Aprendizado de Máquina

Aula 2.2 - Metodologia e Medidas de avaliação

Adriano Rivolli

rivolli@utfpr.edu.br

Especialização em Inteligência Artificial

Universidade Tecnológica Federal do Paraná (UTFPR)
Câmpus Cornélio Procópio
Departamento de Computação

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Conteúdo

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Metologia de avaliação

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Introdução

- Importância dos procedimentos de avaliação
  - Seleção de bons modelos
  - Validade dos resultados
  - Robustez em relação à generalização
- Procedimentos específicos para tarefas supervisionadas

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Divisão dos dados

- **Treino** utilizado para treinar o modelos
- **Validação** utilizado para ajustar os hiperparâmetros e demais configurações (opcional, quando não há ajustes)
- **Teste** utilizado para avaliar o desempenho final

- O uso de **estratificação** visa garantir uma divisão mais representativa

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Amostragem

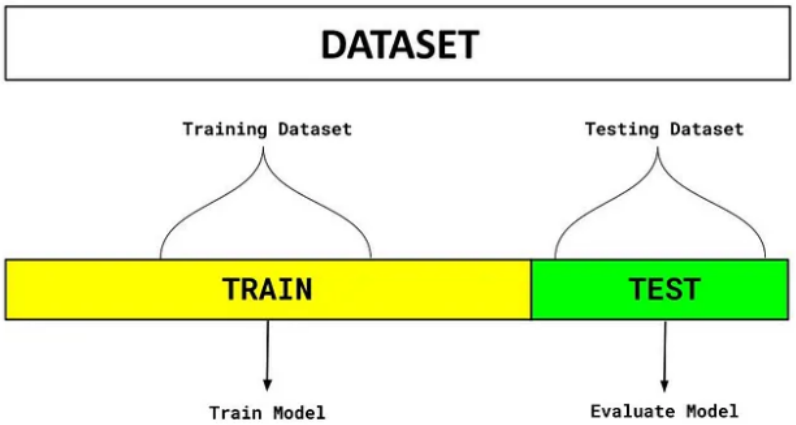
- Como dividir o conjunto de dados disponível?
- Estratégias:
  - *Holdout* (amostragem aleatória)
  - *Bootstarp*
  - *Cross validation* (validação cruzada)
    - *Leave one out*

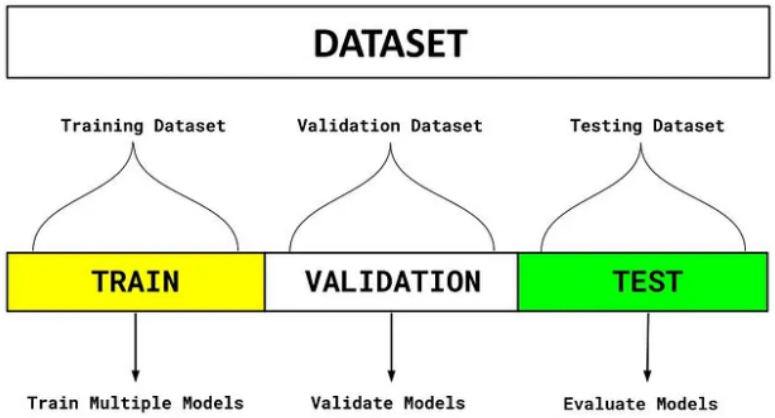
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Holdout

- A mais simples das abordagens
- Divide os dados em proporções de treinamento e teste
  - O conjunto de validação pode ser gerado a partir dos dados de treinamento reaplicando a amostragem
- Apropriado para grandes conjuntos de dados
- Não permite avaliar adequadamente a variância do modelo
  - É possível repetir o mesmo procedimento repetida vezes

Ilustração de Holdout





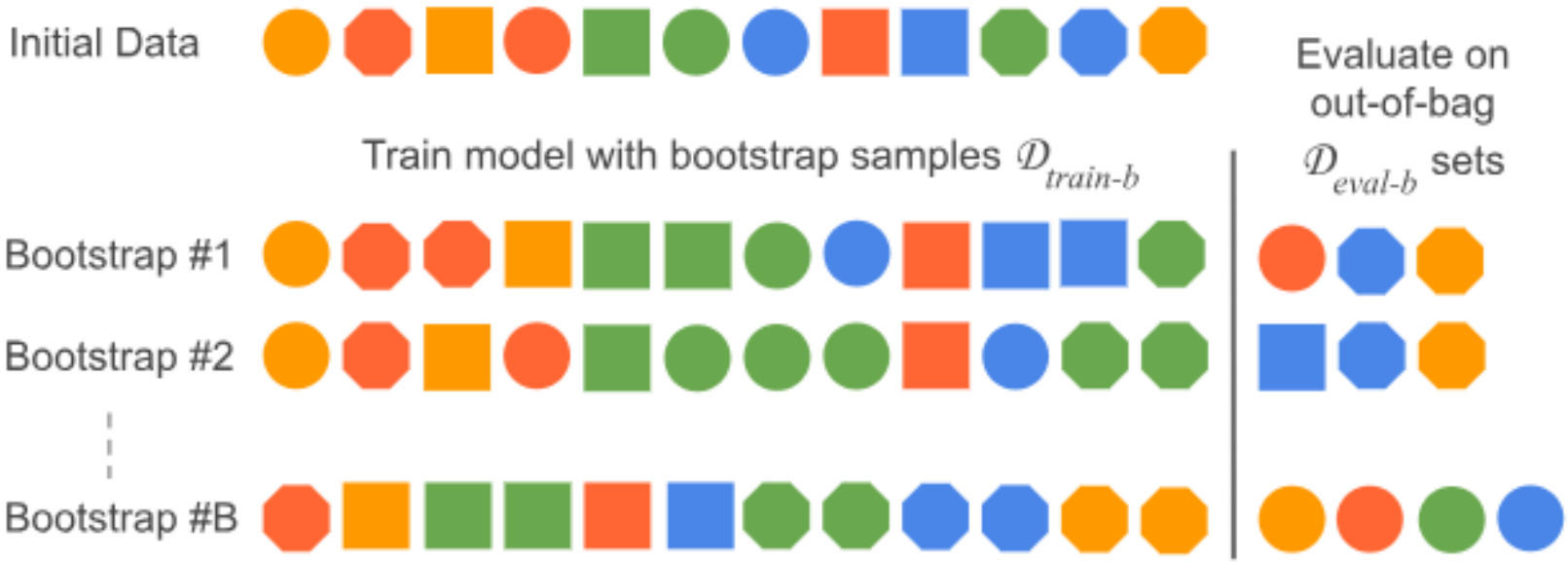
Fonte: https://www.comet.com/site/blog/understanding-hold-out-methods-for-training-machine-learning-models/

Bootstrap

- Neste método os dados são amostrados uniformemente com reposição
  - É gerado um conjunto de treinamento com o mesmo número de instâncias do conjunto original
  - Isso faz com que os dados de treinamento tenham instâncias repetidas

    63% das instâncias originais são selecionadas para o treinamento
- As instâncias de testes são as que não foram utilizadas para o treinamento
- O processo por ser repetido múltiplas vezes
- Não há como

Ilustração de Bootstrap

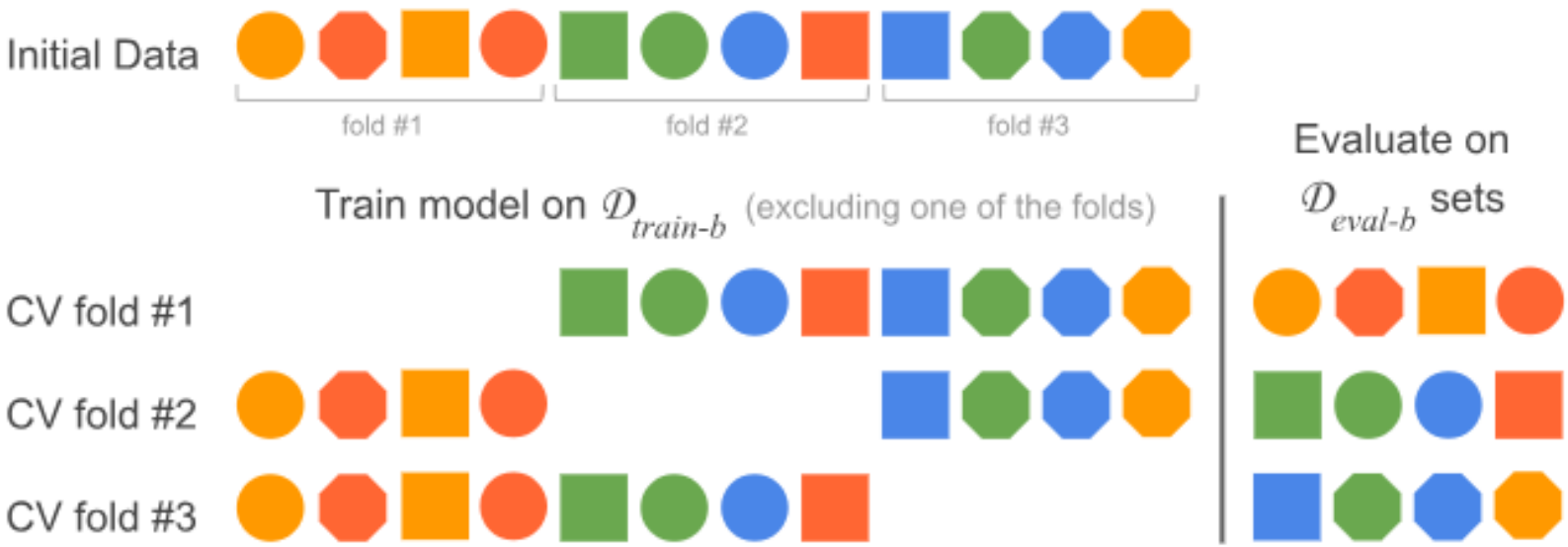


Fonte: http://allmodelsarewrong.github.io/resampling.html

K-fold Cross Validation

- Divide os dados em k grupos de instâncias (mesmo tamanho)
  - 3-folds, 5-folds e 10-folds
- Em cada iteração $k - 1$ grupos são usados para treinamento e o grupo restante para teste
- Sobre a validação:
  - Usa *holdout* no conjunto de treinamento
  - Treinar com $k - 2$ grupos e usar um *fold* para validação
  - Aplicar um novo *k-fold* com o conjunto de treinamento (*nested k-fold*)

Ilustração de k-fold Cross Validation



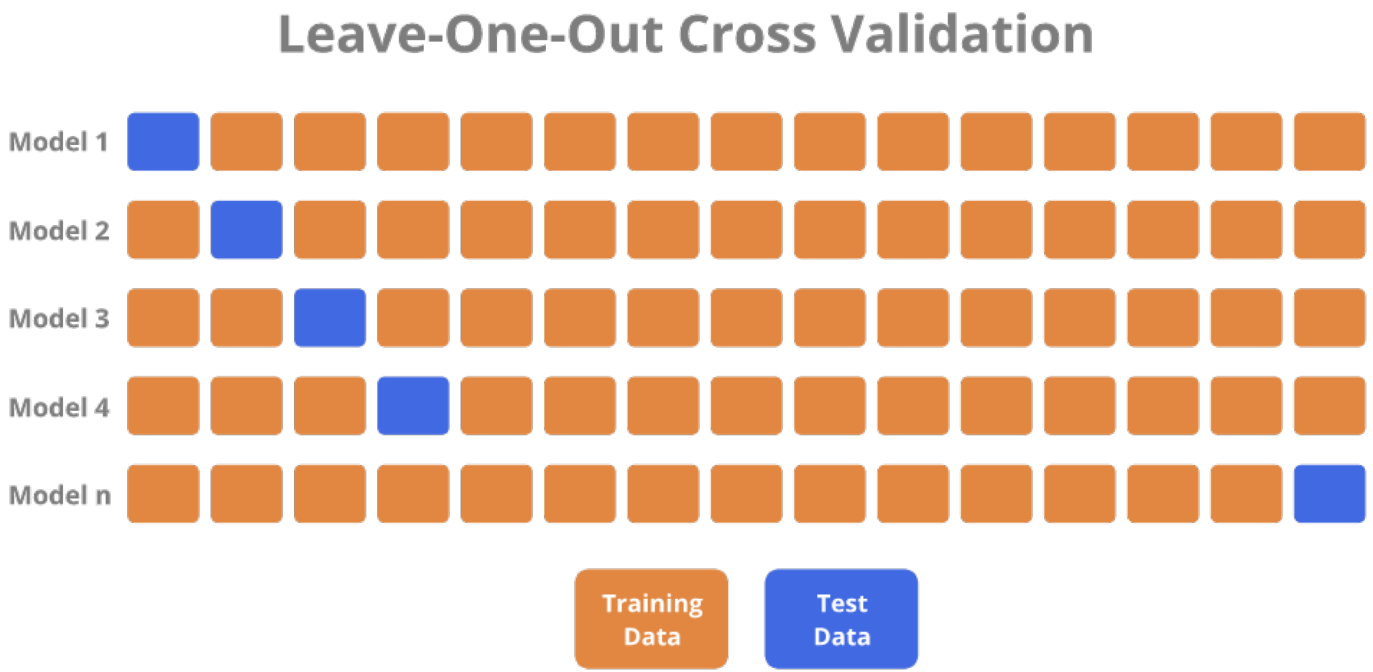
Fonte: http://allmodelsarewrong.github.io/resampling.html

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Leave-one-out

- **Deixe um fora**
  - Apenas uma instância é utilizado para o teste
- *K-fold CV* no qual o número de *folds* é igual ao número de instâncias ($k = |D|$)
- É Geralmente aplicado quando o conjunto de dados é pequeno

Ilustração de Leave-one-out



Fonte: https://medium.datadriveninvestor.com/leave-one-out-cross-validation-32fa248c1739

Medidas de avaliação

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Medidas de avaliação

- Como mensurar o desempenho de um modelo preditivo?
- Diferentes visões sobre os erros/acertos:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R-squared (*Coefficient of Determination*)

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Mean Absolute Error (MAE)

- Média da diferença absoluta do erro
- É fácil de entender e oferece um bom indicativo do desempenho do modelo
- Trata todos os erros igualmente, independente da sua magnitude
- Fórmula:

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

- Média diferença do erro quadrado
- É geralmente como função de custo devido o fato de ser diferenciável
- Penaliza os erros maiores, tornando mais sensível aos *outliers* (ao elevar ao quadrado)
- Fórmula:

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Root Mean Squared Error (RMSE)

- É a raiz quadrada do MSE
- Está na mesma unidade que os dados (atributo alvo)
- É sensível aos grandes erros como MSE
- Fórmula:

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}$$

R-squared (Coefficient of Determination)

- Indica a proporção da variabilidade dos dados explicada pelo modelo
- A medida é indicativa quando está entre 0 e 1
  - 1 é o ajuste perfeito aos dados
  - 0 o modelo não aprendeu nada seria melhor predizer a média dos valores
- Fórmula:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y}_i)^2}$$

Exemplo

Actual Price ($)	Predicted Price ($)
300,000	320,000
400,000	380,000
500,000	520,000
600,000	580,000
700,000	660,000

- MAE = 24000
- MSE = 640000000
- RMSE = 25298
- R2 = 0.968

Actual Price ($)	Predicted Price ($)
300,000	310,000
400,000	390,000
500,000	480,000
600,000	570,000
700,000	660,000

- MAE = 22000
- MSE = 620000000
- RMSE = 24899
- R2 = 0.969

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Testes estatísticos

Teste de hipóteses

- Utilizado para comparar se a diferença de desempenho entre os modelos são estatisticamente significantes
  - Hipótese Nula (H_0) Não há diferença entre os modelos
  - Hipótese alternativa (H_1) há diferença entre os modelos
- Tipos de erros
  - **Tipo I** - A hipótese nula é rejeitada, mas não deveria ser
    - Erro mais problemático
    - Nível de significância $\alpha = 0.05$ ou $\alpha = 0.01$
  - **Tipo II** - A hipótese nula não é correta, mas não é rejeitada pelo teste

Principais testes utilizados em AM

- Wilcoxon signed-rank

  Compara dois modelos em um único conjunto de dados

- Friedman

  Compara múltiplos modelos utilizando diferentes conjuntos de dados

- Nemenyi test

  Pós-teste para determinar quais são as diferenças

- Testes Bayesian

  Gera uma probabilidade da diferença