INF721

2023/2

UFV

Aprendizado em Redes Neurais Profundas

A2: Aprendizado de Máquina

Logística

Avisos

- Aula A1 - Introdução publicada no site [slides, vídeo]

Última aula

- Organização da disciplina
- Visão geral de aprendizado de máquina redes neurais

Plano de Aula

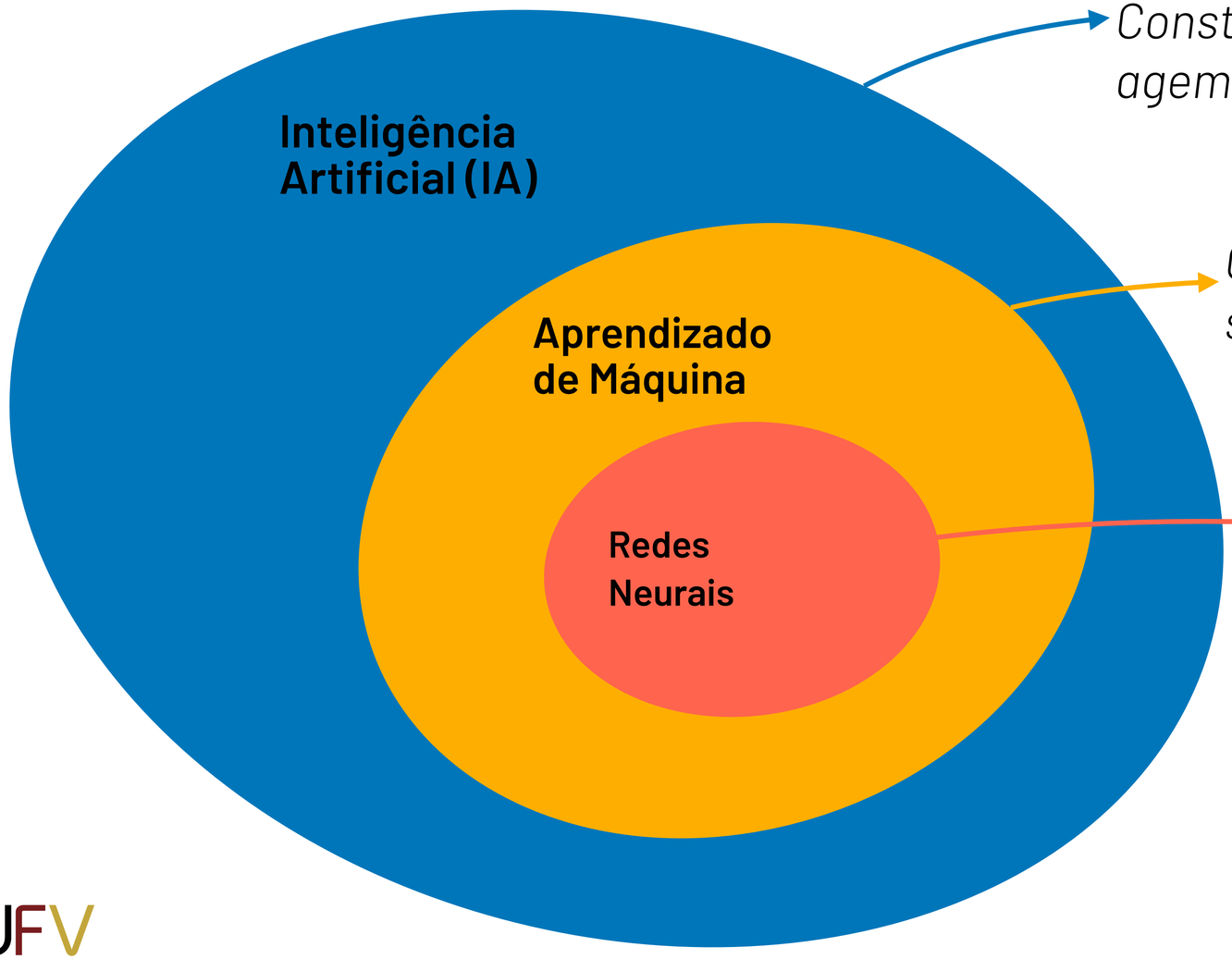
- Aprendizado de Máquina
- Tipos de Aprendizado
- Tipos de Dados
- Espaço de Hipóteses
- Função de Perda
- Generalização

Computação Clássica x Aprendizado de Máquina

Funções programadas explicitamente

Funções encontradas a partir de dados

Inteligência Artificial x Aprendizado de Máquina



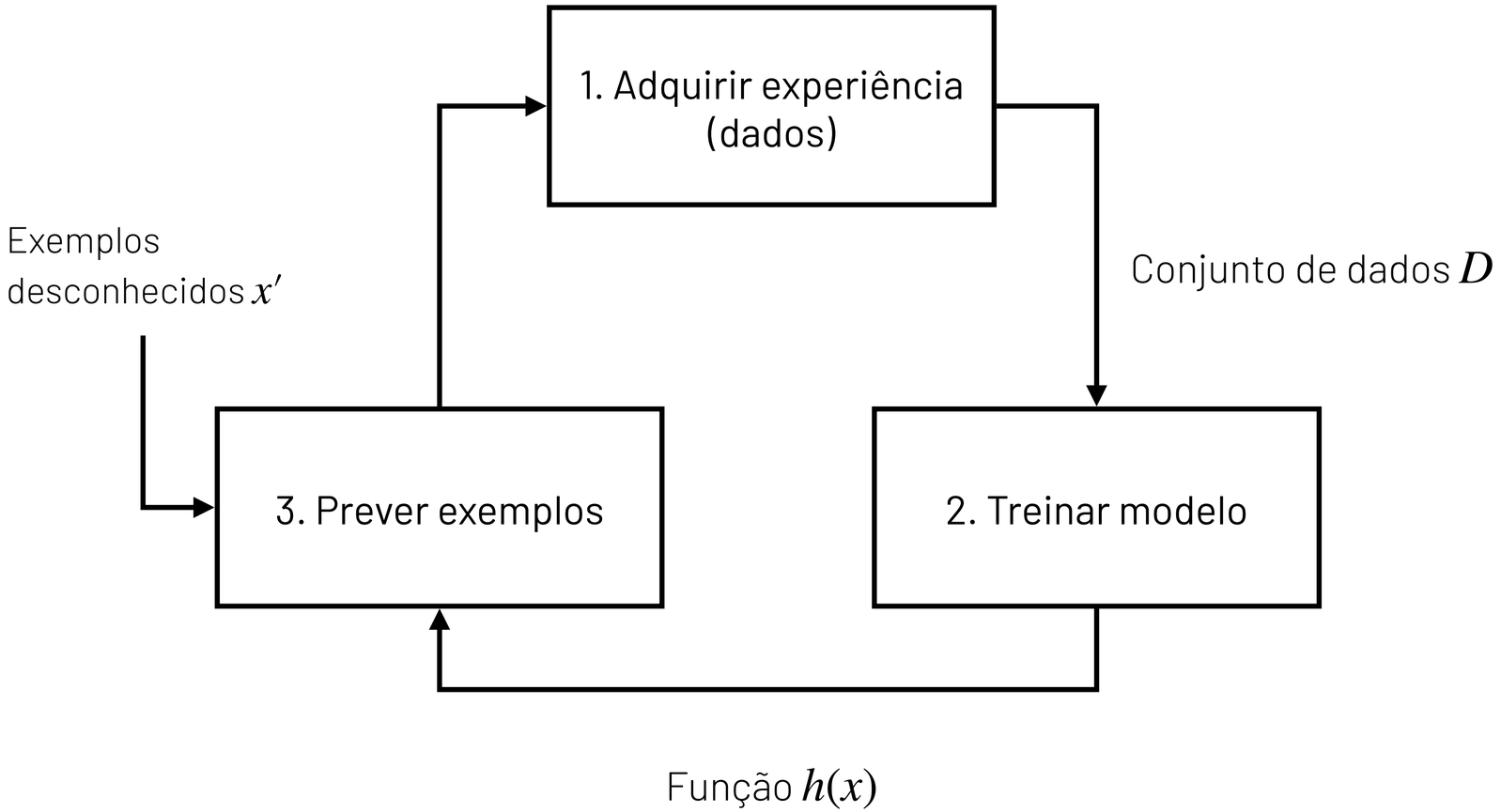
*Construção de **agentes racionais**, isto é, que agem de maneira a alcançar o melhor resultado*

Construção de agentes que melhoram o seu desempenho com novas experiências

Família de técnicas de aprendizado de máquina baseada na otimização de parâmetros de funções compostas.

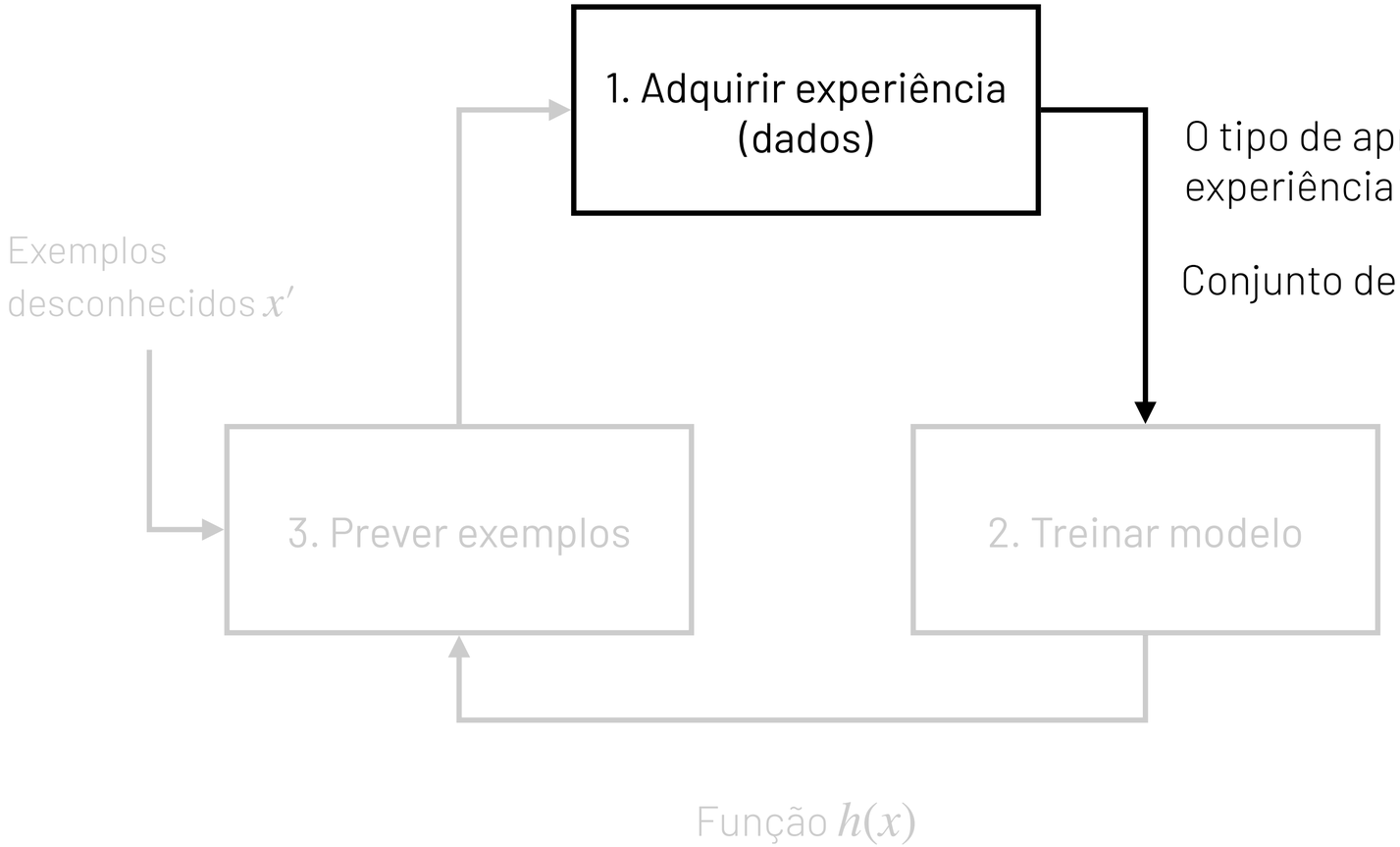
Aprendizado de Máquina

Aprender uma função $h(x)$ a partir de um conjunto de dados D para prever o rótulo de exemplos desconhecidos.



Tipos de Aprendizado

Aprender uma função $h(x)$ a partir de um conjunto de dados D para prever o rótulo de exemplos desconhecidos.



O tipo de aprendizado é definido pelo tipo de experiência (i.e., dados) observada pelo agente.

Conjunto de dados D

Rotulados

- **Aprendizado Supervisionado**

Não-rotulados

- **Aprendizado Não-supervisionado**

Valores de recompensa do ambiente

- **Aprendizado por Reforço**

Aprendizado Supervisionado

Quando todos os *exemplos* do conjunto de dados são pares (x_i, y_i), chamamos o problema de **Aprendizado Supervisionado**.

Formalmente:

$D = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathbb{R}^d \times C$, onde:

- x_i é o vetor de características do $i-$ésimo exemplo
- y_i é o rótulo (ou classe) do i-ésimo exemplo
- $\mathbb{R}^d$ é o espaço de características
- C é o espaço de classes

Exemplos de Aprendizado Supervisionado

Classificação de Imagens de Gatos e Cachorros

$D = \{$

$(x_1 =$  $, y_1 = 1),$

$(x_2 =$  $, y_2 = 1),$

$(x_3 =$  $, y_3 = 0),$

$(x_4 =$  $, y_4 = 0)\}$

- x_i: vetor com os pixels da imagem achatada
- y_i: gato (1) ou cachorro (0)
- d ~ 100.000 - 10M
- $C = \{0, 1\}$

Exemplos de Aprendizado Supervisionado

Classificação de Imagens de Dígitos Escritos Manualmente (MNIST)

$D = \{$

$(x_1 =$  $, y_1 = 0),$

$(x_2 =$  $, y_2 = 1),$

$(x_3 =$  $, y_3 = 5),$

$(x_4 =$  $, y_4 = 8)\}$

- x_i: vetor com os pixels da imagem achatada
- y_i: o valor do dígito da imagem
- d = 784 (28x28)
- C = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}

Exemplos de Aprendizado Supervisionado

Previsão de Preços de Imóveis

$$D = \{$$

$$(x_1 = [72, \text{Centro}, 2]\ , y_1 = 252{,}000),$$

$$(x_2 = [54, \text{Centro}, 1]\ , y_2 = 349{,}999),$$

$$(x_3 = [72, \text{Clélia}, 3]\ , y_3 = 380{,}250),$$

$$(x_4 = [182, \text{Ramos}, 4]\ , y_4 = 640{,}900)\}$$

- x_i: [tamanho (m^2), bairro, número de quartos]
- y_i: preço do imóvel
- $d = 3$
- $C = \mathbb{R}$

Aprendizado Supervisionado

Classificação

Quando o espaço de classes C é um conjunto com K rótulos (discreto e finito), chamamos o problema de **Classificação**.

Classificação Binária

- K = 2 rótulos possíveis: C = {0, 1}
- Exemplo: Classificação de Imagens de Gatos e Cachorros

Classificação Multiclasse

- K > 2 rótulos possíveis: C = {0, 1, 2, ... , K}
- Exemplo: Classificação de Imagens de Dígitos Escritos Manualmente

Aprendizado Supervisionado

Regressão

Quando o espaço de classes $C = \mathbb{R}$ é o conjunto dos reais (contínuo e infinito), chamamos o problema de **Regressão**.

Outros exemplos:

- Previsão de temperatura
- Previsão da nota de INF110 baseado no ENEM
- Regressão de caixa delimitadora

Aprendizado Não-supervisionado

Quando todos os exemplos do conjunto de dados são apenas vetores x_i, **sem rótulos**, chamamos o problema de **Aprendizado Não-supervisionado**.

Formalmente:

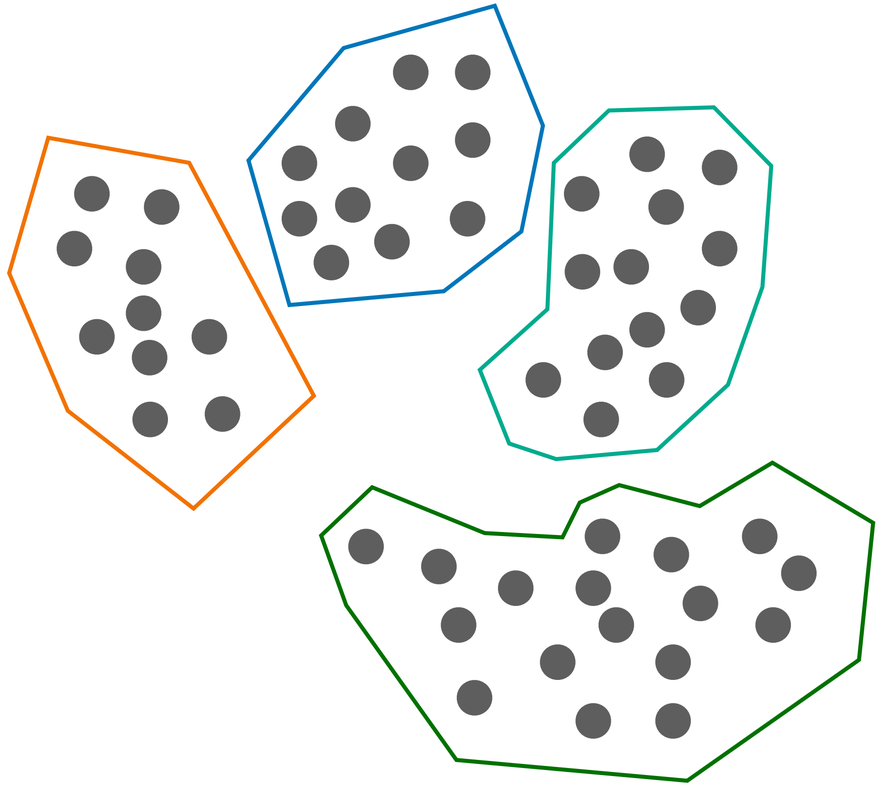
$D = \{x_1, x_2, \cdots, x_n\} \subseteq \mathbb{R}^d$, onde:

- x_i é o vetor de características do $i-$ésimo exemplo
- $\mathbb{R}^d$ é o espaço de características

Exemplos de Aprendizado Não-supervisionado

Agrupamento

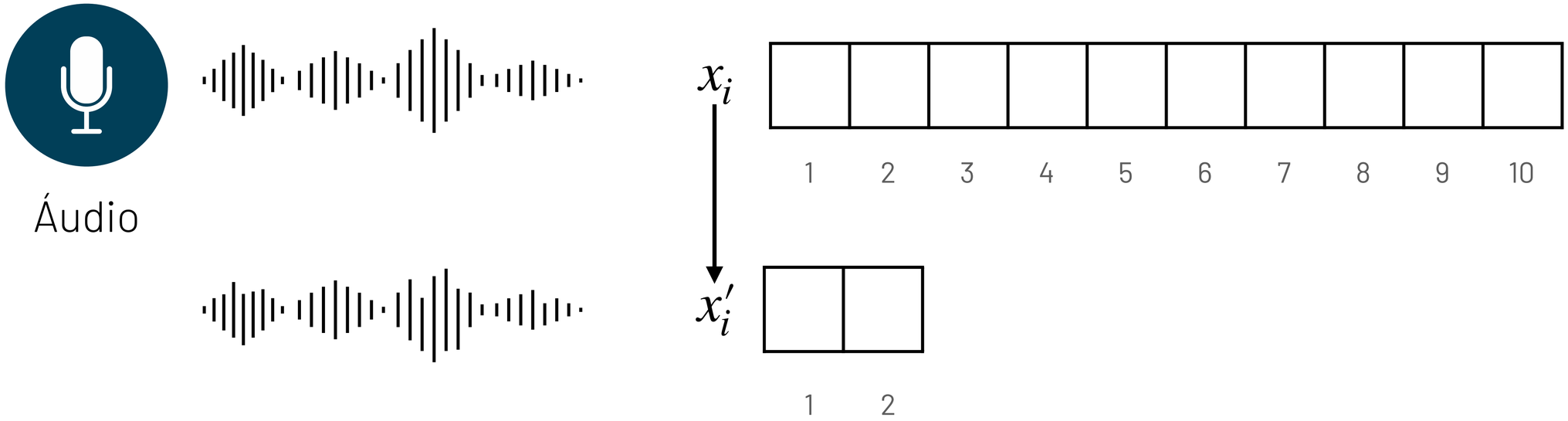
Agrupar os exemplos do conjunto de dados baseado em similaridade



Exemplos de Aprendizado Não-supervisionado

Redução de Dimensionalidade

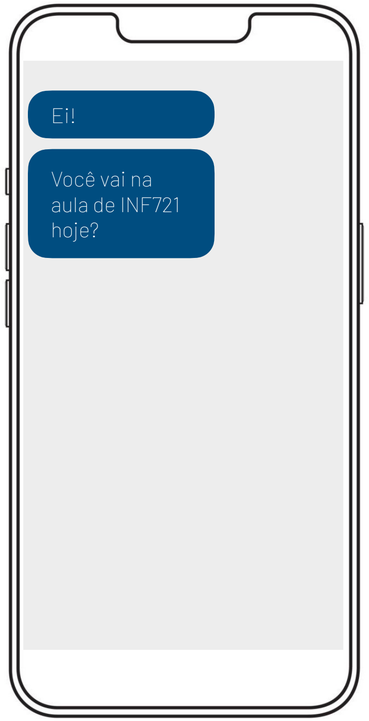
Reduzir a dimensionalidade d dos exemplos do conjunto de dados



Exemplos de Aprendizado Não-supervisionado

Geração de Dados

Inferir a distruibuição que gerou os dados do conjunto de dados



$$P(x_n \mid x_{n-1}, x_{n-2}, \cdots, x_1)$$

Modelo de linguagem

Aprendizado por Reforço

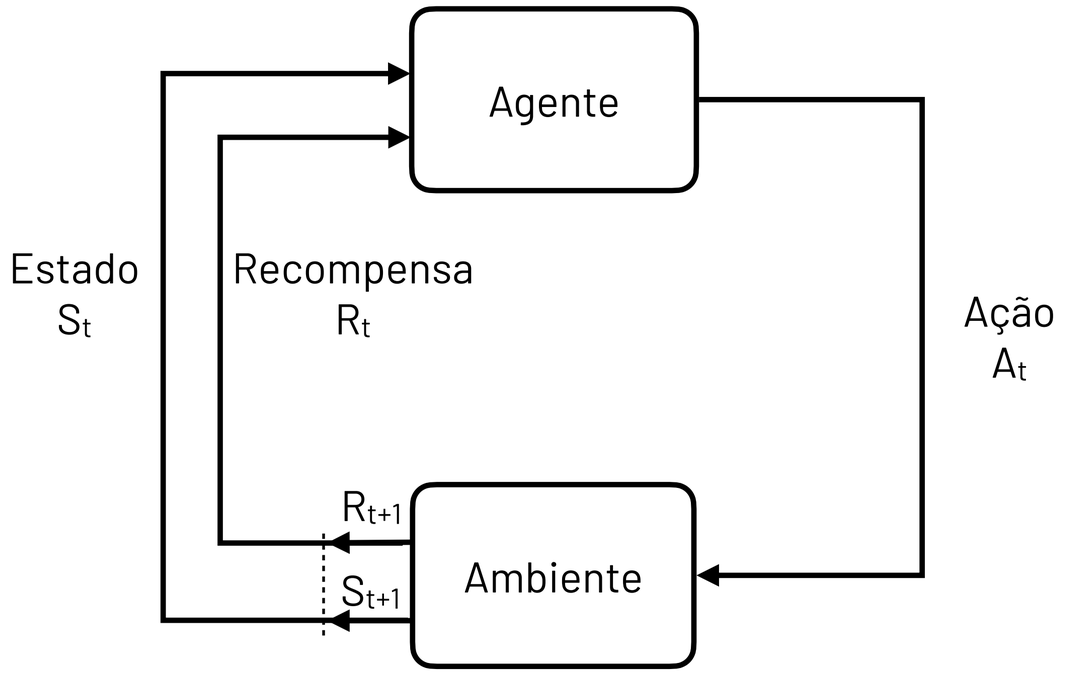
Aprender uma função $\pi(s) = a$ que prevê a *ação* a que um *agente* deve tomar no *estado* s, maximizando as *recompensas* recebidas pelo ambiente

Agente

- Observa uma estado s_t no tempo t
- Produz uma ação a_t no tempo t

Ambiente

- Retorna uma recompensa r_{t+1}
- Gera o próximo estado s_{t+1}



Tipos de Dados

Estruturados (tabulares)

Tamanho	Bairro	# de quartos	...	Preço
72	Centro	2		
54	Centro	1		
...	...	...		...
72	Clélia	3		

Idade	Estado	Ad Id	...	Click
72	MG	93242		1
54	SP	93287		0
...	...	...		...
72	RJ	71244		1

Não-estruturados (não-tabulares)

Imagens

Você vai na aula de INF721 hoje?

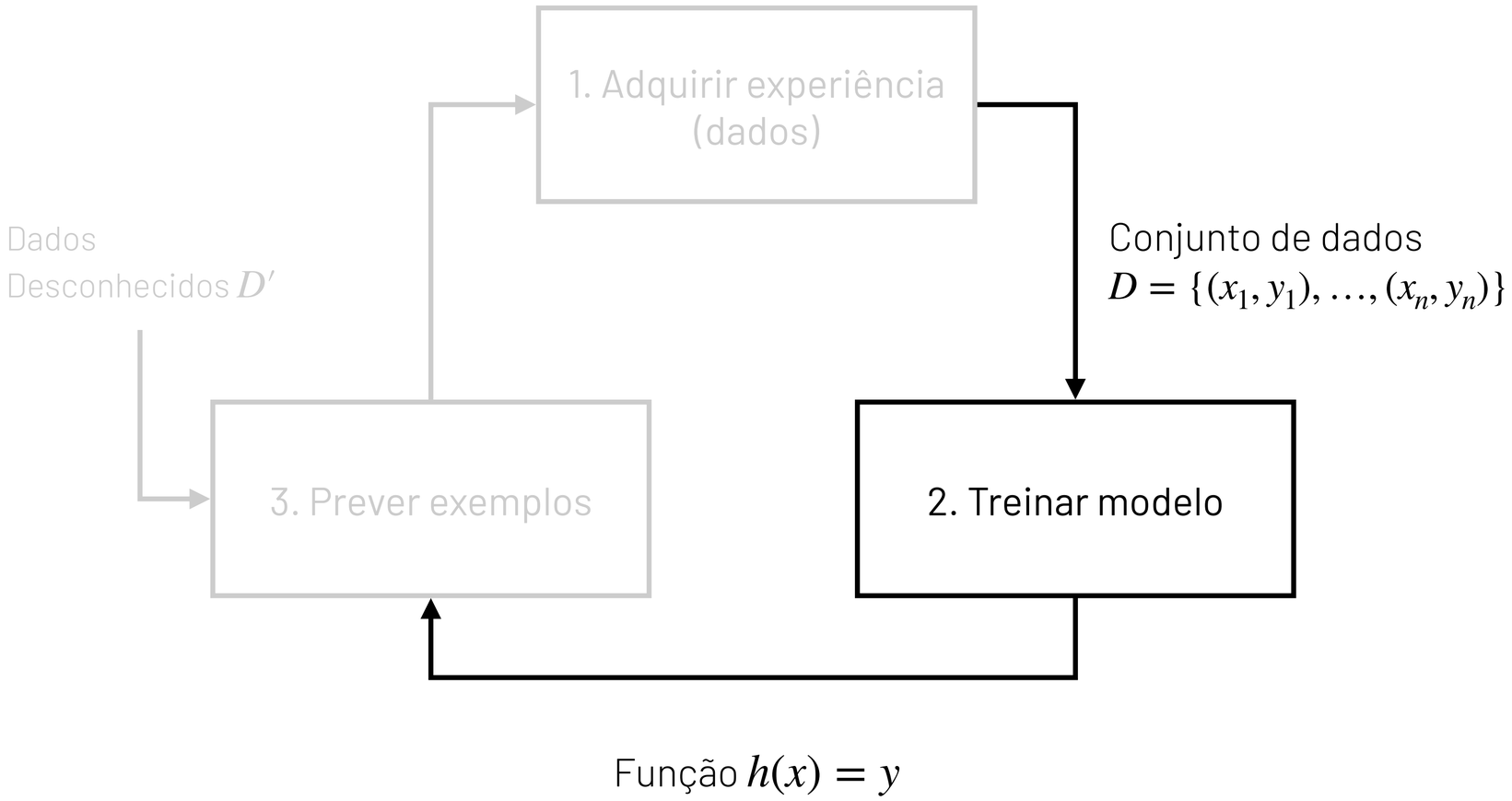
Texto

Áudio

Aprendizado Supervisionado

Aprendizado Supervisionado

Aprender uma função $h(x) = y$ a partir de um conjunto de dados D para prever o rótulo de exemplos desconhecidos.



Objetivo

Formalização

Assumindo que os exemplos $(x_i, y_i) \in D$ são amostrados de uma distribuição desconhecida $P(X, Y)$;

O **objetivo de aprendizado supervisionado** é:

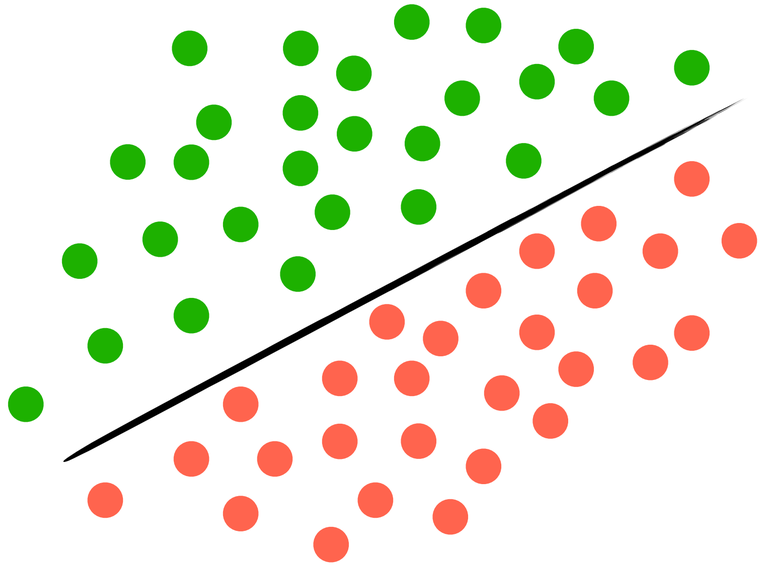
Dado um novo exemplo $(x', y') \notin D$ amostrado de $P(X, Y)$;

Encontrar uma função h a partir de D, tal que $h(x') \approx y'$
(O rótulo previsto $h(x')$ seja aproximadamente y')

Objetivo

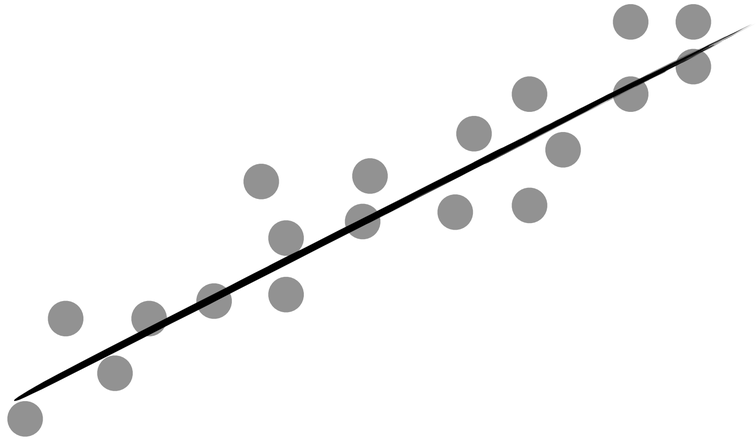
Visualização

Classificação



Encontrar uma função (e.g., linear) que *separa* as classes da melhor forma.

Regressão



Encontrar uma função (e.g., linear) que *passa* pelos pontos da melhor forma.

Treinamento

Treinar um modelo significa encontrar a melhor função $h \in H$ em um espaço específico de funções H.

Para isso, um algoritmo de aprendizado supervisionado precisa:

1. Definir um espaço específico de funções, chamado de **espaço de hipóteses** H;

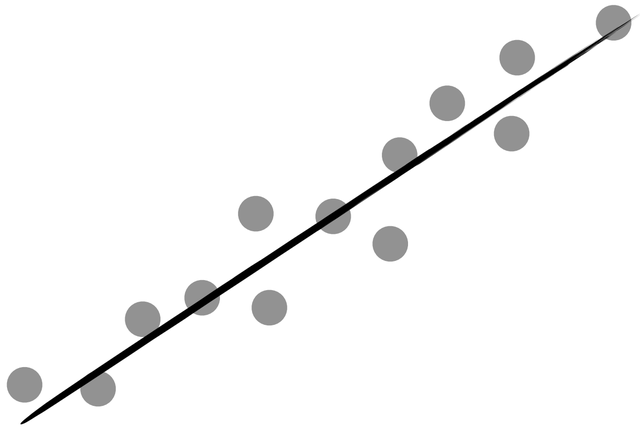
2. Encontrar a melhor função $h \in H$, ou seja, a função que comete menos erros no conjunto de dados, de acordo com uma **função de perda** L.

   Em redes neurais artificiais (e em muitos outros algoritmos), essa etapa é formalizada como um problema de otimização!

Espaço de Hipóteses

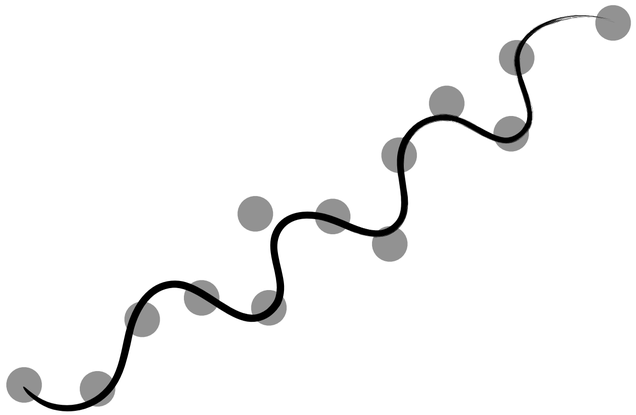
O *espaço de hipóteses* H define o conjunto de funções que um algoritmo de aprendizado supervisionado pode encontrar.

Exemplos:



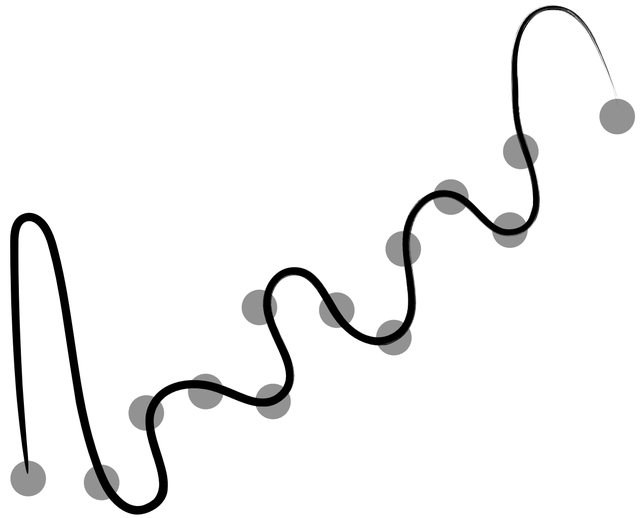
Reta

$$h(x) = w_1 x + w_0$$



Senoide

$$h(x) = w_1 x + sin(w_0 x)$$

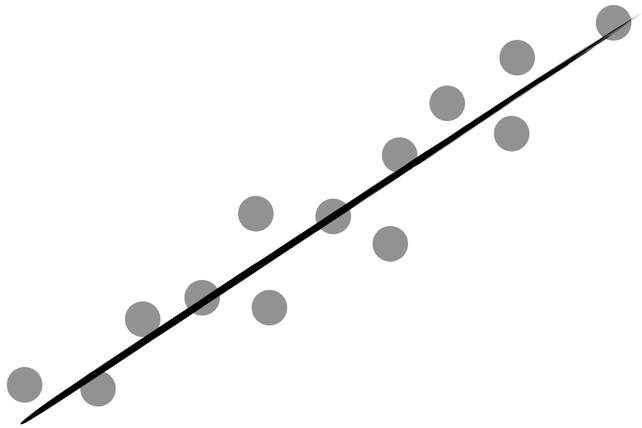


Polinômio de grau 12

$$h(x) = \sum_{i=0}^{12} w_i x^i$$

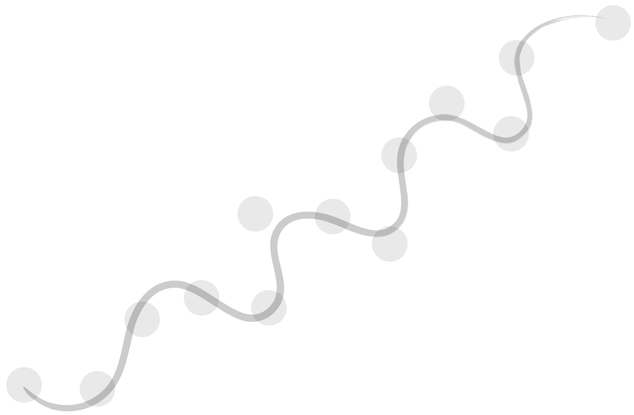
Espaço de Hipóteses

Assumindo, por exemplo, uma reta como hipótese, precisamos ajustar os parâmetros w_1 e w_0 para minimizar o erro no conjunto de dados D.



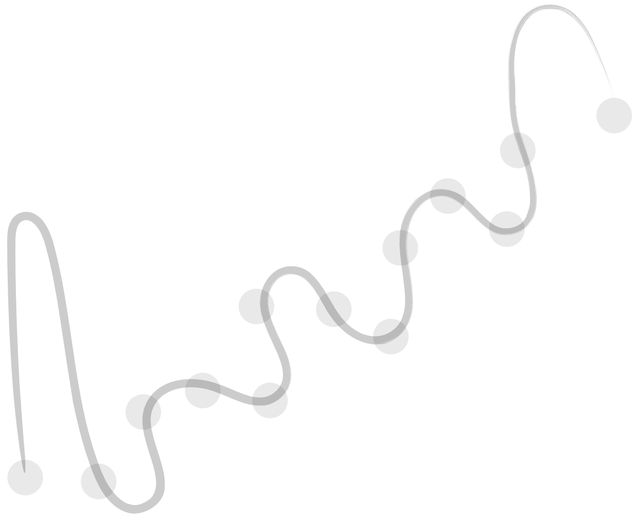
Reta

$$h(x) = w_1 x + w_0$$



Senoide

$$h(x) = w_1 x + sin(w_0 x)$$



Polinômio de grau 12

$$h(x) = \sum_{i=0}^{12} w_i x^i$$

Função de Perda (loss function)

A **função da perda** L avalia uma hipótese $h \in H$ com o conjunto de dados $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$:

- Mede o quão distantes as previsões de $h(x_i)$ estão dos rótulos y_i dos exemplos (x_i, y_i) em D;
- Os valores de perda $L(h)$ são sempre positivos;
- Quanto menor a perda $L(h)$, melhor a hipótese h;
- Uma hipótese com perda $L(h) = 0$ (zero) acerta o rótulo de todos os exemplos em D;
- Tipicamente, a função de perda L é normalizada para que o seu valor seja independente do tamanho n do conjunto de dados.

Exemplos:

- Perda Zero-um
- Perda Quadrática
- Perta Absoluta

Exemplos de Função de Perda

Perda Zero-um

O número de erros que uma hipótese h comete nos exemplos de D.

$$L(h) = \frac{1}{n} \sum_{i=1}^{n} \delta_{h(x_i) \neq y_i} \text{ onde } \delta_{h(x_i) \neq y_i} = \begin{cases} 1, & \text{se } h(x_i) \neq y_i \\ 0, & \text{caso contrário} \end{cases}$$

- Geralmente utilizada para avaliar hipóteses em problemas de classificação;
- Não é utilizada para treinar uma hipótese, pois não é diferenciável.

Exemplos de Função de Perda

Perda Quadrática

A soma do erro quadrático $(h(x_i) - y_i)^2$ da hipótese h nos exemplos de D.

$$L(h) = \frac{1}{n} \sum_{i=1}^{n} (h(x_i) - y_i)^2$$

- Geralmente utilizada para treinar uma hipótese h em problemas de regressão;
- Elevar o erro ao quadrado faz com que exemplos com erros mais altos tenham maior influência no ajuste dos pesos de h.

Exemplos de Função de Perda

Perda Absoluta

A soma do erro absoluto $|h(x_i) - y_i|$ da hipótese h nos exemplos de D.

$$L(h) = \frac{1}{n} \sum_{i=1}^{n} |h(x_i) - y_i|$$

- Geralmente utilizada para treinar uma hipótese h em problemas de regressão;
- Exemplos têm influência uniforme no ajuste dos pesos;
- Adequada para lidar com ruído nos dados (*outliers*).

Generalização

Dado um espaço de hipóteses H e uma função de perda L, queremos encontrar a hipótese $h \in H$:

$$h = argmin_{h \in H} L(h)$$

Se encontrarmos uma hipótese $h \in H$ com baixa perda em D, como saber se ela também terá baixa perda em novos exemplos $(x', y') \notin D$?

Generalização

Subajuste e Sobreajuste

Considere a seguinte função "memorizadora":

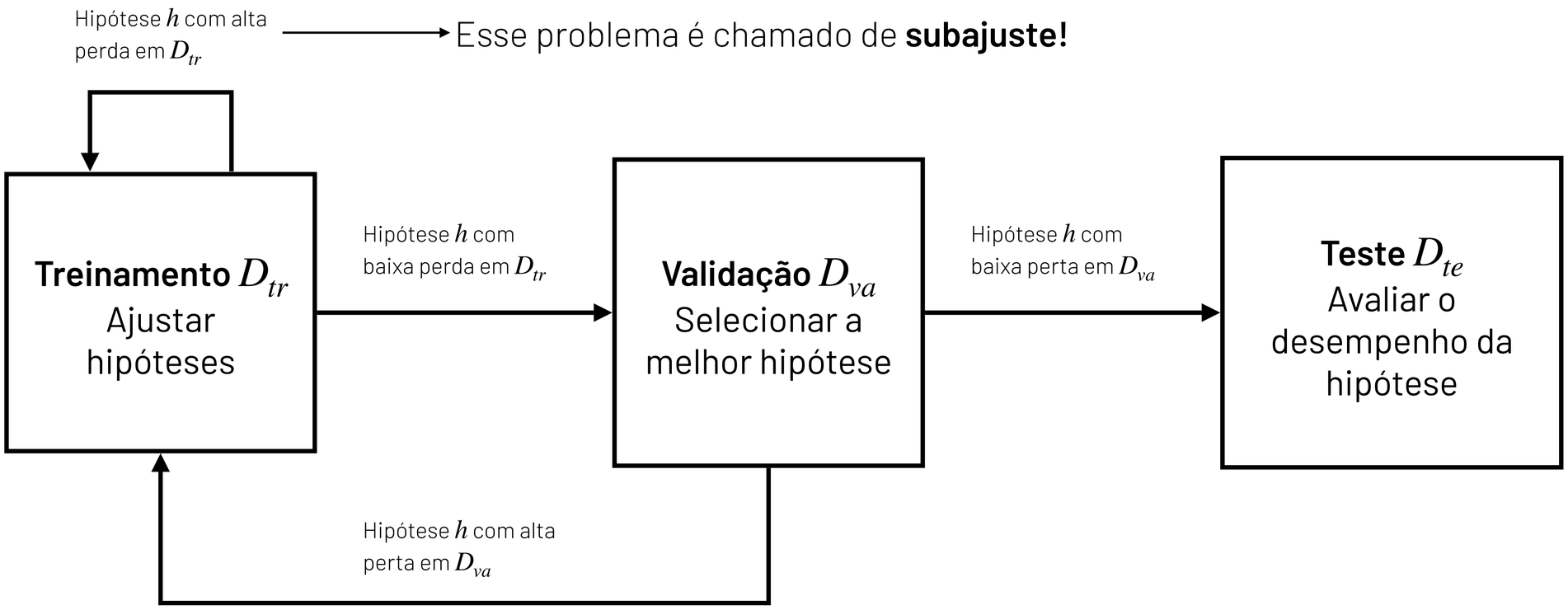
$$h(x) = \begin{cases} y_i, & \text{se } \exists (x_i, y_i) \in D, \text{ tal que, } x = x_i \\ 0, & \text{caso contrário} \end{cases}$$

- Perda 0 nos exemplos de D;
- Perda muito alta em exemplos novos!

Esse problema é chamado de **sobreajuste** (*overfit*)!

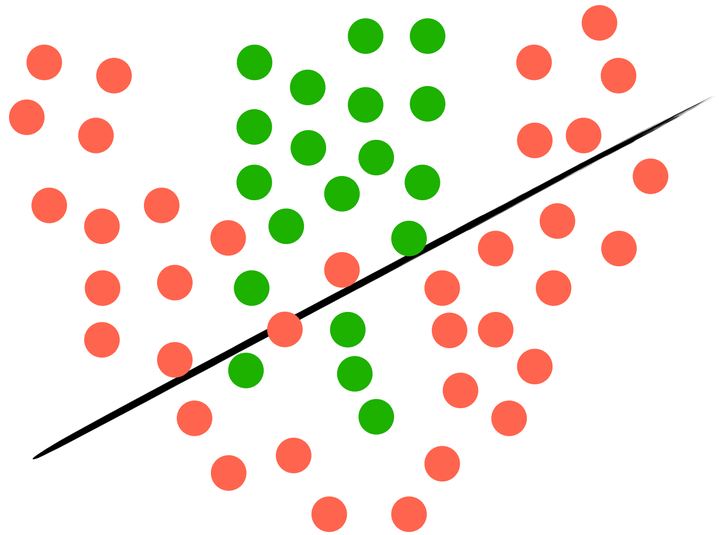
Generalização

Para resolver o problema de sobreajuste, dividimos o conjunto de dados D em três (3) subconjuntos disjuntos D_{tr}, D_{va} e D_{te}:



Subajuste (underfit)

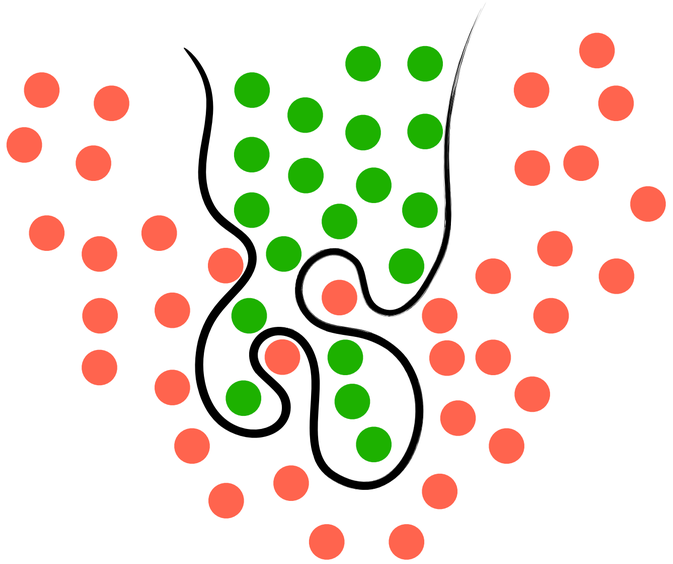
Visualização



Quando a hipótese se ajusta pouco aos dados de treinamento, aprensentando baixo desempenho de previsão tanto no conjunto de treinamento quanto no de teste.

Sobreajuste (overfit)

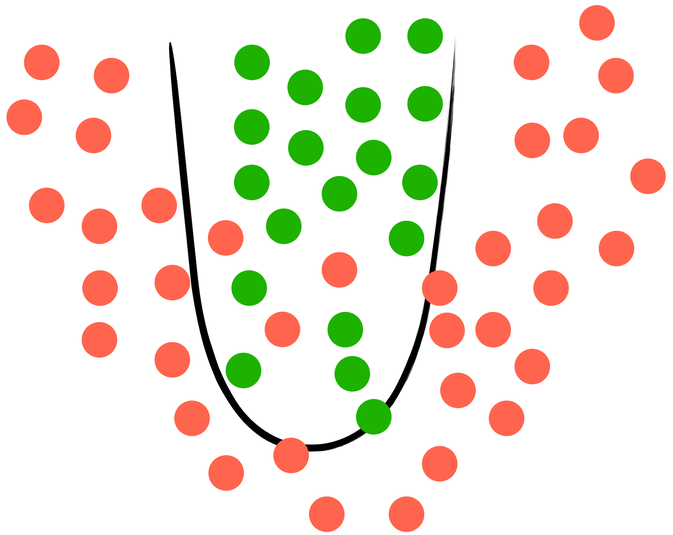
Visualização



Quando a hipótese se ajusta muito aos dados de treinamento, aprensentando alto desempenho de previsão no conjunto de treinamento, mas baixo no conjunto de teste.

Ajuste Adequado

Visualização



Quando a hipótese se ajusta bem aos dados de treinamento, aprensentando alto desempenho de previsão tanto no conjunto de treinamento quanto no de teste.

Generalização

Em aprendizado de máquina, assumimos três premissas sobre o conjunto de dados D:

1. Os exemplos são amostrados de forma **independente e identicamente distribuída (i.i.d)** de $P(X, Y)$;
2. A distribuição $P(X, Y)$ é **estacionária**: não muda ao longo do tempo;
3. Sempre amostramos da **mesma distribuição** $P(X, Y)$, tanto no conjunto de treinamento, quandos nos de validação e teste.

Algoritmos de Aprendizado Supervisionado

Cada algoritmo de aprendizado supervisionado assume uma *hipótese* diferente sobre os dados para definir um espaço de funções H.

- Regressão Linear
- Regressão Logística
- Árvores de Decisão
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Suport Vector Machines (SVMs)
- Redes Neurais

Próxima aula

A3: Regressão Logística

Regressão Logística como uma rede neural para problemas linearmente separáveis.