escola britânica de artes criativas & tecnologia

Profissão: Cientista de Dados

Árvores de Classificação II

Resposta multinomial

Agenda

Árvores de classificação multinomial

Validação cruzada

“There is no greater insult than ‘You’ve created an elegant solution to an irrelevant problem’.”

D. J. Patil (Data Driven)

“Não há insulto maior que ‘você criou uma solução elegante para um problema irrelevante’.”

Introdução:

Resposta multinomial

Definições de impureza

- Gini
- Entropia de Shannon

Índice de Gini

$$I_g(p) = 1 - \sum_{i=1}^{J} p_i^2$$

- Impureza máxima com distribuição uniforme
- Impureza mínima na concentração total

Entropia

$$H = -\sum_{i=1}^{J} p_i \log_2(p_i)$$

Ganho de informação:

$$GI(T, a) = H(T) - H(T|a)$$

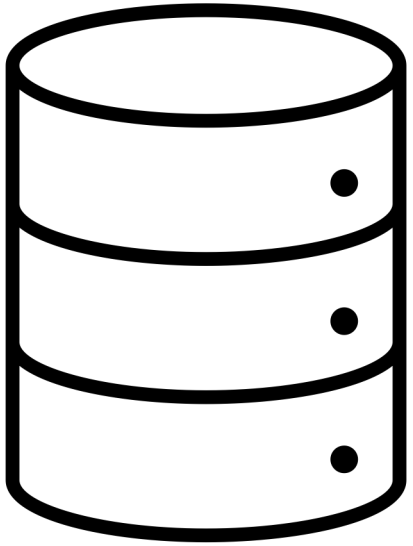
- Impureza máxima com distribuição uniforme
- Impureza mínima na concentração total

Cross validation

É o ato de tentar avaliar como o resultado de um modelo pode ser generalizado para uma população mais ampla.

Treino, Validação e Teste

BASE DE DADOS



Treino — Partição do conjunto de dados utilizada para o desenvolvimento do algoritmo

Validação — Partição utilizada para ajustes no algoritmo (*e.g.*: *post-prunning*)

Teste — Partição utilizada para avaliar o desempenho do algoritmo "na prática"

escola britânica de artes criativas & tecnologia

Profissão: Cientista de Dados

Árvores de Classificação II

Resposta multinomial

Cross Validation

Machine Learning
na veia

Tipos

Exaustivos

- *Leave one out*
- *Leave k out*

Não exaustivos

- *Holdout*
- K-fold
- Sub-amostragem sequencial

Hierárquicos (*nested*)

- K-fold com *holdout*
- K-l-fold

Tipos

Exaustivos

- *Leave one out*
- *Leave k out*

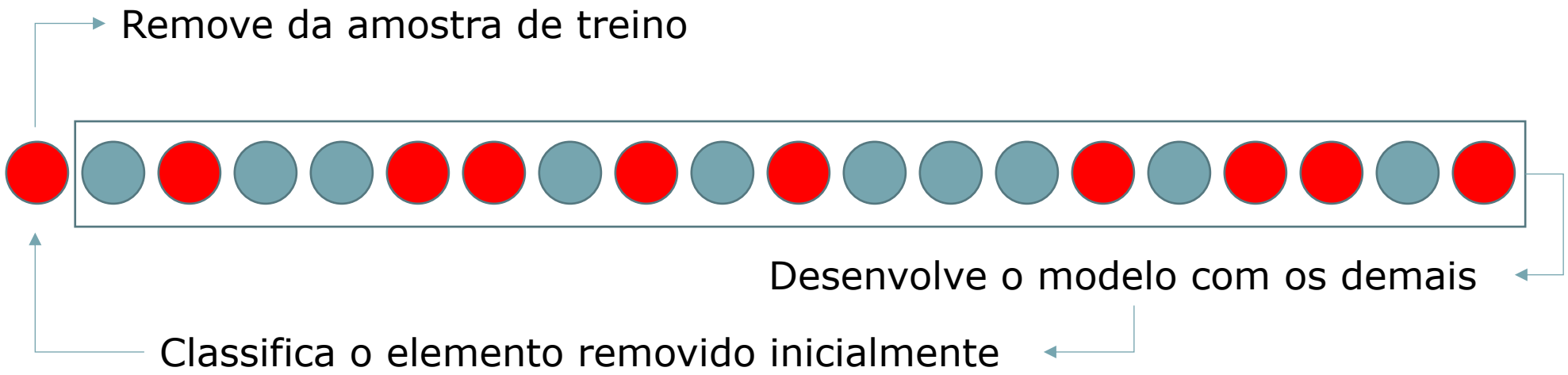
Não exaustivos

- *Holdout*
- K-fold
- Sub-amostragem sequencial

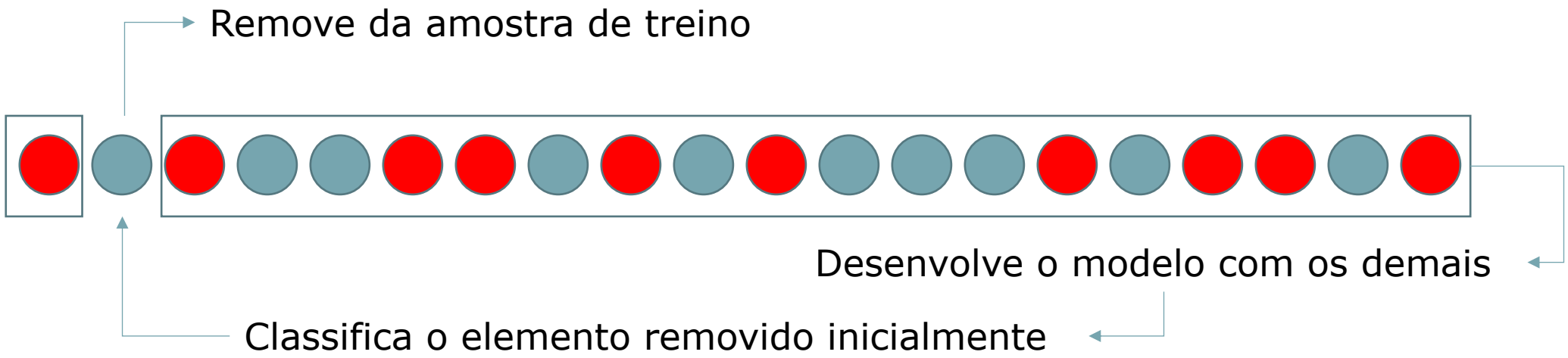
Hierárquicos (*nested*)

- K-fold com *holdout*
- K-l-fold

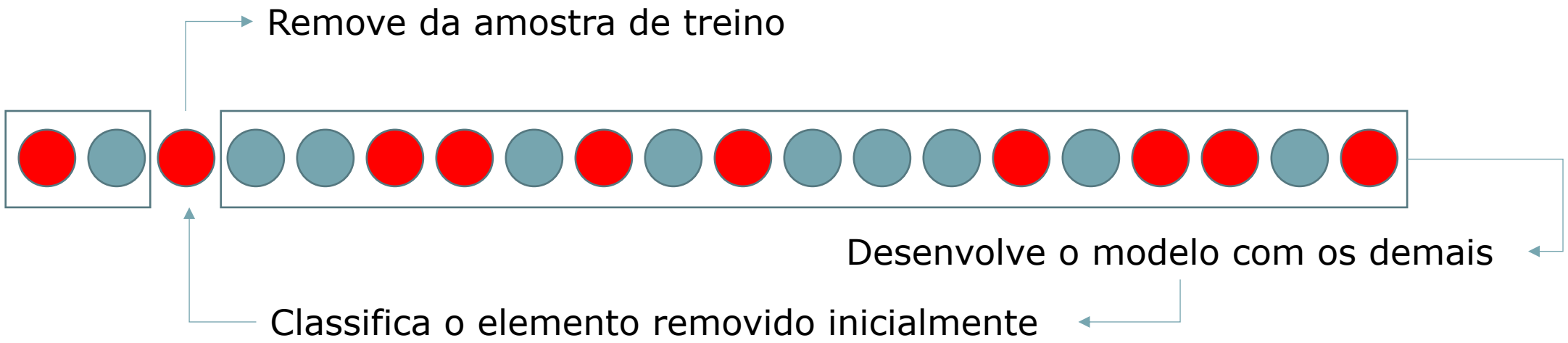
Leave-one-out



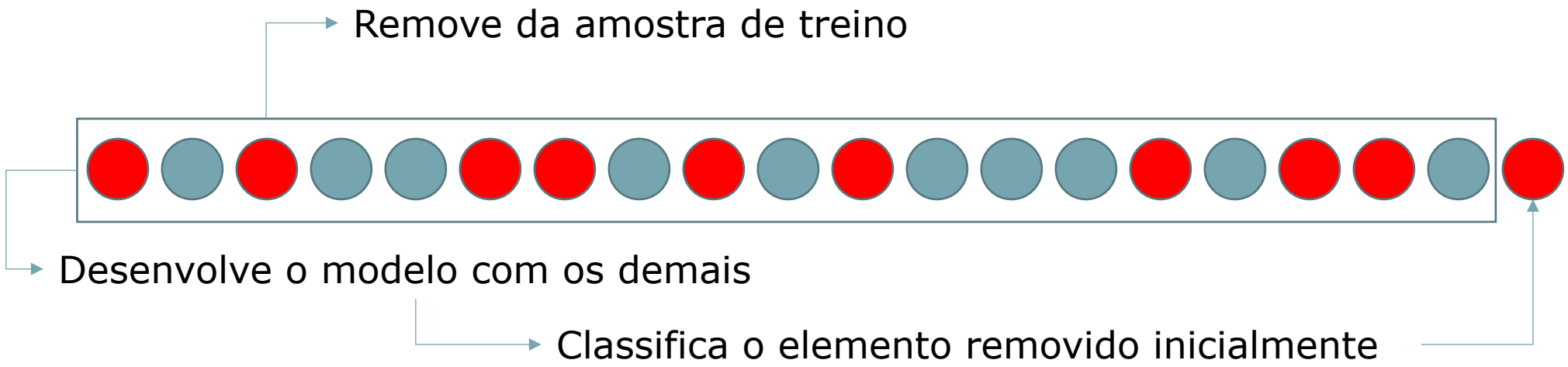
Leave-one-out



Leave-one-out



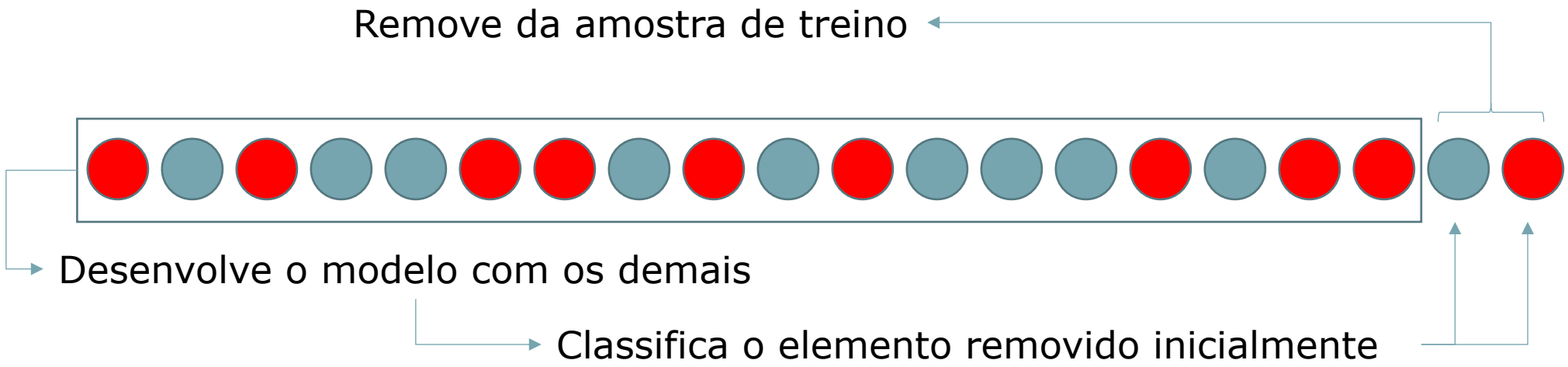
Leave-one-out



- Para cada elemento *i*:
  - Retira o elemento da amostra
  - Desenvolve o modelo M_i com o restante
  - Classifica o elemento *i* com o modelo M_i

- Avalia o modelo com 'predições independentes'

Leave-k-out



Semelhante ao *leave-one-out*, mas retira-se k observações, e utiliza-se todas as combinações N, k a k.

Fica rapidamente intratável.

Tipos

Exaustivos

- *Leave one out*
- *Leave k out*

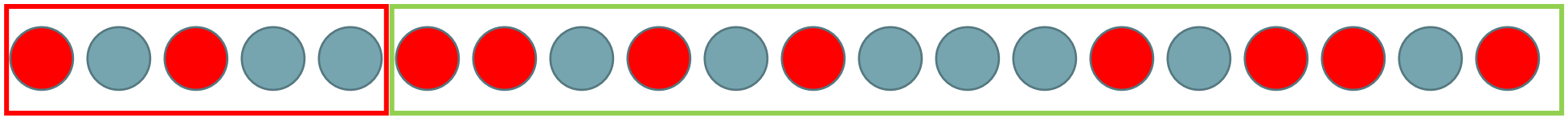
Não exaustivos

- *Holdout*
- K-fold
- Sub-amostragem sequencial

Hierárquicos (*nested*)

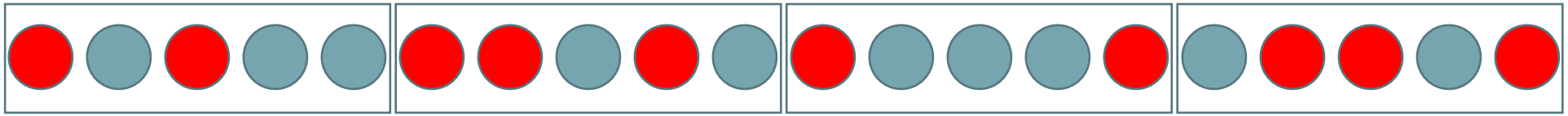
- K-fold com *holdout*
- K-l-fold

Holdout



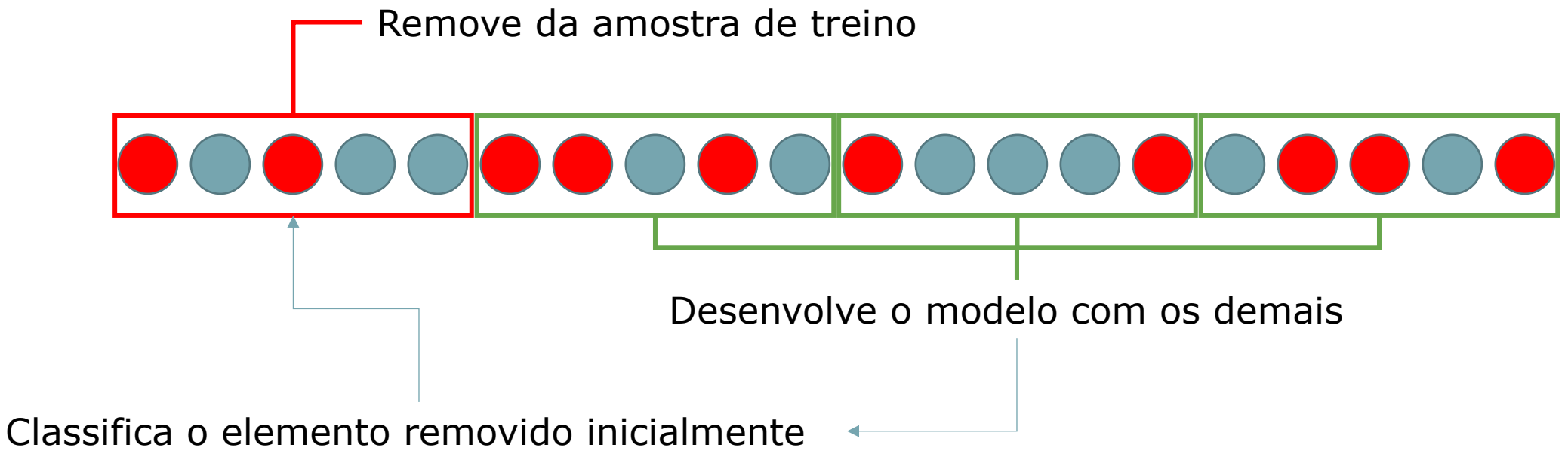
- Dividimos a base em Treinamento e Teste
- Construimos o modelo na base de treinamento
- Avaliamos o modelo na base de Teste

K-fold

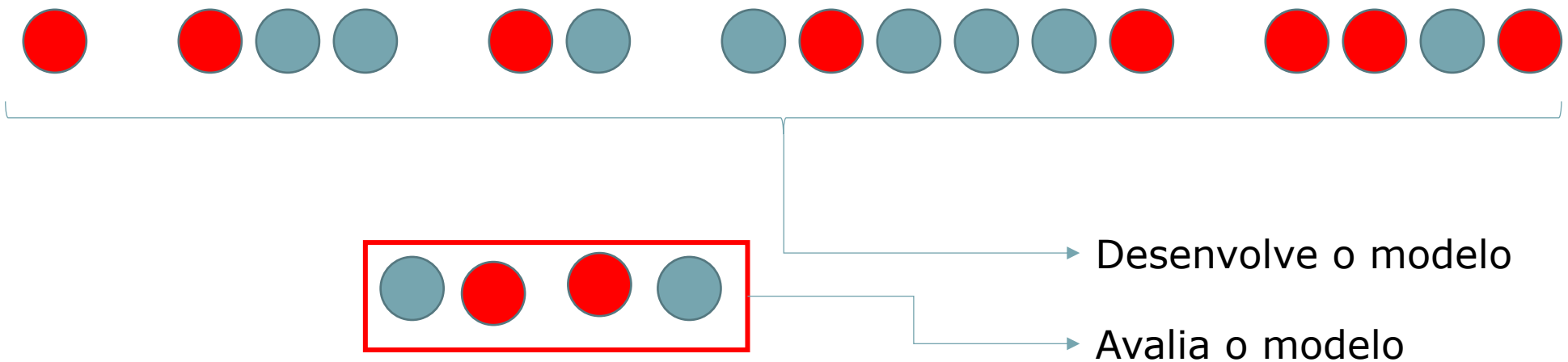


Dividimos a base em *k* subconjuntos chamados *folds*.

K-fold

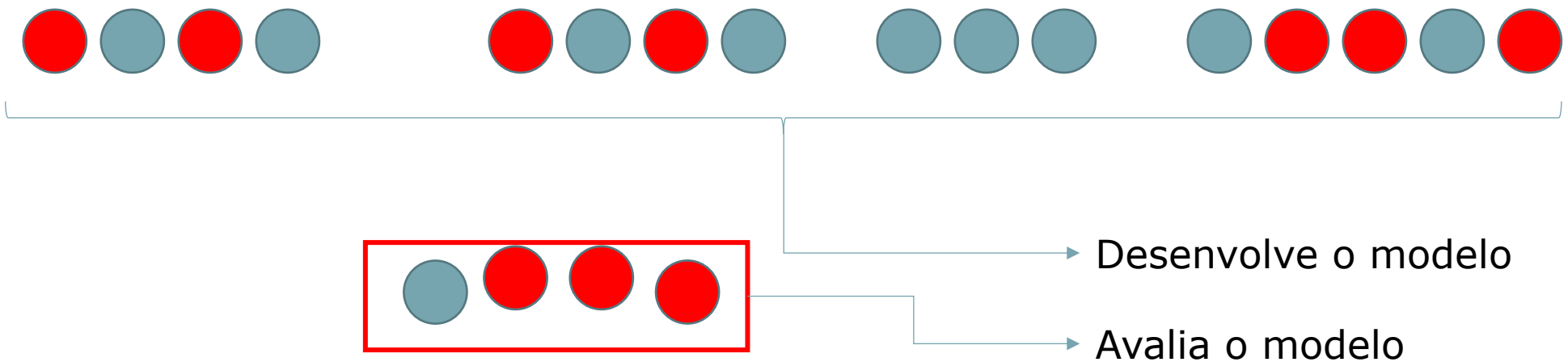

Remove da amostra de treino
Desenvolve o modelo com os demais
Classifica o elemento removido inicialmente

Reamostragem sequencial



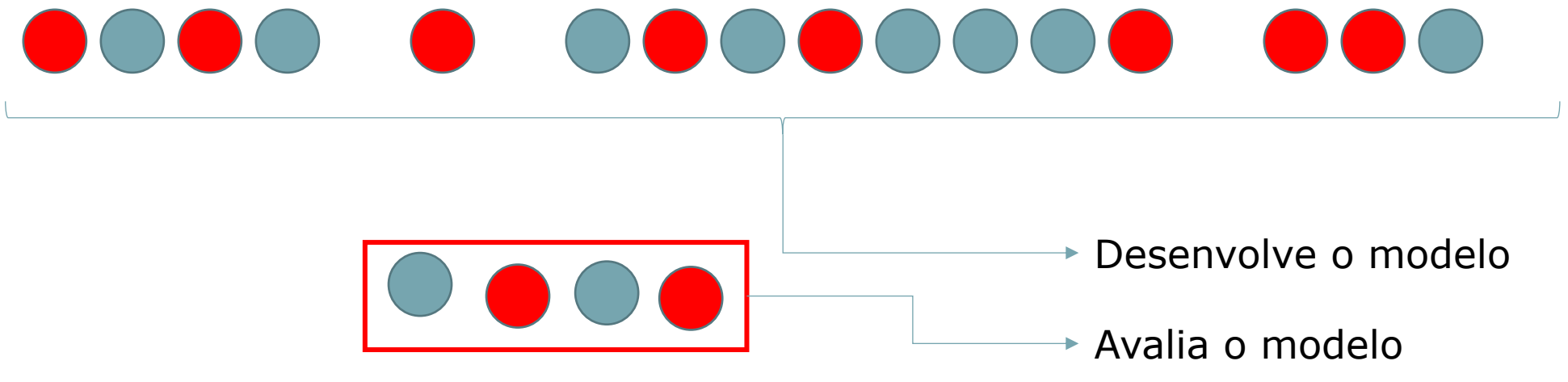
- Avaliação 1

Reamostragem sequencial



- Avaliação 1
- Avaliação 2

Reamostragem sequencial



- Avaliação 1
- Avaliação 2

...

- Avaliação M

Tipos

Exaustivos

- *Leave one out*
- *Leave k out*

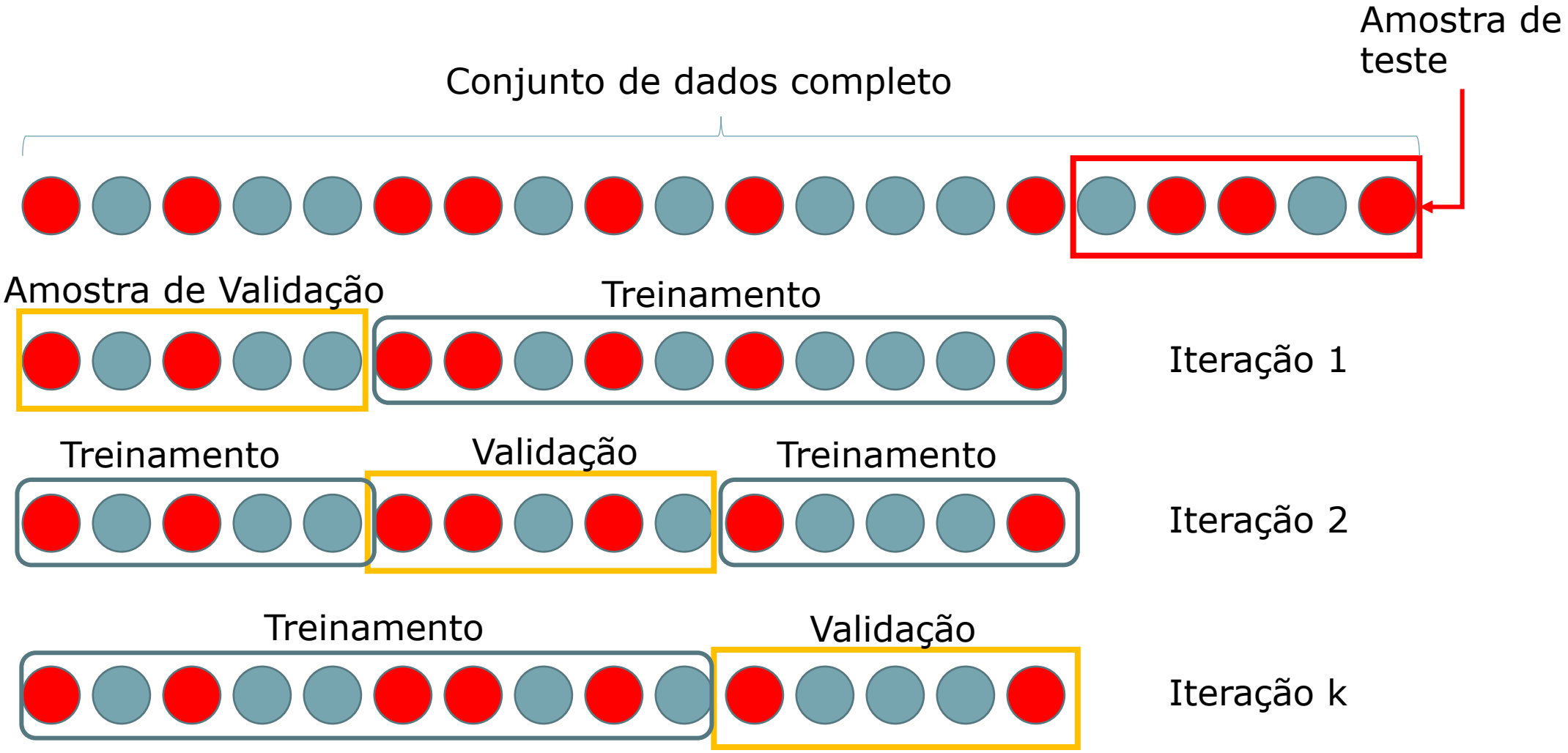
Não exaustivos

- *Holdout*
- K-fold
- Sub-amostragem sequencial

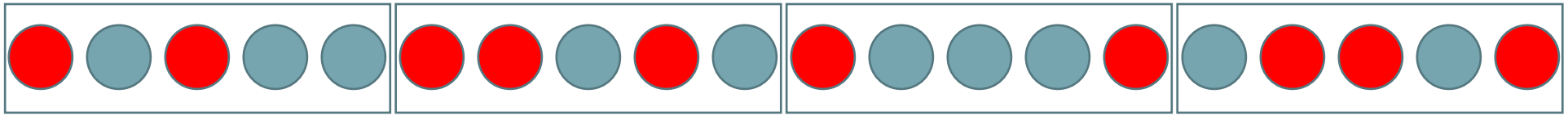
Hierárquicos (*nested*)

- K-fold com *holdout*
- K-l-fold

K-fold com *holdout*



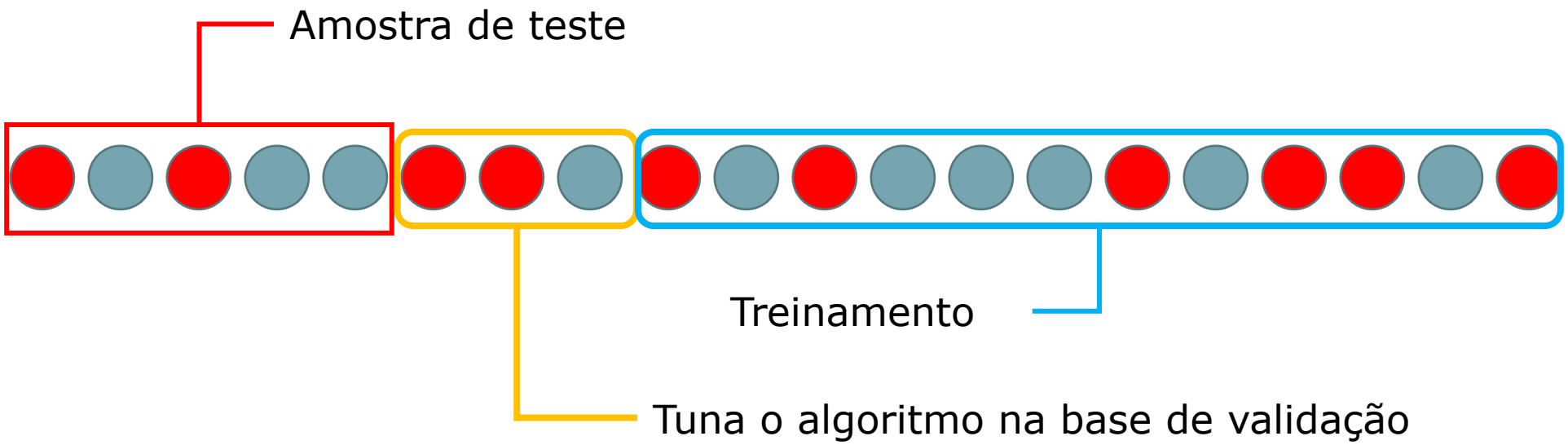
K-l-fold



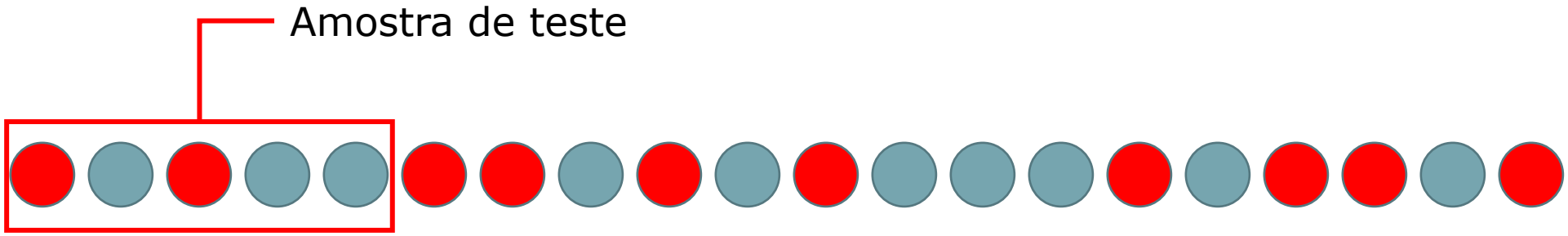
Separamos a base em k subgrupos, para cada um:

- O subgrupo é separado da base como amostra de teste
- Separamos os k-1 subgrupos em novos l subgrupos
- Fazemos um l-fold para ‘tunar’ o algoritmo
- Avaliamos o algoritmo no

K-l-fold


Amostra de teste
Treinamento
Tuna o algoritmo na base de validação

K-l-fold



ALGORITMO

- Separamos a base em k subgrupos, para cada um:
  - Reservar o subgrupo como base de testes
  - Para os demais, dividir novamente em l grupos, para cada um
    - Reservar para base de validação
    - Para os demais, desenvolver o algoritmo e ”tunar” com a base de validação
  - Ajustar novamente o melhor modelo com todos os k-1 grupos
  - Avaliar na base de testes