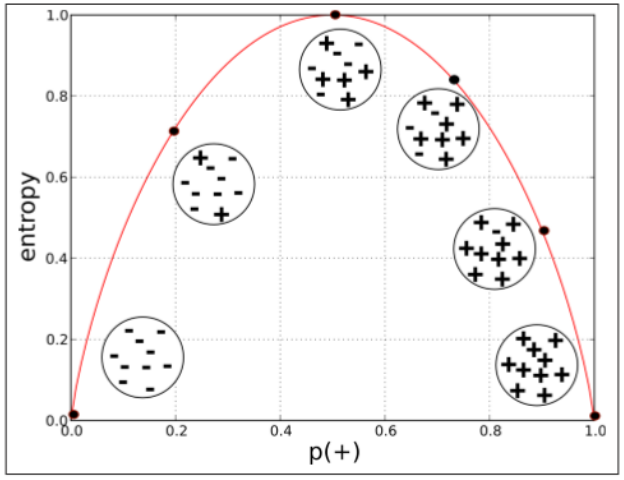
# Regressão Logística

- Mais um dia no MLBB (Machine Learning Bank of Boston) e devido ao aumento na demanda por imóveis na região o mercado de crédito está em alta e é preciso alocar seus empréstimos da melhor maneira possível. O problema é que os pedidos são tantos que estão sobrecarregando os analistas de crédito do banco. Para resolver essa situação você decide desenvolver um modelo para automatizar o processo de aprovação de crédito. Sua mais nova oportunidade de apresentar o modelo é no pedido de crédito feito por um grupo brasileiro de Real Estate que está se expandindo na região.

| Feature | Descrição | Tipo |
|---|---|---|
| GENDER | Gênero | Flag |
| AGE | Idade | Real |
| DEBT | Dívidas | Real |
| MARRIED | Estado Civil | Flag |
| BANK_CUSTOMER | Cliente do Banco | Flag |
| EDUCATION_LEVEL | Nível de Educação (Médio, Superior, etc..) | Categorical |
| ETHNICITY | Etnia | Categorical |
| YEARS_EMPLOYED | Anos de Trabalho | Real |
| PRIOR_DEFAULT | Histórico de Calote/Atraso | Flag |
| EMPLOYED | Situação Empregatícia | Flag |
| CITIZEN | Cidadão USA | Flag |
| ZIPCODE | Localidade | Categorical |
| INCOME | Renda | Real |
| APPROVED | Crédito Aprovado ou Não | Target |

# Árvore de Decisão

- Fácil interpretação, fácil visualização
- Não precisa de variável *dummy* para preditores qualitativos
- Critério de divisão: Ganho de informação - IG (Shannon, 1948)
- Baseado em Entropia como uma medida de desordem
- Desordem: quanto mais mesclado, maior a entropia

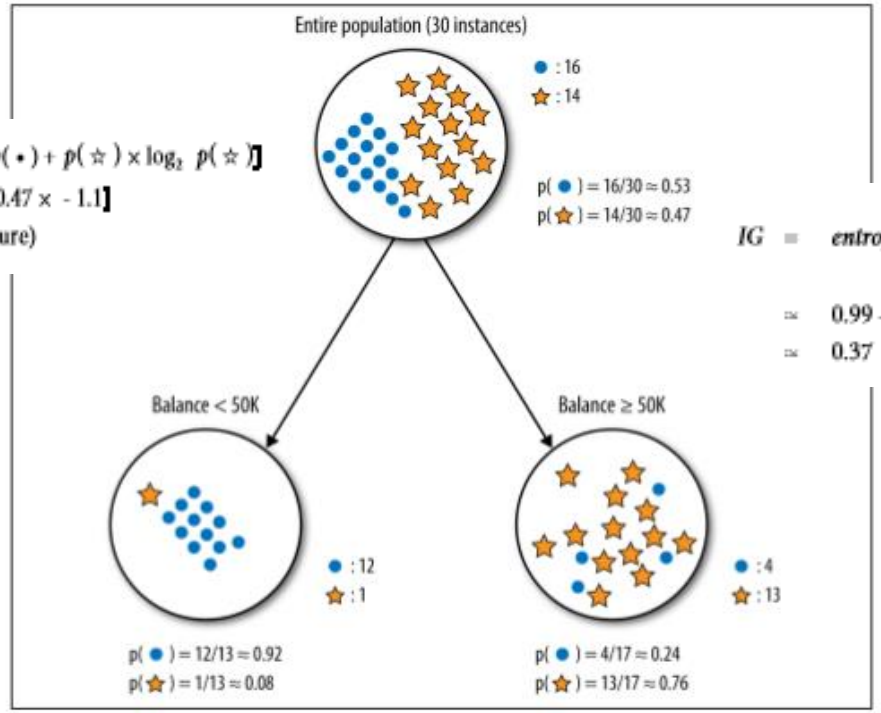



$$entropy = - p_1 \log(p_1) - p_2 \log(p_2) - \cdots$$

$$IG(parent, children) = entropy(parent) - [p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + \cdots]$$

# Árvore de Decisão

- Algoritmo: Se a divisão reduzir a entropia (ou aumenta o ganho de informação) então siga com a divisão




$$entropy(parent) = -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)]$$
$$\simeq -[0.53 \times -0.9 + 0.47 \times -1.1]$$
$$\simeq 0.99 \quad \text{(very impure)}$$

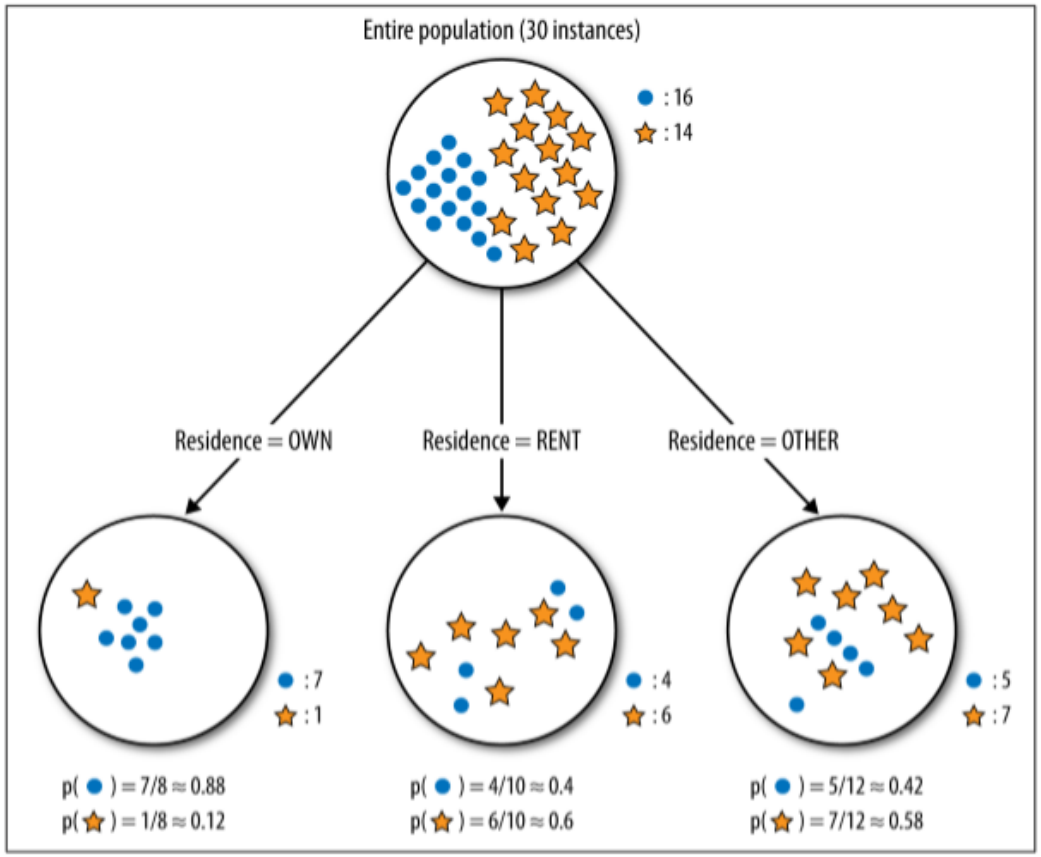
$$IG = entropy(parent) - [p(\text{Balance} < 50K) \times entropy(\text{Balance} < 50K) + p(\text{Balance} \geq 50K) \times entropy(\text{Balance} \geq 50K)]$$
$$\simeq 0.99 - [0.43 \times 0.39 + 0.57 \times 0.79]$$
$$\simeq 0.37$$

$$entropy(Balance < 50K) = -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)]$$
$$\simeq -[0.92 \times (-0.12) + 0.08 \times (-3.7)]$$
$$\simeq 0.39$$

$$entropy(Balance \geq 50K) = -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)]$$
$$\simeq -[0.24 \times (-2.1) + 0.76 \times (-0.39)]$$
$$\simeq 0.79$$

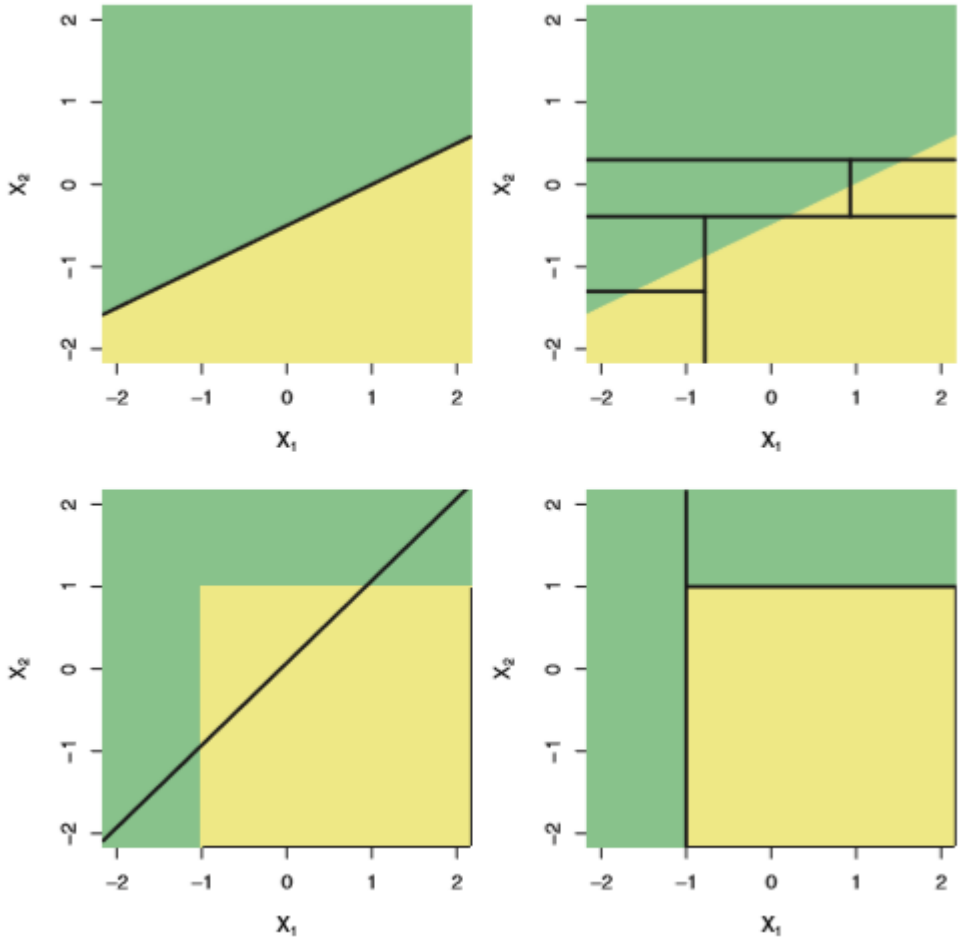
# Árvore de Decisão

- Algoritmo: Se a divisão reduzir a entropia (ou aumenta o ganho de informação) então siga com a divisão




$$entropy(parent) \approx 0.99$$
$$entropy(\text{Residence=OWN}) \approx 0.54$$
$$entropy(\text{Residence=RENT}) \approx 0.97$$
$$tropy(\text{Residence=OTHER}) \approx 0.98$$
$$IG \approx 0.13$$

# Árvore de Decisão x Modelo Linear



# Árvore de Decisão

- Não costumam ter alta acurácia
- Muito voláteis -> uma mudança simples no dado pode mudar completamente a árvore

**Solução**: Combinação

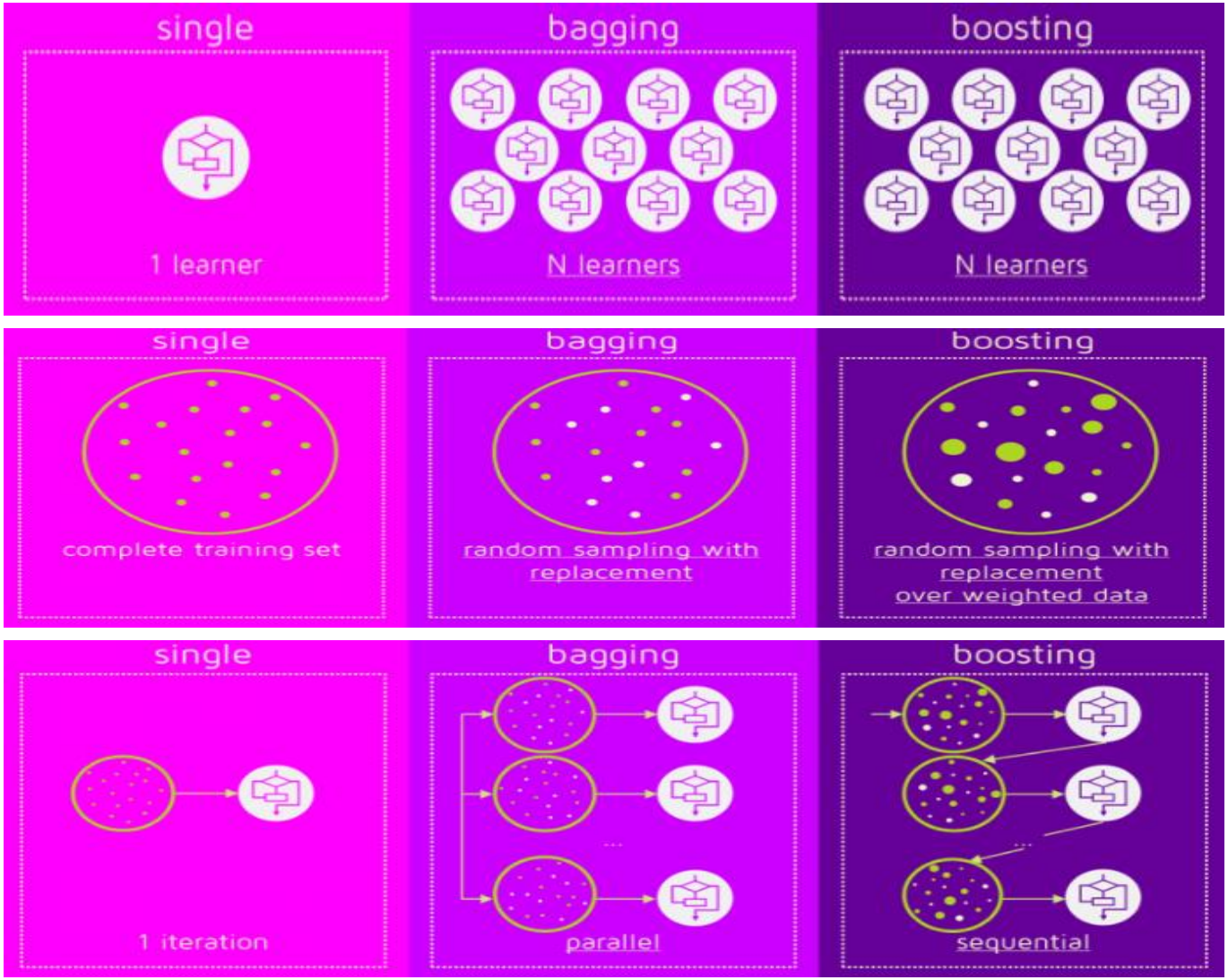
"Set of weak learners are combined to create a strong learner"

**Boosting** -> cria 1 árvore por vez, vai melhorando a próxima utilizando os erros da árvore anterior

**Bagging** (**B**ootstrap **AGG**regat**ING**) -> utiliza n amostras aleatórias de treinamento. Considera a média das árvores

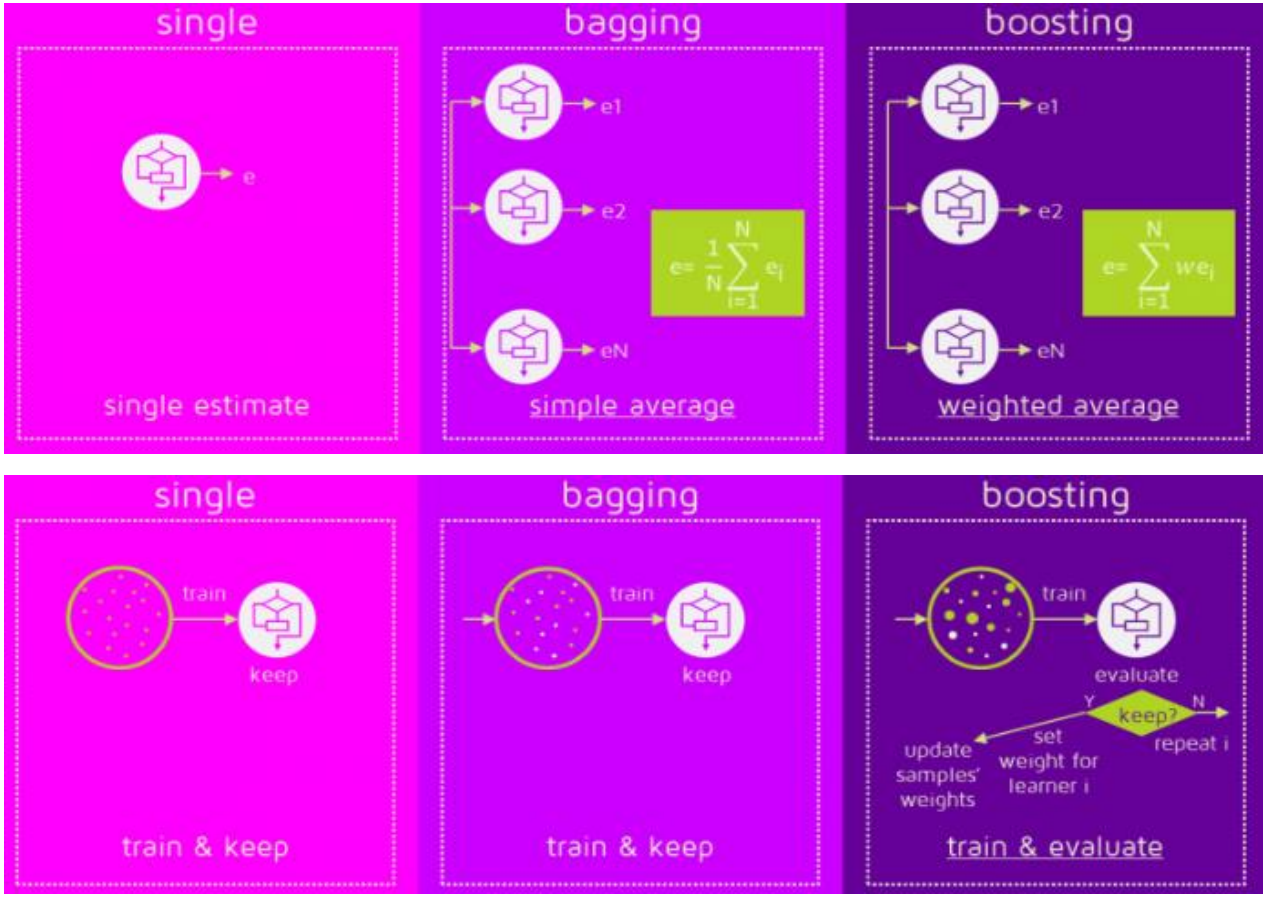
**Random Forest** -> igual Bagging, mas varia a quantidade de features

# Combinação




https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

# Combinação




Variações: AdaBoost, LPBoost, XGBoost, GradientBoost, BrownBoost.

https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

# Árvore de Decisão

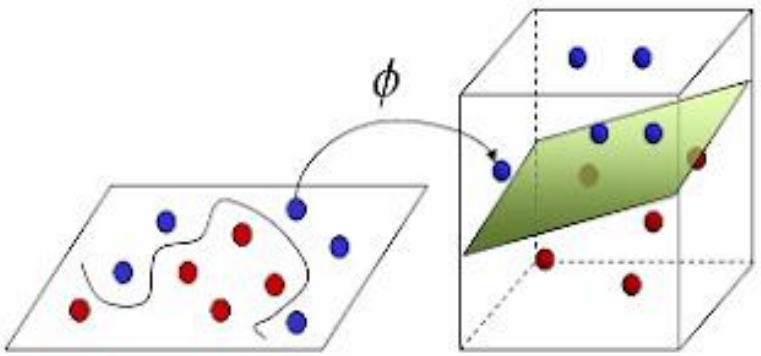
- Com o crescimento urbano acelerado da cidade de Boston e o aquecimento da economia local os serviços públicos estão sobrecarregados. O hospital Boston D'Or decidiu investir em automatizar o processo de direcionamento de pacientes para os especialistas corretos, para isso será desenvolvido um modelo de árvore de decisão baseado em um formulário padrão preenchido pelos paciente.

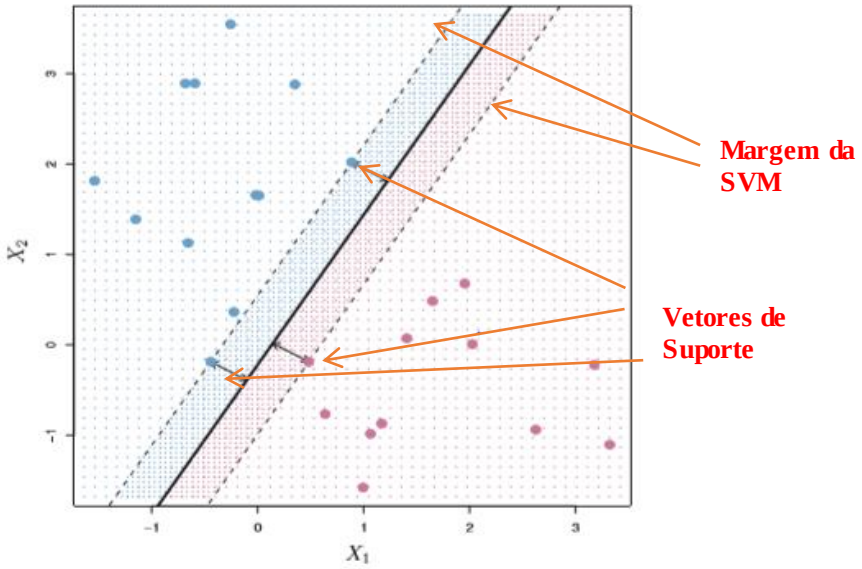
| Feature | Descrição | Tipo |
|---|---|---|
| Q1 | Pergunta 1 de um Questionário Médico | Categorical |
| Q2 | Pergunta 2 de um Questionário Médico | Categorical |
| Q3 | Pergunta 3 de um Questionário Médico | Categorical |
| Q4 | Pergunta 4 de um Questionário Médico | Categorical |
| Q5 | Pergunta 5 de um Questionário Médico | Categorical |
| Q6 | Pergunta 6 de um Questionário Médico | Categorical |
| Q7 | Pergunta 7 de um Questionário Médico | Categorical |
| Q8 | Pergunta 8 de um Questionário Médico | Categorical |
| Categoria | Classificação de Triagem Médica | Target |

# SVM – Máquina de Vetores de Suporte

- Hiperplano

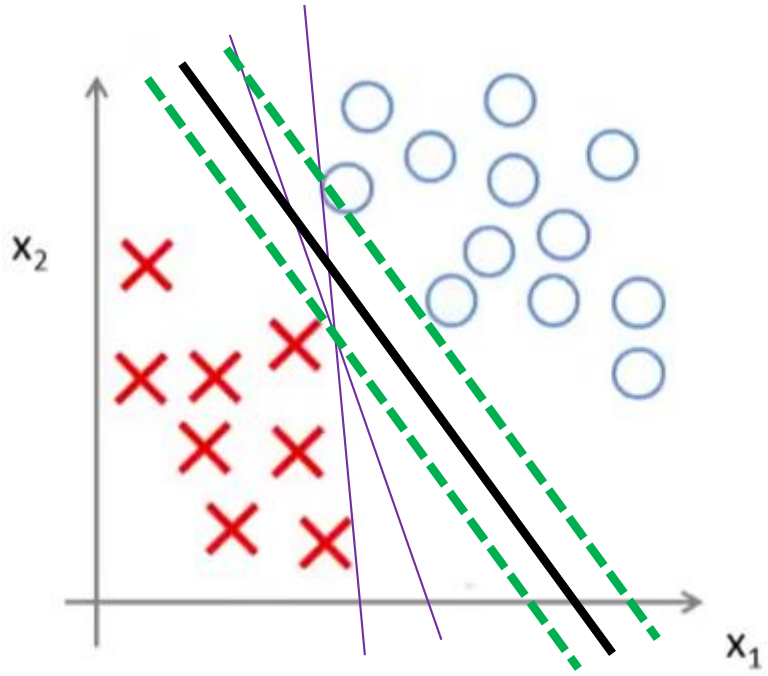
*figura geométrica de curvatura nula em um espaço euclidiano n-dimensional e cuja equação em coordenadas cartesianas é linear.*




O hiperplano depende diretamente dos vetores de suporte, mas não das outras observações: um movimento para qualquer uma das outras observações não afetaria o hiperplano de separação, desde que o movimento da observação não o faça cruzar o limite definido pela margem.

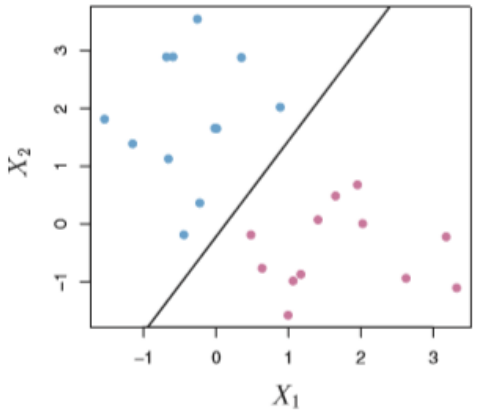
# SVM – Máquina de Vetores de Suporte

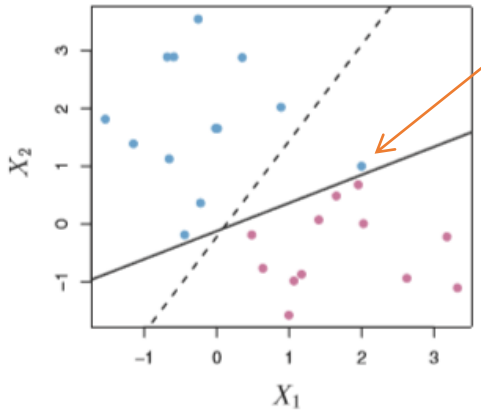



Modelo mais robusto

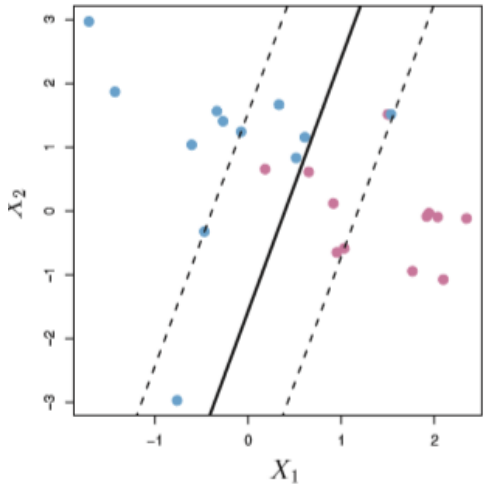
Tenta separar os conjuntos o máximo possível através da margem

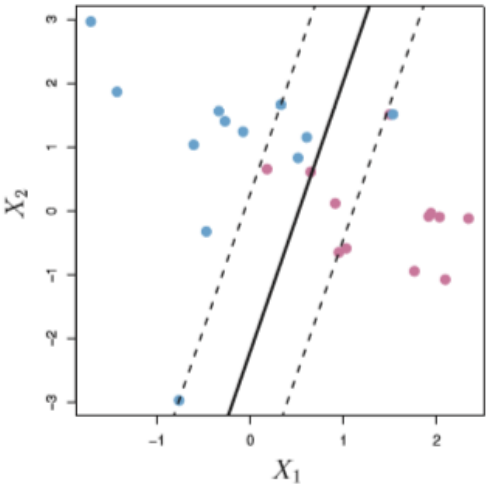
# SVM – Máquina de Vetores de Suporte




Uma única adição da amostra altera significativamente a reta: overfitting

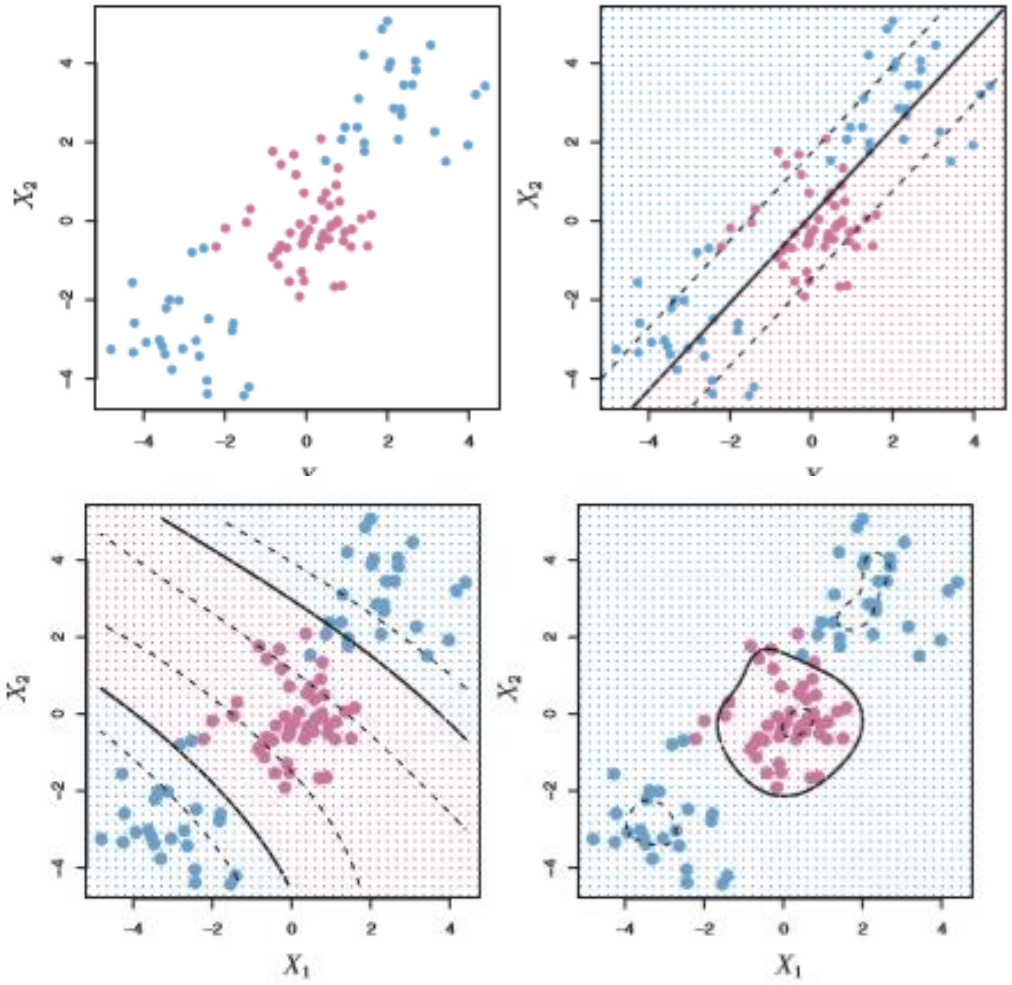



Modelo mais robusto

Tunings diferentes de SVMs
Balanceamento de Bias x Variance

# SVM – Máquina de Vetores de Suporte




SVM utilizando Kernel Polinomial

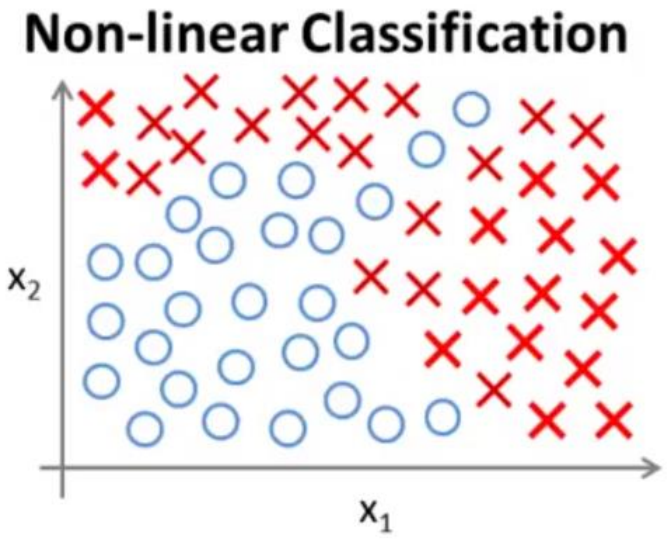
SVM utilizando Kernel Radial

# SVM

- O fundo de investimento Bettinas está investindo em uma nova solução baseada em machine learning para automatizar e aperfeiçoar a classificação do "credit rating" de companhias públicas baseado em um conjunto de métricas fundamentalistas. Assim o fundo pretende assumir o mínimo risco necessário para bater os fundos concorrentes e seu benchmark o CDI.

| Feature | Descrição | Tipo |
|---|---|---|
| P.E | Price to Earning per Share | Real |
| P.B | Price to Book Value Ratio | Real |
| EPS | Earnings per Share | Real |
| P.S | Price to Sales Ratio | Real |
| US | US Listed ? | Flag |
| PEG | Price to Earnings Growth Ratio | Real |
| RATING | Security Score | Target |

# Redes Neurais – Por que usar ?




Regressão Logística multi polinomial

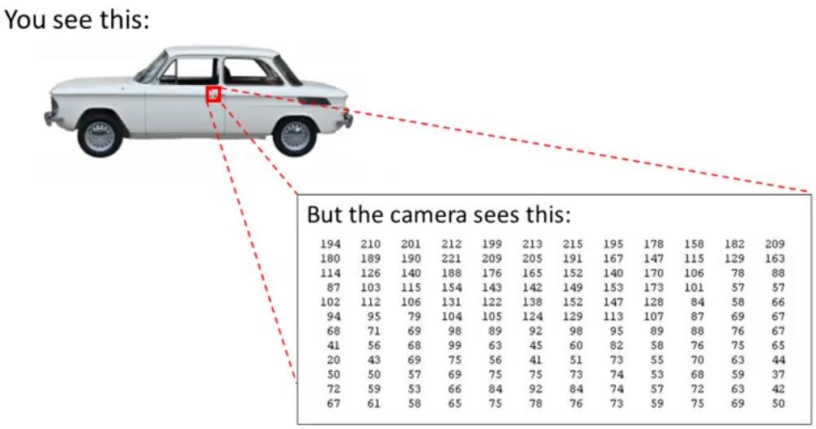
$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^3 x_2 + \theta_6 x_1 x_2^2 + \dots)$$

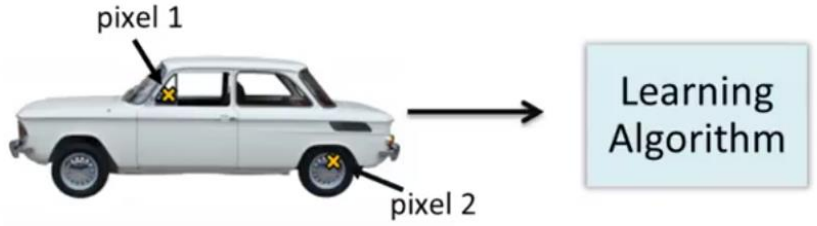
E se tivermos mais do que somente X1 e X2 ? (mundo real)

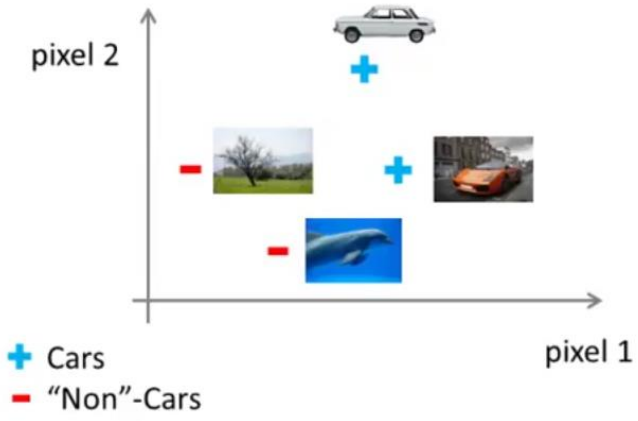
Para 100 deles, chegaríamos a mais de 5000 termos !!!

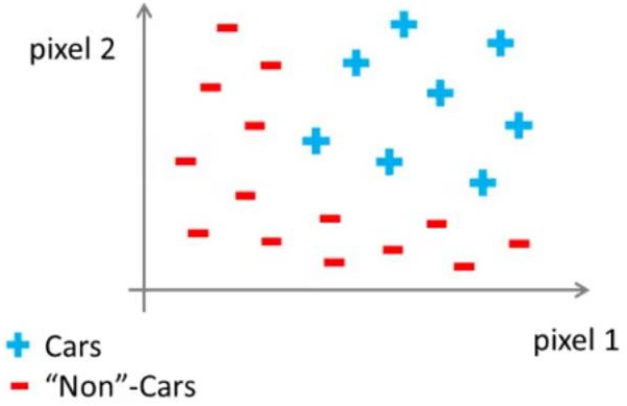
https://www.coursera.org/learn/machine-learning/

# Redes Neurais – Por que usar ?



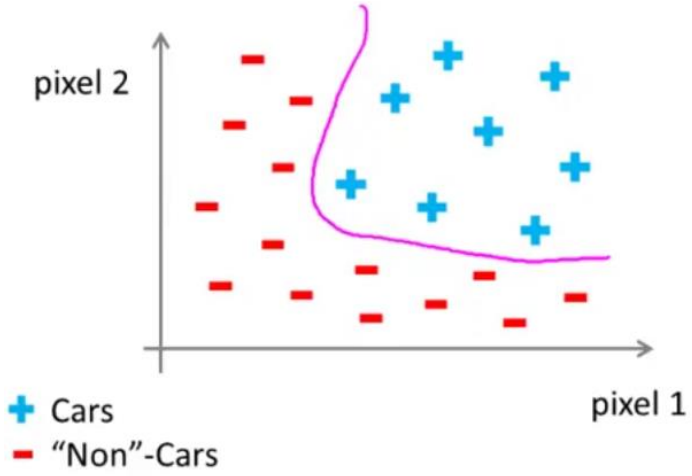




https://www.coursera.org/learn/machine-learning/

# Redes Neurais – Por que usar ?




50 x 50 pixel images→ 2500 pixels

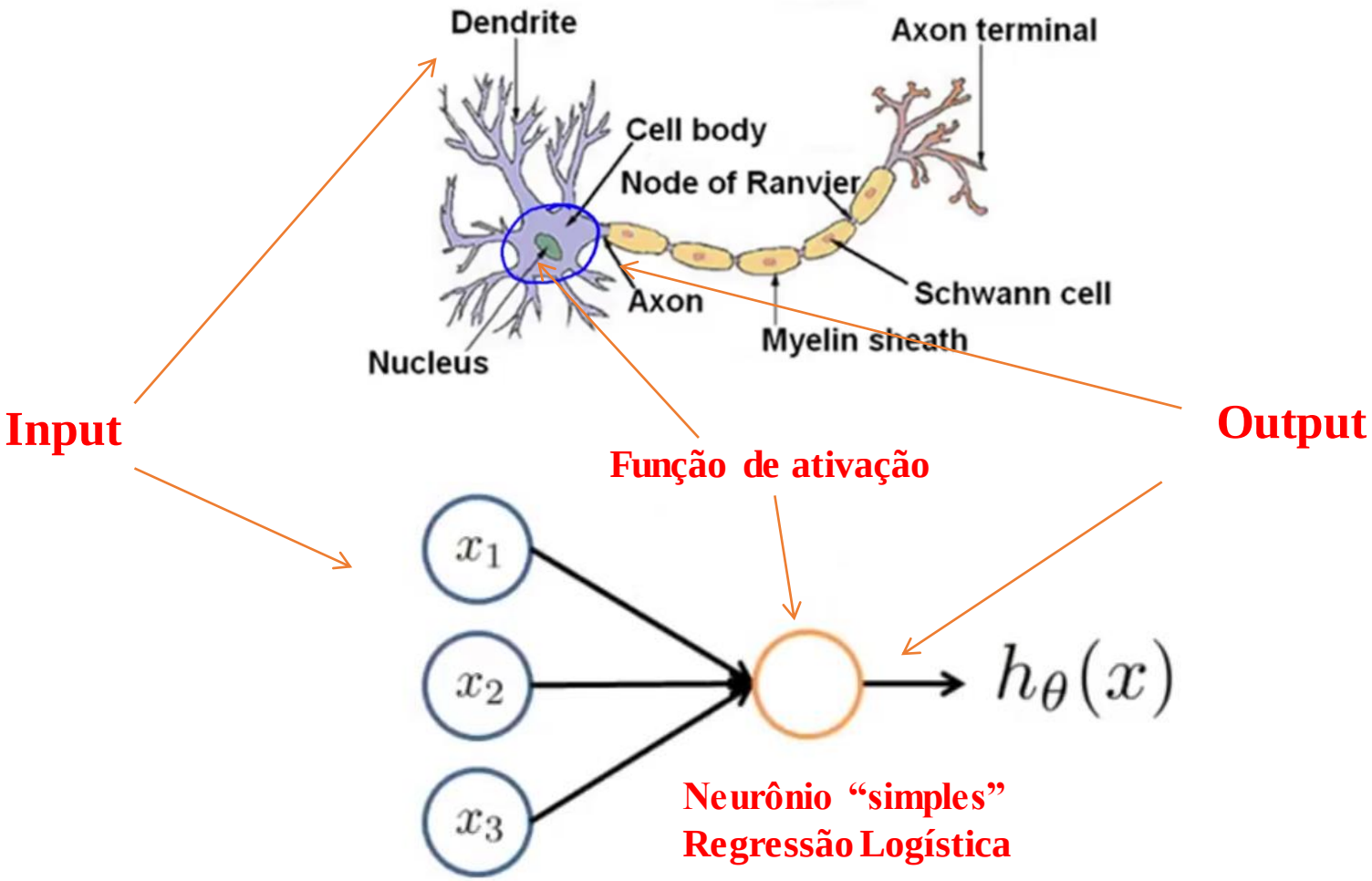
$n = 2500$ (7500 if RGB)

$$x = \begin{bmatrix} \text{pixel 1 intensity} \\ \text{pixel 2 intensity} \\ \vdots \\ \text{pixel 2500 intensity} \end{bmatrix}$$

Quadratic features ($x_i \times x_j$): ≈3 million features

https://www.coursera.org/learn/machine-learning/

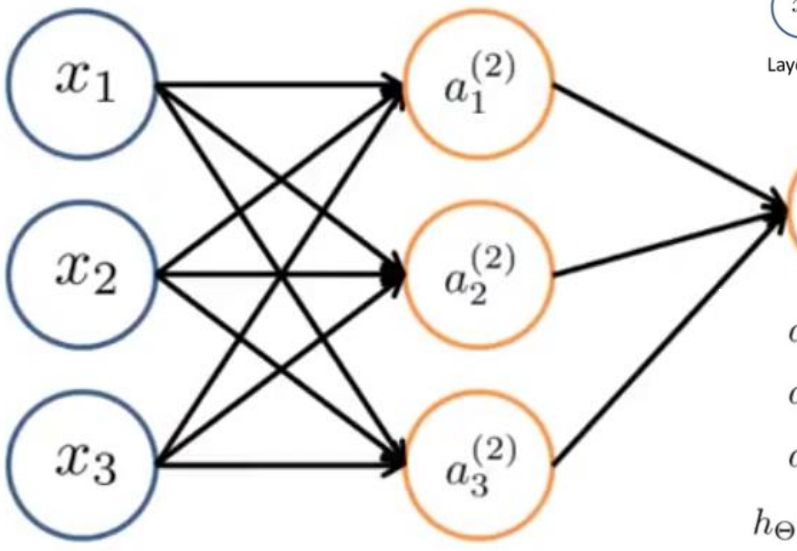
# Redes Neurais



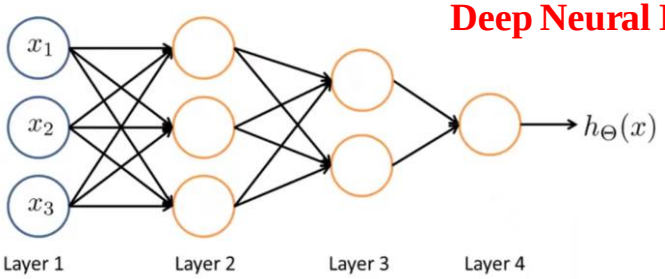

https://www.coursera.org/learn/machine-learning/

# Redes Neurais

## Rede de Neurônios




## Deep Neural Networks




$$a_1^{(2)} = g(\Theta_{10}^{(1)} x_0 + \Theta_{11}^{(1)} x_1 + \Theta_{12}^{(1)} x_2 + \Theta_{13}^{(1)} x_3)$$
$$a_2^{(2)} = g(\Theta_{20}^{(1)} x_0 + \Theta_{21}^{(1)} x_1 + \Theta_{22}^{(1)} x_2 + \Theta_{23}^{(1)} x_3)$$
$$a_3^{(2)} = g(\Theta_{30}^{(1)} x_0 + \Theta_{31}^{(1)} x_1 + \Theta_{32}^{(1)} x_2 + \Theta_{33}^{(1)} x_3)$$
$$h_\Theta(x) = a_1^{(3)} = g(\Theta_{10}^{(2)} a_0^{(2)} + \Theta_{11}^{(2)} a_1^{(2)} + \Theta_{12}^{(2)} a_2^{(2)} + \Theta_{13}^{(2)} a_3^{(2)})$$

**Forward Propagation**

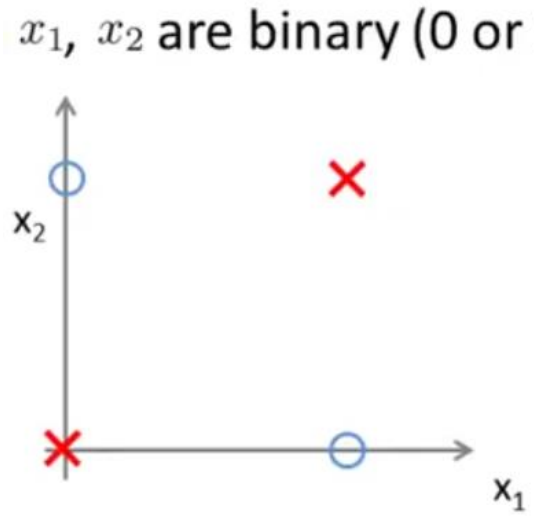
**Input Layer**

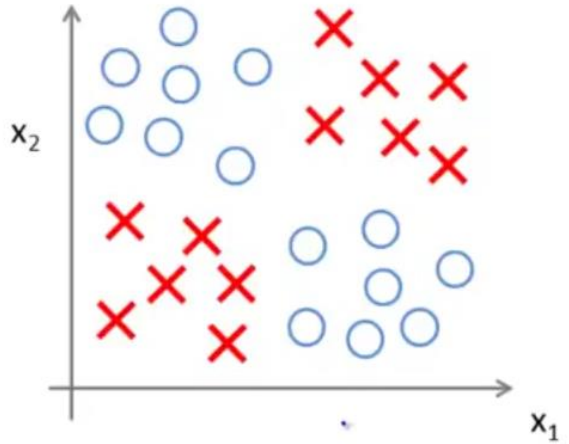
**Hidden Layer**

**Output Layer**

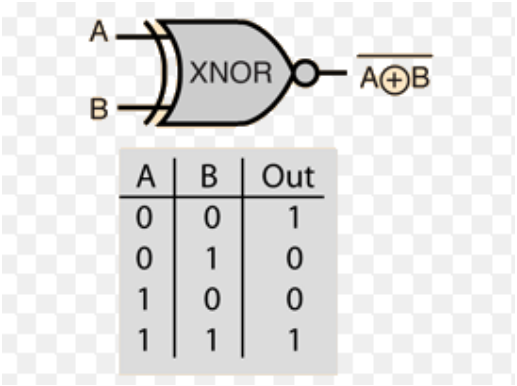
https://www.coursera.org/learn/machine-learning/

# Redes Neurais




**x1 XNOR x2**

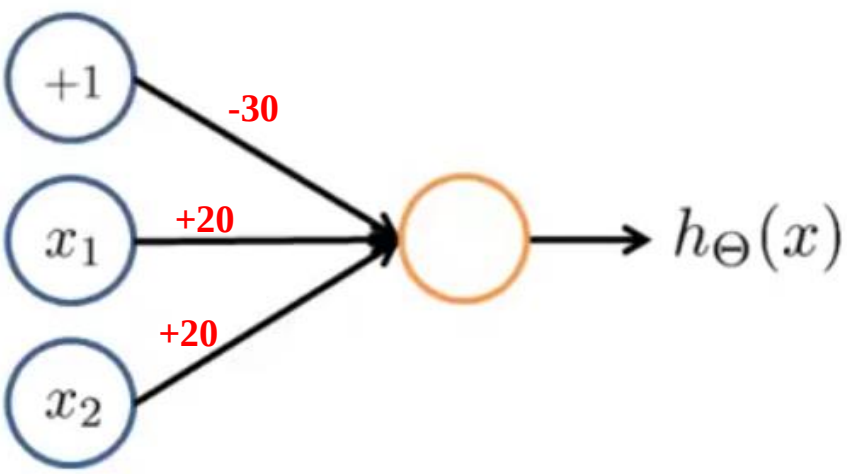


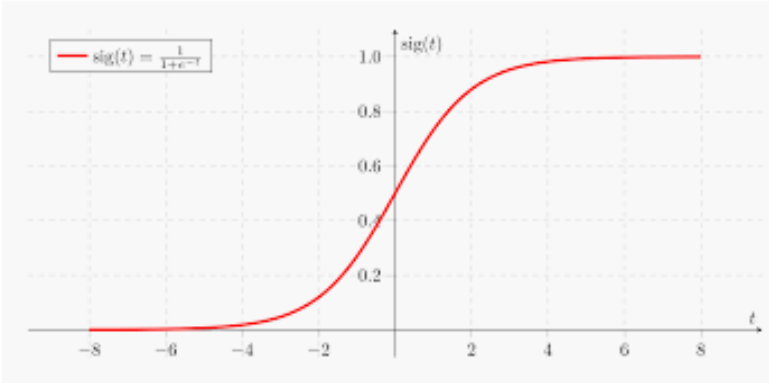

| A | B | Out |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

https://www.coursera.org/learn/machine-learning/

# Redes Neurais

y = x1 AND x2




h(x) = g (-30 + 20*x1 + 20*x2)

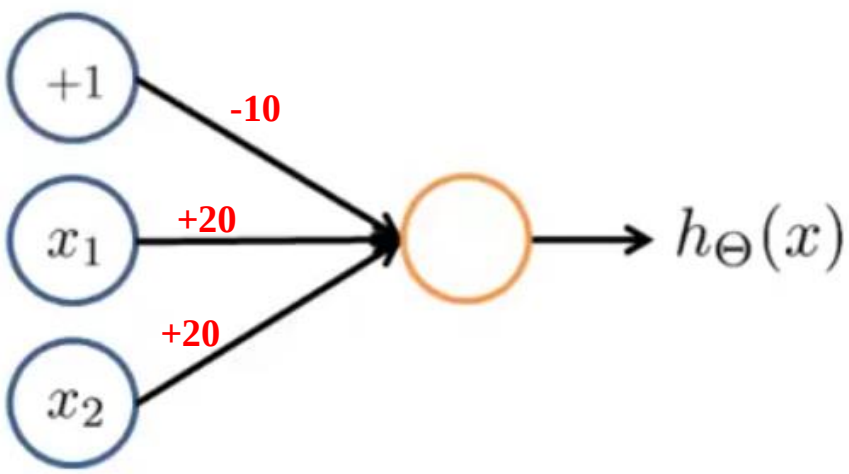
| x1 | x2 | h(x) |
|---|---|---|
| 0 | 0 | g(-30) -> 0 |
| 0 | 1 | g(-10) -> 0 |
| 1 | 0 | g(-10) -> 0 |
| 1 | 1 | g(10) -> 1 |

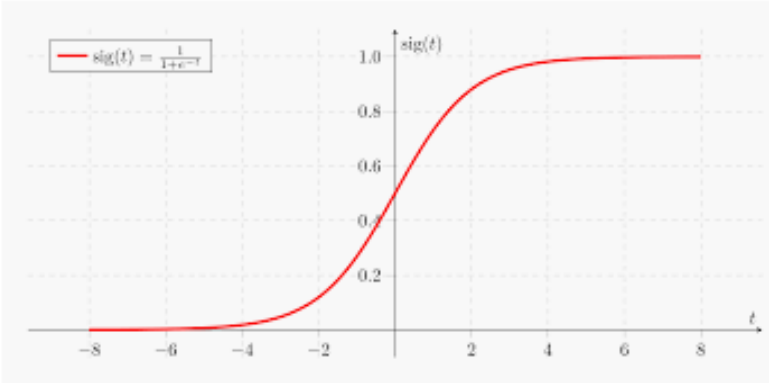
h(x) = x1 AND x2

https://www.coursera.org/learn/machine-learning/

# Redes Neurais

y = x1 OR x2




h(x) = g (-10 + 20*x1 + 20*x2)

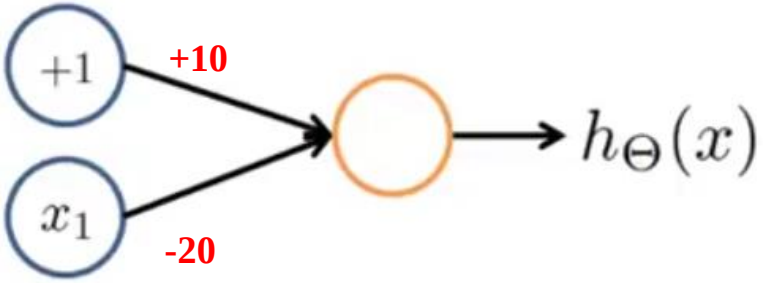
| x1 | x2 | h(x) |
|---|---|---|
| 0 | 0 | g(-10) -> 0 |
| 0 | 1 | g(10) -> 1 |
| 1 | 0 | g(10) -> 1 |
| 1 | 1 | g(30) -> 1 |

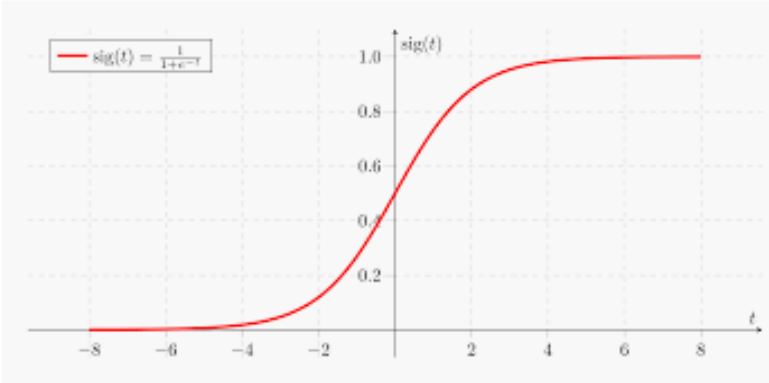
h(x) = x1 OR x2

https://www.coursera.org/learn/machine-learning/

# Redes Neurais

NOT X1



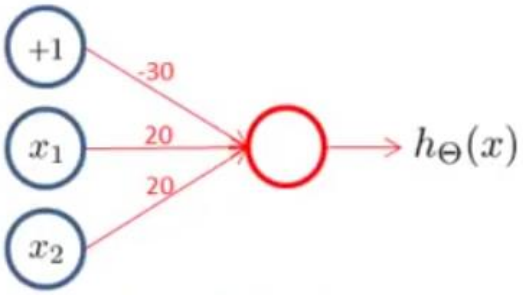

h(x) = g (10 + 20*x1)

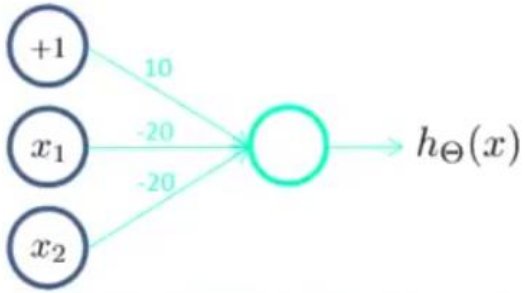
| x1 | h(x) |
|---|---|
| 0 | g(10) -> 1 |
| 1 | g(-10) -> 0 |

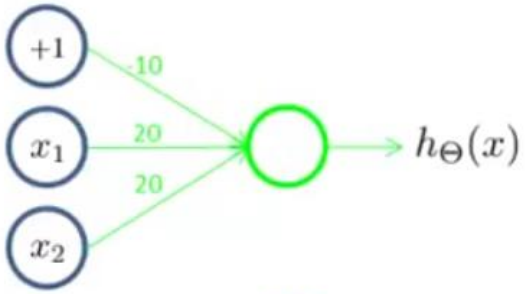
h(x) = NOT x1

https://www.coursera.org/learn/machine-learning/

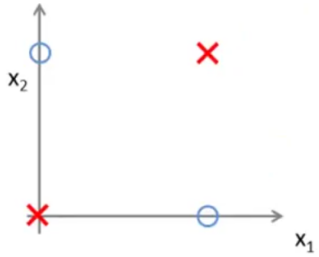
# Redes Neurais

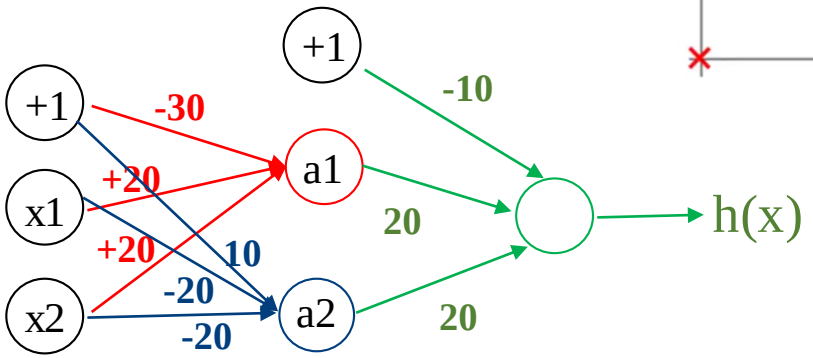



X1 XNOR X2




| x1 | x2 | a1 | a2 | h(x) |
|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 1 |

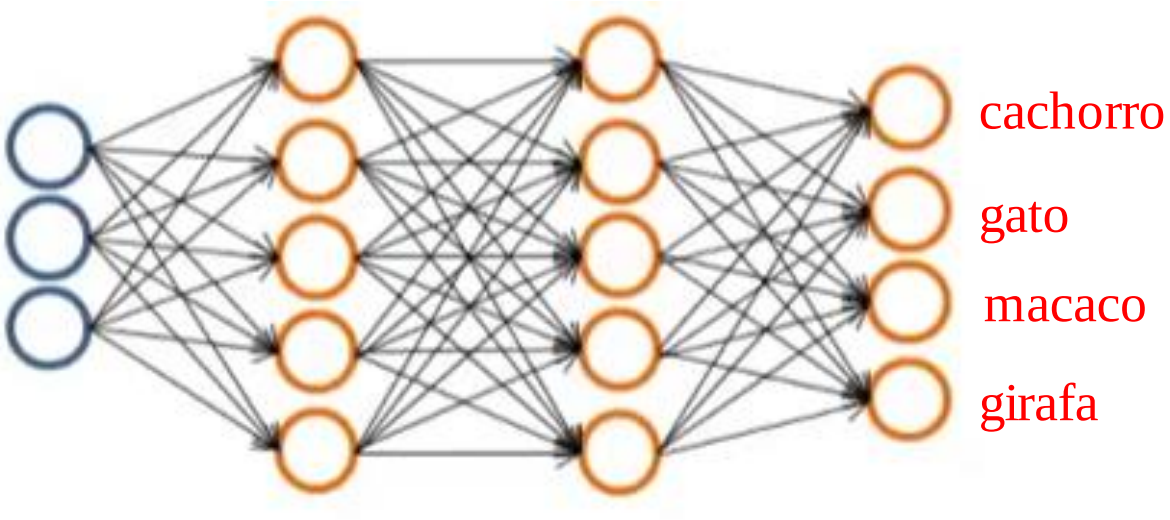
h(x) = X1 XNOR X2

https://www.coursera.org/learn/machine-learning/

# Redes Neurais

E se objetivo da predição for uma classe multi-variada ?

Ex: definir se a imagem é um cachorro, gato, macaco ou girafa ?

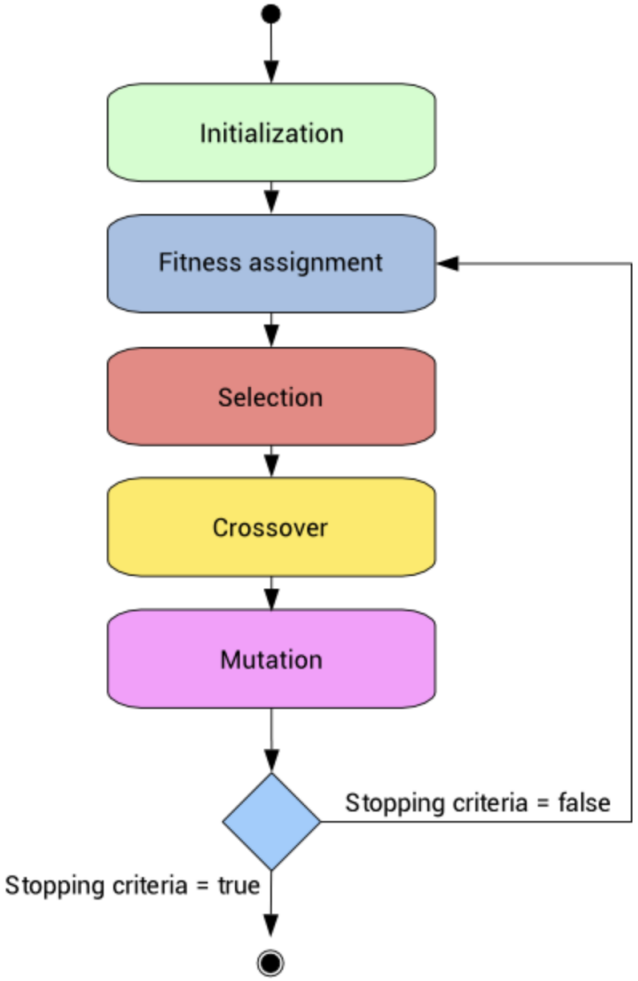


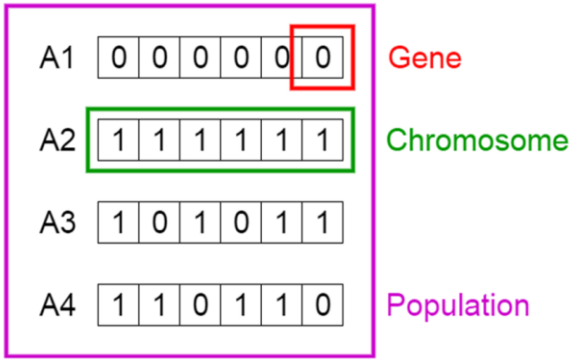

https://playground.tensorflow.org/

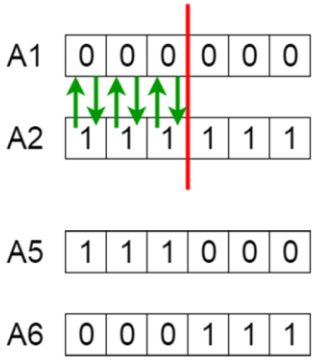
# Redes Neurais

- Com o recente sucesso do Machine Learning Bank of Boston no setor de crédito, o seu velho concorrente Goldman Data decide focar em captar novos clientes para se capitalizar. Para isso é preciso que o processo de abertura de conta seja o mais breve possível, assim surge a idéia de que o cadastramento de todos os dados seja feito baseado apenas no envio de uma foto do RG. Para isso você recebe a responsabilidade de desenvolver um OCR (Optical Character Recognition) para reconhecer dígitos e adicioná-los no formulário de cadastro do cliente.

# Algoritmo Genético (GA)



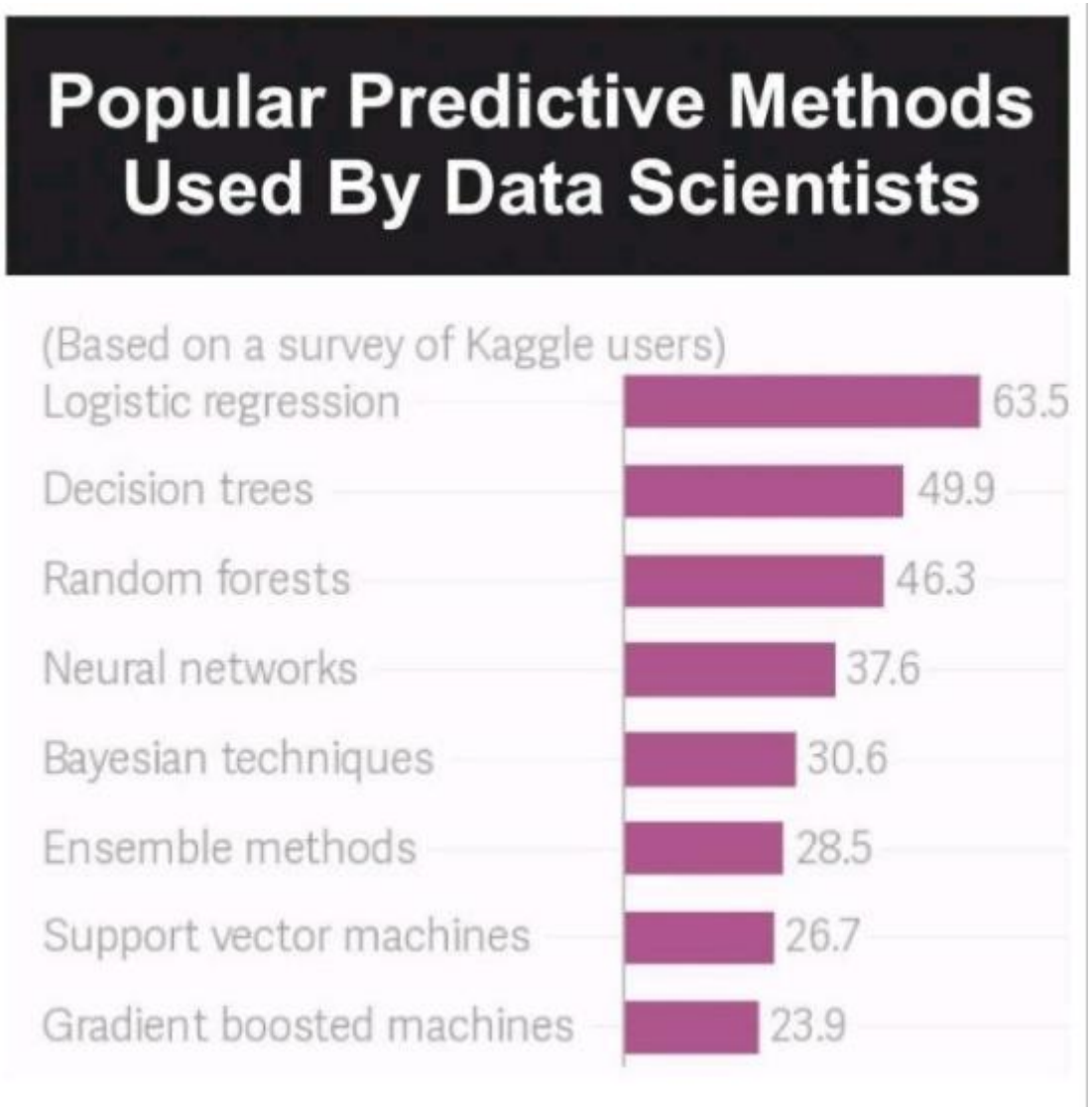



# Algoritmo Genético (GA)

- O grupo de e-commerce Amazonia está investido em soluções de data-driven marketing para customizar a experiência dos usuários em seu site. Para ser o mais eficiente possível o grupo de Data Scientist's da Amazonia decide desenvolver um modelo junto com a equipe de marketing para classificar a faixa de renda de um usuário qualquer. Assim na página inicial irão aparecer produtos de uma natureza mais premium ou genéricos.

| Feature | Descrição | Tipo |
|---|---|---|
| AGE | Idade | Real |
| WORKCLASS | Tipo de Emprego (Funcionário Público, Empresa Privada, ...) | Real |
| FNLWGT | "Final Weight" | Real |
| EDUCATION | Nível de Educação | Categorical |
| EDU_NUM | Anos de Educação | Real |
| MARITAL_STATUS | Estado Civíl | Categorical |
| OCCUPATION | Ocupação | Categorical |
| RELATIONSHIP | Relação Familiar | Categorial |
| RACE | Raça | Categorical |
| SEX | Gênero | Categorical |
| CAPITAL_GAIN | Ganhos de Capital Recentes | Real |
| CAPITAL_LOSS | Perdas de Capital Recentes | Real |
| HOURS_PER_WEEK | Horas de Trabalho por Semana | Real |
| COUNTRY | País de Origem | Categorical |
| INCOME | Faixa de Renda | Target |


Popular Predictive Methods Used By Data Scientists
(Based on a survey of Kaggle users)
Logistic regression
63.5
Decision trees
49.9
Random forests
46.3
Neural networks
37.6
Bayesian techniques
30.6
Ensemble methods
28.5
Support vector machines
26.7
Gradient boosted machines
23.9


CATHY O'NEIL
TED

# Case Final

- A multinacional de varejo Waldata está querendo expandir a sua presença na américa latina e por isso decide firmar uma parceria com a FGV para desenvolver um modelo preditivo do valor de vendas. Além disso a companhia decide apostar em um segundo modelo de análise de sentimentos baseado em câmeras escondidas nas vitrines que capturam o movimentos dos olhos. Assim a rede varejista pretende melhorar suas projeções de fluxo de caixa e otimizar a distribuição de seus produtos por departamentos.

- Para desenvolver seu modelo você irá realizar as seguintes tarefas:
  1) Importar os datasets RETAIL_1 e RETAIL_2 disponíveis em https://github.com/rbarsotti/MBA/archive/master.zip
  2) Importar os datasets para o ambiente R.
  3) Fazer uma exploração detalhada dos dados. (Distribuições, valores faltantes etc ..)
  4) Dividir as bases em 70% para treino e 30% para teste do modelo. (Utilize sempre seed(314))
  5) Testar modelos de classificação para a análise de sentimentos:
     1) Regressão Logística, Árvores de Decisão, SVM e Redes Neurais
  6) Testar modelos de regressão para o valor de vendas das lojas:
     1) Regressão Linear, Árvore de Decisão, e Redes Neurais
  7) Validar a performance dos modelos ($R^2$ & Matriz de Confusão)
  8) Fazer o "scoring" dos modelos para os dados nas respectivas bases de teste