Machine Learning

Seja bem-vindo!

Data Science Academy

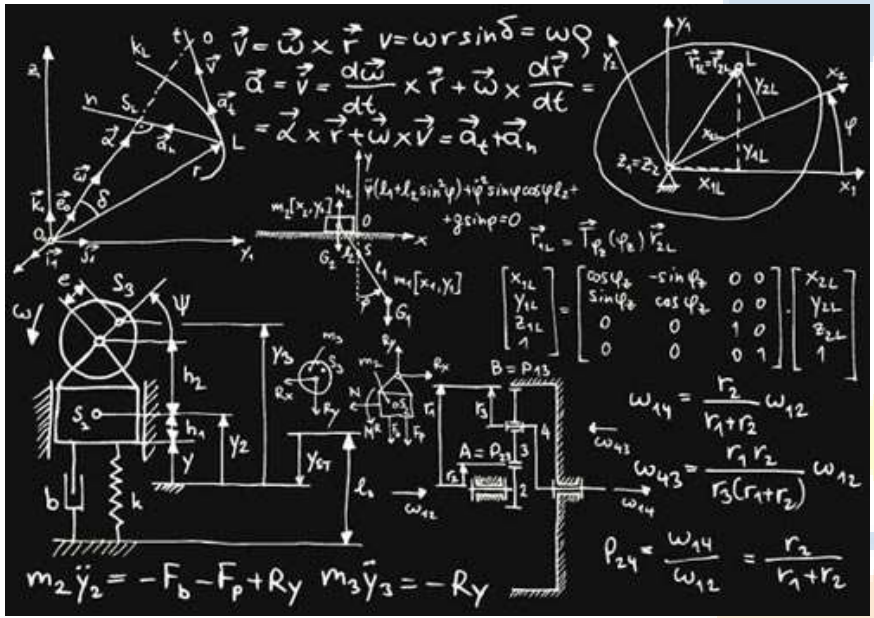
Como Funciona a Aprendizagem de Máquina

Data Science Academy

Processo de Aprendizagem

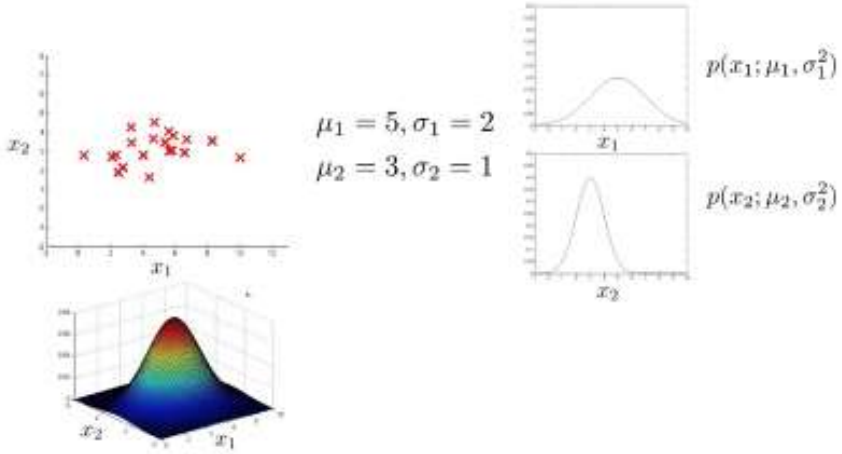
Como os algoritmos aprendem

Processo de Aprendizagem



Processo de Aprendizagem

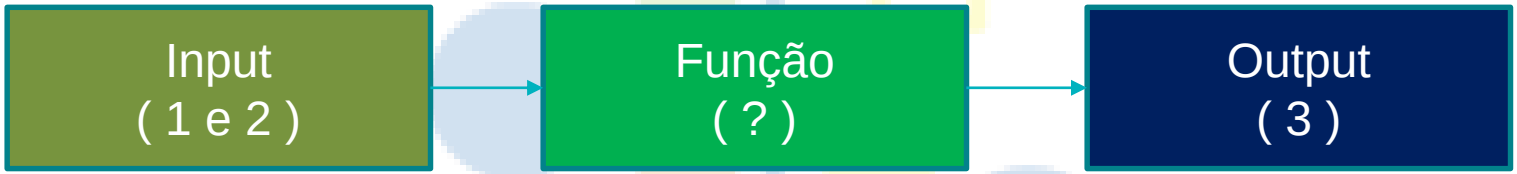
Função alvo – f(x)



Processo de Aprendizagem

Processo de Aprendizagem

Machine Learning



E como um algoritmo encontra a função matemática que descreve este relacionamento?

Isso é o que vamos responder ao longo dos próximos vídeos!

Data Science Academy

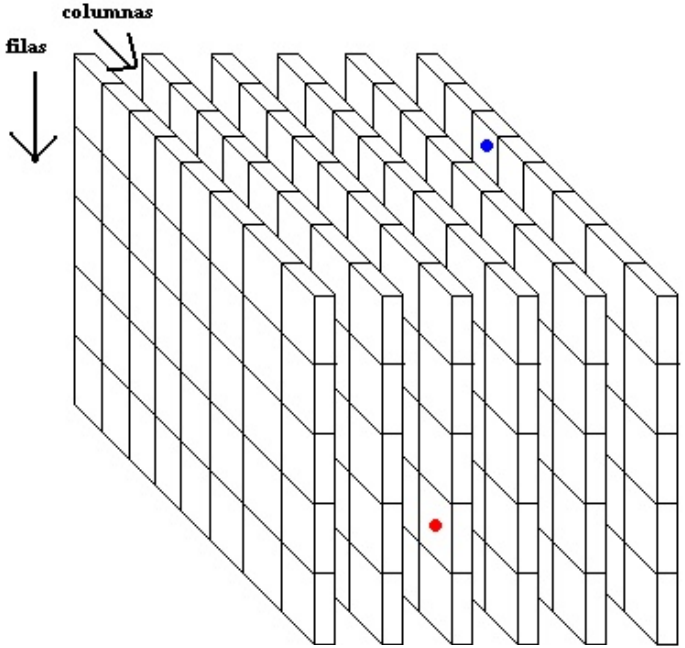
Processo de Aprendizagem

Um componente chave do processo de aprendizagem é a generalização

Se um algoritmo de Machine Learning não for capaz de generalizar uma função matemática que faça previsões sobre novos conjuntos de dados, ele não está aprendendo nada e sim memorizando os dados, o que é bem diferente.

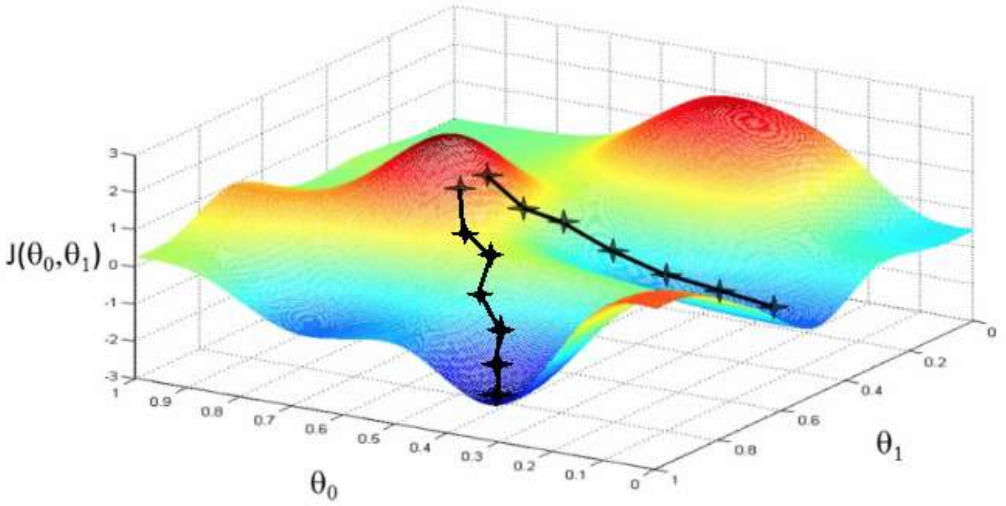
E para poder generalizar a função que melhor resolve o problema, os algoritmos de Machine Learning se baseiam em 3 componentes:

- Representação
- Avaliação
- Otimização

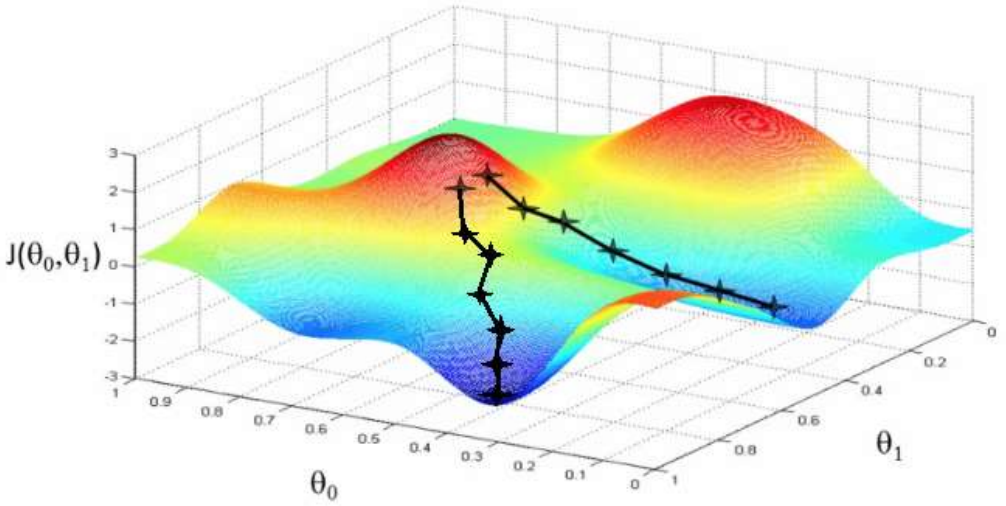


Os algoritmos de aprendizagem possuem diversos parâmetros internos

Otimização



Espaço de Hipótese



Nenhum algoritmo único ou uma combinação de algoritmos é 100% preciso o tempo todo.

Pelo menos não ainda!!

Big Data é uma grande mistura de dados.

Um bom algoritmo de Machine Learning deve ser capaz de distinguir os sinais e mapear as funções alvo de forma eficiente.

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

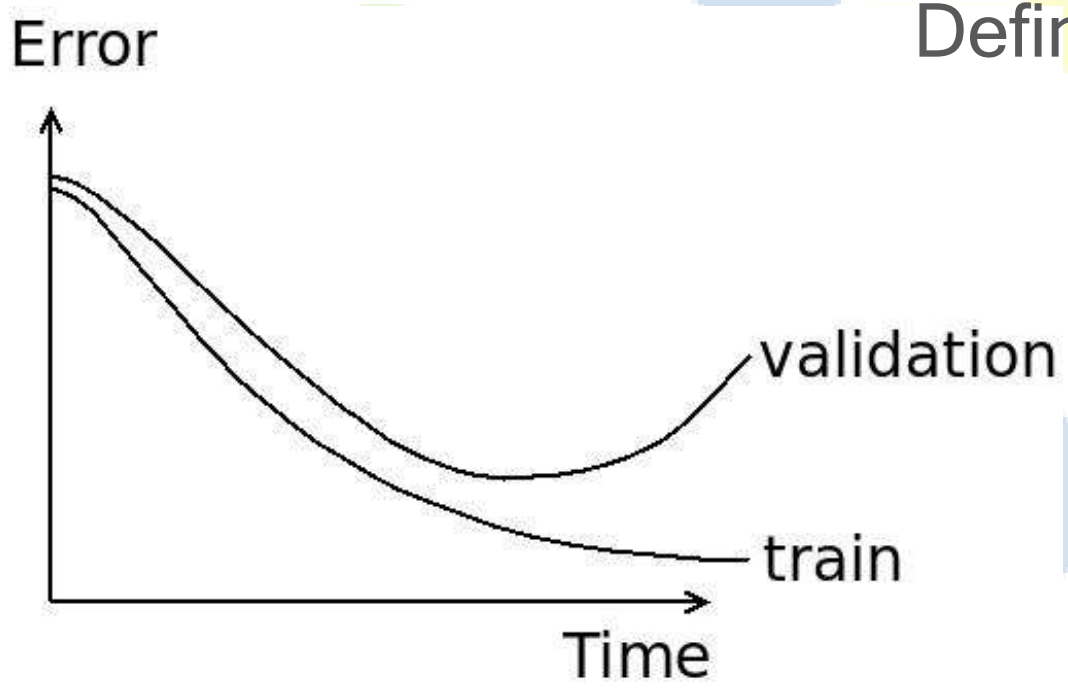
Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Cost Function

Definindo o Erro



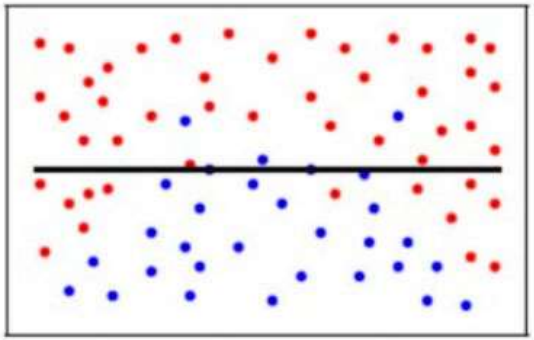
Cost Function → Nível de erro

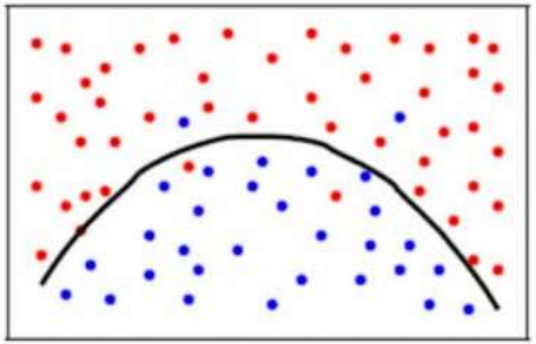
Underfitting x Overfitting

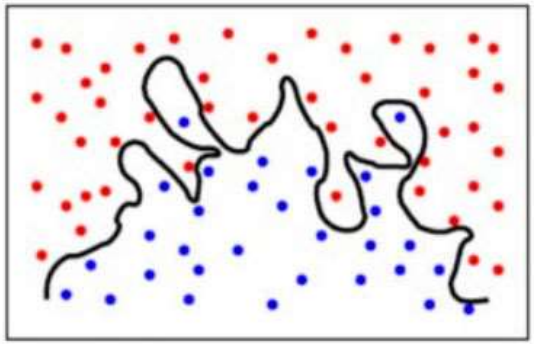
Underfitting

Ok

Overfitting







Simplicidade

Complexidade

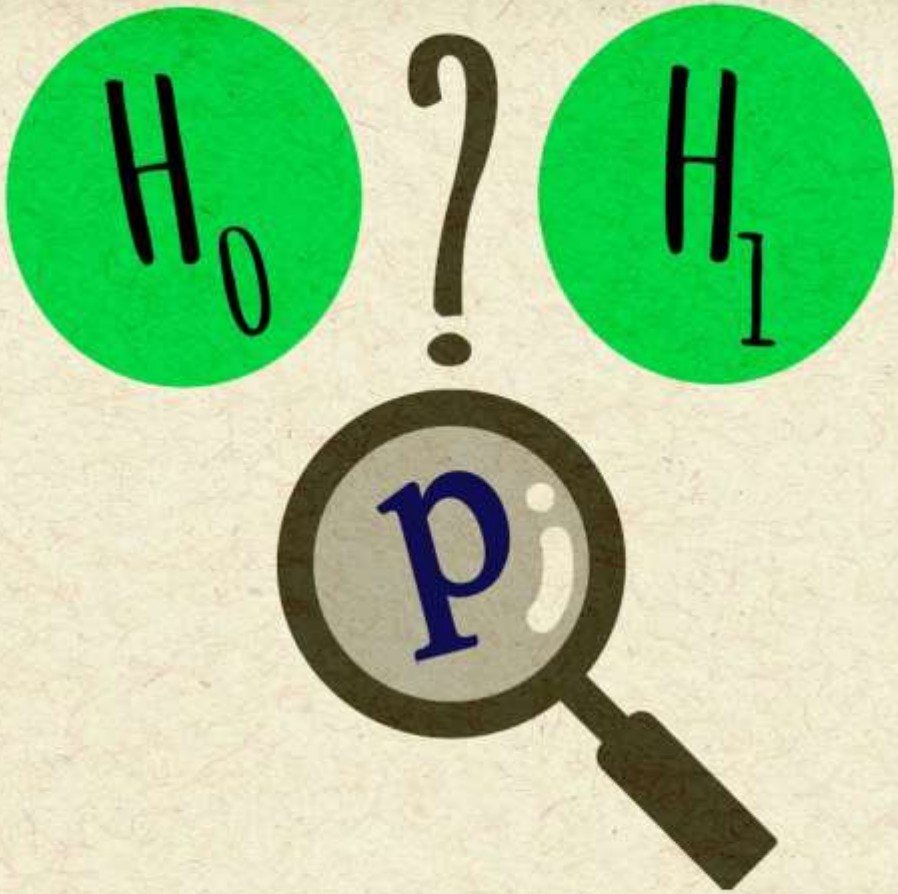
Ponto Ideal

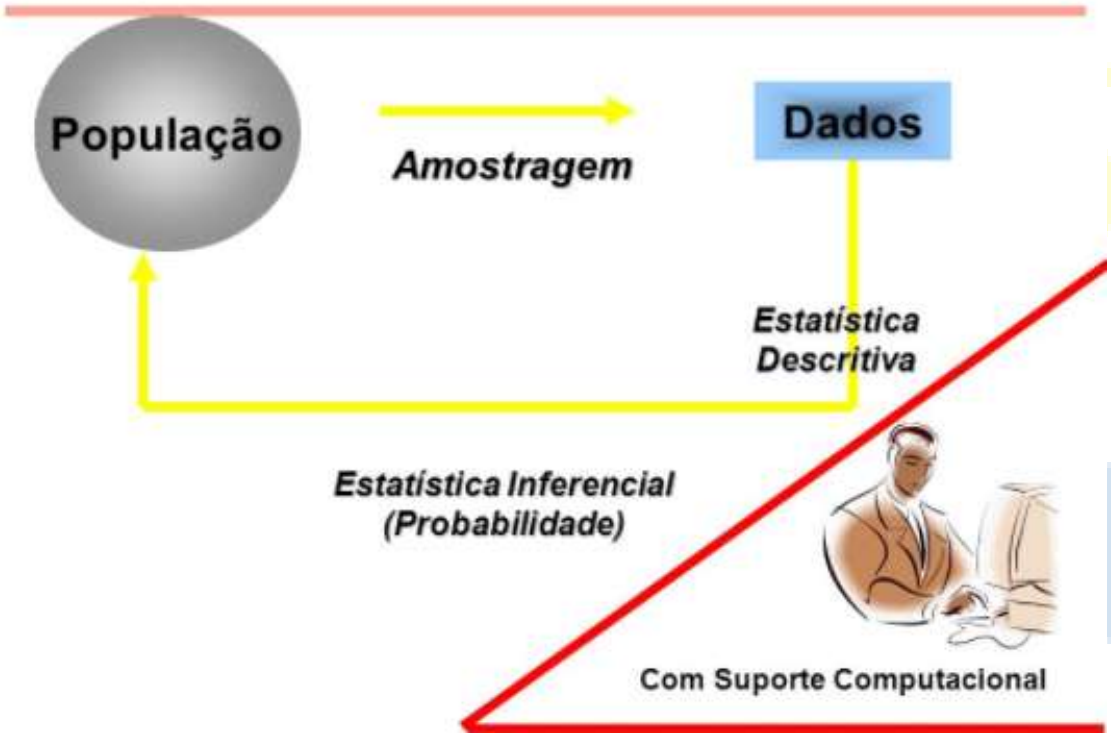
Data Science Academy

Teste de Hipóteses

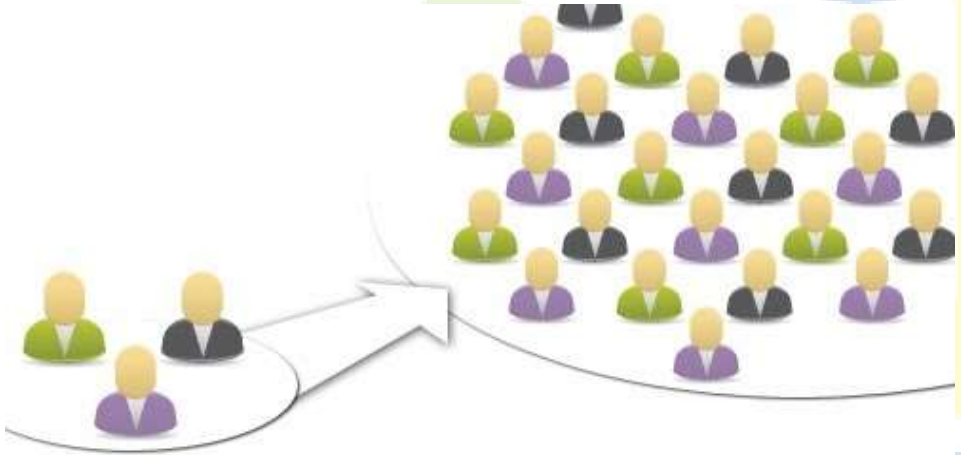
Data Science Academy

TESTES DE HIPÓTESE





Inferência Estatística



Formulada uma determinada hipótese particular é necessário coletar dados e com base nestes dados decide-se então sobre a validade ou não da hipótese.

Mas o que é exatamente uma hipótese?

Uma hipóteses estatística é uma afirmação sobre o parâmetro, ou parâmetros, da distribuição de probabilidades de uma característica, X, de uma população.

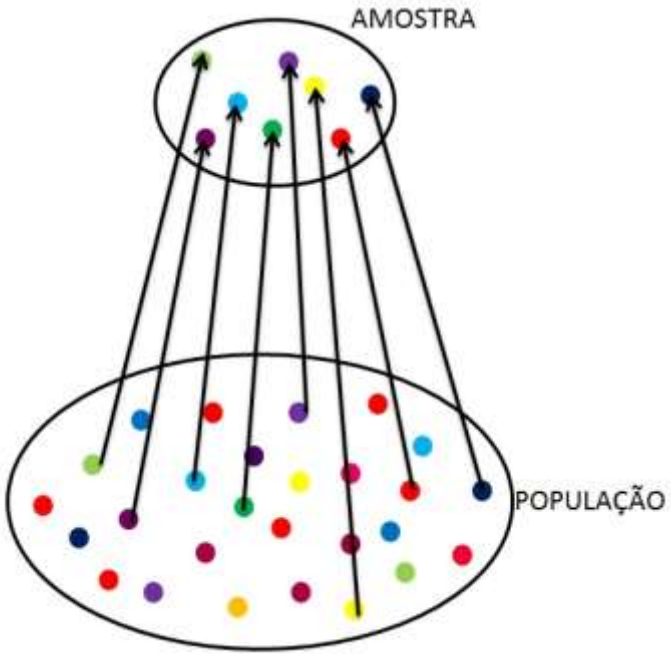
Mas o que é exatamente uma hipótese?

- H0 = Hipótese Nula
- H1 = Hipótese Alternativa

Hipótese Nula
H0

Hipótese Alternativa
H1

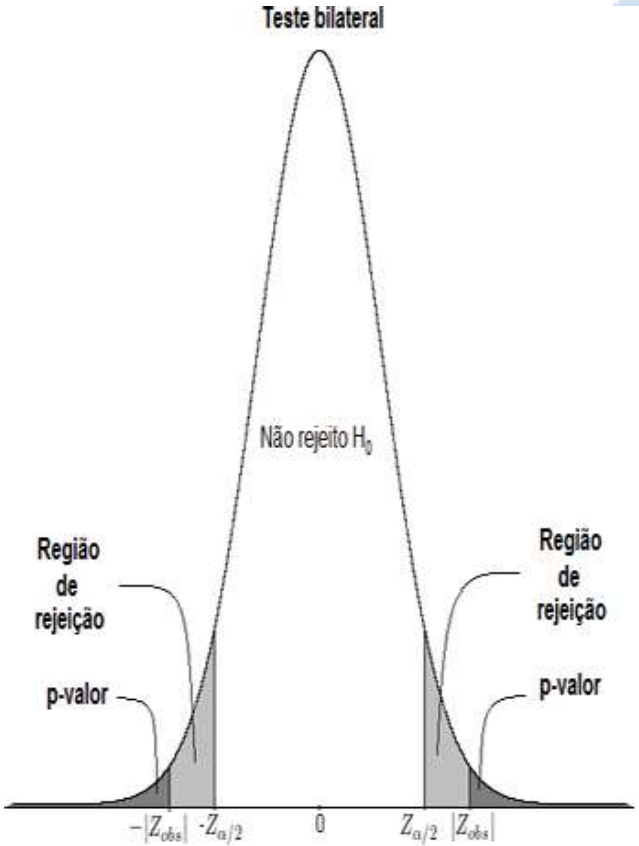
Quando as hipóteses são formuladas sobre os parâmetros do modelo probabilístico da população o Teste de Hipóteses é chamado de **Paramétrico**. Quando as hipóteses são formuladas sobre outras características do modelo o Teste é chamado de **Não Paramétrico**.

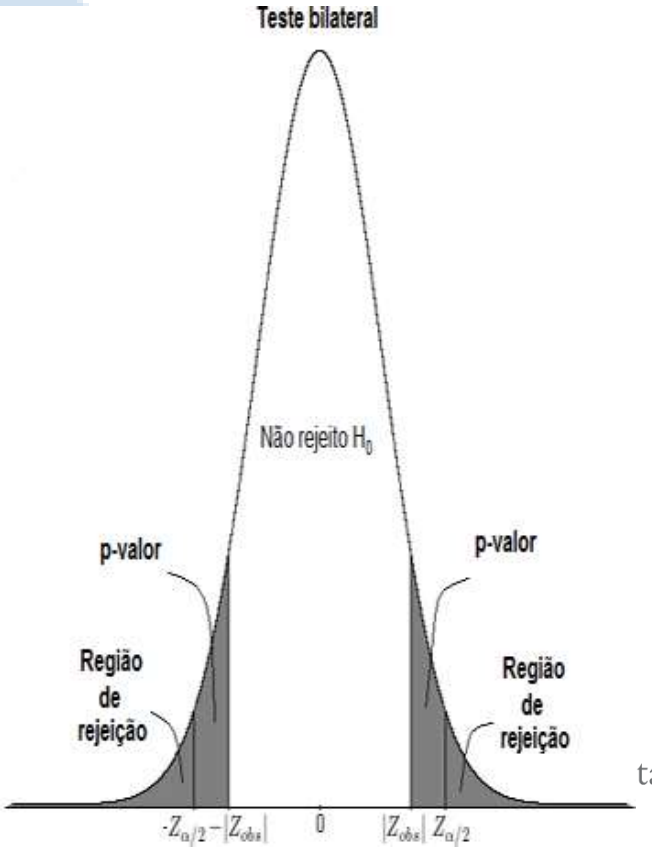


Para testar a hipótese é coletada uma amostra aleatória representativa da população, sendo calculadas as estatísticas necessárias para o teste.

Se a diferença entre o que foi observado na amostra e o que era esperado (sob a condição da hipótese verdadeira) não for SIGNIFICATIVA a hipótese será **aceita**

Se a diferença entre o que foi observado na amostra e o que era esperado (sob a condição da hipótese verdadeira) for SIGNIFICATIVA a hipótese será **rejeitada**.

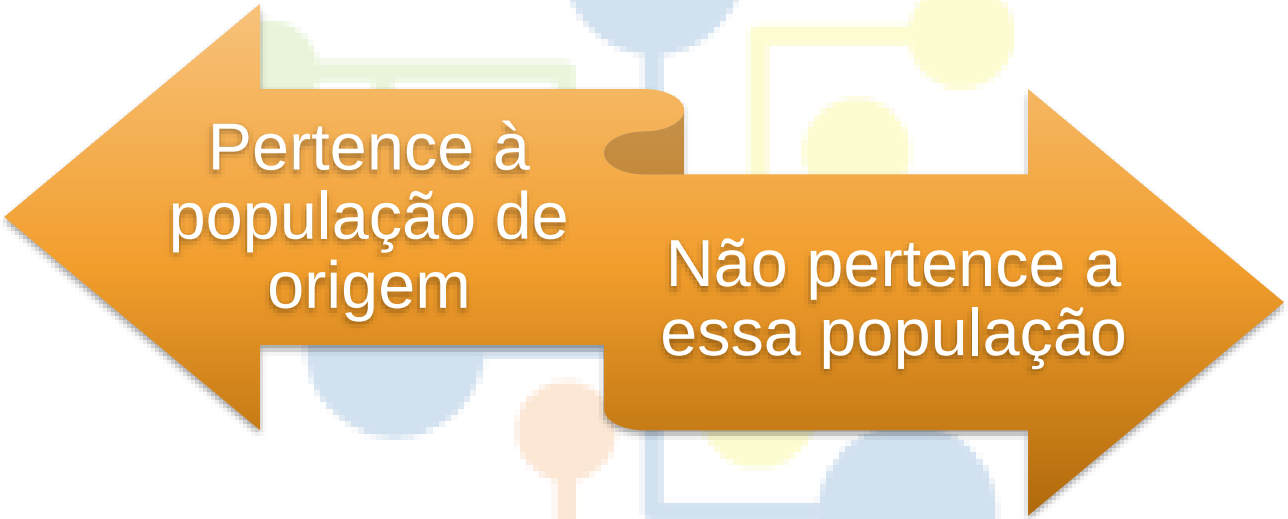

Teste bilateral
Não rejeito H_0
Região de rejeição
p-valor
Região de rejeição
p-valor
$-|Z_{obs}|$
$-Z_{\alpha/2}$
0
$Z_{\alpha/2}$
$|Z_{obs}|$


Teste bilateral
Não rejeito H_0
p-valor
Região de rejeição
p-valor
Região de rejeição
$-Z_{\alpha/2}$
$-|Z_{obs}|$
0
$|Z_{obs}|$
$Z_{\alpha/2}$

H0 = O réu é inocente

H1 = O réu é culpado

O Nível de Significância será a probabilidade assumida de rejeitar H0 sendo H0 verdadeira.


Pertence à população de origem
Não pertence a essa população

Tipos de Testes Paramétricos

Testes Unilaterais

H0 : parâmetro = valor de teste

H1 : parâmetro < valor de teste

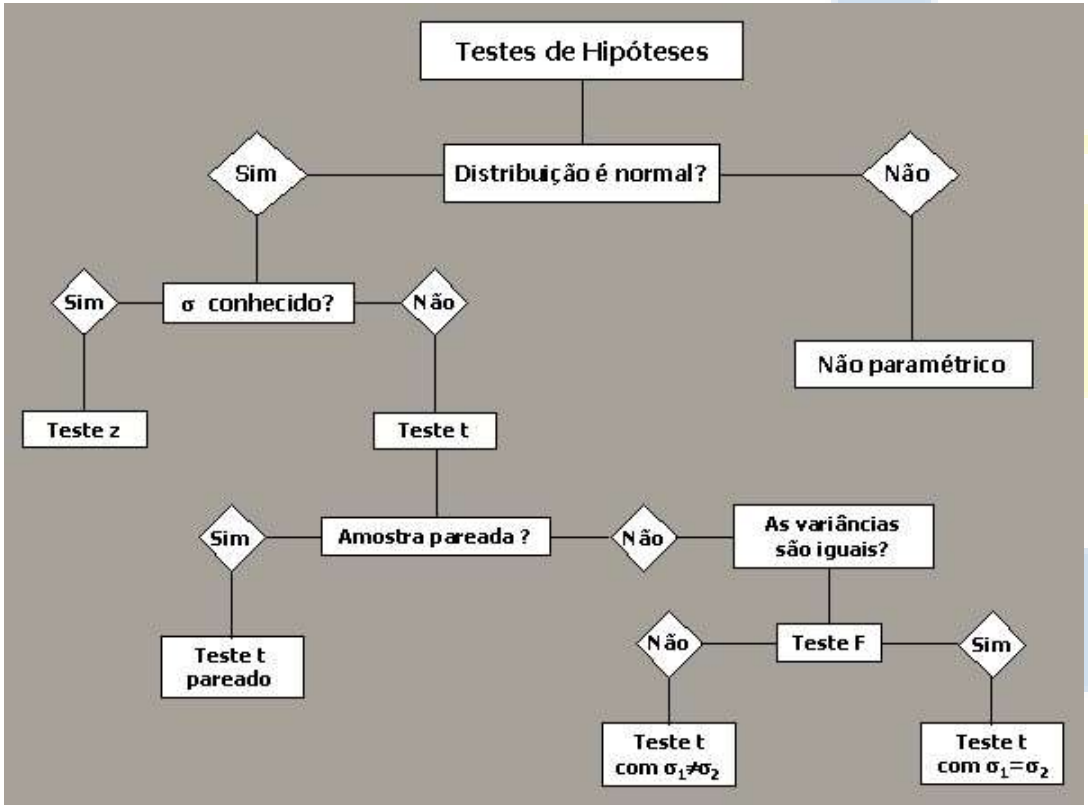
H0 : parâmetro = valor de teste

H1 : parâmetro > valor de teste

Testes Bilaterais

H0 : parâmetro = valor de teste

H1 : parâmetro ≠ valor de teste



Testes Estatísticos

Testes Estatísticos

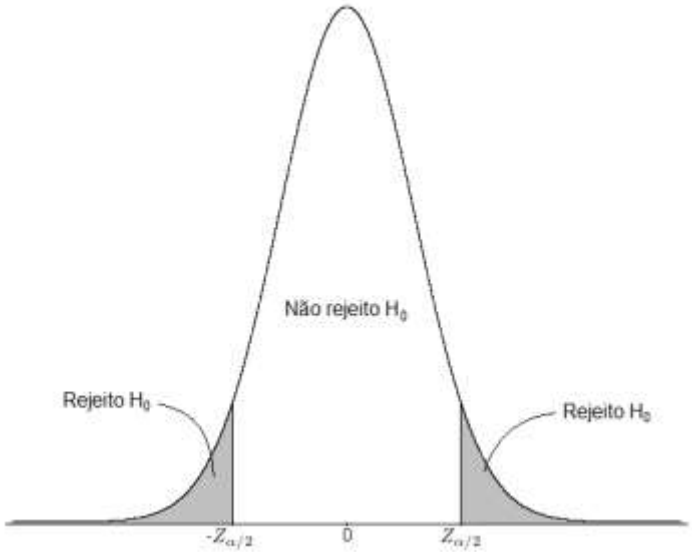
Variáveis Quantitativas

Paramétricos		Não-Paramétricos	
Não-pareadas	Pareadas	Não pareadas	Pareadas
2 amostras	**2 amostras**	**2 amostras**	**2 amostras**
Teste t (Student)	Teste t (Student)	Mann-Whitney X^2 (2x2) Exato (Fisher)	Wilcoxon Mc Nemar
Mais de 2	Mais de 2	Mais de 2	Mais de 2
ANOVA	ANOVA	Kruskal Wallis X^2 (mxn)	Cochram Friedman

Testes Estatísticos

1. Definir a hipótese de igualdade ou hipótese nula (H0).
2. Escolher a prova estatística (com o modelo estatístico associado) para tentar rejeitar H0.
3. Definir o nível de significância (α) e um tamanho de amostra (n).
4. Determinar a distribuição amostral da prova estatística sob a hipótese de nulidade.
5. Definir a região de rejeição.
6. Calcular o valor da prova estatística, utilizando os valores obtidos na(s) amostra(s).

Se o valor da prova estatística estiver na região de rejeição, rejeitamos então a hipótese nula, senão a decisão será que a hipótese nula não poderá ser rejeitada ao nível de significância determinado.



Mas o que é esse tal nível de significância?

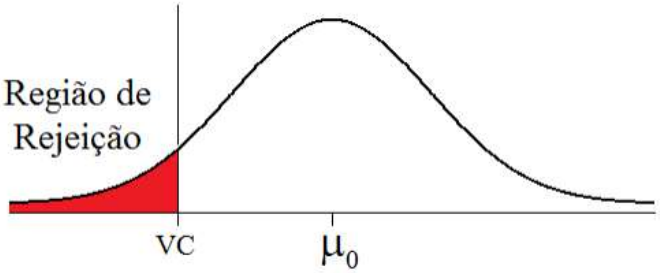
P = 0,05 e P = 0,01, ou seja, 5% e 1% respectivamente

Hipótese Nula:

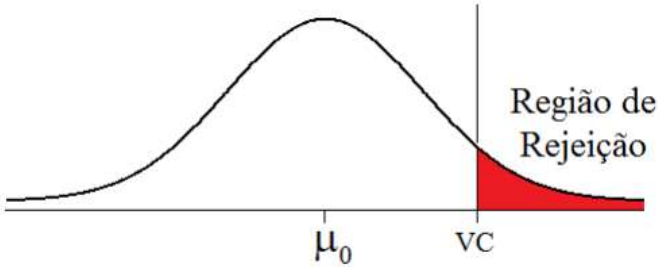
H0: O lote atende as especificações
H0: $\mu = 4$

Hipótese Alternativa

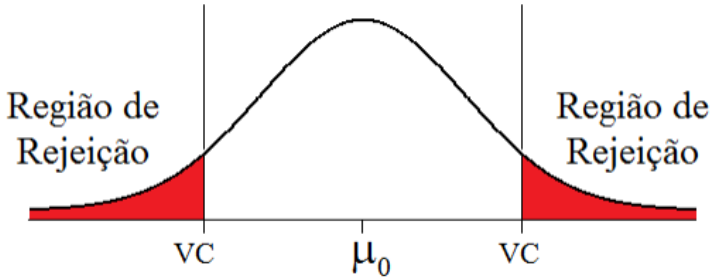
H1: O lote não atende as especificações
H1: $\mu \neq 4$



Região de rejeição para o teste unicaudal para a média (cauda inferior)



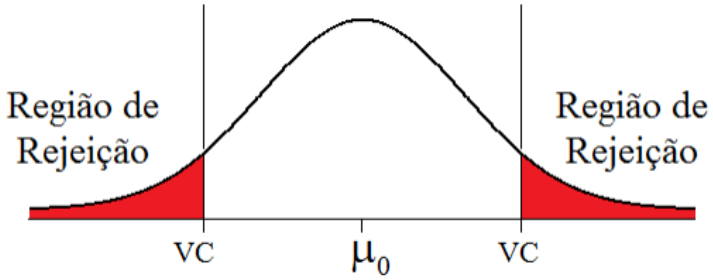
Região de rejeição para o teste unicaudal para a média (cauda superior)



Região de rejeição para o teste bicaudal para a média

*VC = Valor Crítico

Região Crítica ou Região de Rejeição é o conjunto de valores assumidos pela variável aleatória ou estatística de teste para os quais a hipótese nula é rejeitada.



“Testamos a hipótese nula, no sentido em que, supondo-a verdadeira, procuramos chegar a uma conclusão que nos leve à sua rejeição.”

Tipos de Erros

Erro Tipo I

Rejeitar H0 quando de fato H0 é verdadeiro.

Erro Tipo II

Não rejeitamos H0 quando de fato H0 é falsa.

Tipos de Erros

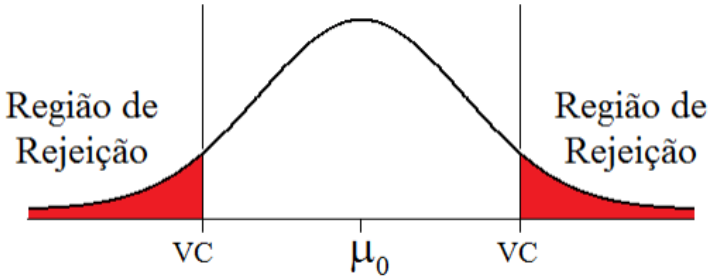
Decisão	Situação	
	H0 Verdadeira	H1 Falsa
Não Rejeitar H0	Decisão Correta	Erro Tipo II
Rejeitar H0	Erro Tipo 1	Decisão Correta

$$\alpha = \mathbb{P}(\text{Erro do tipo I}) = P(\text{rejeitar } H_0 \text{ dado } H_0 \text{ verdadeira});$$

$$\beta = \mathbb{P}(\text{Erro do tipo II}) = P(\text{aceitar } H_0 \text{ dado } H_0 \text{ falsa}).$$

Tipos de Erros

Decisão	Situação	
	H0 Verdadeira	H1 Falsa
Não Rejeitar H0	Decisão Correta	Erro Tipo II
Rejeitar H0	Erro Tipo 1	Decisão Correta



1) Identifique o parâmetro de interesse

média

2) Estabeleça H_0 e H_a

$H_0: \mu = 3,5$; $H_a: \mu \neq 3,5$

3) Estabeleça o nível de significância α que determinará a região de rejeição

$\alpha = 0,05$

Estabeleça uma estatística apropriada de teste.
É o que determina o teste!

Estabeleça uma estatística apropriada de teste.
É o que determina o teste!

Estatística de Teste é a estatística amostral, cujo valor baseado nos dados será utilizado para a tomada de decisão a respeito da hipótese nula. Está associada à distribuição de probabilidade do estimador do parâmetro que se deseja testar.

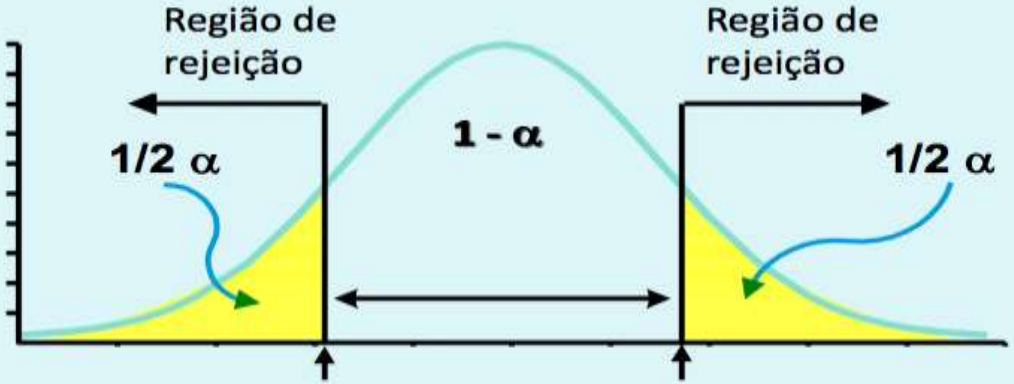
No teste para uma média por exemplo, utilizam-se as estatísticas Z ou t:

$$Z = \frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}}, \quad \text{se a variância populacional é conhecida}$$

$$t = \frac{(\bar{X} - \mu)}{s / \sqrt{n}}, \quad \text{se a variância populacional não é conhecida}$$

Decida de H_0 deve ou não ser rejeitada.

Precisamos estabelecer o valor crítico.



A hipótese nula será rejeitada se:

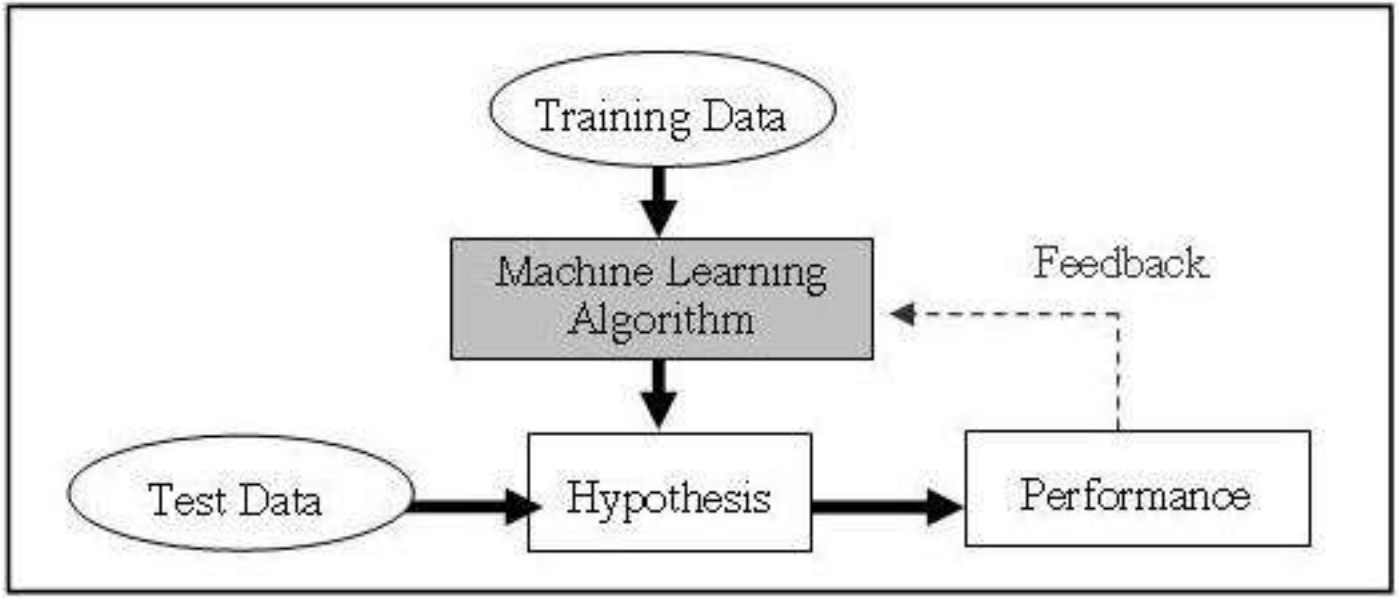
$$Z_0 > z_{\alpha/2}$$

ou

$$Z_0 < -z_{\alpha/2}$$

E falharemos em rejeitar se:

$$-z_{\alpha/2} < Z_0 < z_{\alpha/2}$$

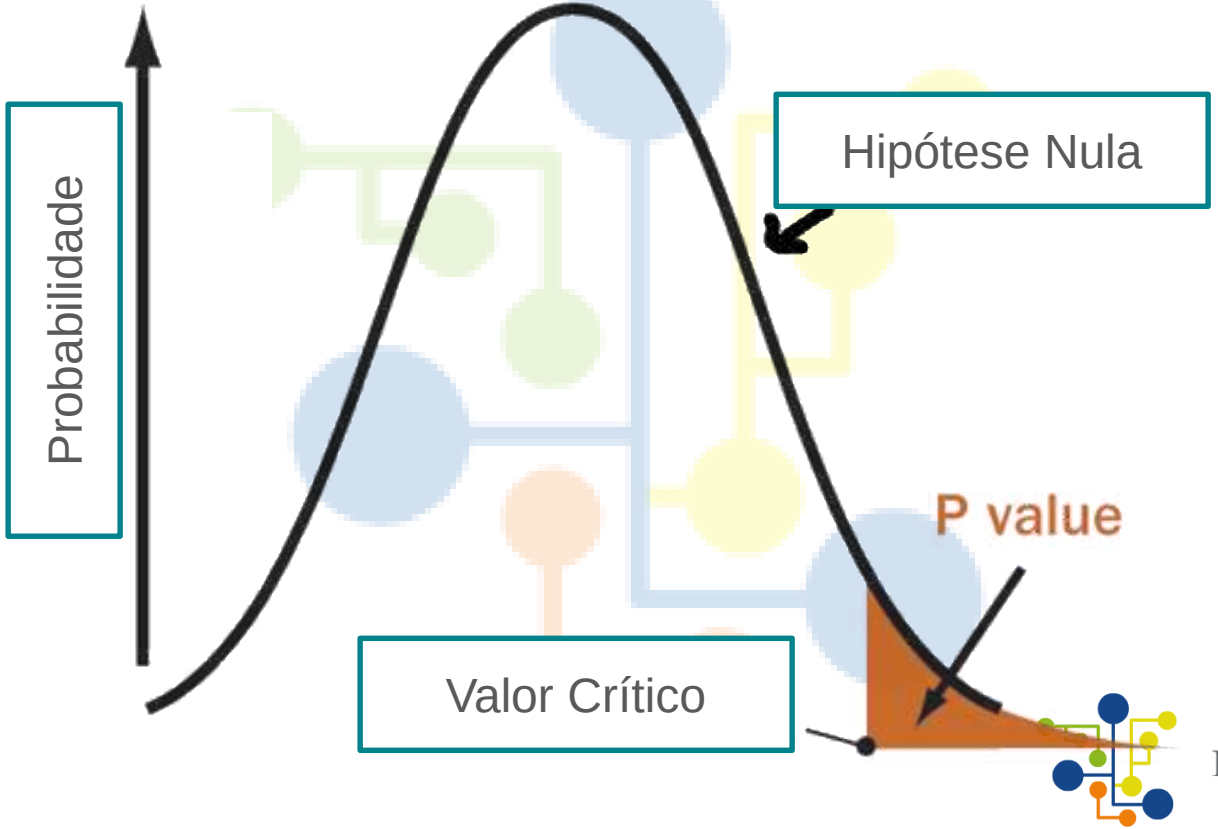

Training Data
Machine Learning Algorithm
Feedback
Test Data
Hypothesis
Performance

Valor-p
(p-value)

Valor-p
(p-value)

O valor-p (p-value) é o valor de significância observado

De um ponto de vista prático, podemos afirmar que o valor-p representa a chance ou a probabilidade do efeito (ou da diferença) observada ser devido ao **acaso** e não aos fatores que estão sendo estudados/analisados.


Probabilidade
Hipótese Nula
P value
Valor Crítico

Valor-p

Se p-valor $\geq \alpha$, NÃO rejeita H_0

Se p-valor $< \alpha$, REJEITA H_0

Média do Tratamento A > Média do Tratamento B

Valor-p = 0,3

O que não é o valor-p?

- Valor-p e nível de significância não são sinônimos. O valor-p é sempre obtido de uma amostra, enquanto o nível de significância é geralmente fixado antes da coleta dos dados.
- O valor-p não é a probabilidade da hipótese nula ter sido enganosamente rejeitada.
- O valor-p não é a probabilidade da hipótese nula de um teste ser verdadeira.
- O valor-p não é a probabilidade de um dado resultado ter sido obtido de um "acaso".
- A magnitude do valor-p não indica o tamanho ou a importância de um efeito observado.

Data Science Academy

Processo de Aprendizagem

Data Science Academy

Você já viu um algoritmo de Machine Learning?

Data Science Academy

Componentes do Processo de Aprendizagem

Elementos do Processo de Aprendizagem

- Um padrão existe
- Não há um único modelo matemático que explique esse padrão
- Dados estão disponíveis

Componentes do Processo de Aprendizagem

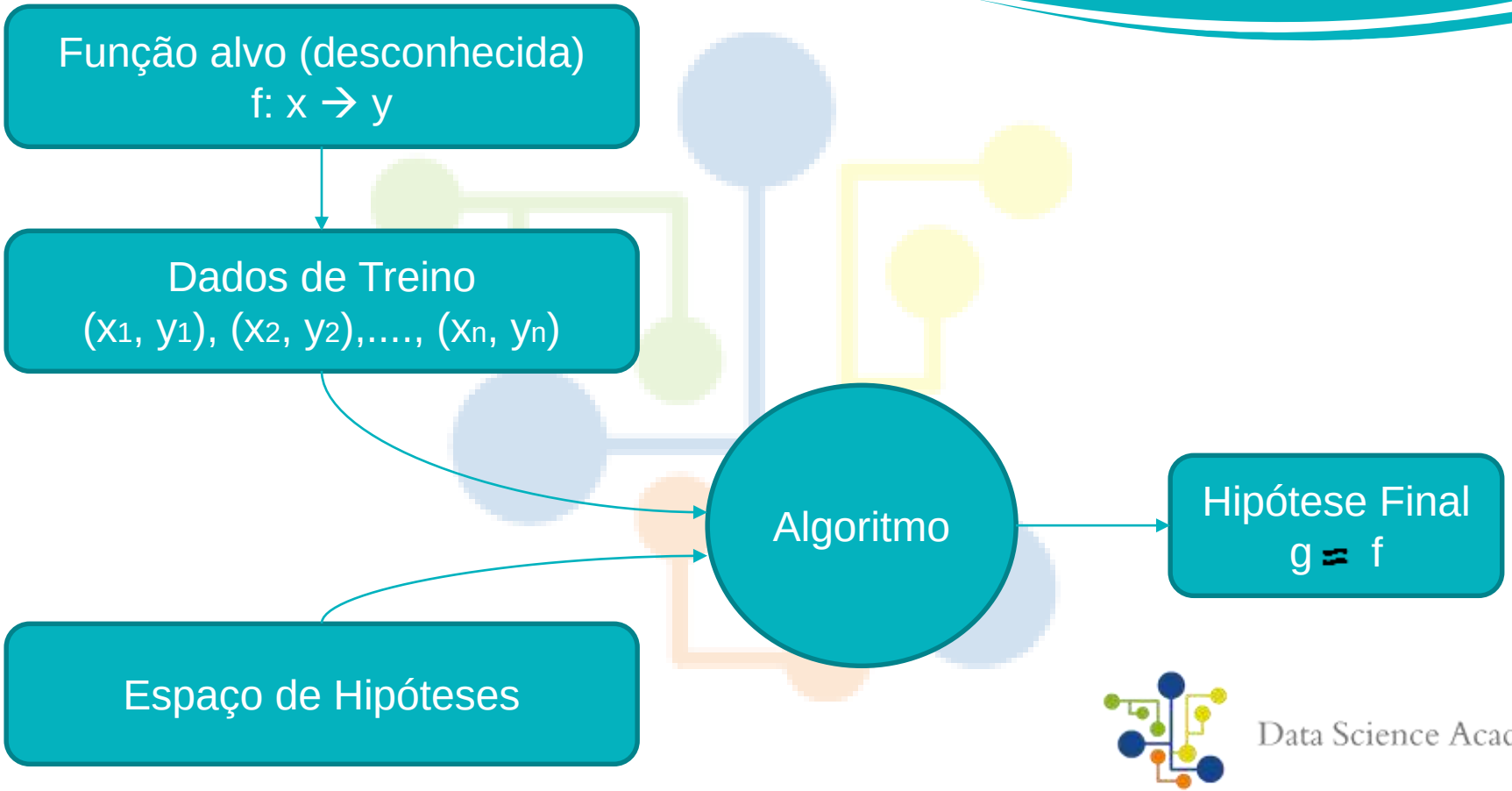
Aprovação de Crédito

Aprovação de Crédito de um Indivíduo

Atributo	Valor
Sexo	Masculino
Idade	34
Salário Mensal	R$ 18.000,00
Anos no Emprego Atual	3
Anos de Residência	7
Saldo Bancário	R$ 32.671,94

Componentes do Processo de Aprendizagem

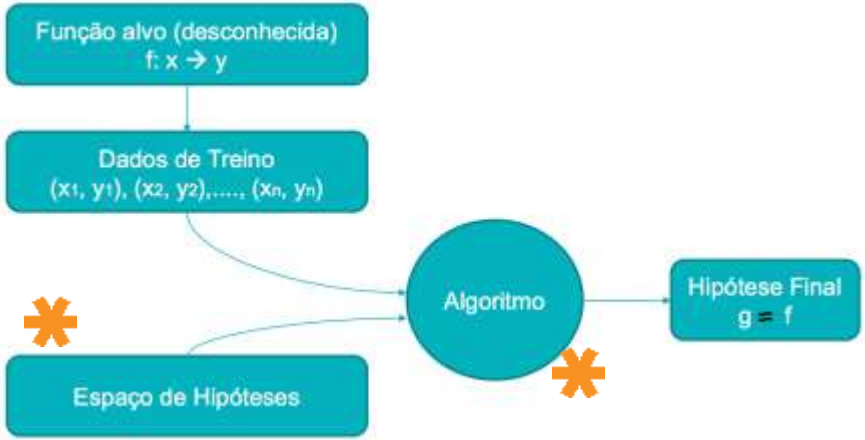
Input	x	{Dados do cliente}
Output	y	{Decisão → Crédito: Sim/Não}
Função alvo	$f: x \rightarrow y$	{Representação do relacionamento} {Fórmula matemática desconhecida}
Dados	$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$	{Dados históricos}
Hipótese	$g: x \rightarrow y$	{Fórmula a ser usada}


Função alvo (desconhecida)
f: x → y
Dados de Treino
(x1, y1), (x2, y2),...., (xn, yn)
Algoritmo
Hipótese Final
g ≈ f
Espaço de Hipóteses

Data Science Academy

Modelo de Aprendizagem

Data Science Academy



- Espaço de Hipóteses

$$\mathcal{H} = \{h\} \qquad g \in \mathcal{H}$$

- Algoritmo de Aprendizagem

Espaço de Hipóteses	+	Algoritmo de Aprendizagem	=	Modelo de Aprendizagem
Redes Neurais Support Vector Machines		Back Propagation Programação Quadrática		

O Espaço de Hipóteses contém os recursos com os quais podemos trabalhar. O Algoritmo de Aprendizagem recebe os dados e navega pelo Espaço de Hipóteses a fim de encontrar a melhor hipótese que gera o resultado desejado.

Modelo de Aprendizagem

Input → $X = (x_1, x_2, ..., x_d)$ → Vetor de atributos do indivíduo

$$\sum_{i=1}^{d} w_i x_i$$

$$\sum_{i=1}^{d} w_i x_i$$

Weight (Peso)

Input → X = (x_1, x_2, ..., x_d)

Crédito é **aprovado** se $\sum_{i=1}^{d} w_i x_i$ > threshold

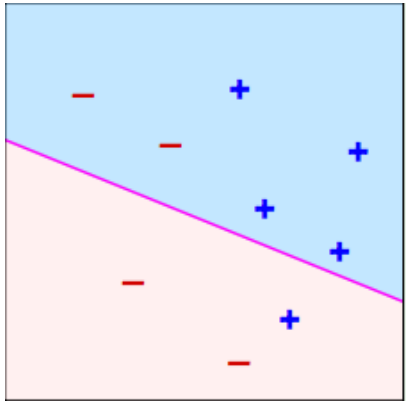
Crédito é **negado** se $\sum_{i=1}^{d} w_i x_i$ < threshold

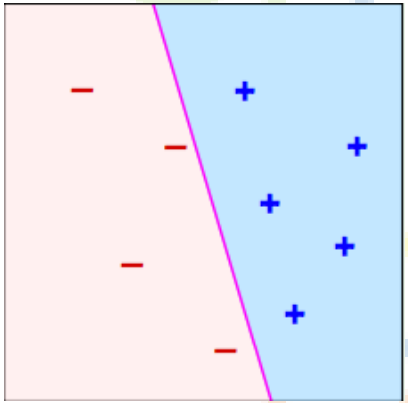
Fórmula que Define as Hipóteses no Espaço de Hipóteses

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - \text{threshold}\right)$$

As diferentes combinações weight/threshold vão formar diferentes hipóteses

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - \text{threshold}\right)$$





Dados Linearmente Separáveis

Algoritmo de Aprendizagem $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x})$

Dados de Treino $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Erro de Classificação $\text{sign}(\mathbf{w}^T\mathbf{x}_n) \neq y_n$

Ajuste $\mathbf{w} \leftarrow \mathbf{w} + y_n\mathbf{x}_n$

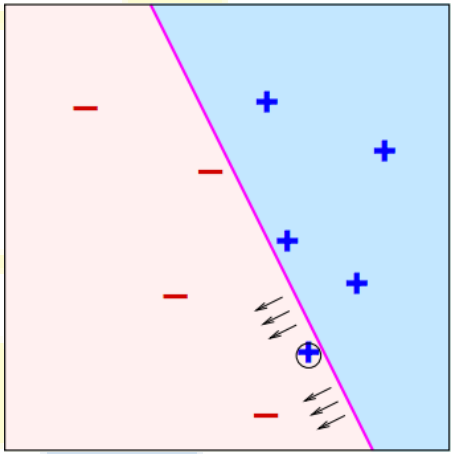
Iteração 1

$$\mathbf{w} \leftarrow \mathbf{w} + y_n \mathbf{x}_n$$

Iteração 2

$$\mathbf{w} \leftarrow \mathbf{w} + y_n \mathbf{x}_n$$

.
.
.
.



Se os dados forem linearmente separáveis, o algoritmo fará diversas iterações até encontrar a linha que realmente separa as duas classes

Data Science Academy

Machine Learning é Aprendizado a partir de Dados

Data Science Academy

Elementos do Processo de Aprendizagem

Um padrão existe

Não há um único modelo matemático que explique esse padrão

Dados estão disponíveis

Elementos do Processo de Aprendizagem

Um padrão existe

Não há um único modelo matemático que explique esse padrão

Dados estão disponíveis

Componentes do Processo de Aprendizagem

Input	Output	Função alvo	Dados	Hipótese
				

Data Science Academy

Cost Function
(Função de Custo)

Data Science Academy

Aprendizagem Supervisionada

Coleção de vetores de atributos $\{x_i\}, i = 1, n$

Coleção de respostas observadas $\{y_i\}, i = 1, n$

Queremos construir uma área de respostas (espaço de hipóteses)

h(x)

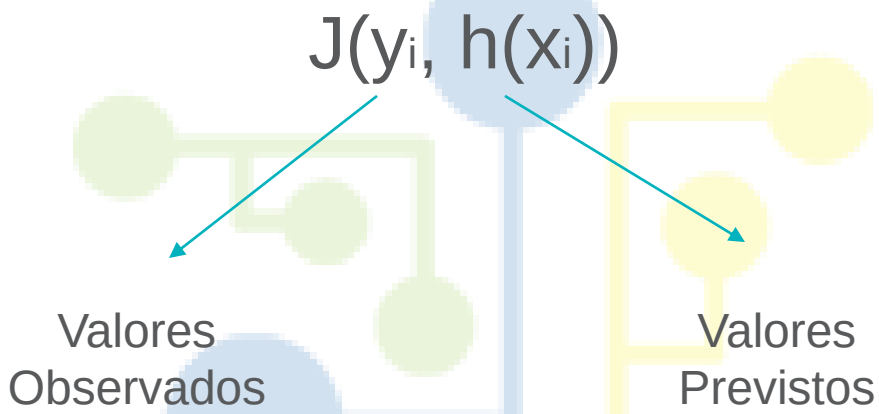
Como sabemos se os valores de *h(x)* são bons ou ruins?

Cost Function

Descreve quão bem resposta na área de respostas (espaço de hipóteses) se encaixa no conjunto de dados que está sendo analisado

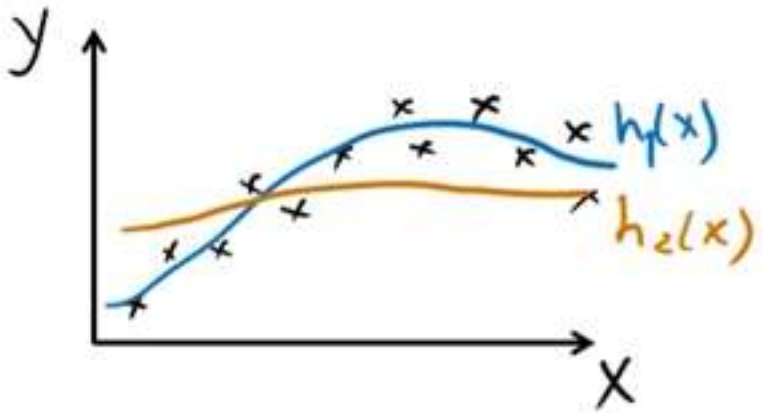
$h(x)$

$J(y_i, h(x_i))$



A Cost Function é um número que melhor representa a relação entre esses elementos.

Cost Function



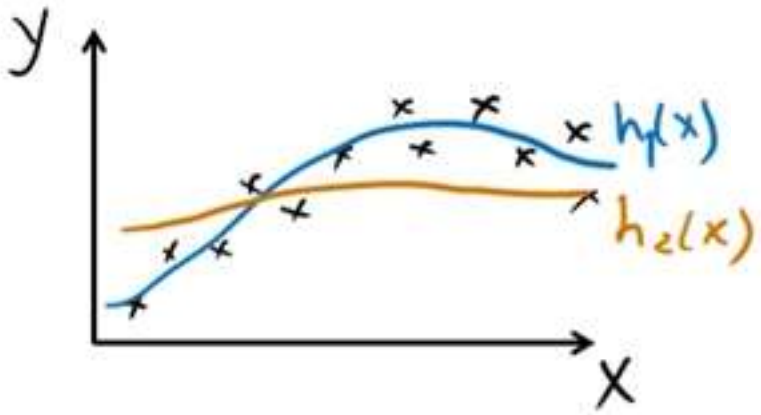
J_1

$J_2 > J_1$

Valores menores da Cost Function significam um melhor ”fit”

O objetivo do algoritmo de ML é aprender um modelo que minimize os erros

Cost Function



J_1

$J_2 > J_1$

Um dos objetivos em Machine Learning é construir *h(x)* de modo que o valor de J seja minimizado

Em problemas de regressão, *h(x)* é normalmente interpretada diretamente como a resposta a ser prevista

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

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Comparando uma previsão contra o seu valor real, usando uma cost function, determinamos o nível de erro do algoritmo.

Comparando uma previsão contra o seu valor real, usando uma cost function, determinamos o nível de erro do algoritmo.

Por ser uma formulação matemática, a cost function expressa o nível de erro em uma forma numérica. A cost function transmite o que é realmente importante e significativo para seus propósitos com o algoritmo de aprendizagem.

Data Science Academy

Gradiente Descendente

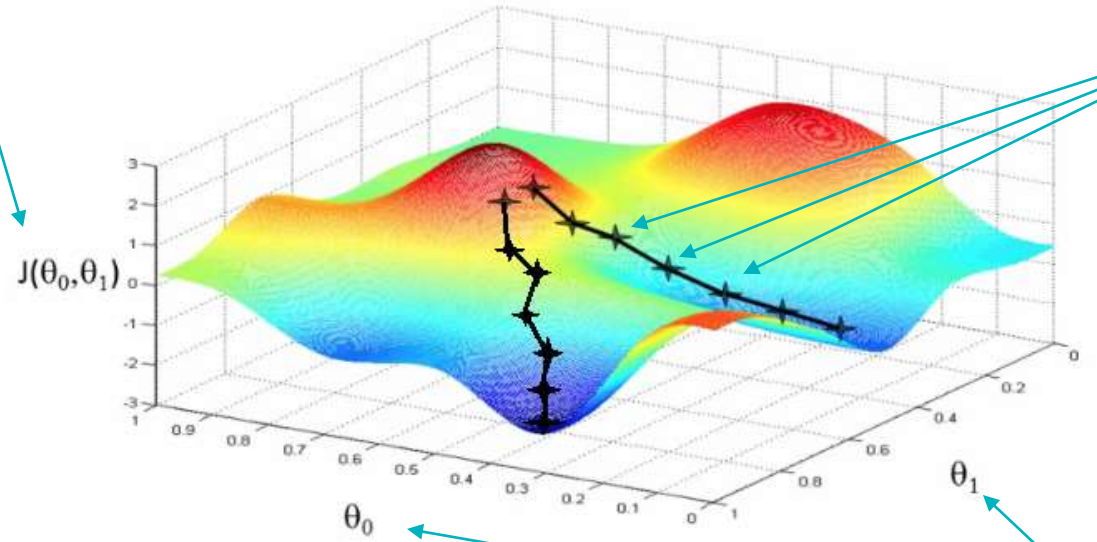
Data Science Academy

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

Objetivo → $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Cost Function

Gradient Descent



Quanto melhor os valores de parâmetros, menor o valor de J.

Parâmetros da Cost Function

Gradiente Descendente é ideal quando os parâmetros não podem ser calculados analiticamente (por exemplo, usando álgebra linear) e devem ser pesquisados por um algoritmo de otimização.

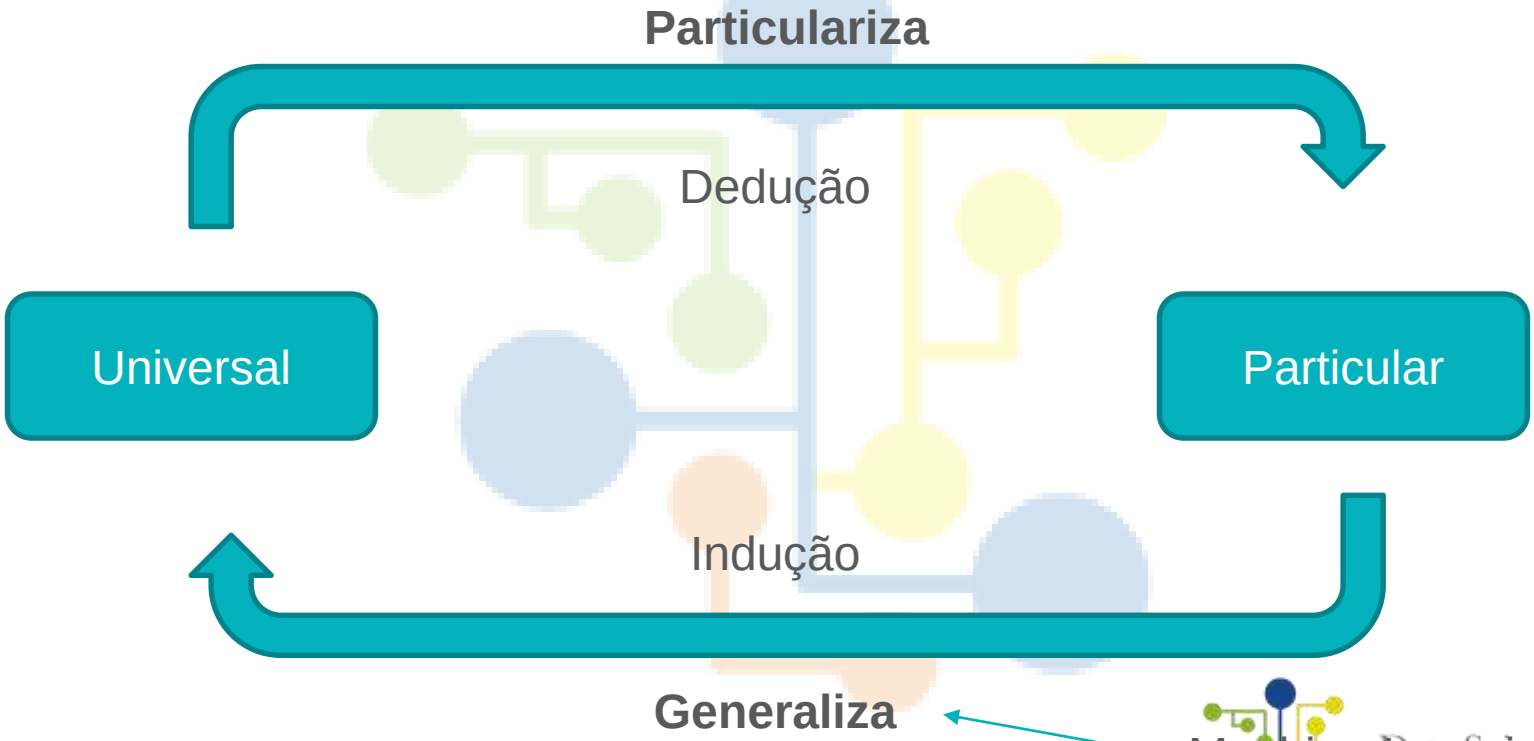
Data Science Academy

Overfitting x Underfitting

Data Science Academy

Aprendizagem Supervisionada

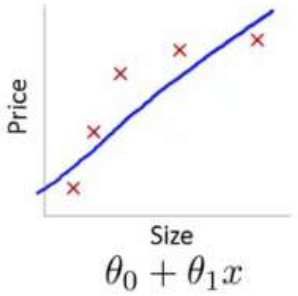
$$Y = f(X)$$



Machine Learning

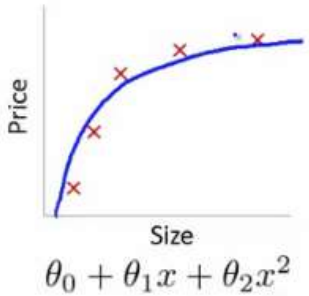
Generalização refere-se a quão bem os conceitos aprendidos por um modelo de aprendizado de máquina se aplicam a exemplos específicos não vistos pelo modelo durante o processo de aprendizado

Overfitting e underfitting são as duas maiores causas de mau desempenho dos algoritmos de aprendizagem da máquina



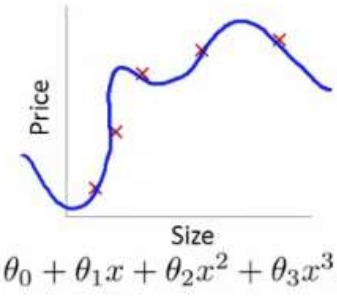
$\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

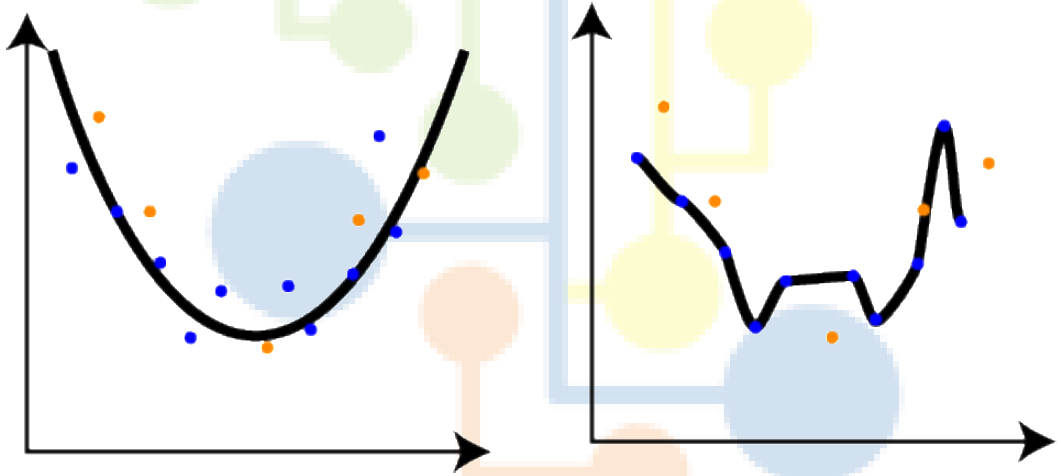


$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

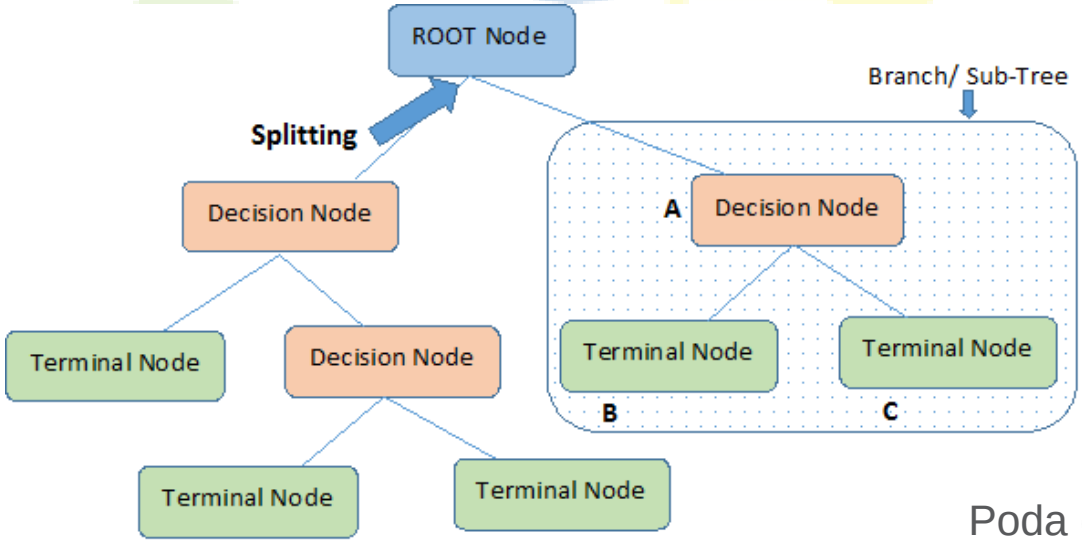
High variance
(overfit)

Se soubéssemos a forma da função-alvo, poderíamos usá-la diretamente para fazer previsões, ao invés de tentar aprender uma aproximação a partir de amostras de dados

Overfitting

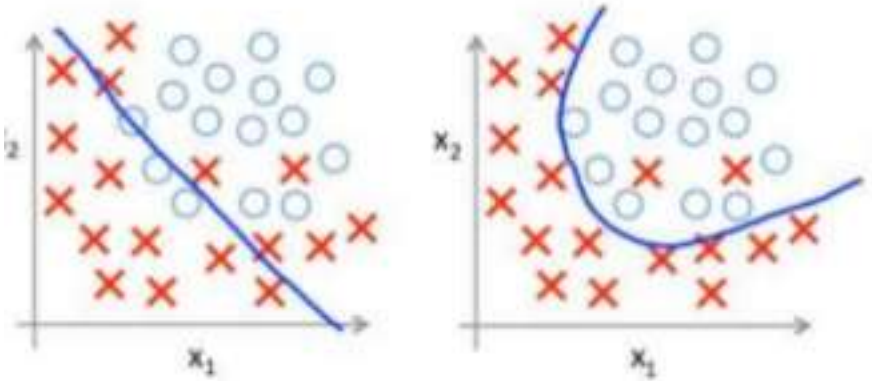


Overfitting

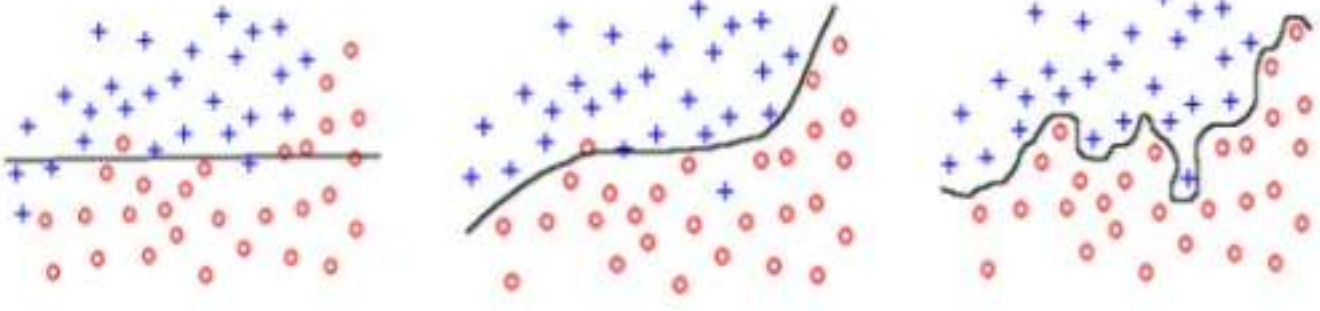


Poda (pruning) da árvore de decisão para evitar overfitting.

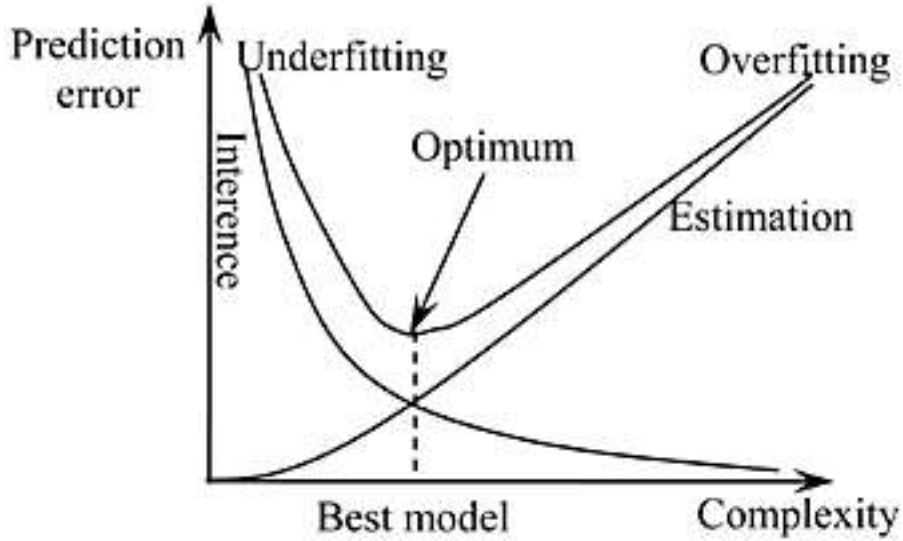
Underfitting



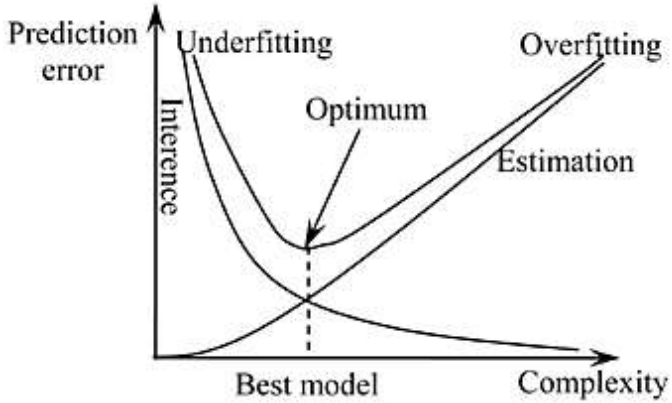
Good Fit



Good Fit


Prediction error
Underfitting
Overfitting
Optimum
Inference
Estimation
Best model
Complexity

Good Fit

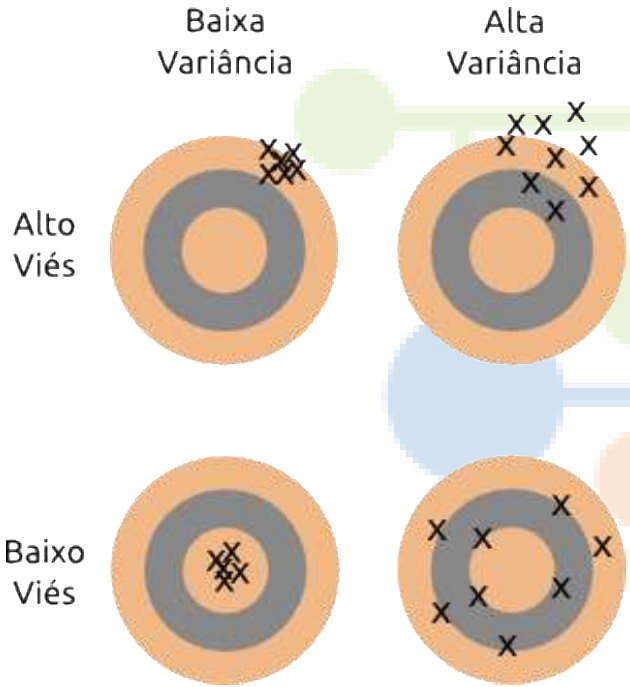


- Reamostragem
- Conjuntos de Dados de Validação

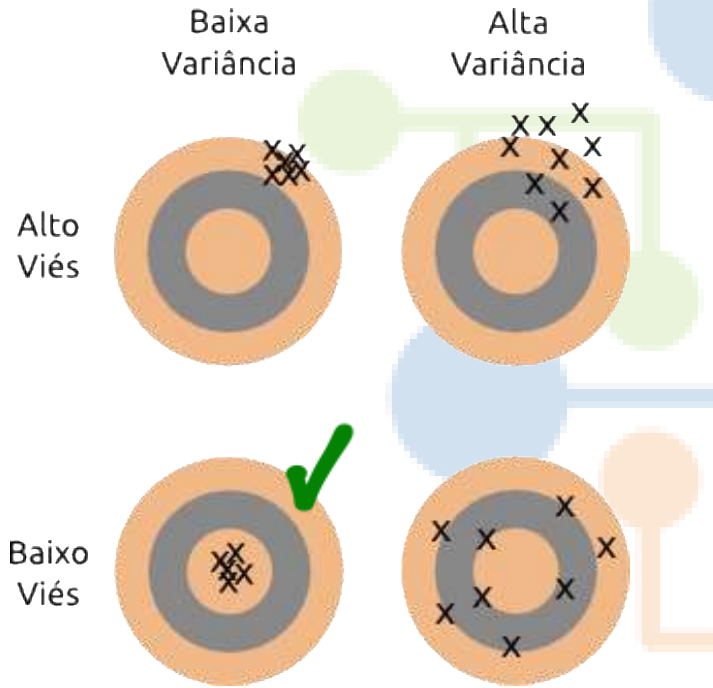
Data Science Academy

Bias (Viés) e Variância

Data Science Academy



Separando o erro de generalização em viés (*bias*) e variância (*variance*)



Viés é a tendência do modelo aprender consistentemente uma generalização incorreta

Variância é a tendência de se aprender fatos aleatórios independentemente do sinal real

Erro de Previsão de um Modelo

$$E[(y - \hat{f}(x))^2] = Bias[\hat{f}(x)]^2 + Var[\hat{f}(x)] + \sigma^2$$

Bias

$$Bias[\hat{f}(x)] = E[\hat{f}(x) - f(x)]$$

Variance

$$Var[\hat{f}(x)] = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2$$


TRAINING / TEST

Utilizar um modelo complexo que é capaz de reduzir consideravelmente o erro de previsão no dataset de treino, mas ao mesmo tempo não é tão generalizável a ponto de apresentar um bom resultado no dataset de teste

Viés

Variância

Utilizar um modelo simples que é bem generalizável, mas não reduz consideravelmente o erro de previsão no dataset de treino

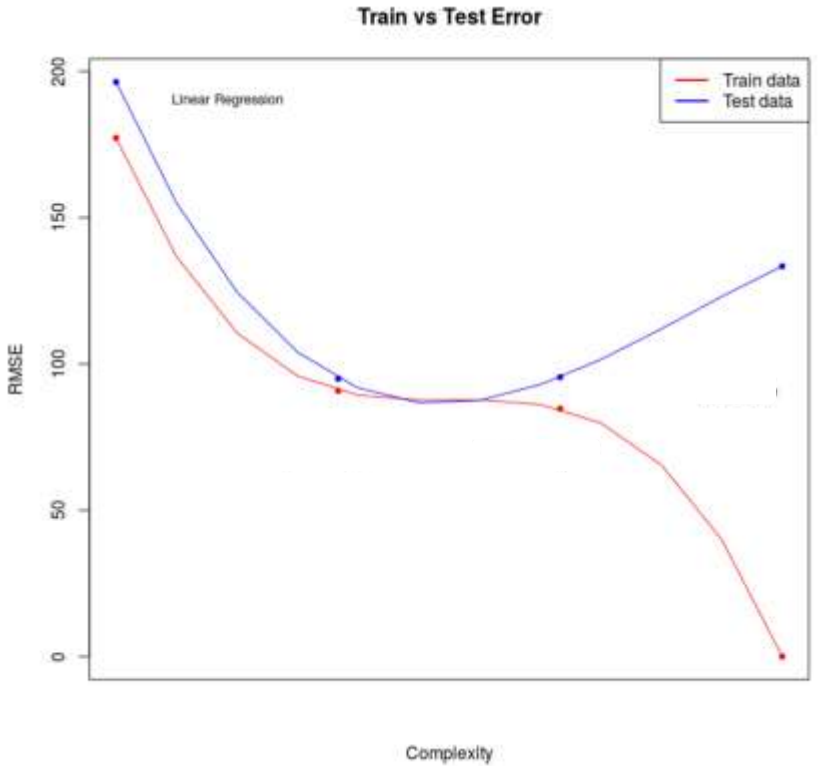
Viés

Variância

Tradeoff

Modelos mais simples tem viés alto mas variância baixa (underfitting)

Modelos mais complexos tem viés baixo mas variância alta (overfitting)



A tarefa essencial de previsão é selecionar um modelo que se aproxime do ponto mínimo da curva de erro do dataset de teste

Data Science Academy

Overfitting x Underfitting

Data Science Academy

Data Science Academy

Overfitting x Underfitting

Data Science Academy

Data Science Academy

Overfitting x Underfitting

Data Science Academy

Data Science Academy

Obrigado