Aprendizado de Máquina

Aula 3: Partição de dados

André C. P. L. F de Carvalho
ICMC/USP
andre@icmc.usp.br

Tópicos

- Desempenho de algoritmos/modelos
- Desempenho preditivo
- Partição dos dados
- Amostragem
- Reamostragem

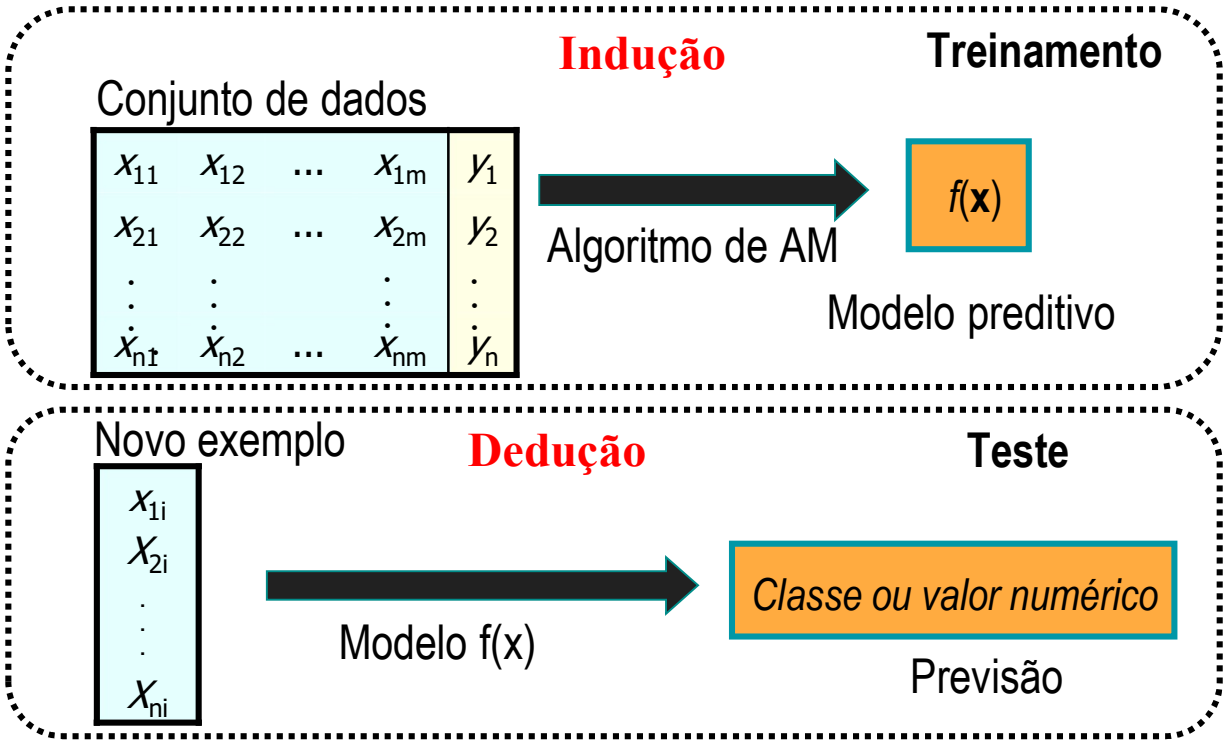
Introdução

- Procedimentos experimentais e avaliação de desempenho
  - Diferente para tarefas descritivas e preditivas
  - Este módulo tratará de tarefas preditivas
    - Classificação
    - Fácil adaptar para regressão

Avaliação de algoritmos/modelos

- Erro
  - Desempenho preditivo
- Custo
  - Tempo de processamento
  - Memória necessária
- Interpretabilidade
- Medidos para algoritmo e/ou modelo

Tarefa preditiva



Desempenho preditivo

- Depende da tarefa a ser resolvida:
  - Classificação: considera taxa de exemplos incorretamente classificados
    - Ex.: Acurácia
  - Regressão: considera diferença entre valor previsto e valor correto
    - Ex.: R^2
- Média dos erros obtidos em diferentes execuções de um experimento

Desempenho preditivo

- Comparação de algoritmos
  - Algoritmo que gera melhor(es) modelos
  - Deve ser justo para os algoritmos investigados
    - Pode variar valores de hiperparâmetros
    - Mesmos subconjuntos de dados para todos
    - Mesmo número de modelos avaliados por todos
      - Mesmos recursos para todos

Desempenho preditivo

- Comparação de modelos
  - Melhor modelo gerado pelo mesmo algoritmo
  - Pode variar subconjuntos de dados usados
  - Deve variar valores dos hiperparâmetros
  - Pode variar conformação dos dados
    - Amostragem
    - Partição
    - Seleção de atributos

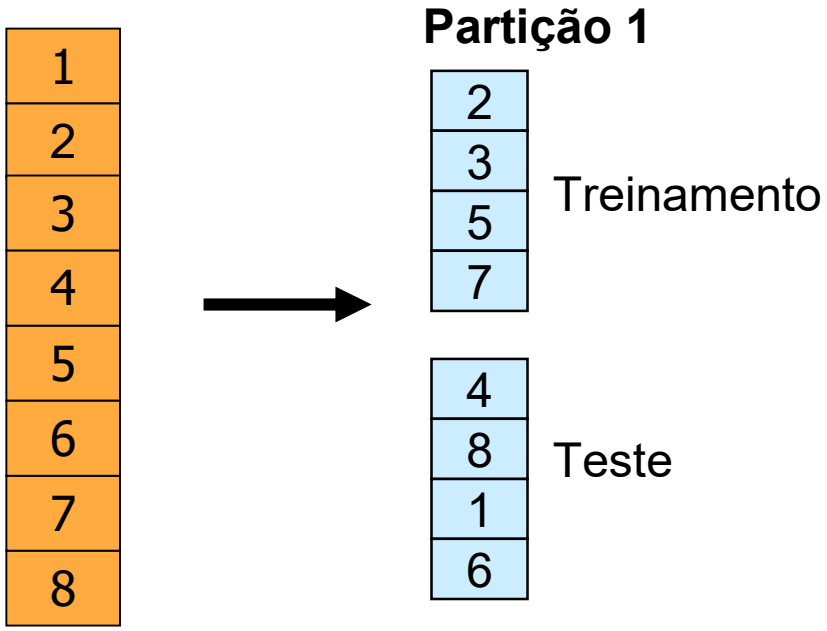
Desempenho preditivo

- Principal objetivo em tarefas de classificação:
  - Classificação correta de novos exemplos
    - Errar o mínimo possível
      - Minimizar taxa de erro para novos exemplos
- Geralmente não é possível medir com exatidão essa taxa de erro para novos exemplos
  - Deve ser estimada com duas amostras do conjunto de dados original
    - Utilizar uma amostra A (treinamento) para Induzir um modelo
    - Utilizar uma amostra B (teste, que simula situação em que existem novos exemplos)

Partição de dados

- Permite melhor estimativa do desempenho de um modelo ou algoritmo
    - Treinamento (validação) e teste
- Procedimentos
    - Amostragem única
        - *Hold-out*
    - Re-amostragem

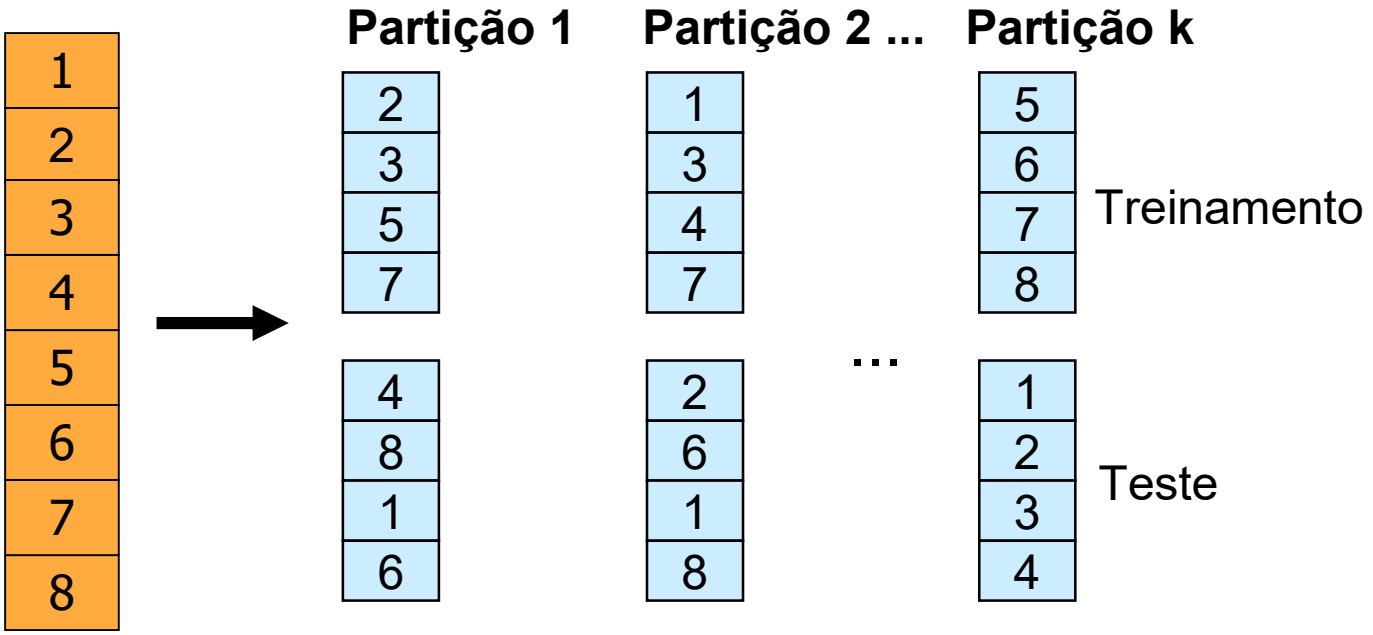
Hold out



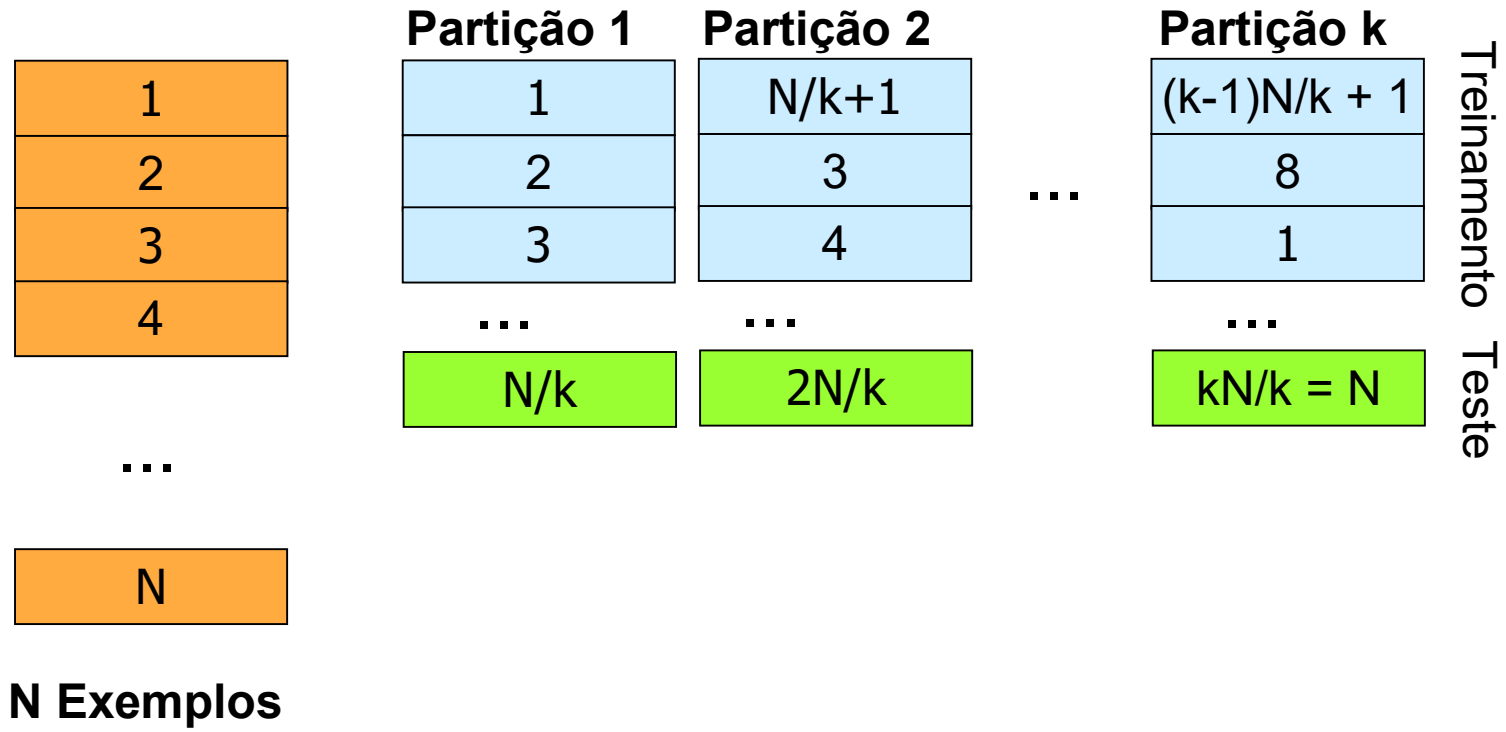
Métodos de reamostragem

- Amostragem única é pouco confiável
- Gerar várias partições para conjuntos de treinamento (validação) e teste
- *Reamostragem*
    - *Random subsampling*
    - *K-fold Cross-validation*
        - *Leave-one-out*
    - *Bootstrap (ou Bootstrapping)*

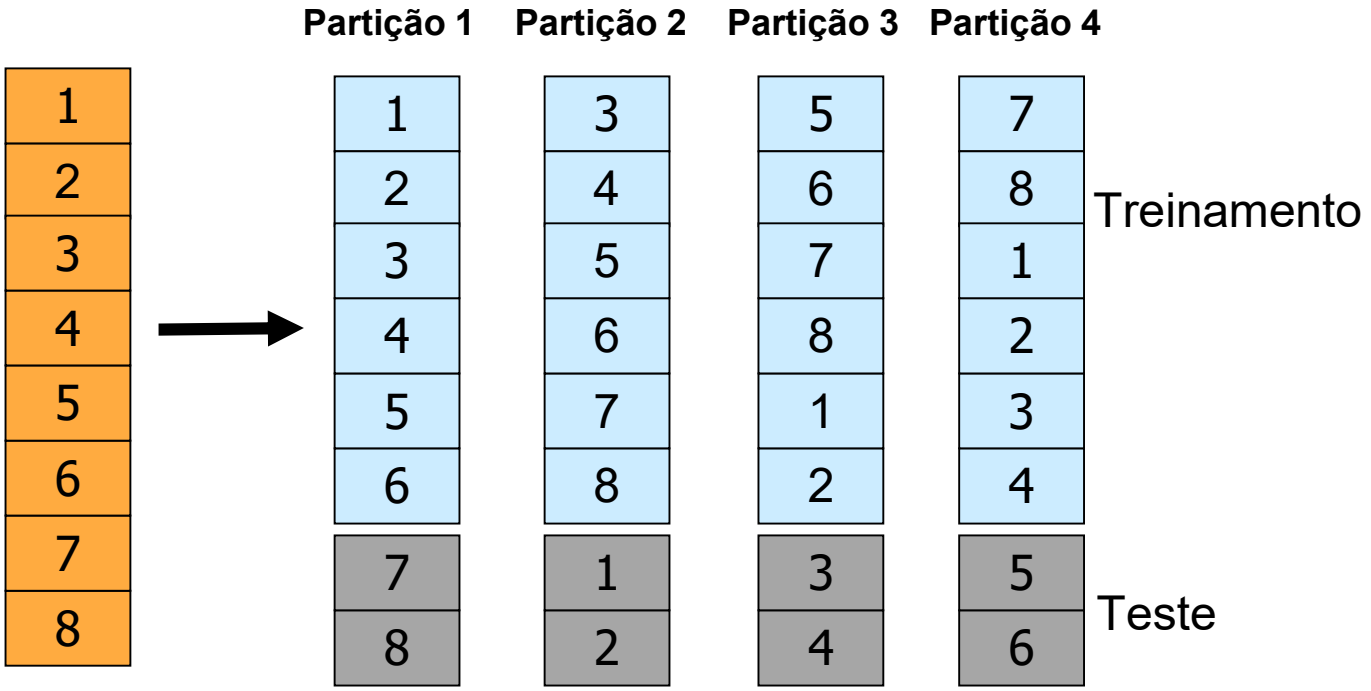
Random subsampling



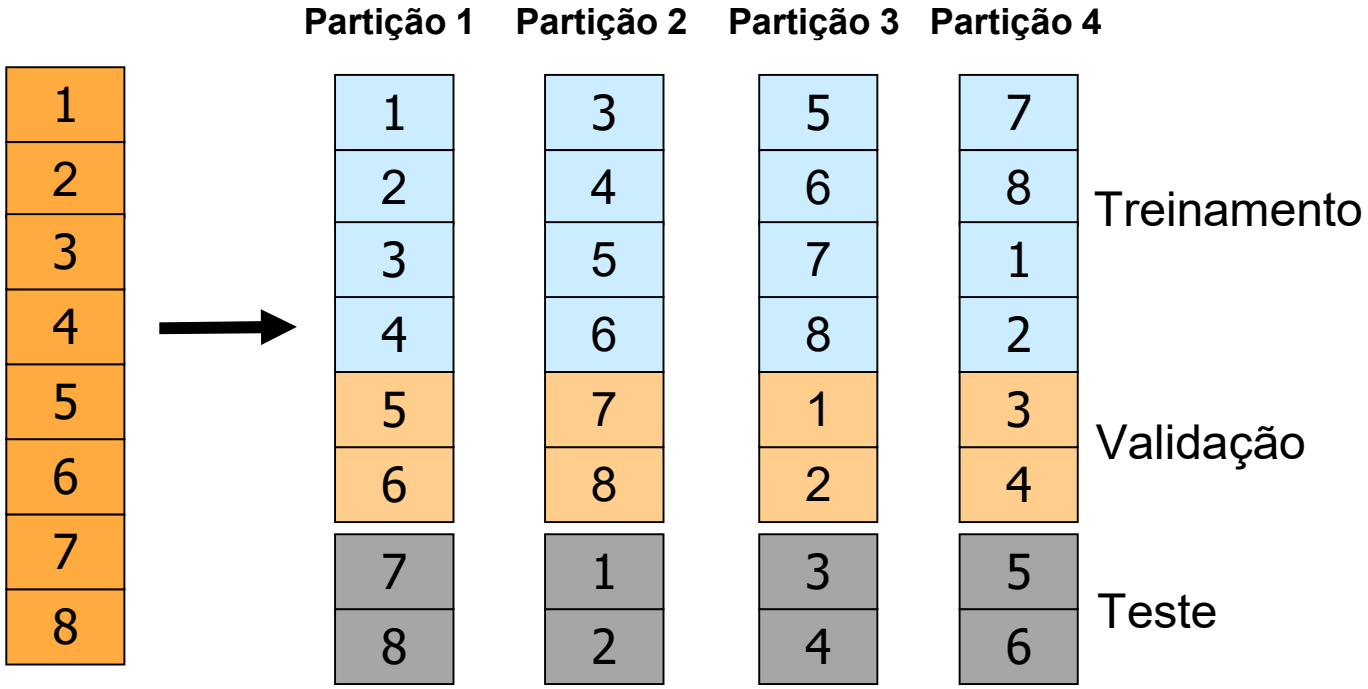
k-fold cross-validation



4-fold cross-validation



4-fold cross-validation



Leave-one-out

- Tende à taxa de erro verdadeira
- Custo computacionalmente elevado para conjuntos grandes
    - Geralmente utilizado para pequenos conjuntos de dados
    - 10-fold cross validation aproxima leave-one-out
- Resultado é a média dos N experimentos
- Variância tende a ser elevada

5 x 2 Cross-validation

- Conjuntos de treinamento e teste com mesmo tamanho

Seja um conjunto de N exemplos
Para i = 1 até 5
Dividir N aleatoriamente em duas metades
Usar metade 1 para treinamento e metade 2 para teste
Usar metade 2 para treinamento e metade 1 para teste

Bootstrap

- Estocástico, com diversas variações
  - Alguns exemplos podem não participar do treinamento
- Variação mais simples:
  - Amostragem com reposição
    - Cada partição é uma amostra aleatória com reposição do conjunto total de exemplos
      - Conjunto de treinamento têm o mesmo número de exemplos do conjunto total
      - Exemplos que restarem são utilizados para teste

Bootstrap

- Se conjunto original tem N exemplos
    - Amostra de tamanho N tem ≈ 63,2% dos exemplos do conjunto de dados original
- Processo é repetido k vezes
    - Resultado final é a média dos k experimentos

Bootstrap

- Estima incerteza de um algoritmo
    - *K-fold cross-validation* é mais usado para estimar acurácia preditiva
        - Seleção de algoritmos/modelos
- Tende a ter menor variância e ser mais pessimista que *k-fold cross-validation*

Considerações Finais

- Desempenho preditivo
- Avaliação do desempenho
  - Erro
  - Tempo de resposta
  - Memória
  - Representação
- Partições do conjunto de dados

Final da Apresentação