

ICC204 - Aprendizagem de Máquina e Mineração de Dados

# Erro empírico e generalização



Prof. Rafael Giusti  
[rgiusti@icomp.ufam.edu.br](mailto:rgiusti@icomp.ufam.edu.br)

# Agenda

- Revisão e formalização dos conceitos
- Generalização
- Aprendizado viciado
- Dilema viés-variância
- Teorema NFL (*no free-lunch*)
- Combatendo o aprendizado viciado

# Agenda

- Revisão e formalização dos conceitos
- Generalização
- Aprendizado viciado
- Dilema viés-variância
- Teorema NFL (*no free-lunch*)
- Combatendo o aprendizado viciado

# Linguagem de descrição dos dados

- A amostra é tipicamente descrita através de uma coleção de **exemplos**  $E_i = (\mathbf{x}_i, y_i)$ 
  - $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iM})$  contém os valores do exemplo  $E_i$  para os atributos  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M$
  - Esses valores **caracterizam** os exemplos
  - O valor  $y_i$  caracteriza o **conceito** associado aos exemplos
    - O atributo  $\mathbf{Y}$  é denominado **rótulo**

# Linguagem de descrição dos dados

- É importante que os atributos sejam significativos e representativos do conceito a ser aprendido
  - Deve ser necessário distinguir exemplos de acordo com suas características
- É ideal que os atributos sejam objetivos
  - Atributos objetivos: idade, altura, profissão, velocidade, concentração de glicose etc.
  - Não objetivos: beleza, inteligência, humor etc.

# Linguagem de descrição dos dados

- A coleção de exemplos  $E_i$  é tipicamente representada como uma **tabela atributo-valor**

	$X_1$	$X_2$	$\dots$	$X_M$	$Y$
$E_1$	$x_{11}$	$x_{12}$	$\ddots$	$x_{1M}$	$y_1$
$E_2$	$x_{21}$	$x_{22}$	$\ddots$	$x_{2M}$	$y_2$
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$	$\vdots$
$E_N$	$x_{N1}$	$x_{N2}$	$\ddots$	$x_{NM}$	$y_N$

# Modalidades de AM

- Nem sempre todos os exemplos possuem valores para todos os atributos
  - Atributos podem ser **ausentes** ou **desconhecidos**
  - Seja por custo, falhas no processo ou por definição do problema
- Quando características são ausentes, os exemplos *podem* estar **mal representados**
  - Alguns indutores não admitem valores ausentes

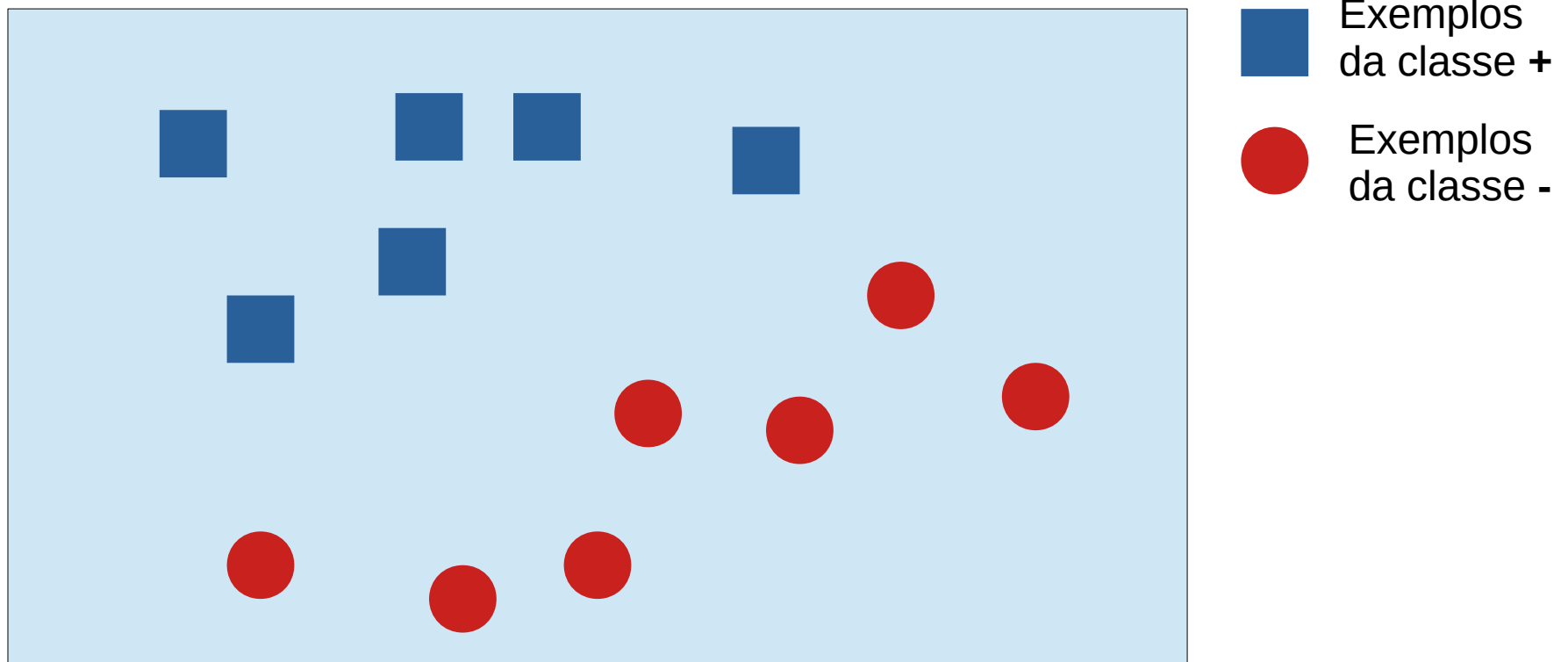
# Modalidades de AM

- O rótulo pode ser ausente por definição
  - Aprendizado supervisionado: todos os exemplos possuem um valor para o rótulo
  - Aprendizado não supervisionado: nenhum exemplo possui um valor para o rótulo
  - Aprendizado semissupervisionado: nem todos os exemplos possuem valores para o rótulo



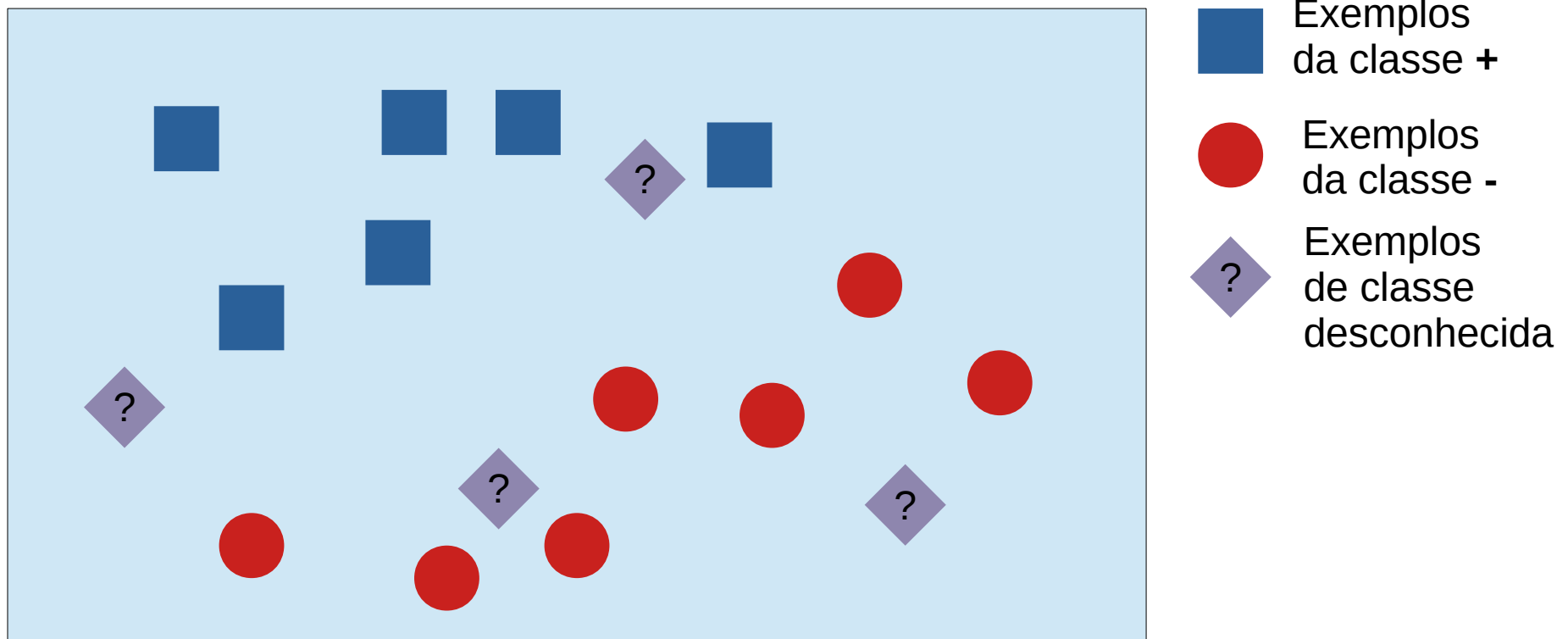
# Modalidades de AM

- No aprendizado **supervisionado**, os exemplos de treinamento são **rotulados** de acordo com o conceito que desejamos aprender



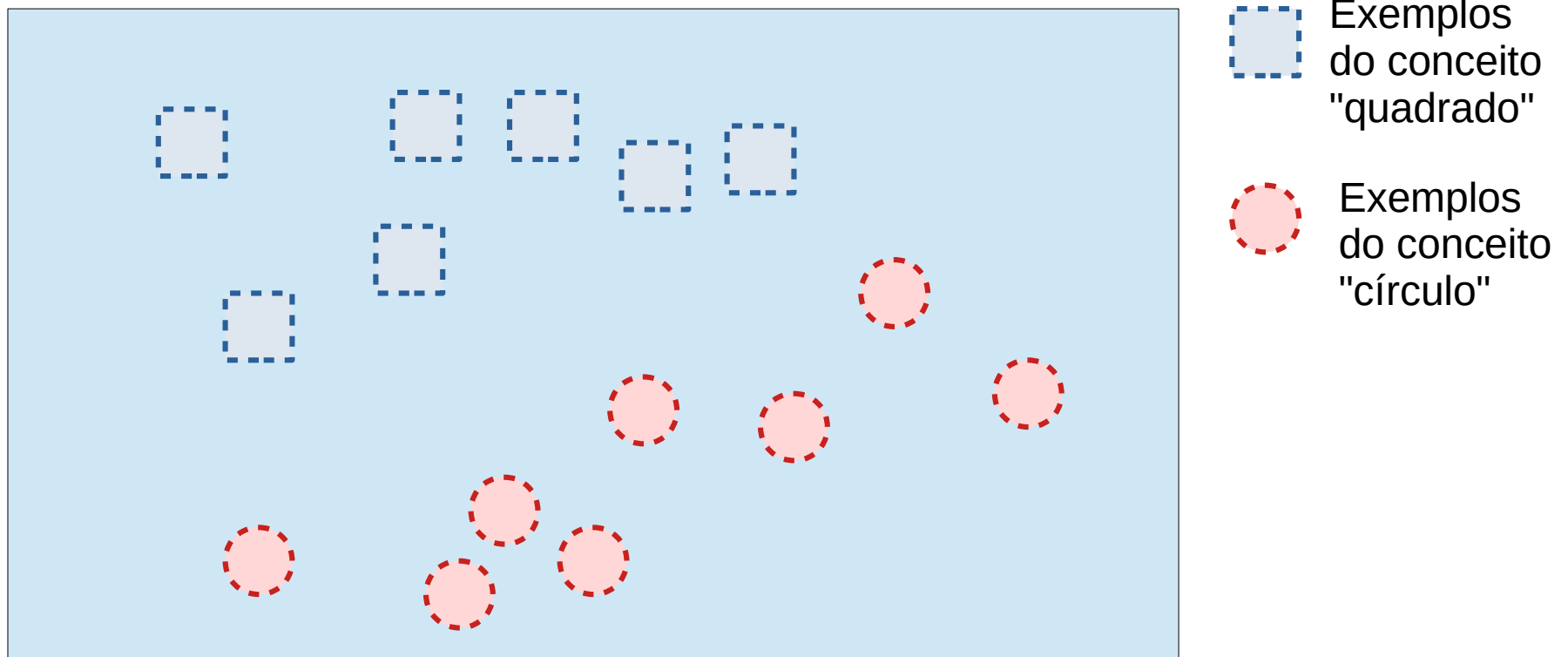
# Modalidades de AM

- Um conceito muito semelhante é o aprendizado **semisupervisionado**, no qual **nem todos** os exemplos possuem rótulos conhecidos



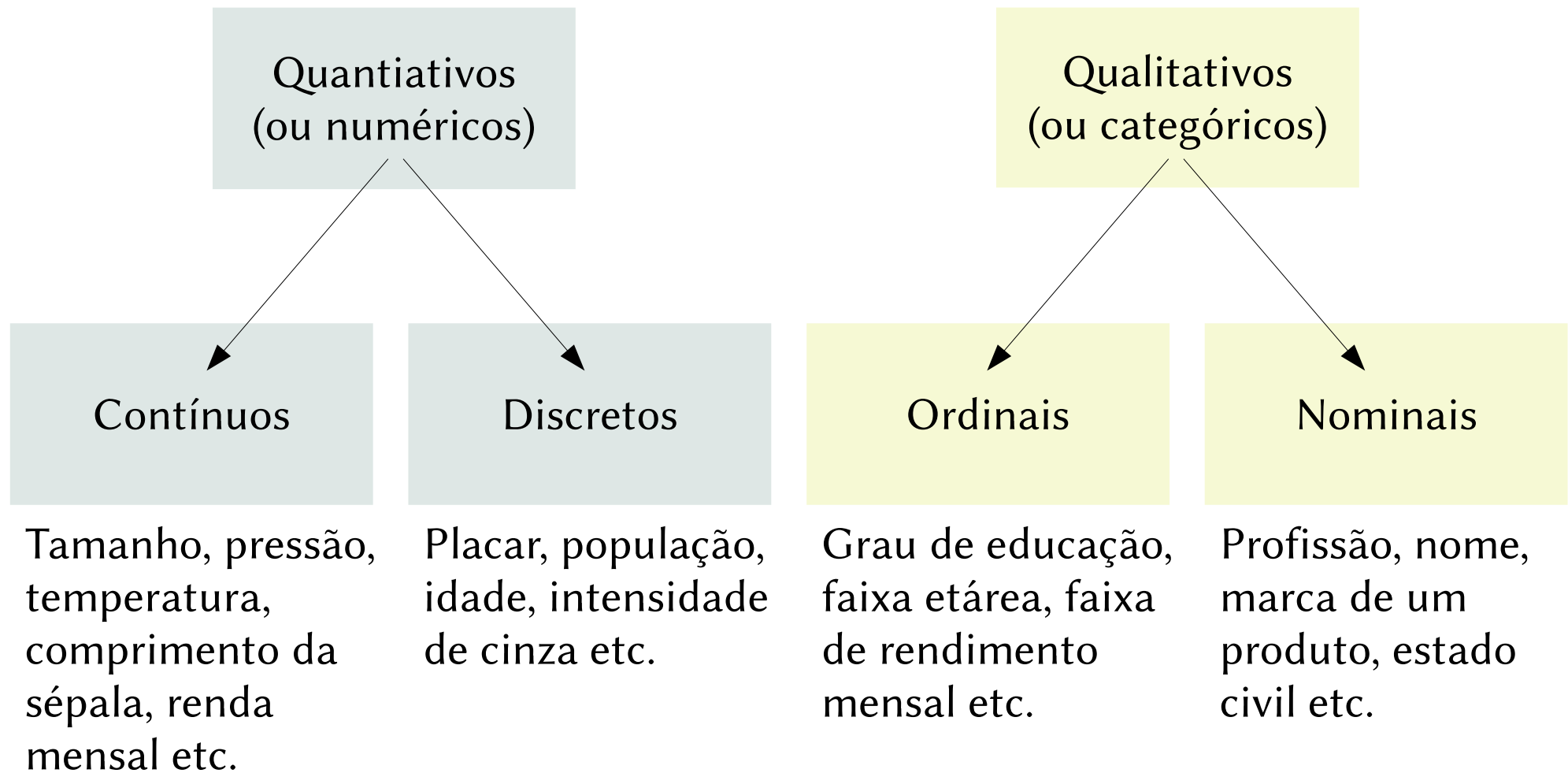
# Modalidades de AM

- Já no aprendizado **não supervisionado**, o conceito existe, mas os exemplos **não são rotulados**; o objetivo é encontrar uma estrutura nos dados



# Tipos de atributos

- Atributos podem ser



# Agenda

- Revisão e formalização dos conceitos
- Generalização
- Aprendizado viciado
- Dilema viés-variância
- Teorema NFL (*no free-lunch*)
- Combatendo o aprendizado viciado

# Generalização

- Em todos os casos, o objetivo é obter um modelo que **generalize** o conceito
- Intuitivamente, generalizar significa
  - Ser bom para exemplos nunca antes vistos, mesmo tendo sido induzido a partir de uma amostra relativamente pequena dos dados
- De modo geral, é bastante difícil definir quando um modelo generalizou bem

# Generalização

- Generalização parece ser uma consequência inevitável da hipótese fundamental do aprendizado indutivo
  - *Qualquer hipótese que aproxime razoavelmente a função-conceito para um conjunto suficientemente grande de exemplos de treinamento também irá aproximar razoavelmente bem a função-conceito para exemplos nunca observados.*

# Generalização

- Generalização parece ser uma consequência inevitável da hipótese fundamental do aprendizado indutivo
  - *Qualquer hipótese que aproxime razoavelmente a função-conceito para um conjunto suficientemente grande de exemplos de treinamento também irá aproximar razoavelmente bem a função-conceito para exemplos nunca observados.*

**Só que..... como se garante que o conjunto é suficientemente grande?**



# Generalização

- Generalização parece ser uma consequência inevitável da hipótese fundamental do aprendizado indutivo
  - *Qualquer hipótese que aproxime razoavelmente a função-conceito para um conjunto suficientemente grande de exemplos de treinamento também irá aproximar razoavelmente bem a função-conceito para exemplos nunca observados.*

**Só que..... o que significa aproximar razoavelmente?**

# Erro empírico

- Para verificar se estamos próximos dessas condições, utilizaremos **estimadores**
- O **erro empírico** é o estimador mais simples para qualquer modelo
  - É o erro cometido pelo modelo quando avaliado sob os mesmos exemplos em que foi treinado
  - Vejamos um exemplo para um regressor linear

# Erro de um regressor linear

- Um modelo de regressão linear é da forma

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M$$

- Seu erro pode ser estimado, para um conjunto de referência  $T = \{T_1, T_2, \dots, T_N\}$ ,  $T_i = (x_i, t_i)$  como

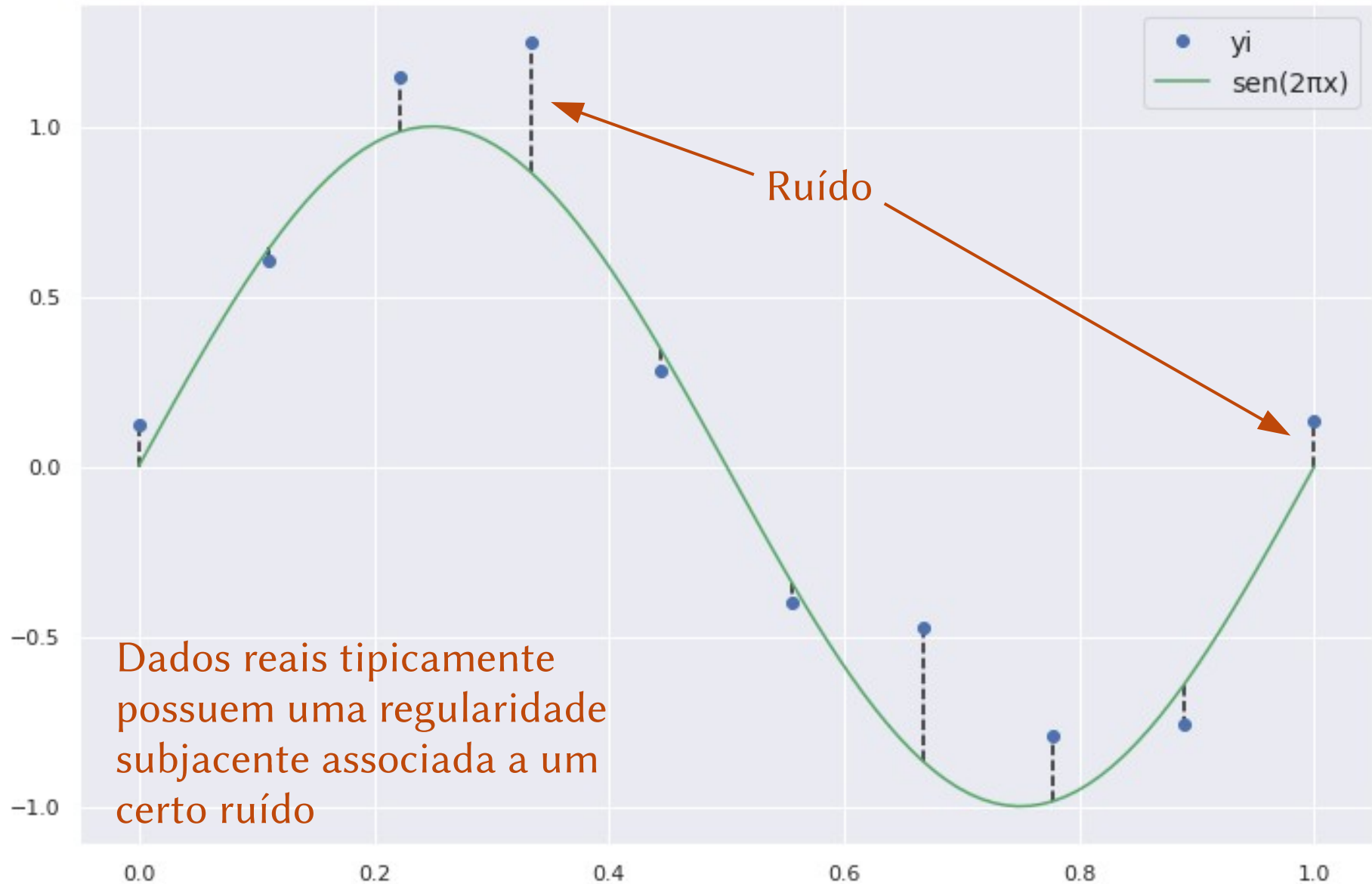
$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N [y(x_i, \mathbf{w}) - t_i]^2$$

Essa equação é denominada **função de perda** do regressor linear.

# Erro de um regressor linear

- Podemos ilustrar esse conceito com um experimento
  - Vamos gerar um pequeno conjunto de treinamento (10 pontos)
  - Cada ponto segue uma estrutura regular afetada por um pequeno ruído
    - $x = \sin(2\pi x) + X$
    - $X \sim \mathcal{N}(\mu, \sigma^2)$  é uma variável aleatória que segue uma distribuição gaussiana

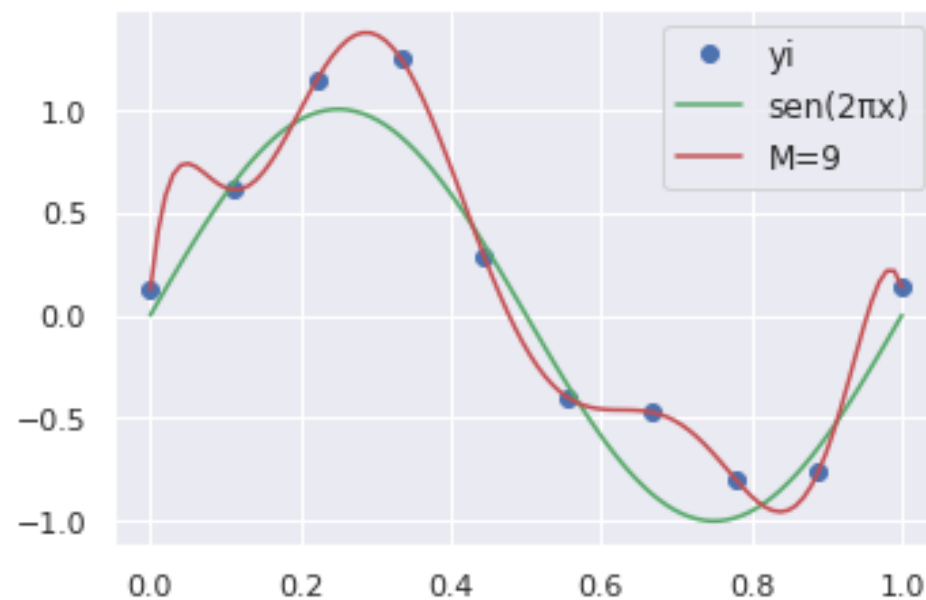
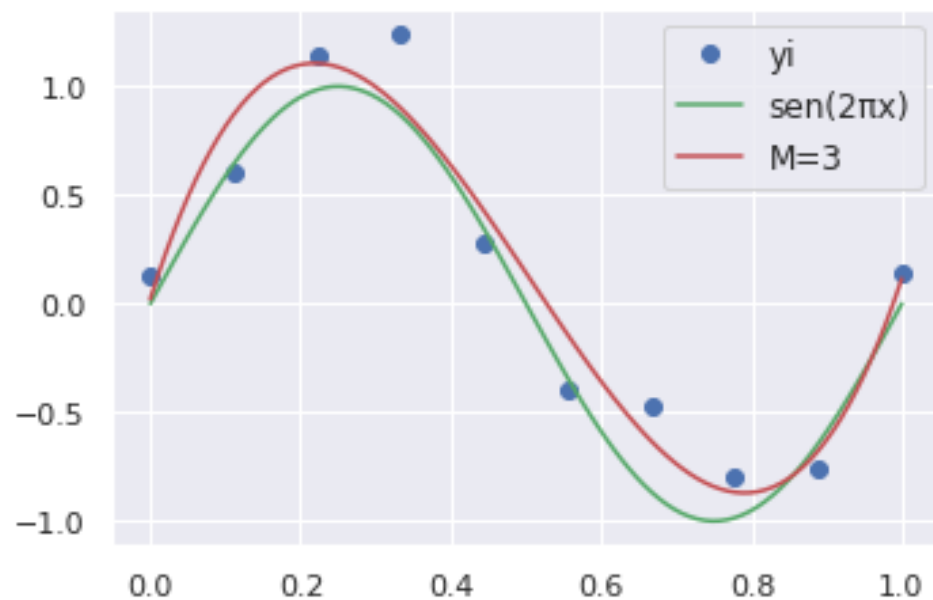
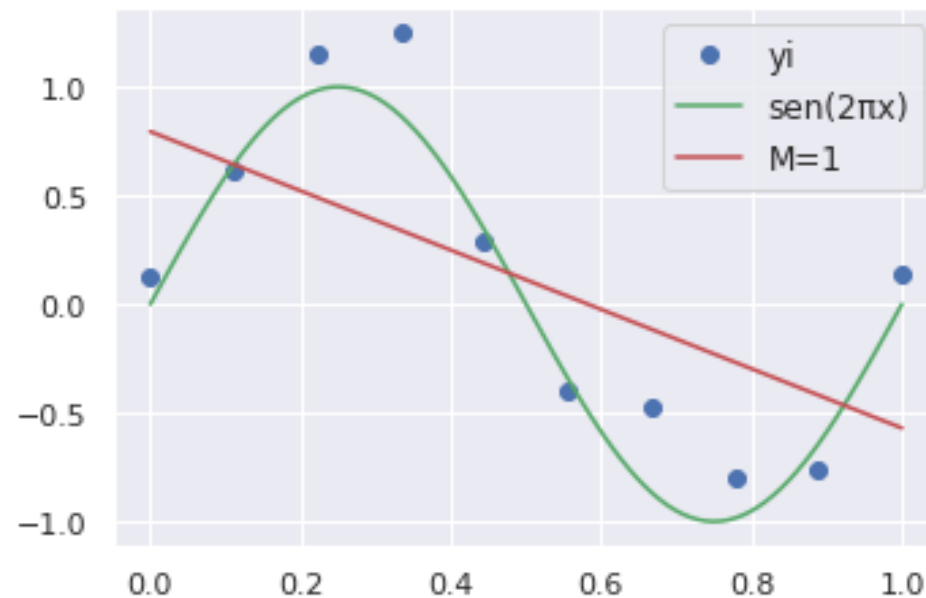
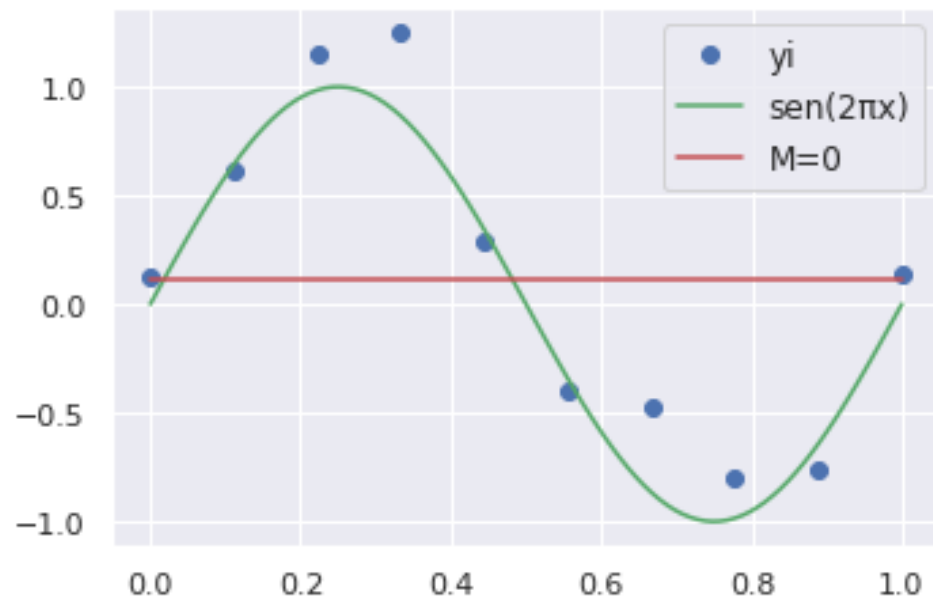
# Erro de um regressor linear



# Erro de um regressor linear

- O modelo ideal para uma amostra de treinamento é o conjunto de parâmetros  $\mathbf{w}^*$  que minimiza a função de perda para essa amostra
  - Existem vários métodos para minimizar  $E(\mathbf{w})$
  - Como a relação  $\mathbf{w} \cdot \mathbf{x}$  é linear, existe uma fórmula fechada
- O número de coeficientes em  $\mathbf{w}$  e  $\mathbf{x}$  depende do hiperparâmetro  $M$ , que estabelece a **complexidade** do modelo

# Complexidade do regressor linear



# Complexidade do regressor linear

- Notamos que, quanto mais complexo o modelo, mais fielmente ele aproxima os exemplos de treinamento
- Podemos mensurar essa aproximação em função do **erro quadrático médio**

$$\text{RMSE} = \sqrt{\frac{2E(\mathbf{w})}{N}}$$



# Complexidade do regressor linear

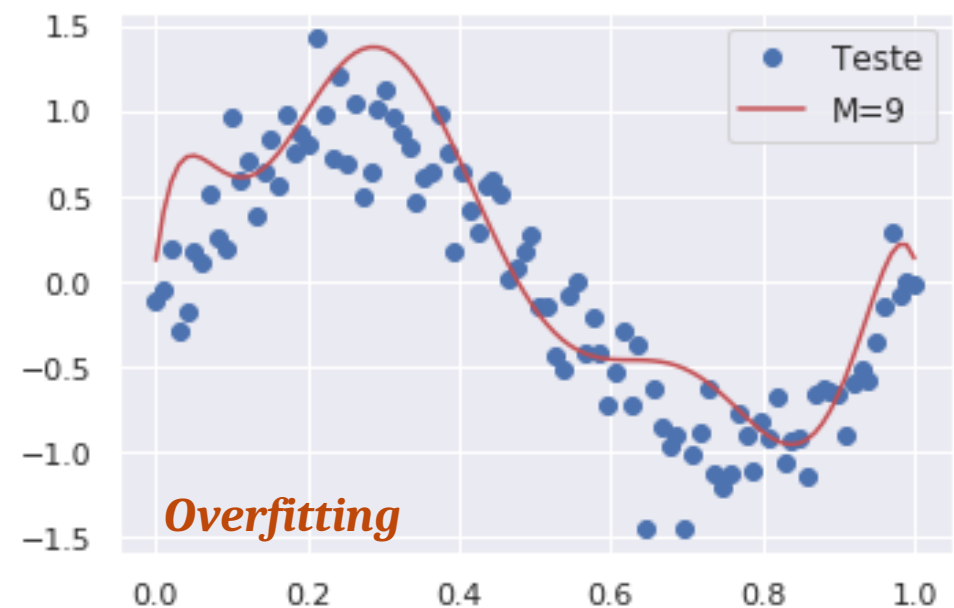
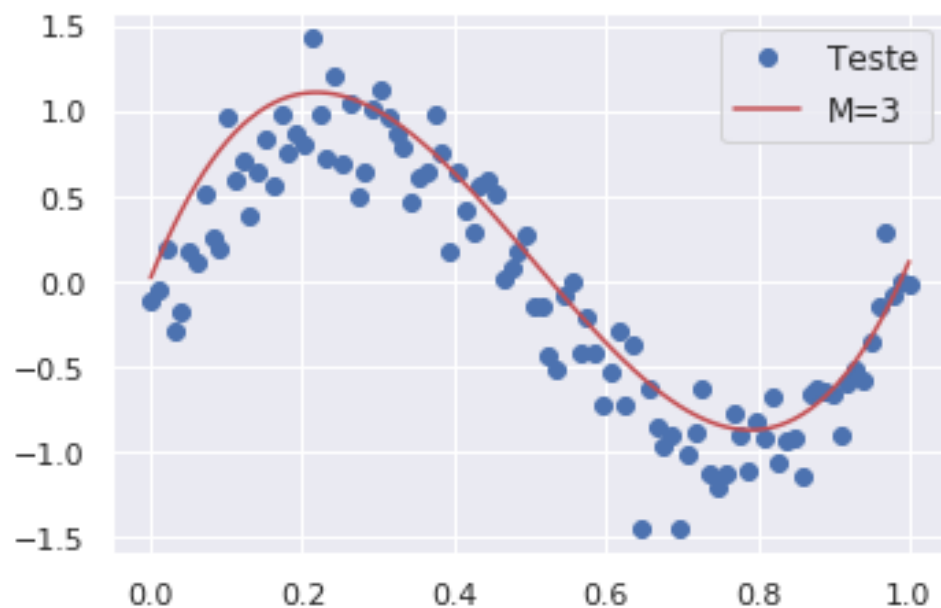
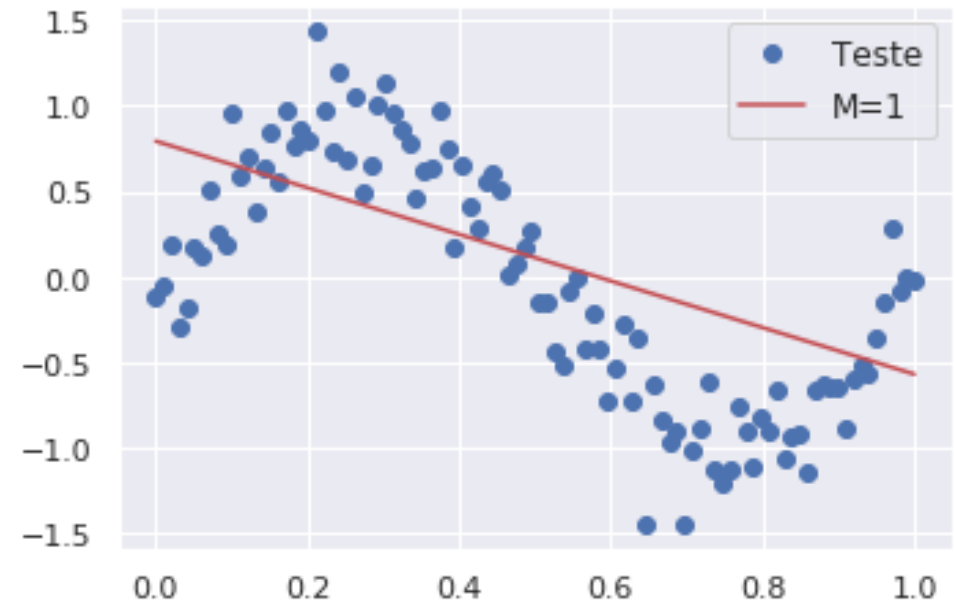
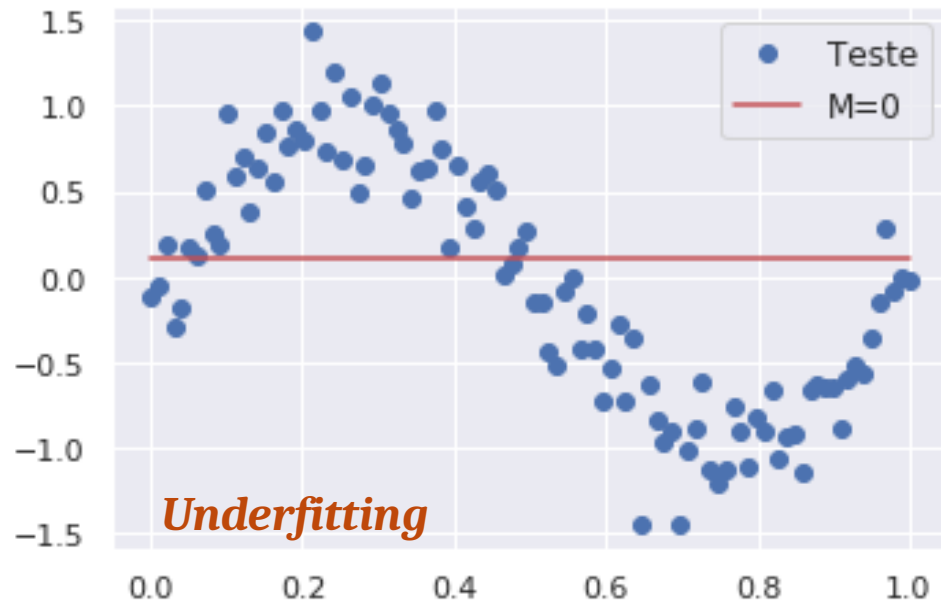
- O erro empírico, em função do RMSE, diminui conforme a complexidade do modelo aumenta
- Isso significa que o modelo é bom?
- Podemos definir  $M=9$ , colocar o modelo em ambiente de produção e \$\$\$?

M	RMSE
0	0,6941
1	0,5405
2	0,5399
3	0,1809
4	0,1804
5	0,1684
6	0,0897
7	0,0884
8	0,0008
9	0

# Complexidade do regressor linear

- Podemos simular o que aconteceria com esses modelos caso fossem utilizados no mundo real gerando um novo conjunto
- Conjunto de teste (100 pontos)
  - $x = \sin(2\pi x) + \mathbf{X}$
  - $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$  é uma variável aleatória que segue uma distribuição gaussiana

# Complexidade do regressor linear



# Agenda

- Revisão e formalização dos conceitos
- Generalização
- **Aprendizado viciado**
- Dilema viés-variância
- Teorema NFL (*no free-lunch*)
- Combatendo o aprendizado viciado

# Aprendizado viciado

- O aprendizado viciado ocorre quando o modelo não consegue generalizar o conceito
  - Muitas vezes caracterizado por um erro empírico baixo e um erro de generalização elevado
  - Pode ocorrer por superajuste ou sub-ajuste do modelo aos dados
  - Ou por uma experiência de aprendizado excessiva

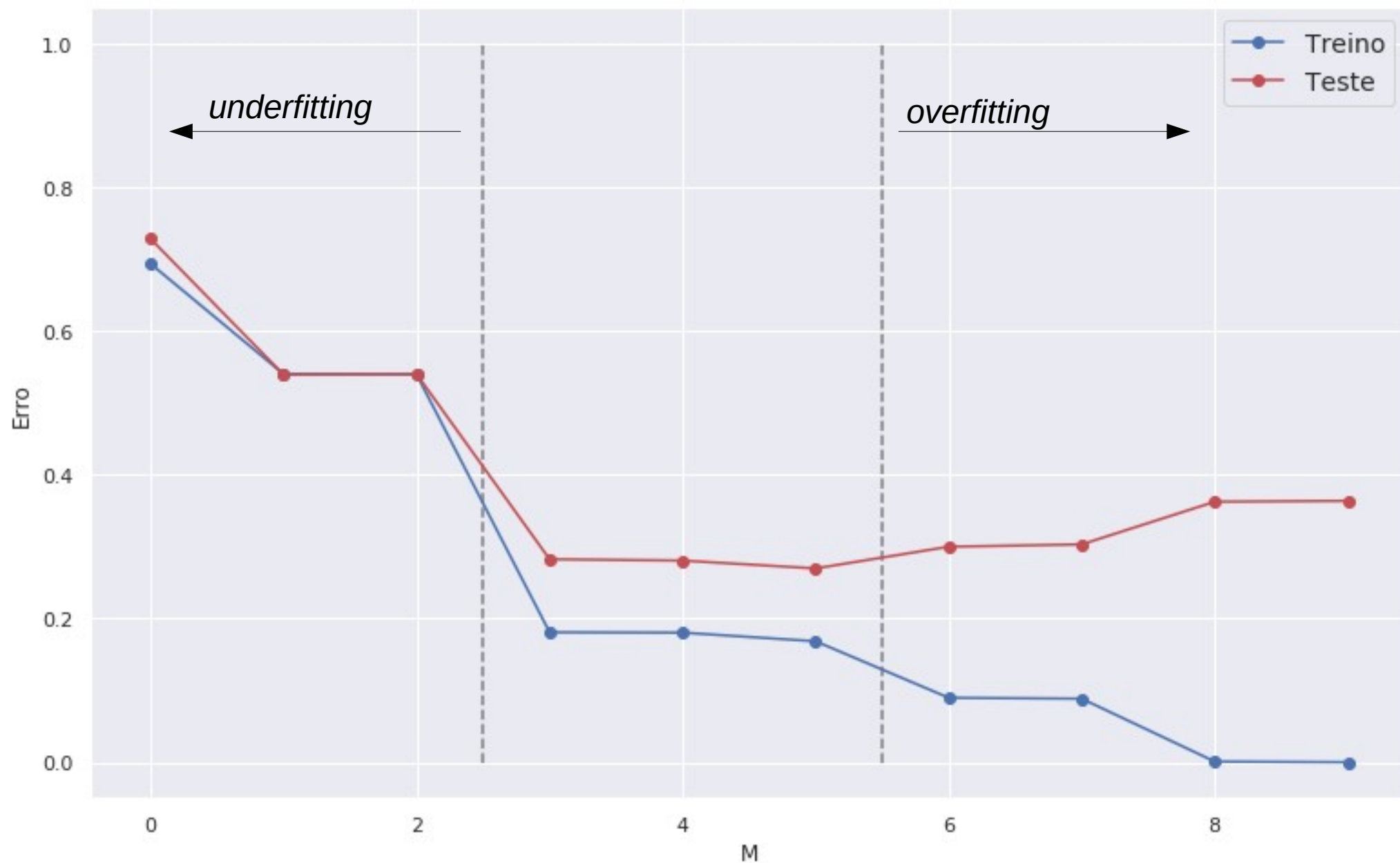
# Superajuste e sub-ajuste

- O modelo está **sub-ajustado** aos dados (*underfitting*) quando ele não consegue alcançar a complexidade necessária para representar nem mesmo os exemplos de treinamento
- O modelo está **superajustado** aos dados (*overfitting*) quando ele é complexo demais para os dados
  - Erro empírico baixo, erro de generalização alto

# Overfitting e underfitting

- Nosso regressor linear sofre *underfitting* com  $M=1$  porque uma reta é simples demais para representar dados espalhados periodicamente
- Já com  $M=9$  o regressor sofre *overfitting* porque tem liberdade de mais para se adaptadr aos dados
  - De fato,  $M=9$  provê 10 graus de liberdade
  - O modelo pode ter coeficientes que lhe permitem "passar" **exatamente** sobre os 10 pontos de treinamento

# Overfitting e underfitting

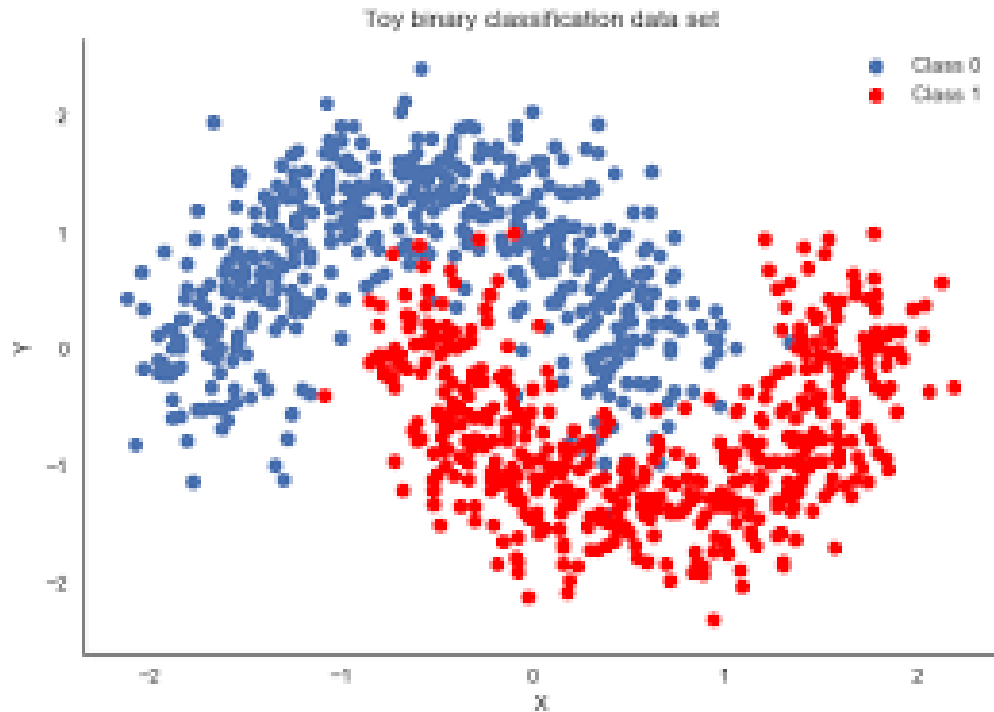




# Motivos do aprendizado viciado

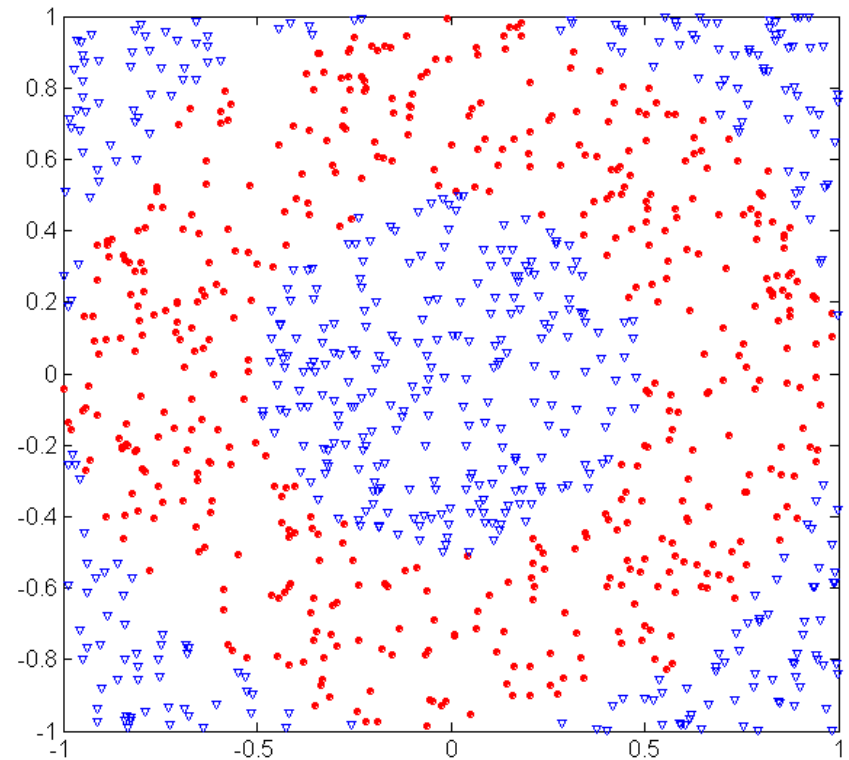
- Amostras inadequada
  - Poucos exemplos ou **má distribuição** das classes
  - Excesso de ruído
- Atributos inadequados
  - Mal escolhidos
  - Número excessivo (**maldição da dimensionalidade**)
- Função de aproximação muito complexa
- Treinamento excessivo

# Má distribuição das classes



**Caso 1:** Elevada probabilidade de ser uma boa representação dos dados: alta similaridade intra-classe e baixa similaridade inter-classe.

**Caso 2:** Pode conduzir a padrões muito dispersos para as classes => pouco valor estatístico



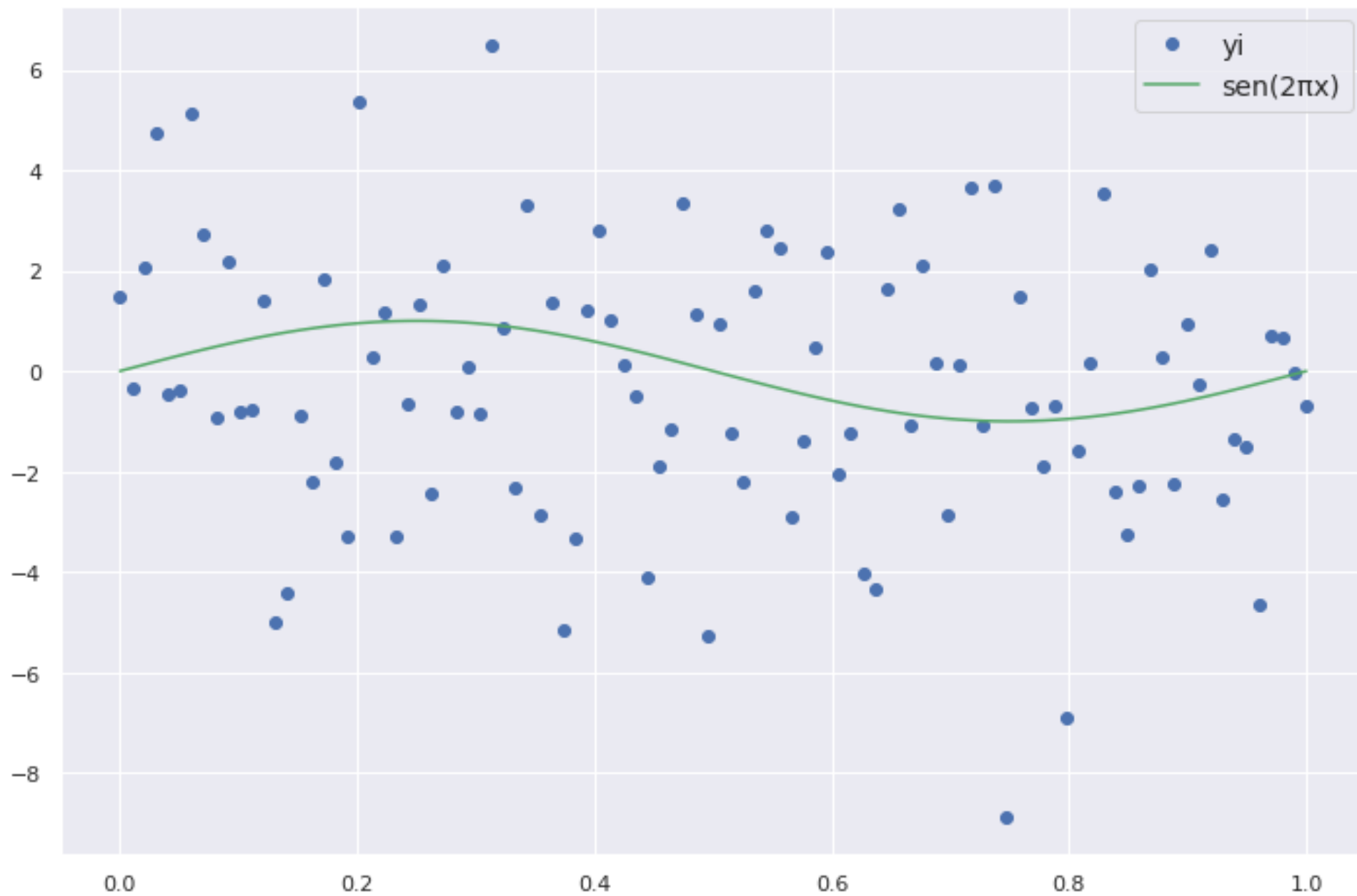
# Falta de representatividade

<b>Nome</b>	<b>Temp. do Corpo</b>	<b>Nasc. Uterino</b>	<b>4 Pernas</b>	<b>Hiberna</b>	<b>Mamífero?</b>
Salamandra	fria	não	sim	sim	Não
“Guppy”	fria	sim	não	não	Não
Águia	quente	não	não	não	Não
“Poorwill”	quente	não	não	sim	Não
Ornitorrinco	quente	não	sim	sim	Sim

# Falta de representatividade

Nome	Temp. do Corpo	Nasc. Uterino	4 Pernas	Hiberna	Mamífero?
Salamandra	fria	não	sim	sim	Não
“Guppy”	fria	sim	não	não	Não
Águia	quente	não	não	não	Não
“Poorwill”	quente	não	não	sim	Não
Ornitorrinco	quente	não	sim	sim	Sim
<i>Professorium inventatus</i>	quente	sim	não	não	?

# Ruído



# Maldição da dimensionalidade

- A maldição da dimensionalidade é um fenômeno que ocorre quando a dimensão dos dados aumenta excessivamente
- Em síntese: muitos atributos
  - Quando o número de dimensões aumenta demais, o conceito de distância perde o significado
  - Precisamos de **muitos** exemplos para caracterizar o conceito
- Vamos retomar esse ponto em aulas posteriores

# Treinamento excessivo

- Observado principalmente em modelos de redes neurais
- Também é causado por um superajuste do modelo aos exemplos de treinamento
  - Conjunto muito grande de exemplos com pequena variação intra-classe
  - Ou com muitas iterações de treinamento

# Agenda

- Revisão e formalização dos conceitos
- Generalização
- Aprendizado viciado
- Dilema viés-variância
- Teorema NFL (*no free-lunch*)
- Combatendo o aprendizado viciado



# Dilema viés/variância

- O erro de generalização pode ser decomposto em
  - **Viés**: diz respeito às preferências do modelo
    - Viés elevado implica que o modelo tem pouca flexibilidade em se ajustar aos exemplos de treinamento
  - **Variância**: diz respeito à capacidade do modelo se ajustar à variância dos exemplos de treinamento
    - Variância elevada implica que o modelo é muito afetado por ruído nos dados

# Dilema viés/variância

- Viés e variância estão sempre **acoplados**
  - Quando se reduz o viés, a tendência é que a variância aumente
  - Quando se reduz a variância, a tendência é que o viés aumente
  - Existe um **compromisso** entre viés e variância
- Precisamos selecionar corretamente os hiperparâmetros e a experiência de aprendizado
  - **Como???**

# Agenda

- Revisão e formalização dos conceitos
- Generalização
- Aprendizado viciado
- Dilema viés-variância
- Teorema NFL (*no free-lunch*)
- Combatendo o aprendizado viciado

# *No free lunch*

- Formalmente, é um conceito que se aplica a sistemas de otimização
  - Sua versão para sistemas de otimização é um pouco mais complexa do que empregamos em problemas de aprendizado indutivo
  - A "variante" que estudamos também se aplica a diversas outras áreas da computação

# No free lunch

- No aprendizado indutivo, o teorema do *no free lunch* diz que não é possível aprender um conceito sem levar em conta o contexto do problema e do usuário
  - Um modelo não pode ser melhor que todos os outros em todos os casos
  - Um algoritmo não pode ser mais rápido que todos os outros em todas as instâncias
  - Um algoritmo não pode usar menos memória que todos os outros para todas as instâncias

# No free lunch

- Uma intuição
  - Imagine um algoritmo  $f(\mathbf{x}) = \mathbf{y}$  que retorna uma versão comprimida da imagem  $\mathbf{x}$
  - Suponha que esse algoritmo é o *melhor do mundo* para qualquer tipo imagem
  - Dada uma imagem  $\mathbf{x}$ , a razão de compressão do algoritmo  $f$  será maior que qualquer outro algoritmo  $g$  para a mesma imagem

# No free lunch

- Uma intuição
  - Agora suponha um algoritmo  $F$  (super  $f$ ) que funciona da seguinte forma para esta foto:

$F(\mathbf{x})$ :

se  $\mathbf{x}$  é a imagem de Lenna:

**retorne** o bit 0

**senão:**

**retorne** o bit 1 seguido da  
sequência de bits  $f(\mathbf{x})$



# No free lunch

- Uma intuição
  - A taxa de compressão de  $F$  será muito menor que  $f$  para a imagem de Lenna
  - Se todos os exemplos forem igualmente prováveis, então o desempenho de  $f$  será, na média, maior que  $F$
  - Entretanto, se o conjunto de teste for apenas imagens de Lenna,  $F$  será melhor



## *No free lunch* em AM

- Não se pode comparar dois modelos livre de contexto
- Sem saber qual é a separação ideal ou a distribuição exata dos dados, ao constatar que o modelo  $f_a(\mathbf{x}, \mathbf{w}_1)$  parece melhor que o modelo  $f_b(\mathbf{x}, \mathbf{w}_2)$ 
  - Como podemos determinar se isso se deve ao tipo do modelo, aos parâmetros obtidos ou aos dados que utilizamos para os testes?

# No free lunch em AM

- Conclusões
  - Todas as afirmações do tipo " $f_a(\mathbf{x}, \mathbf{w}_1)$  parece melhor  $f_b(\mathbf{x}, \mathbf{w}_2)$ " são relevantes para o domínio de aplicação investigado e não para todos os problemas
  - Quanto mais ricos os dados e maior o número de algoritmos testados, melhores serão as conclusões
  - Conclusões tiradas sobre modelos mais simples tendem a ser mais confiáveis

**Navalha de Occam**

# Agenda

- Revisão e formalização dos conceitos
- Generalização
- Aprendizado viciado
- Dilema viés-variância
- Teorema NFL (*no free-lunch*)
- Combatendo o aprendizado viciado

# Como combater o AM viciado?

- Não é possível garantir que o modelo aprendido será capaz de generalizar o conceito
- Mas podemos adotar estratégias para tentar generalizar mais
  - Obter mais dados ajuda até certo ponto
  - A partir desse ponto, os novos exemplos são representações redundantes do conceito
  - Outras técnicas incluem regularização, adotar um conjunto de validação e *ensembles*

# Regularização

- Alguns modelos admitem **regularização**
- Se observarmos os valores dos parâmetros para diferentes valores de  $M$ , veremos que  $||\mathbf{w}||$  tende a aumentar
- O modelo tem muita liberdade para se ajustar aos dados

	M=0	M=1	M=3	M=9
$w_0$	0,19	0,82	0,31	0,35
$w_1$		-1,27	7,99	232,27
$w_2$			-25,43	-5.321,83
$w_3$			17,37	48.568,31
$w_4$				-231.639,30
$w_5$				640.042,26
$w_6$				-1.061.800,52

# Regularização

- A regularização restringe o modelo "punindo" valores elevados dos parâmetros
- Isso é obtido através de uma pequena modificação da função de perda

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N [y(x_i, \mathbf{w}) - t_i]^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Os termos  $\frac{1}{2}$  não são estritamente necessários, mas simplificam o cálculo de  $\mathbf{w}^*$

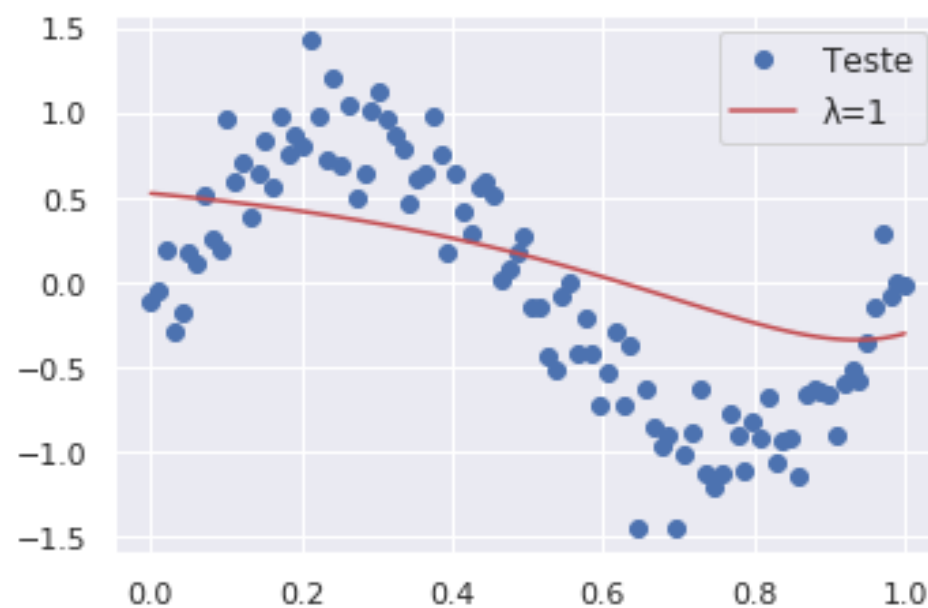
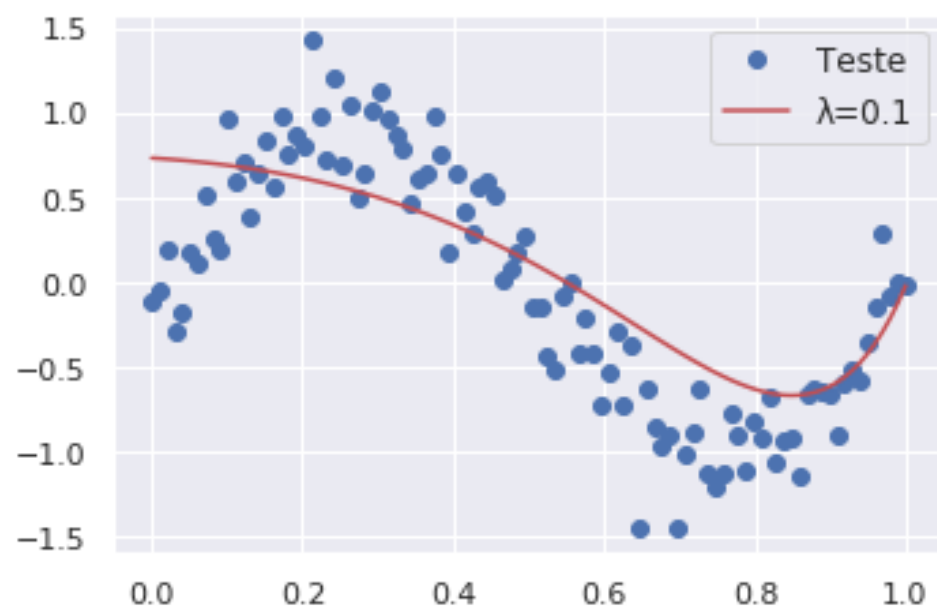
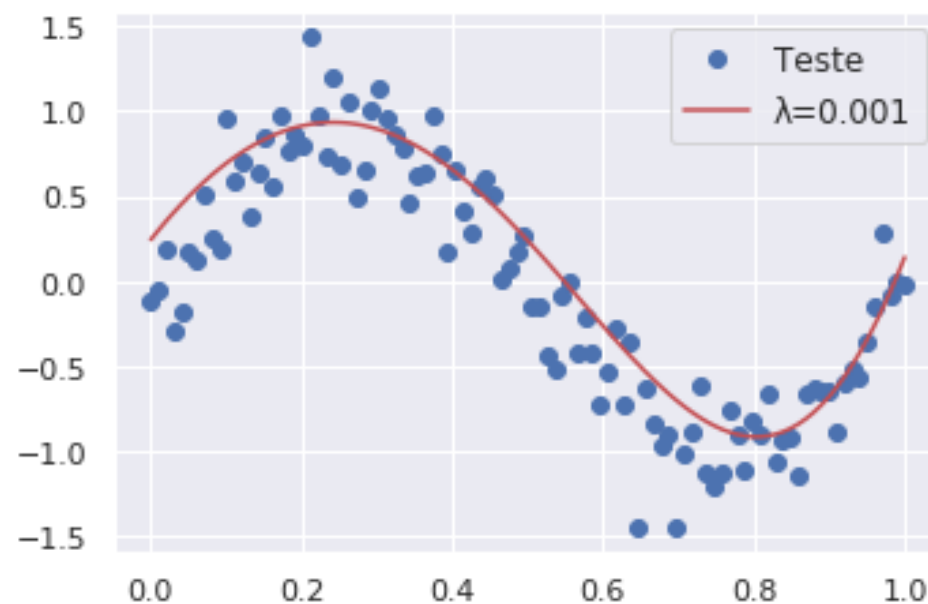
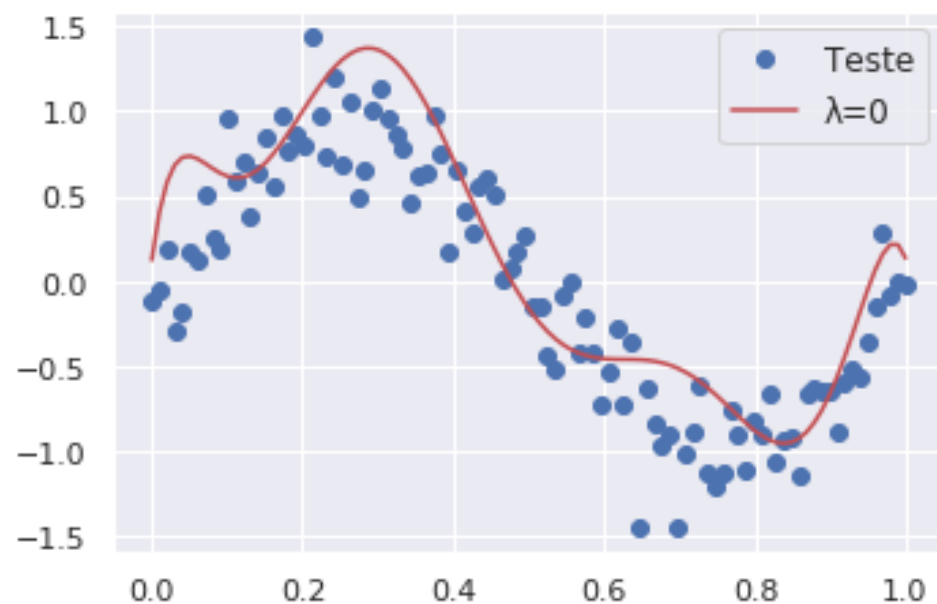
# Regularização

- Expandindo o regularizador, temos

$$||\mathbf{w}||^2 = \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + w_2^2 + \dots + w_M^2$$

- Portanto a minimização de  $\tilde{E}(\mathbf{w})$  não apenas minimiza o erro empírico, mas também os coeficientes de  $\mathbf{w}$

# Regularização



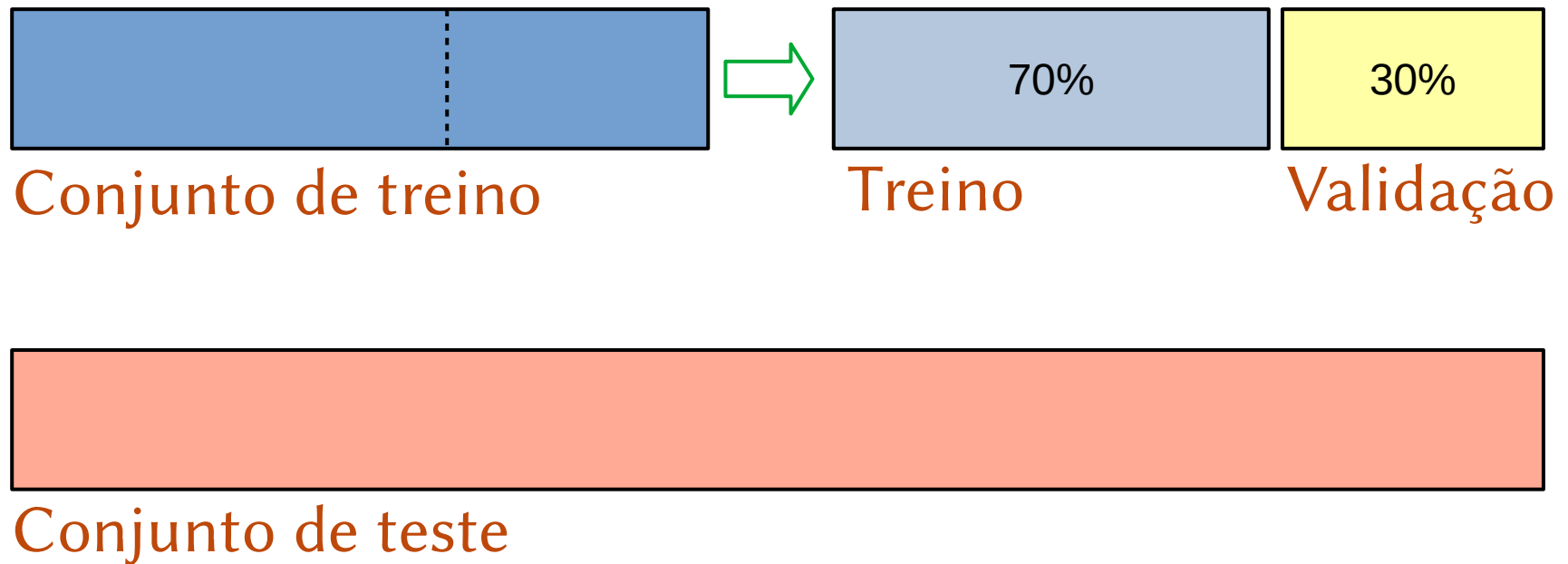


# Regularização

- Consequências da regularização
  - O modelo é desencorajado a privilegiar uma variável de entrada  $x_i$  em detrimento de outras
  - A variância aumenta e a possibilidade de *overfitting* diminui
  - Em contrapartida, o modelo regularizado possui um hiperparâmetro a mais

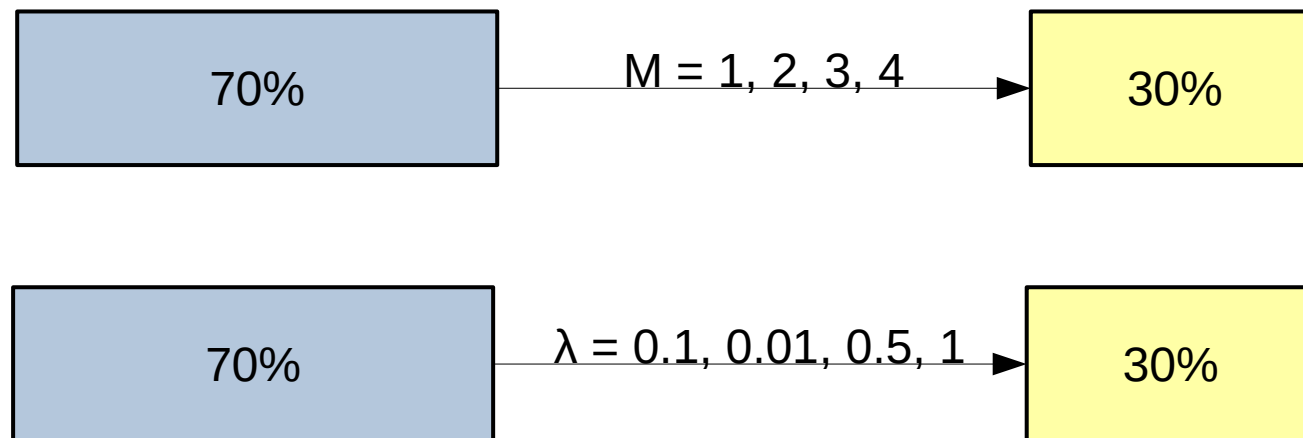
# Validação

- Podemos separar uma porção dos dados de treinamento para **validação**



# Validação

- O conjunto de validação pode ser utilizado para testar diferentes configurações de hiperparâmetros



- A melhor configuração de hiperparâmetros pode ser utilizada para gerar o modelo final

# Classificação

- Um classificador é um modelo na forma

- $f(x_i) = \hat{y}_i$

sendo que  $y_i \in \Omega = \{\Omega_1, \Omega_2, \dots, \Omega_c\}$  é o espaço de classes associado à função-conceito

- Quando  $|\Omega| = 2$ , tem-se classificação **binária**
  - Quando  $|\Omega| > 2$ , tem-se classificação **multiclasse**
    - Não confundir com classificação **multirrótulo**

# Classificação

- A função de perda de um classificador é

$$L(y, \hat{y}) = L(y, f(\mathbf{x}, \mathbf{w}))$$

sendo

$$L(y, \hat{y}) = \begin{cases} 0, & \text{se } y = \hat{y} \\ 1, & \text{se } y \neq \hat{y} \end{cases}$$

# Classificação

- Portanto o erro empírico do classificador é

$$E(f) = \sum_{i=1}^N \frac{L(y_i, f(\mathbf{x}_i, \mathbf{w}))}{N}$$

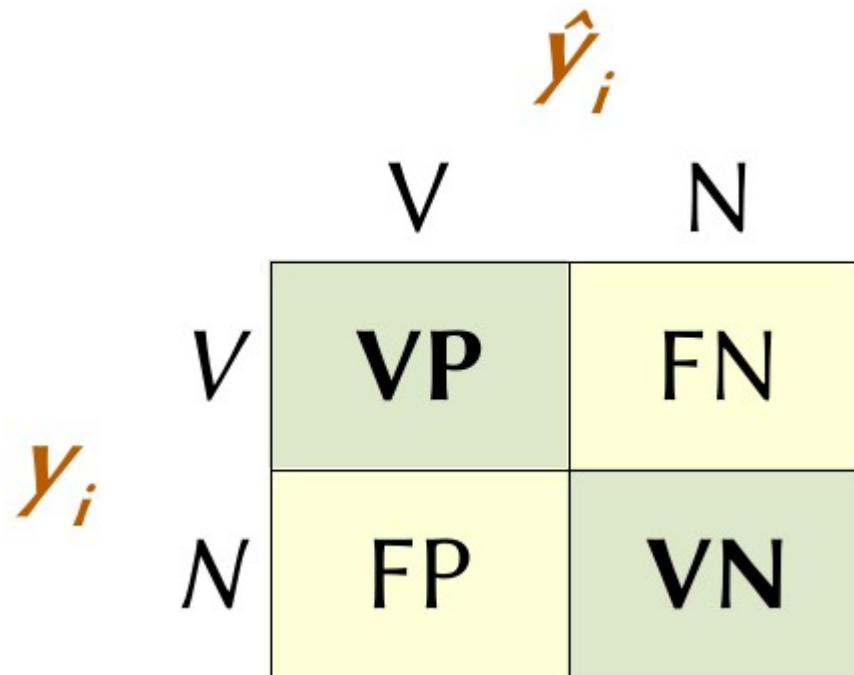
# Matriz de confusão

- Para problemas de decisão binária, podemos considerar o desempenho do classificador analisando quatro casos

Caso	Classe real	Classe inferida
Verdadeiro positivo	Positiva	Positiva
Falso negativo	Positiva	Negativa
Falso positivo	Negativa	Positiva
Verdadeiro negativo	Negativa	Negativa

# Matriz de confusão

- Dados  $y_i$  e  $\hat{y}_i$  para um conjunto de teste, podemos contar os casos e encontrar a **matriz de confusão**



A confusion matrix diagram showing the relationship between actual labels ( $y_i$ ) and predicted labels ( $\hat{y}_i$ ). The matrix is a 2x2 grid with rows labeled  $y_i$  (V, N) and columns labeled  $\hat{y}_i$  (V, N). The cells contain the counts: VP (True Positive), FN (False Negative), FP (False Positive), and VN (True Negative). The VP and VN cells are shaded green, while the FN and FP cells are shaded yellow.

		$\hat{y}_i$	
		V	N
$y_i$	V	VP	FN
	N	FP	VN



# Matriz de confusão

- Dados  $y_i$  e  $\hat{y}_i$  para um conjunto de teste, podemos contar os casos e encontrar a **matriz de confusão**

$y_i$	$\hat{y}_i$	Caso
1	0	
1	1	
1	1	
1	0	
1	0	
0	0	
0	1	
0	0	
0	1	
0	0	
0	1	

		$\hat{y}_i$	
		V	N
$y_i$	V		
	N		

0: negativo; 1: positivo

# Matriz de confusão

- Dados  $y_i$  e  $\hat{y}_i$  para um conjunto de teste, podemos contar os casos e encontrar a **matriz de confusão**

$y_i$	$\hat{y}_i$	Caso
1	0	FN
1	1	VP
1	1	VP
1	0	FN
1	0	FN
0	0	VN
0	1	FP
0	0	VN
0	1	FP
0	0	VN
0	1	FP

		$\hat{y}_i$	
		V	N
$y_i$	V	2	3
	N	3	3

0: negativo; 1: positivo

# Matriz de confusão

- Com base na matriz de confusão podemos estabelecer diferentes métricas de qualidade de um modelo

- Erro

$$\text{Err}(f) = \frac{\text{FP} + \text{FN}}{N} = 1 - \text{Acc}(f)$$

- Acurácia

$$\text{Acc}(f) = \frac{\text{VP} + \text{VN}}{N} = 1 - \text{Err}(f)$$

# Matriz de confusão

- Nem sempre erro e acurácia são formas adequadas de testar a qualidade de um modelo
- Podemos utilizar a matriz de confusão para estabelecer diversas medidas de qualidade do modelo
  - Precisão

$$\text{Prec}(f) = \frac{VP}{VP + FP}$$

# Matriz de confusão

- Para problemas de decisão multiclasse, a matriz de confusão é facilmente extensível

	$\Omega_1$	$\Omega_2$	$\hat{y}_i$ $\Omega_3$	...	$\Omega_c$
$\Omega_1$					
$\Omega_2$					
$y_i$ $\Omega_3$					
$\vdots$					
$\Omega_c$					

# Matriz de confusão

- Para problemas de decisão multiclasse, a matriz de confusão é facilmente extensível

				$\hat{y}_i$		
		$\Omega_1$	$\Omega_2$	$\Omega_3$	...	$\Omega_c$
$y_i$	$\Omega_1$	<b>V</b>			...	
	$\Omega_2$		<b>V</b>		...	
	$\Omega_3$			<b>V</b>	...	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$\Omega_c$				...	<b>V</b>

Na diagonal principal estão os exemplos corretamente classificados

# Matriz de confusão

- Para problemas de decisão multiclasse, a matriz de confusão é facilmente extensível

		$\Omega_1$	$\Omega_2$	$\Omega_3$	...	$\Omega_c$
$\Omega_1$	<b>V</b>	F	F	...	F	
$\Omega_2$	F	<b>V</b>	F	...	F	
$\Omega_3$	F	F	<b>V</b>	...	F	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$\Omega_c$	F	F	F	...	<b>V</b>	

No restante da matriz estão os exemplos incorretamente classificados

# Matriz de confusão

- Para problemas de decisão multiclasse, a matriz de confusão é facilmente extensível

$\hat{y}_i$

	$\Omega_1$	$\Omega_2$	$\Omega_3$	...	$\Omega_c$	
$\Omega_1$	V	F	F	...	F	Acurácia
$\Omega_2$	F	V	F	...	F	
$\hat{y}_i$ $\Omega_3$	F	F	V	...	F	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$\Omega_c$	F	F	F	...	V	



# Matriz de confusão

- Para problemas de decisão multiclasse, a matriz de confusão é facilmente extensível

$\hat{y}_i$

	$\Omega_1$	$\Omega_2$	$\Omega_3$	...	$\Omega_c$	
$\Omega_1$	V	F	F	...	F	Erro
$\Omega_2$	F	V	F	...	F	
$y_i$ $\Omega_3$	F	F	V	...	F	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$\Omega_c$	F	F	F	...	V	

# Validação e teste

- Nas próximas aulas, estudaremos
  - Como validar modelos utilizando diferentes métricas
  - Como comparar modelos utilizando teorias de probabilidade e estatística
    - *Spoiler*: métodos de amostragem, validação cruzada e testes de hipótese