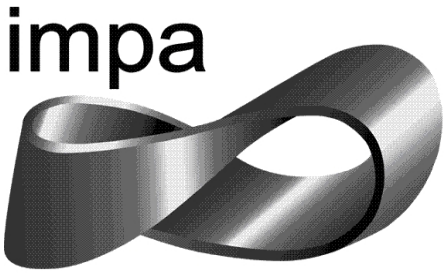
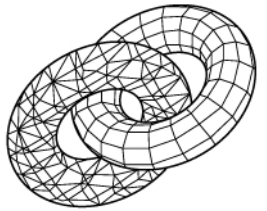
Ciência de Dados Aplicada

Aula 9: Overfitting





VisgrafLab

Revisão de termos importantes

Representação de características

Amostras (X)

Valor-alvo (y)

Conjunto de treinamento e testes

Treinamento é o processo
de estimar os parâmetros
do modelo

Avaliação

fruits

	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	0.60
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	mandarin	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn	172	7.4	7.0	0.89
10	1	apple	braeburn	166	6.9	7.3	0.93
11	1	apple	braeburn	172	7.1	7.6	0.92
12	1	apple	braeburn	154	7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69
15	1	apple	golden_delicious	156	7.7	7.1	0.69
16	1	apple	golden_delicious	156	7.6	7.5	0.67
17	1	apple	golden_delicious	168	7.5	7.6	0.73
18	1	apple	cripps_pink	162	7.5	7.1	0.83

Classificação

- Valor-alvo é discreto

- Tipos de problemas

    Binária

    Multi-Classe

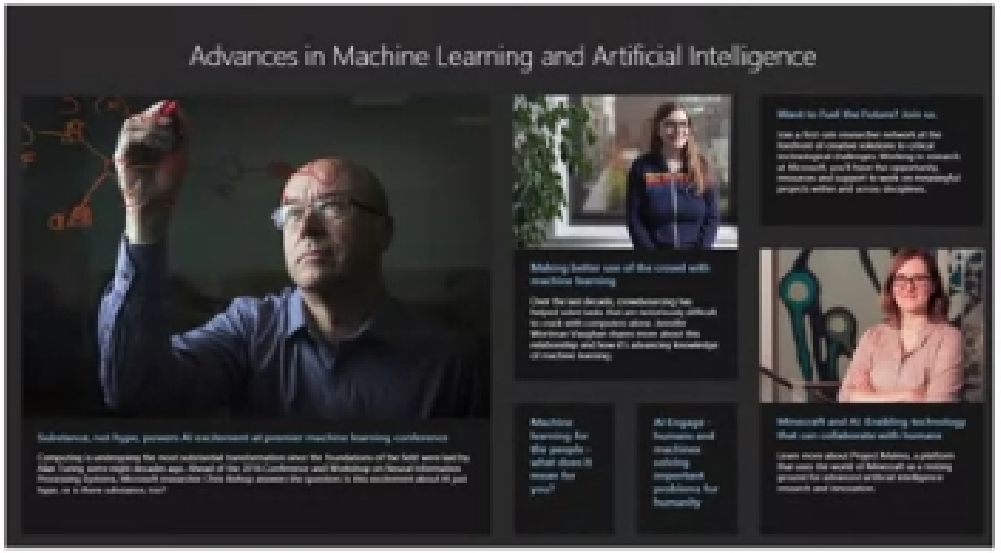
    Multi-Label

Regressão

- Valor-alvo é contínuo

- Geralmente os Algoritmos de classificação possuem variações para cada tipo de problema

Classificação Multi-Label



Classes

0.83 computers

0.20 science

0.08 society/issues

0.07 reference

0.03 reference/education

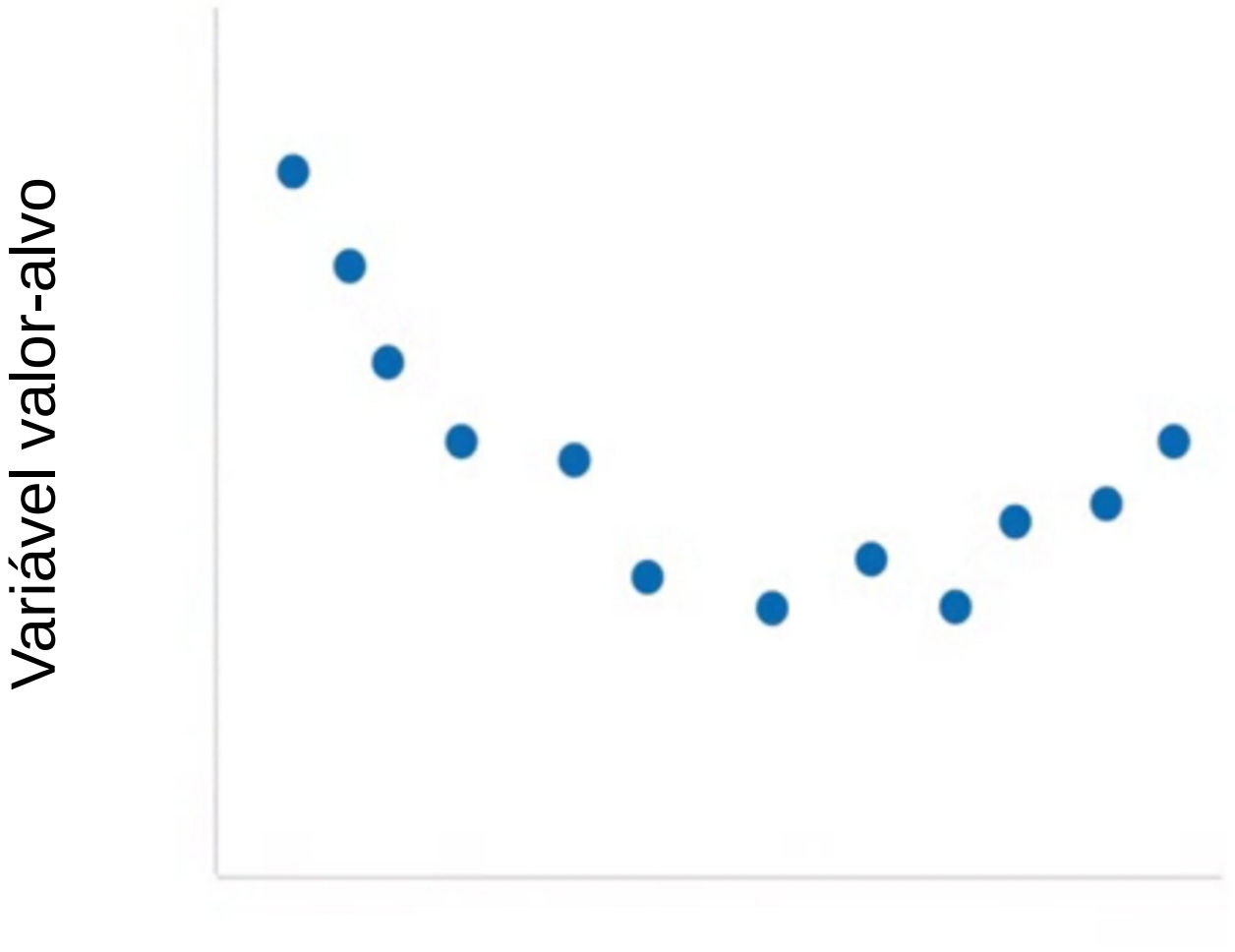
Generalização, Overfitting e Underfitting

Quantidade de dados utilizados para treinar o algoritmo

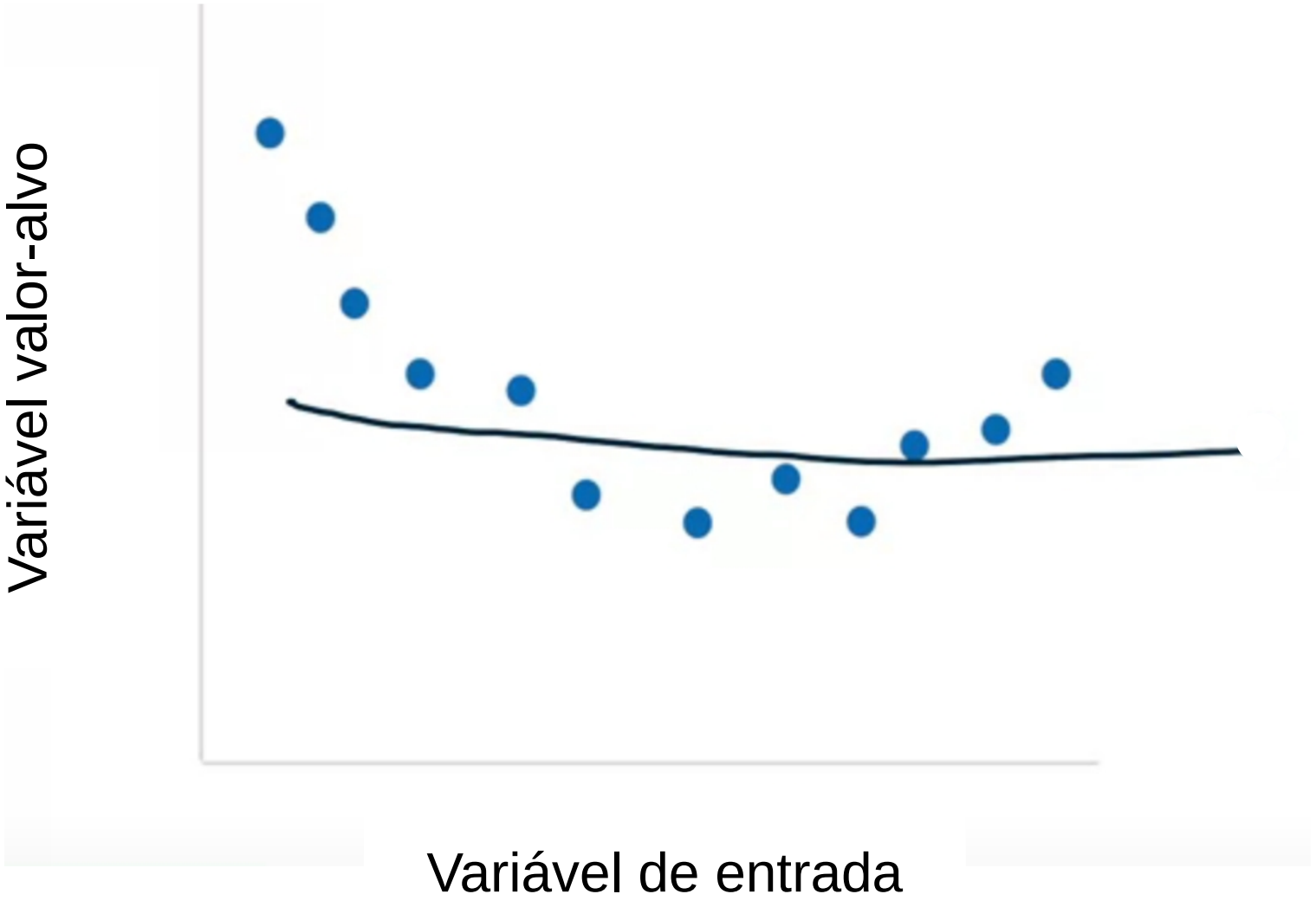
É o problema principal quando estamos utilizando algoritmos de aprendizado de máquina

Tendência global dos dados de treinamento

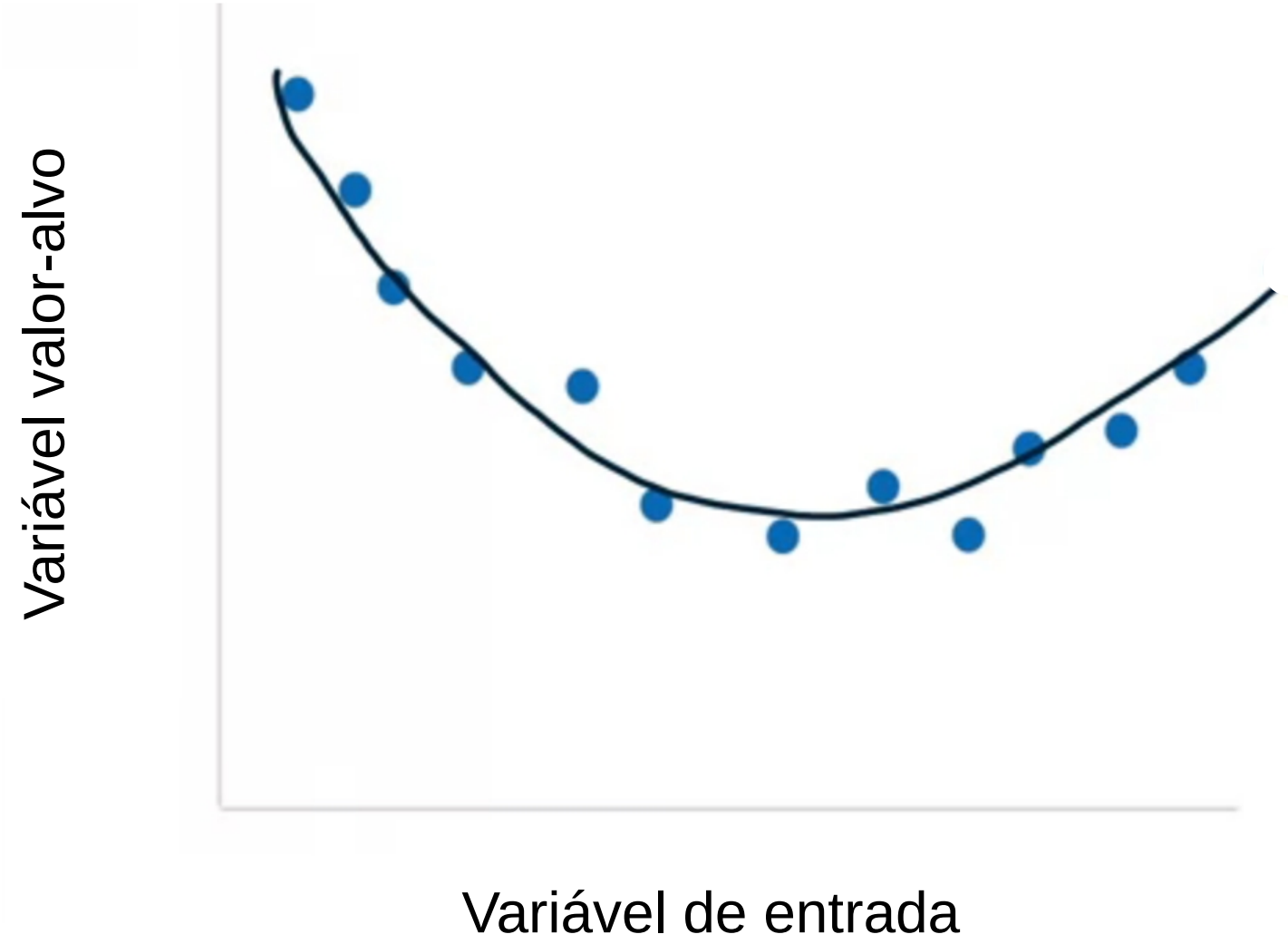
Overfitting na Regressão



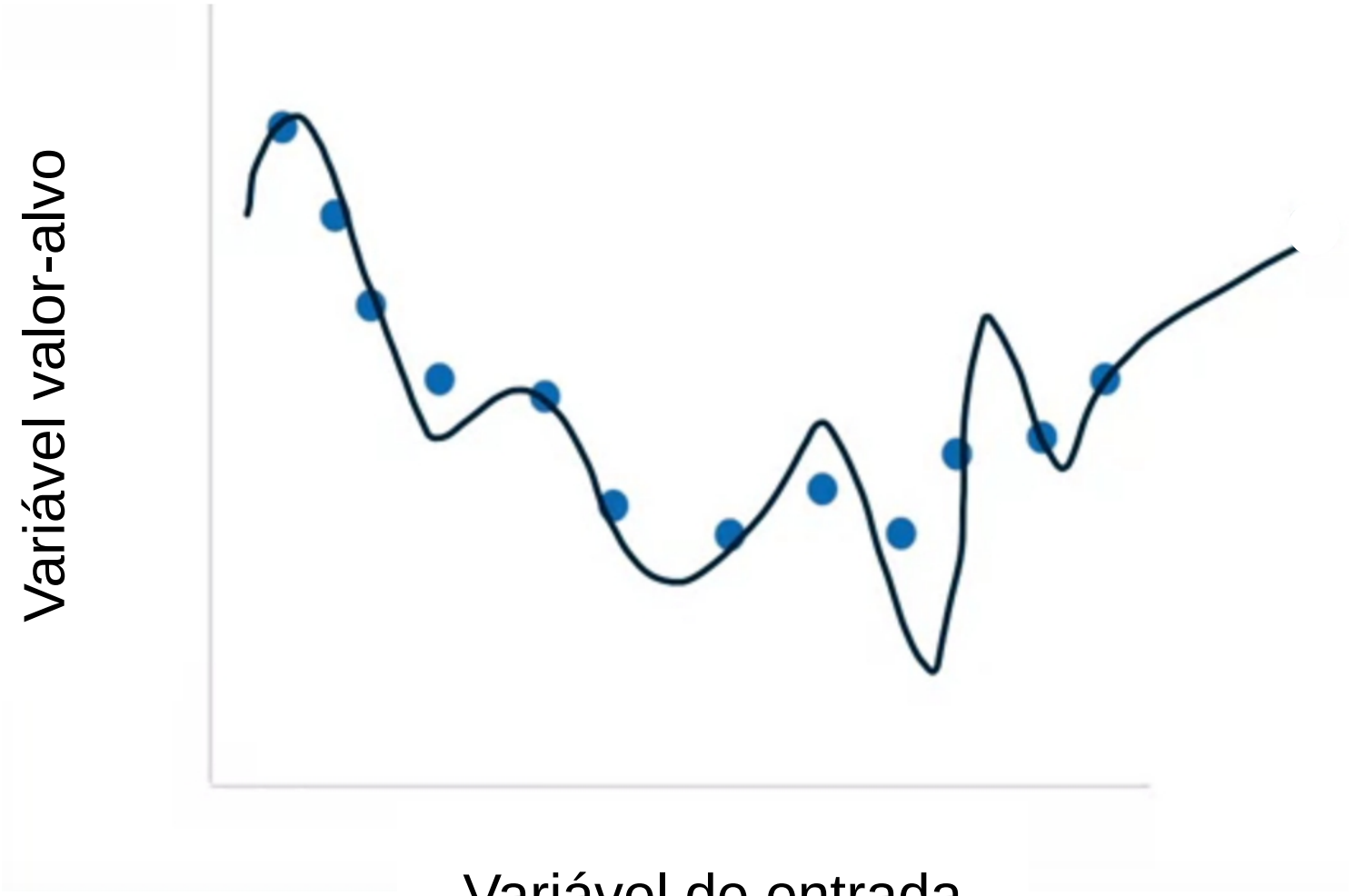
Underfitting



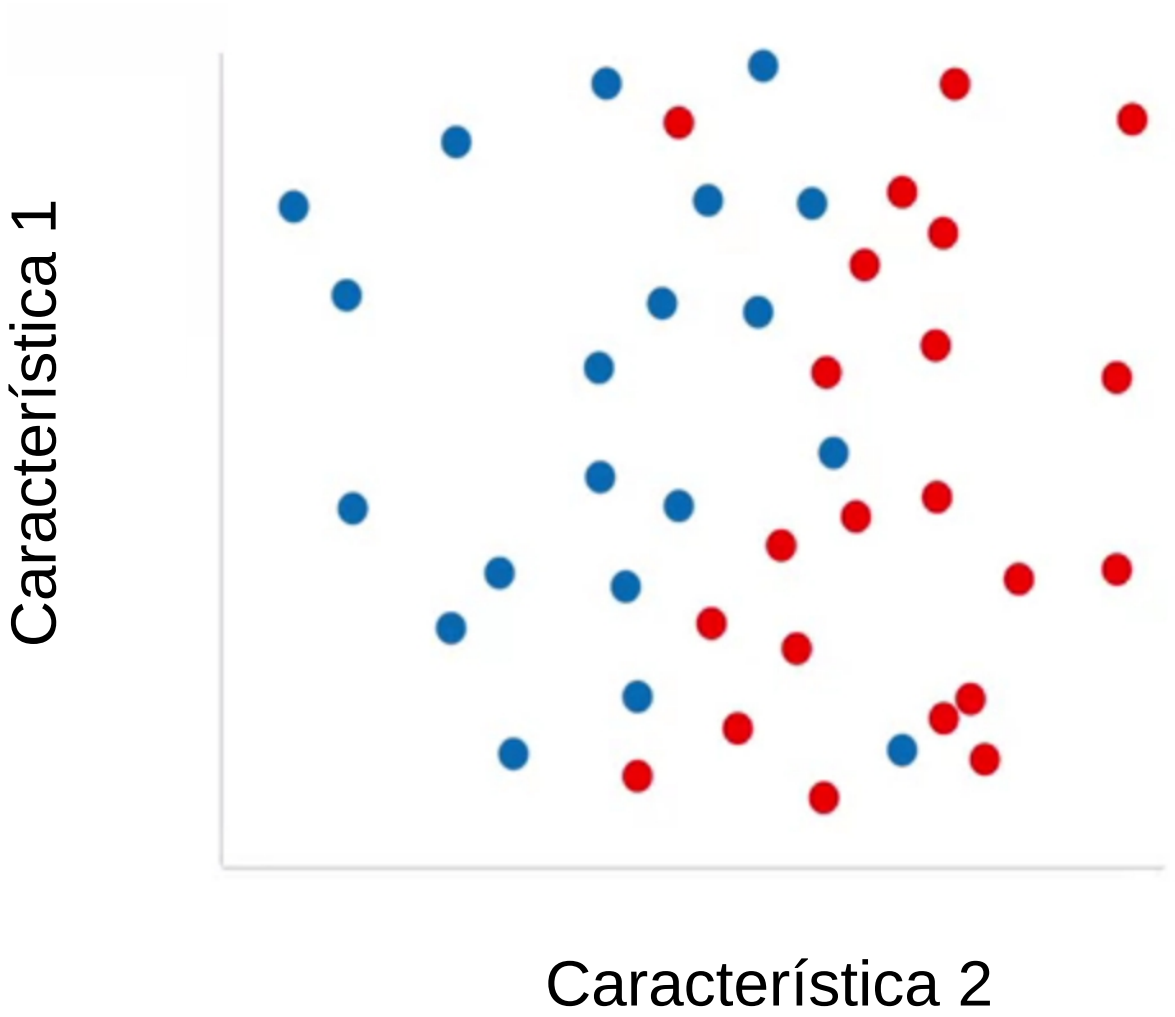
Modelo Ideal



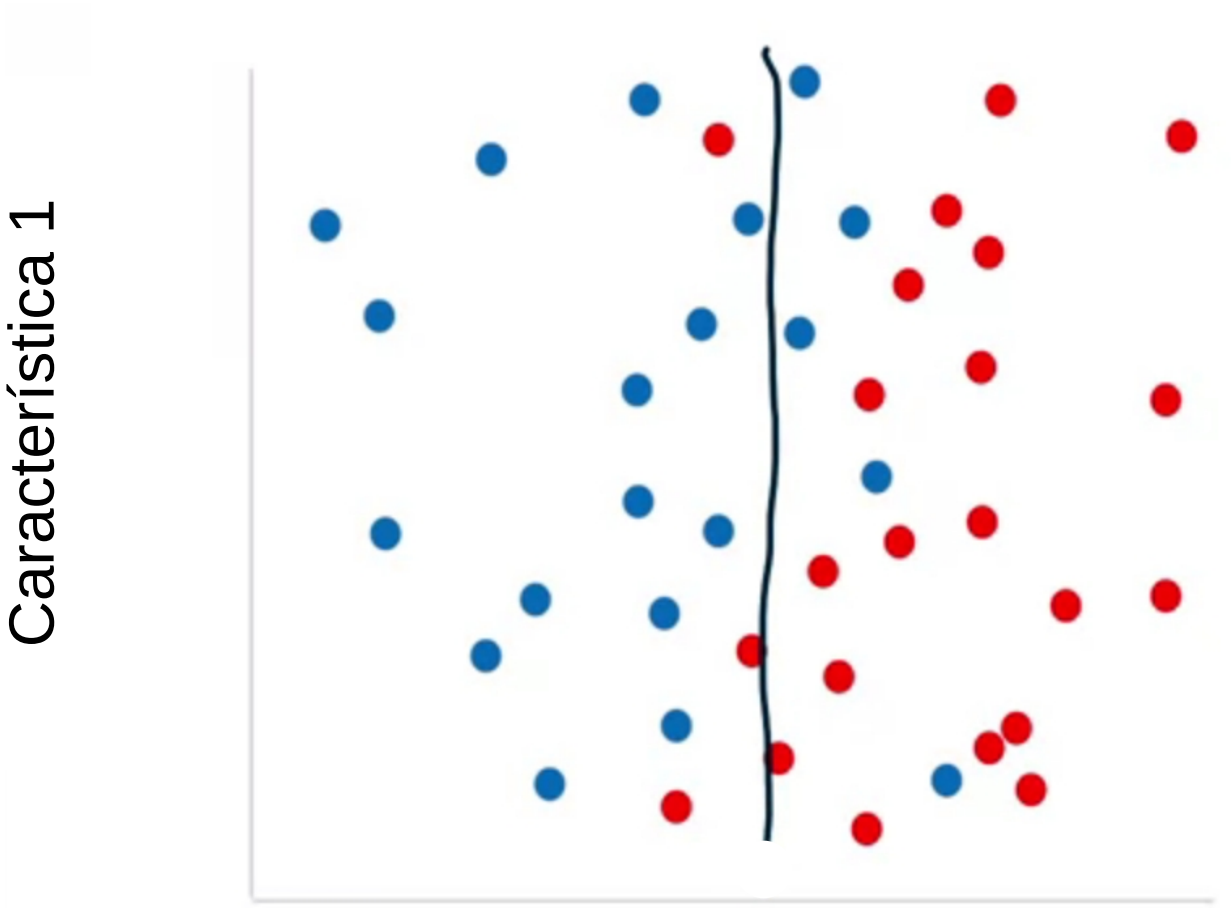
Overfitting



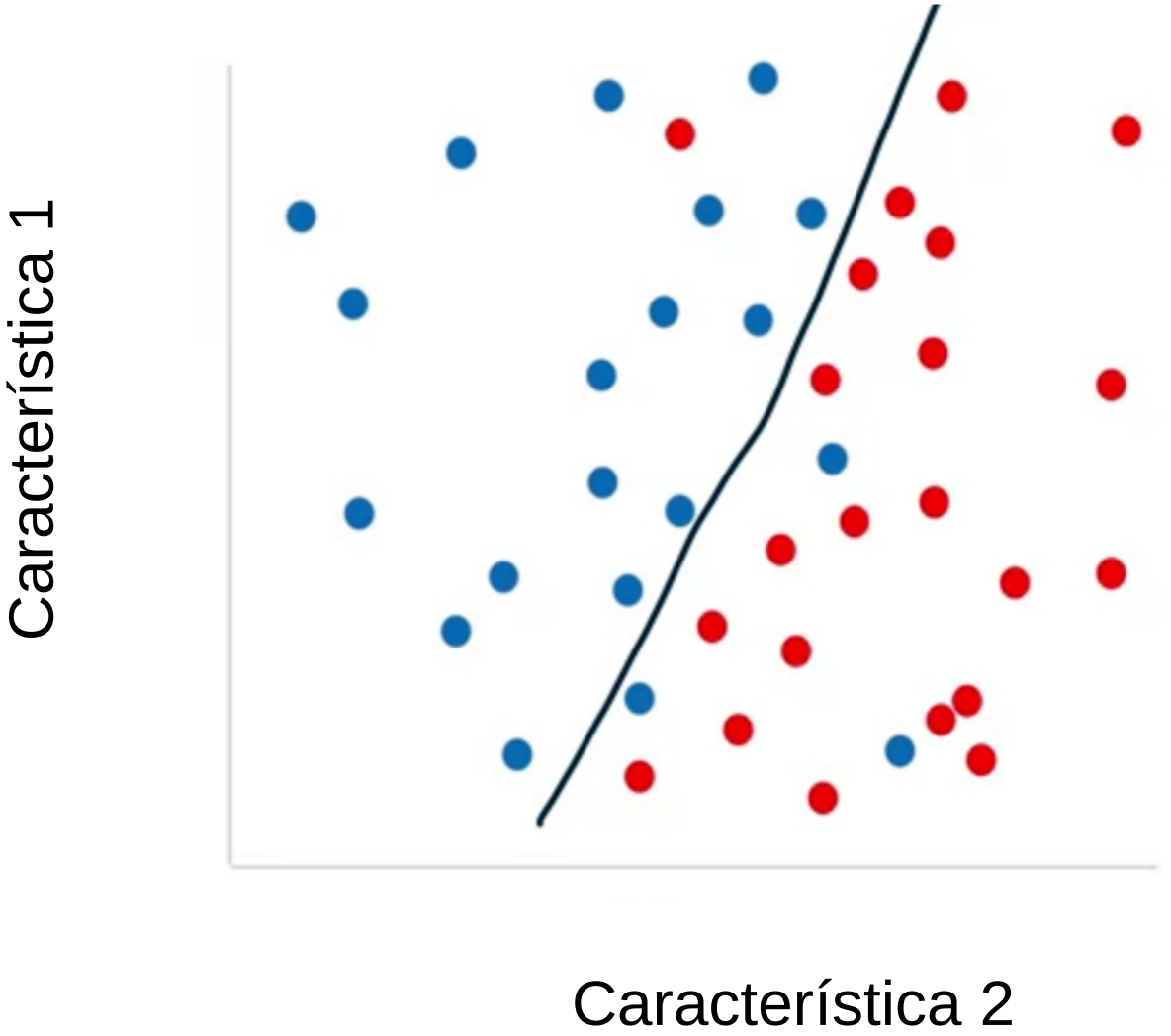
Overfitting na Classificação



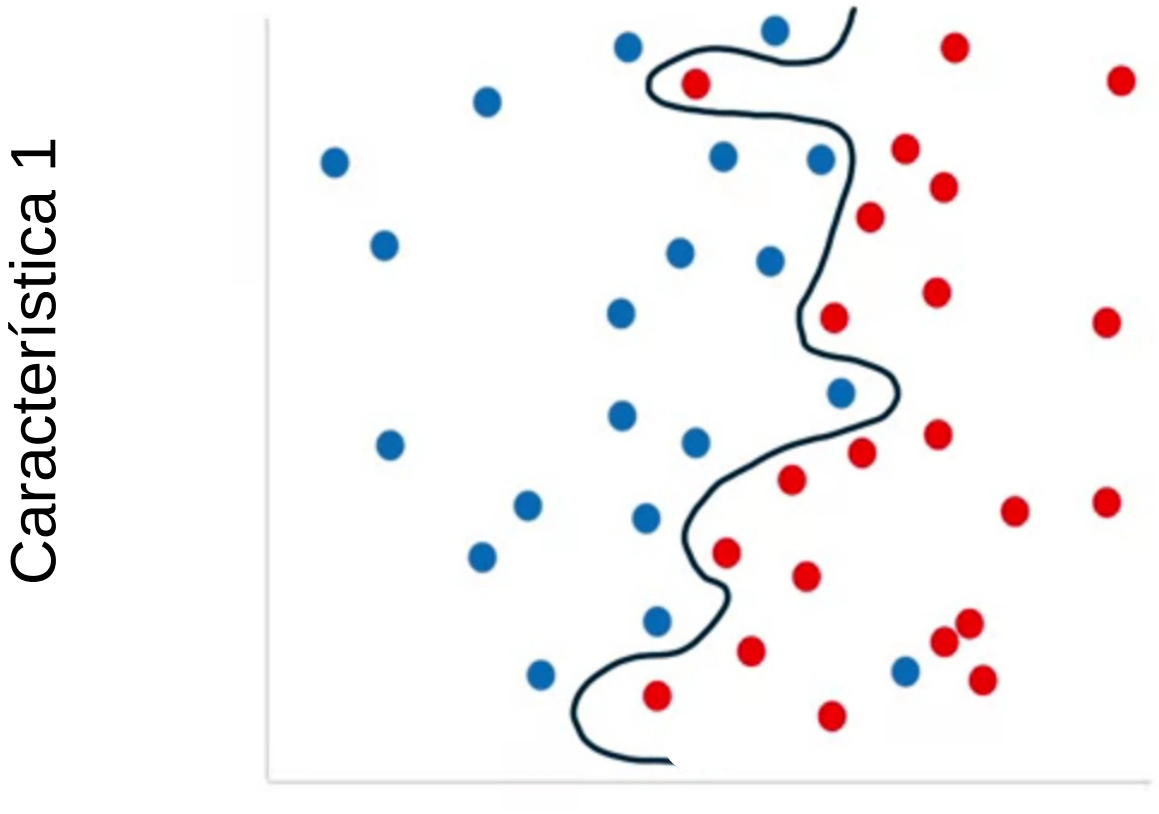
Underfitting



Modelo Ideal



Overfitting



Overfitting