Aprendizado de Máquina

Aula 5: Experimentos e amostragem de dados

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

Tópicos a serem abordados

- Planejamento de experimentos
- Avaliação de desempenho de algoritmos/modelos
- Desempenho preditivo
- Partição dos dados
- Amostragem
- Reamostragem

Desempenho preditivo

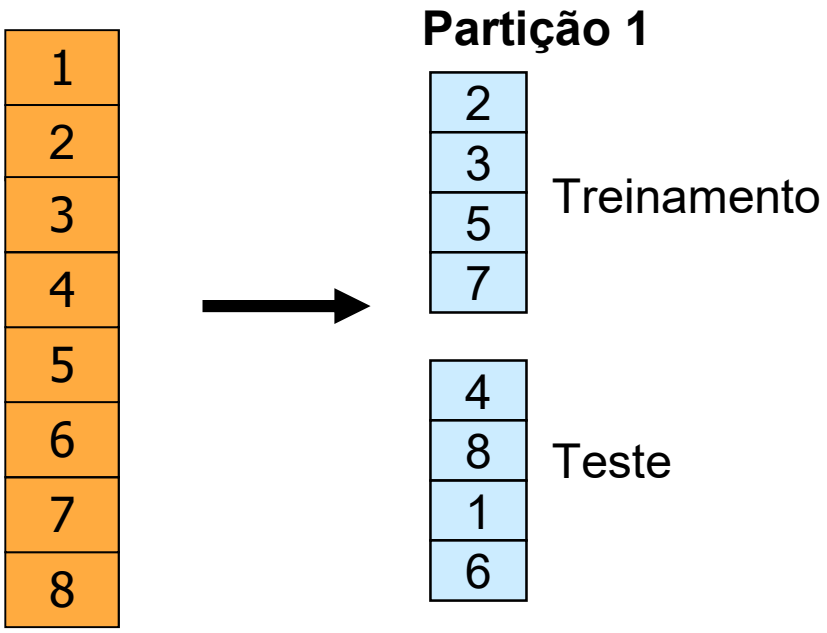
- Principal objetivo em tarefas de classificação:
  - Classificação correta de novos exemplos
    - Errar o mínimo possível
      - Minimizar taxa de erro para novos exemplos
- Geralmente não é possível medir com exatidão essa taxa de erro para novos exemplos
  - Deve ser estimada utilizando duas amostras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

Partição de dados

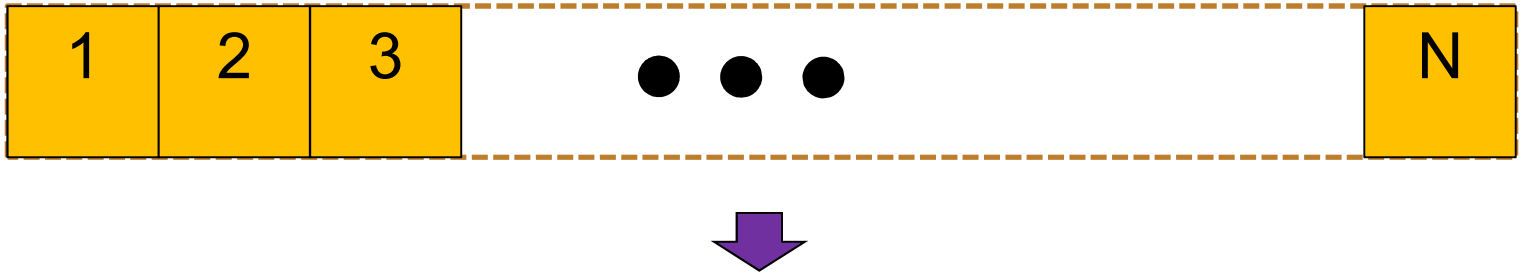
- Permite melhor estimativa do desempenho de um modelo ou algoritmo
  - Treinamento (validação) e teste
- Procedimentos
  - Amostragem única
    - *Hold-out*
  - Várias amostragens
    - Re-amostragem

Hold out

- Geralmente 50% para treino e 50% para teste
- Outras divisões também são usadas
  - 66,6% e 33,3%
  - 75% e 25%



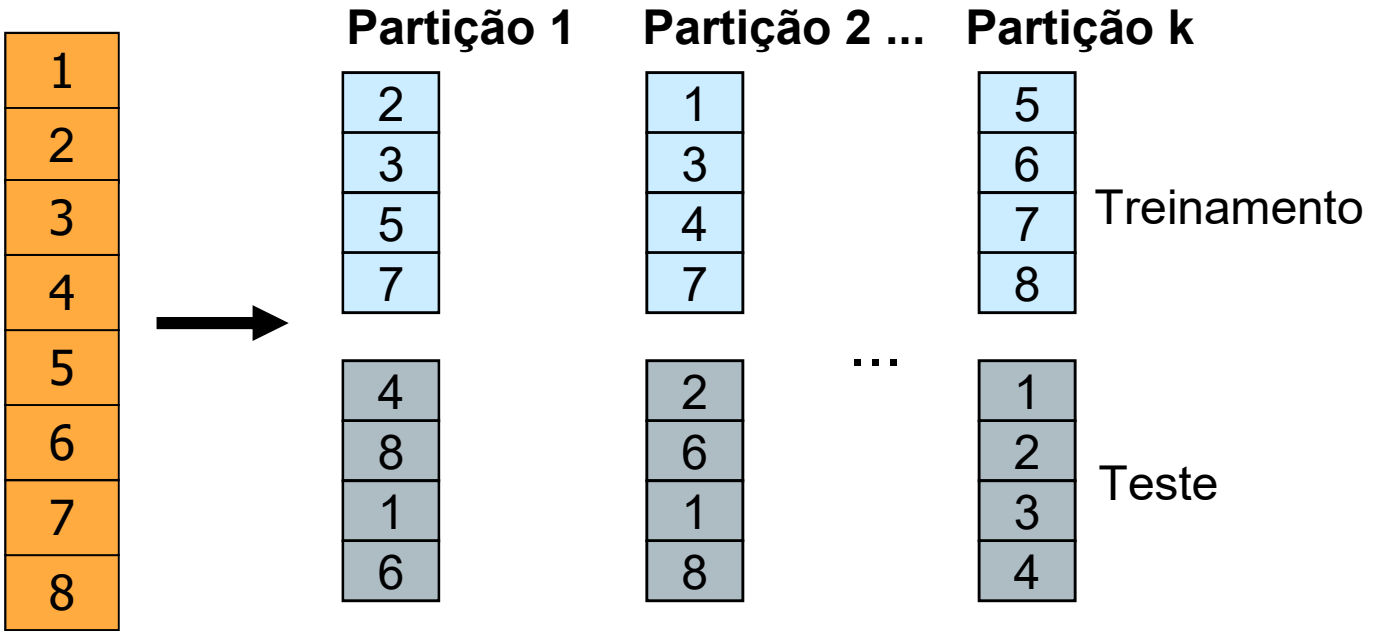
Hold out



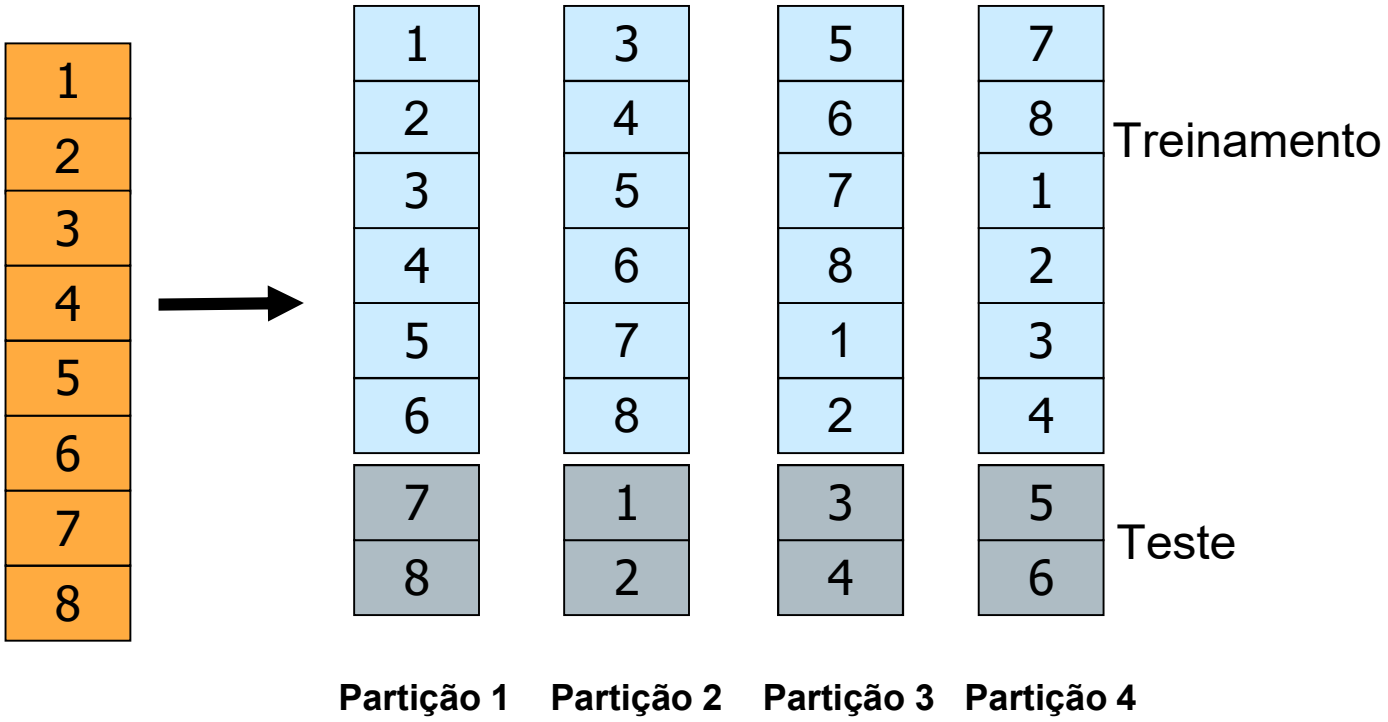
Hold out

- Amostragem única é pouco confiável
  - Sorte (ou azar) na definição das amostras
- Para ter um resultado mais confiável, gerar várias partições para conjuntos de treinamento (validação) e teste
  - *Reamostragem*
    - *Random subsampling*
    - *K-fold Cross-validation*
      - *Leave-one-out*
    - *Bootstrap (ou Bootstrapping)*

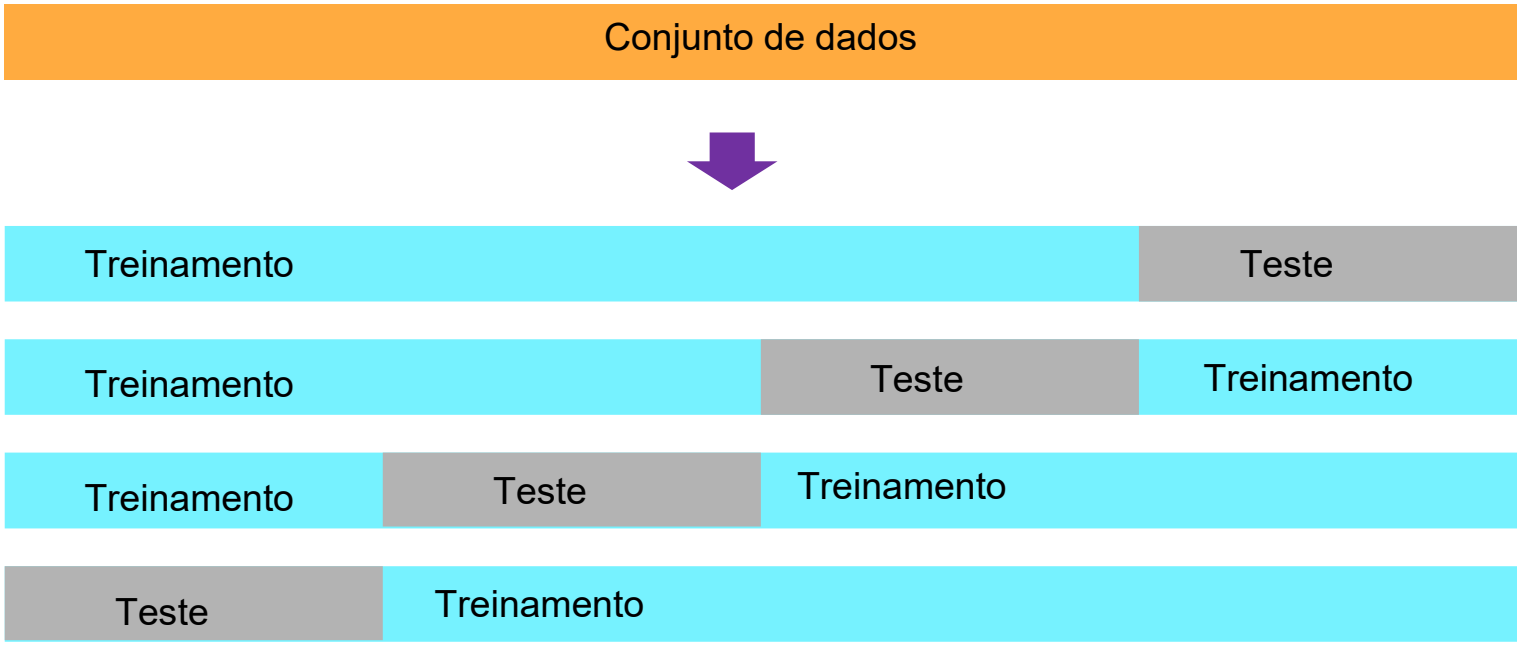
Random subsampling



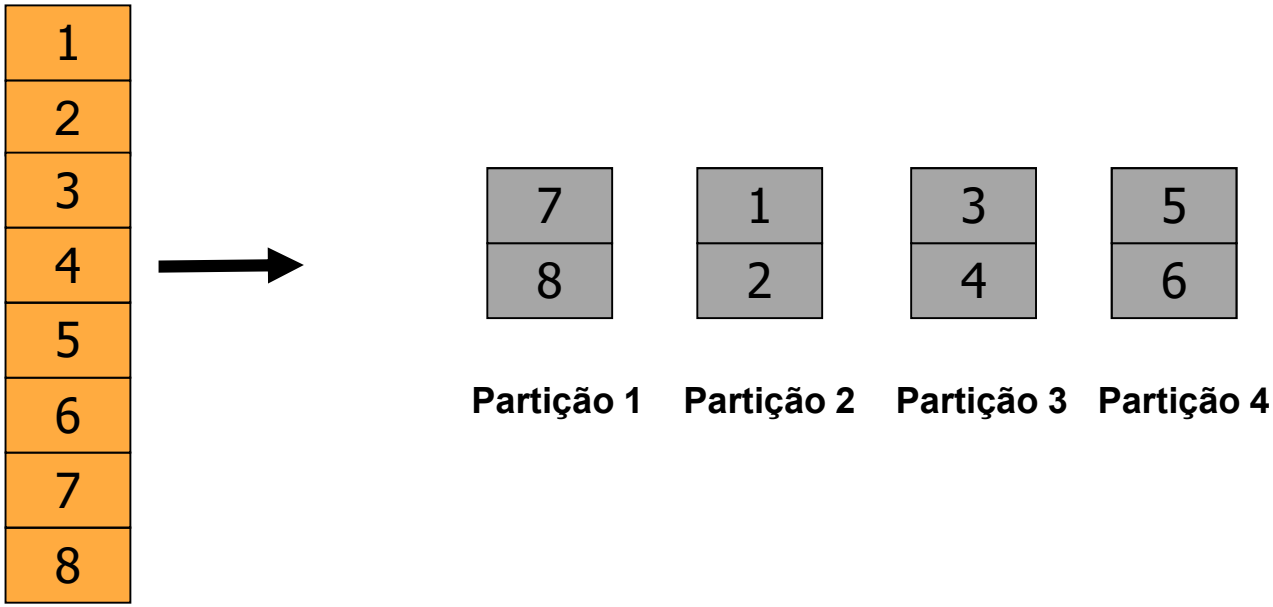
4-fold cross-validation



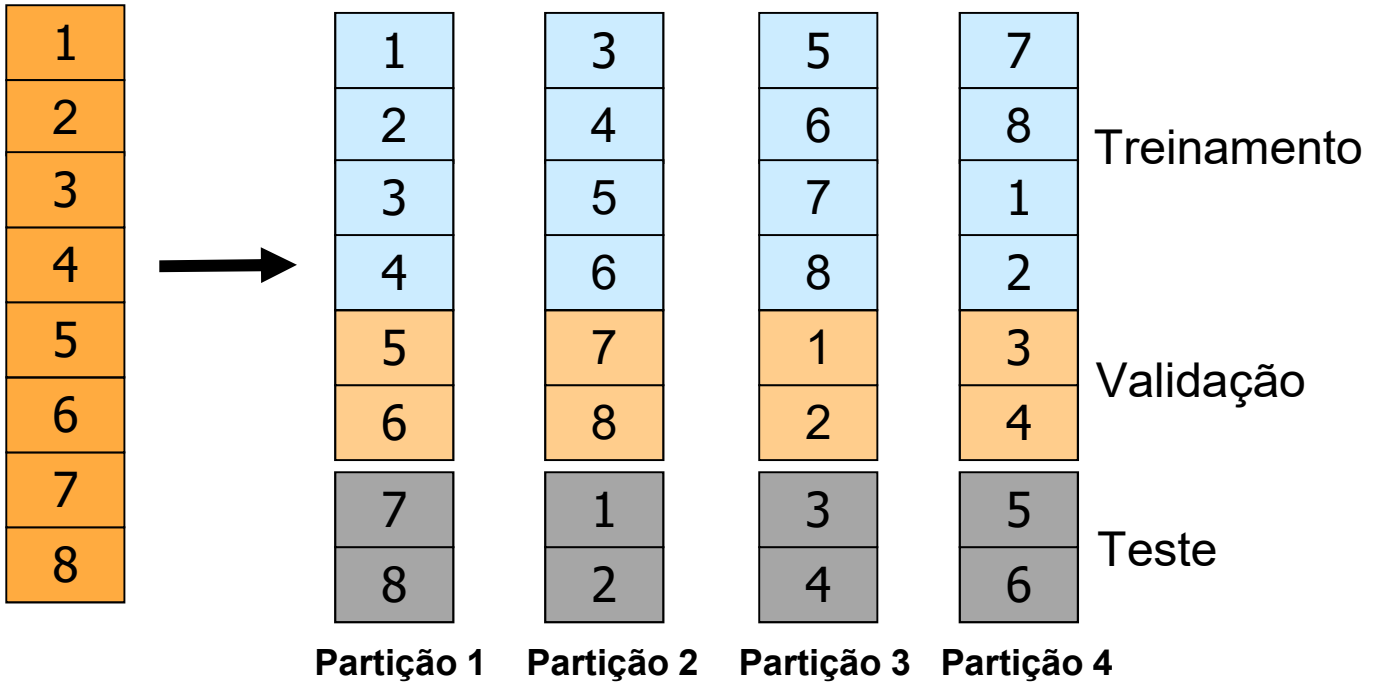
4-fold cross-validation



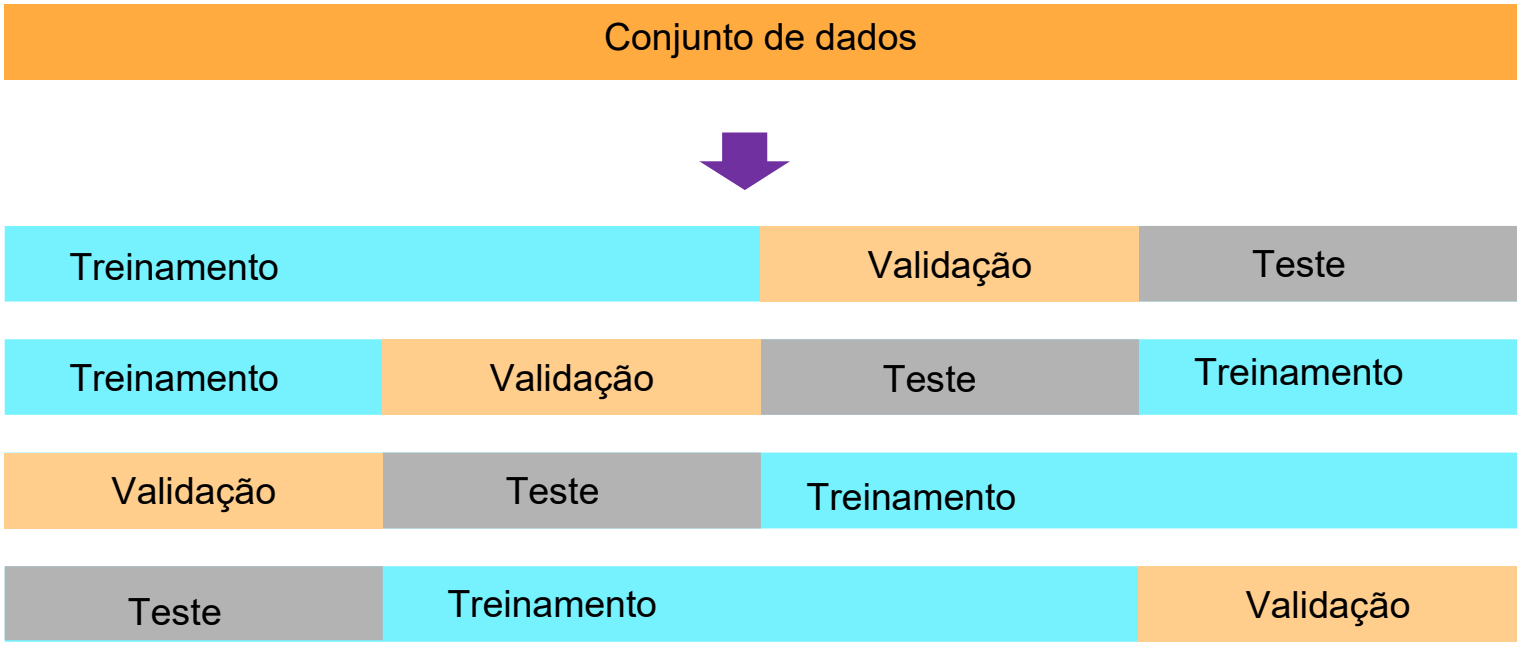
4-fold cross-validation



4-fold cross-validation com dados de validação



4-fold cross-validation com dados de validação



Leave-one-out

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

Bootstrap

- Estocástico, com diversas variações
  - Alguns objetos podem não participar do processo de 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

Considerações Finais

- Estimativa de desempenho de modelos preditivos
    - Para dados novos
- Não é possível predizer
- Mas é possível estimar
    - Simulando dados de teste
    - Particionando o conjunto de dados
        - Várias alternativas
        - Custo computacional e proximidade da estimativa

Final da Apresentação