MBA USP ESALQ

Encontro para resolução de exercícios adicionais: Árvores e Ensemble Models

Prof. Dr. Wilson Tarantin Junior

MBA USP ESALQ

A responsabilidade pela idoneidade, originalidade
e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização.

Lei nº 9610/98

Modelos de classificação ou regressão?

Depende da variável dependente (Y)

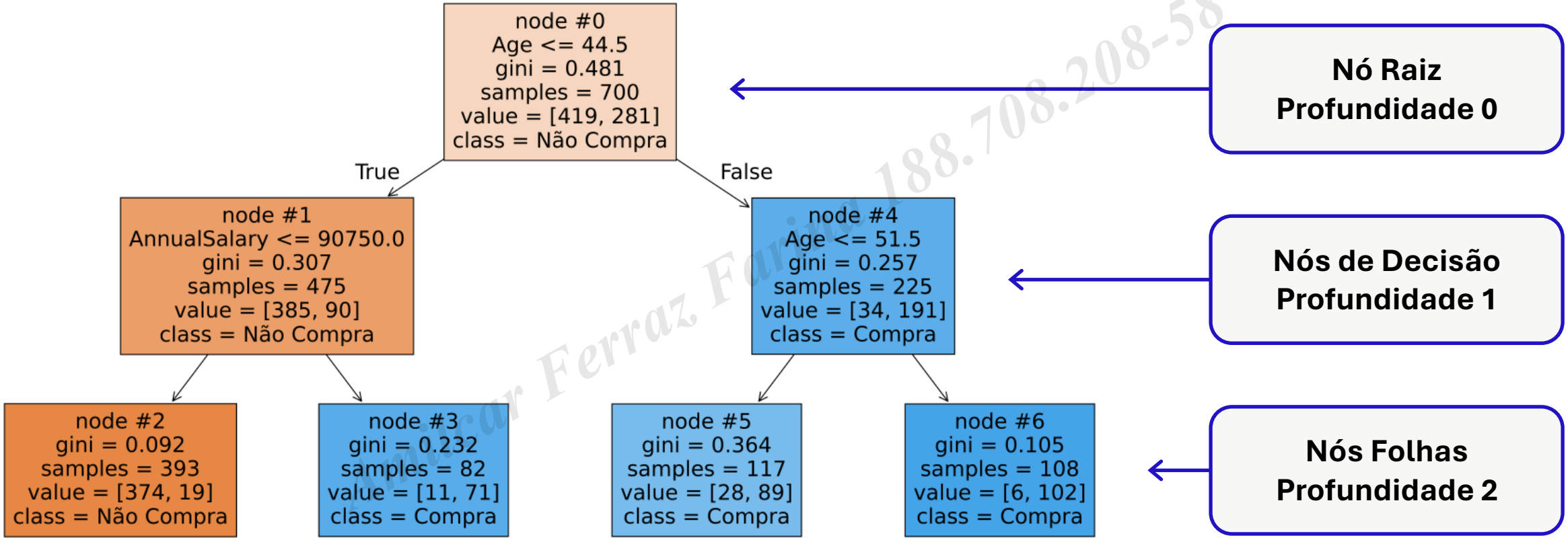
- Modelos de machine learning com a finalidade de classificação ou regressão:
  - Quando Y for categórica → classificação
  - Quando Y for métrica → regressão
- As árvores de decisão servem de base para as *Random Forests* e para o *Boosting*
- Podem apresentar boa capacidade preditiva mesmo em bancos de dados complexos
  - Porém, é necessário cuidado para evitar o *overfitting*!

Árvores de Decisão

Modelo baseado no particionamento recursivo dos dados

- Seleciona-se a variável que melhor separa os dados em 2 grupos
- Em seguida, o mesmo processo é aplicado separadamente em cada subgrupo gerado
- Tal processo é repetido até que atinja-se os fatores limitantes do modelo, isto é, aqueles estabelecidos pelos hiperparâmetros
- Como o algoritmo seleciona a variável e o limiar para o split do nó?
  - É a combinação variável + limiar que gera a divisão com a máxima redução da impureza (ex.: Gini) ou do erro de previsão (ex.: MSE)

Árvores de Decisão



Ensemble Models

O que são os ensemble models?

- São modelos que **agregam (agrupam)** as predições feitas por um grupo de preditores com o intuito de obter melhores resultados (melhores previsões) comparativamente ao resultado obtido em um preditor individual

- A ideia fundamental é que as respostas advindas de muitos modelos é melhor do que a resposta de um modelo único, sendo assim, o erro de previsão pode ser reduzido e a capacidade de generalização ainda pode ser obtida (evitando o *overfitting*)

- O conceito de ensemble pode ser aplicado em modelos com Y métrica ou categórica

Ensemble Models: Random Forest

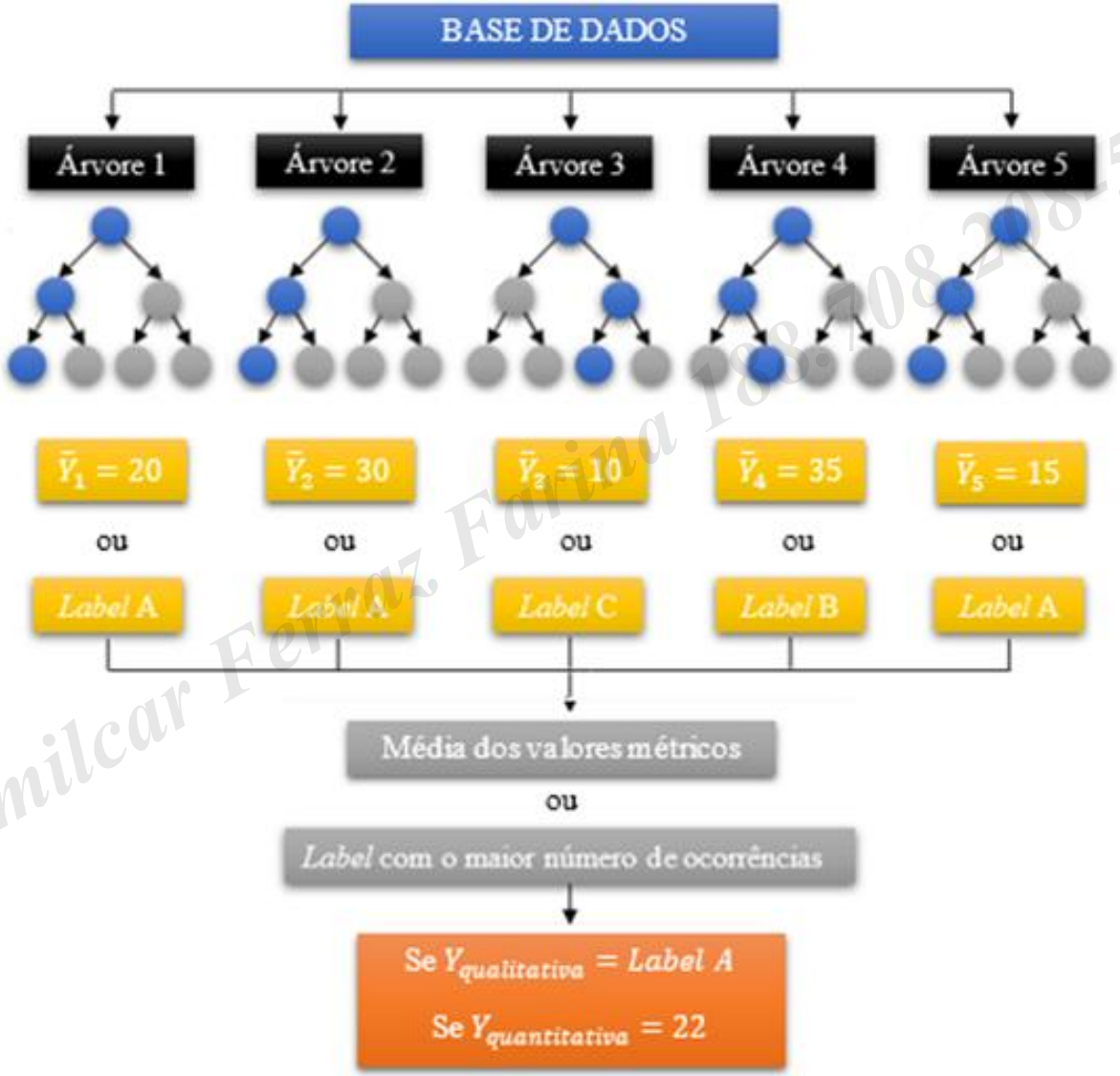
Random Forests são um caso particular de *bagging* (que é um tipo de *ensemble*)

- *Bagging*: método que treina o mesmo algoritmo em diferentes conjuntos aleatórios da base de dados: utiliza múltiplas amostras aleatórias com reposição geradas dos dados de treino; após os modelos serem treinados em cada amostra aleatória, uma previsão agregada é feita agrupando-se as previsões dos modelos individuais

- *Random Forests* são agrupamentos de árvores de decisão treinadas pelo método de *bagging* em que, além da amostragem aleatória da base de dados, realiza a amostragem aleatória das variáveis preditoras que serão utilizadas nos particionamentos dos nós

  - Para variável Y métrica: agregação pela média das previsões
  - Para variável Y categórica: agregação pela categoria mais frequentemente predita

Ensemble Models: Random Forest

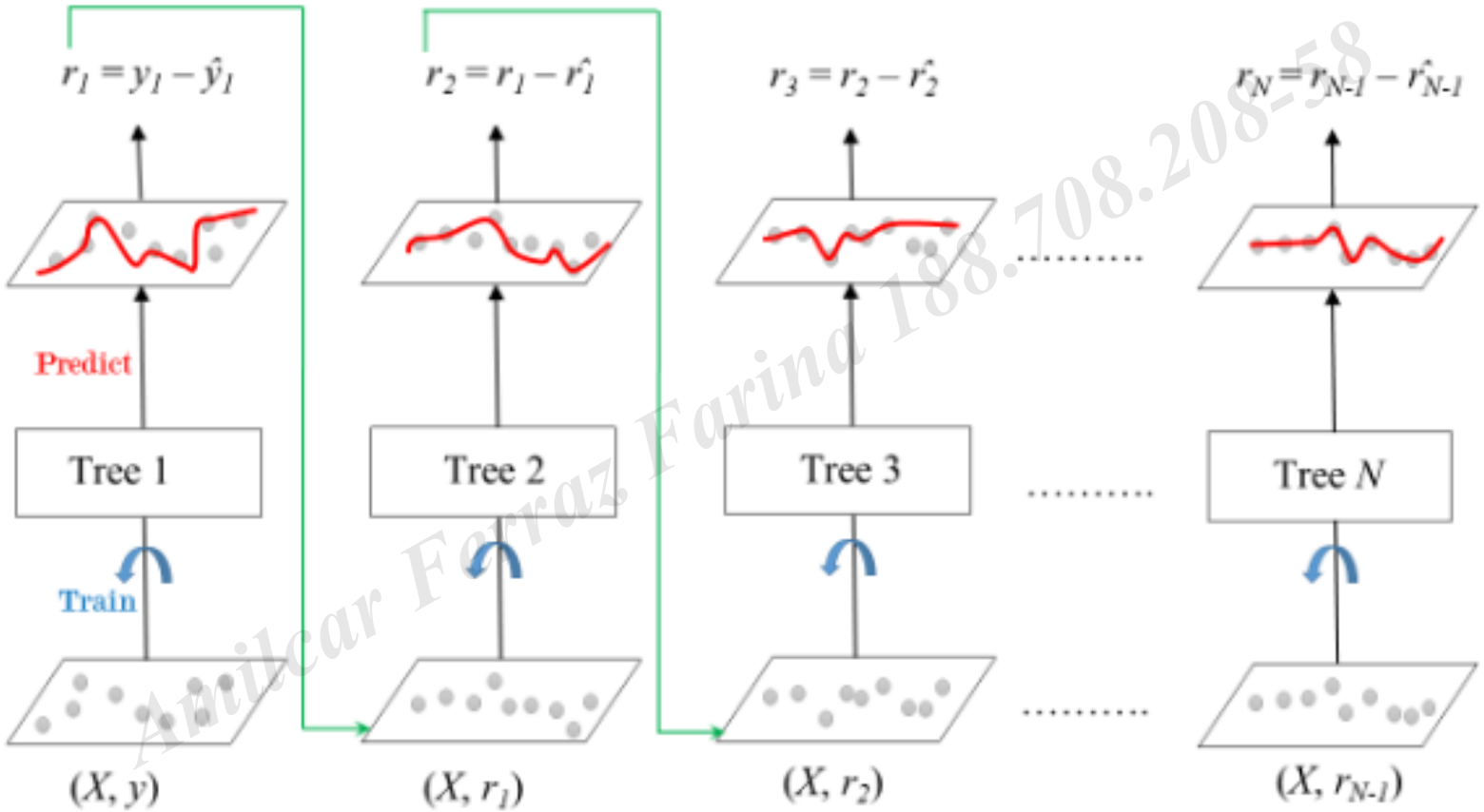


Ensemble Models: Gradient Boosting

Boosting são algoritmos que treinam modelos em sequência

- *Gradient Boosting*: agrupamento de modelos em que o novo modelo busca corrigir os erros cometidos pelo modelo antecessor
- Frequentemente, o modelo utilizado como base é a árvore de decisão
- O objetivo do *gradient boosting* é que o modelo que vem na sequência melhore a predição em relação ao modelo anterior ajustando erros residuais dele; se a cada novo modelo os erros residuais forem diminuindo, as primeiras estimativas erram mais do que as últimas
  - Ao agregar as predições de todos os modelos individuais, a previsão final tende a ser melhor

Ensemble Models: Gradient Boosting



Fonte: https://www.geeksforgeeks.org/ml-gradient-boosting/

Dados de treino e teste

Treinamento e avaliação do modelo

- Ao utilizar dados de treino e teste, o objetivo é desenvolver o modelo na amostra de treino e, em seguida, avaliar sua capacidade preditiva na amostra de teste

  - É fundamental quando os modelos de machine learning têm propensão ao *overfitting*
  - Amostra de teste: avaliar a capacidade de generalização em dados "desconhecidos" pelo modelo

- Algumas características importantes nos dados de treinamento:

  - Quantidade suficiente de dados: amostras muito pequenas podem prejudicar a generalização
  - Dados representativos da população que deseja estudar e generalizar
  - Dados de qualidade (sem erros de mensuração, outliers, ruídos) e com variáveis relevantes

Overfitting

Ajuste excessivo aos dados de treino

- Se for deixado sem restrições, o modelo fica livre e pode se aderir aos dados de treino
  - Na amostra de treinamento, a capacidade preditiva torna-se bastante elevada
  - No entanto, na amostra de teste a capacidade preditiva fica bastante reduzida
- Quando ocorre *overfitting*, não há generalização dos resultados e o modelo não consegue fazer boas previsões para novas observações (aquelas que não foram usadas para treinar o modelo)
  - Uma solução para evitar o *overfitting* é reduzir a liberdade do modelo na amostra de treinamento
  - É comum utilizar hiperparâmetros que regulam o modelo treinado

Hiperparâmetros

Elementos responsáveis pelo ajuste do modelo

- Os ajustes de hiperparâmetros são responsáveis por reduzir o *overfitting* e também por melhorar a qualidade das estimativas e capacidade preditiva

- Cada modelo de machine learning tem seus hiperparâmetros específicos, mas os modelos baseados em árvores de decisão têm como mais comuns:

  - Profundidade máxima das árvores, número mínimo de observações em dado nó para que seja realizada sua divisão, número mínimo de observações em um nó folha, a própria quantidade de árvores estimadas (se for ensemble), taxa de aprendizagem nos modelos *boosting*, dentre outros...

Grid Search

Escolha de hiperparâmetros

- Os hiperparâmetros podem ser selecionados por meio de diversas simulações para encontrar uma combinação ótima para seus valores com o objetivo de aperfeiçoar o modelo

- A seguir, o passo a passo ilustra a realização de um Grid Search:

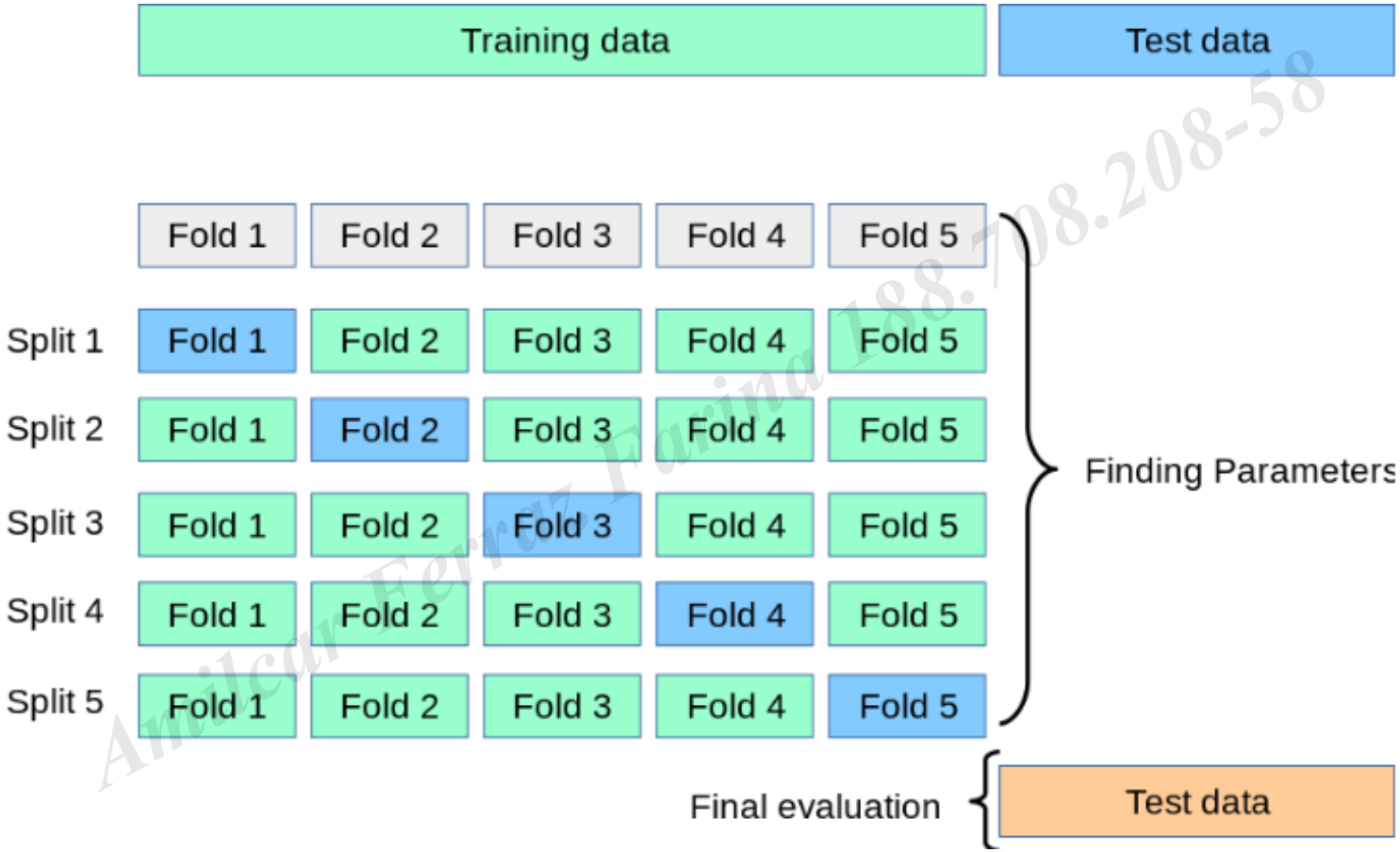
  - Escolher os hiperparâmetros que serão ajustados e seus valores desejados
  - É criada uma lista com todas as combinações de valores dos hiperparâmetros
  - Gerar os modelos com cada uma das combinações de hiperparâmetros
  - Definir uma medida de avaliação para os modelos (exemplos: acurácia ou MSE)
  - Atribui-se a melhor combinação de hiperparâmetros ao modelo com a melhor avaliação

Cross-Validation

Técnica aplicada com o intuito de evitar o *overfitting*

- A *cross-validation* é uma técnica em que a amostra de treino é dividida em partes menores (chamadas de *k folds*)
  - Por exemplo: é comum usar a divisão em 5 partes (*folds*)
- *K – 1* partes são utilizadas para treino (estimação com ajustes de hiperparâmetros) e a outra parte fica para avaliação dos resultados
  - As partes treino/avalição vão se alternando entre elas
- A validação cruzada é aplicada mesmo já existindo a divisão inicial treino/teste

Cross-Validation



Fonte: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation

Critérios de avaliação dos modelos: classificação

Como identificar a qualidade das estimativas?

- **Acurácia**: percentual de classificações corretas em relação ao total de observações analisadas, isto é, trata-se da eficiência geral do modelo

- **Sensibilidade**: percentual de classificações corretas considerando apenas as observações que são evento. Esta medida também é chamada de Recall

- **Especificidade**: percentual de classificações corretas considerado apenas as observações que não são evento

  - Para uma variável Y com 2 categorias, pode-se estabelecer um *cutoff* para classificação, isto é, um ponto de corte acima do qual a observação é classificada como “evento“ e abaixo é “não evento”

Critérios de avaliação dos modelos: classificação

Matriz de confusão

Sensibilidade ↓ Especificidade ↓

		Observado (Real)	
		Sim	Não
Predito (Modelo)	Sim	VP	FP
	Não	FN	VN

$$Acuracia = \frac{VP + VN}{(VP + FN + VN + FP)}$$

$$Sensibilidade = \frac{VP}{(VP + FN)}$$

$$Especificidade = \frac{VN}{(VN + FP)}$$

$$Precision = \frac{VP}{(VP + FP)}$$

$$F1\ Score = 2x\frac{Precision\ x\ Recall}{Precision + Recall}$$

Critérios de avaliação dos modelos: classificação

Curva ROC

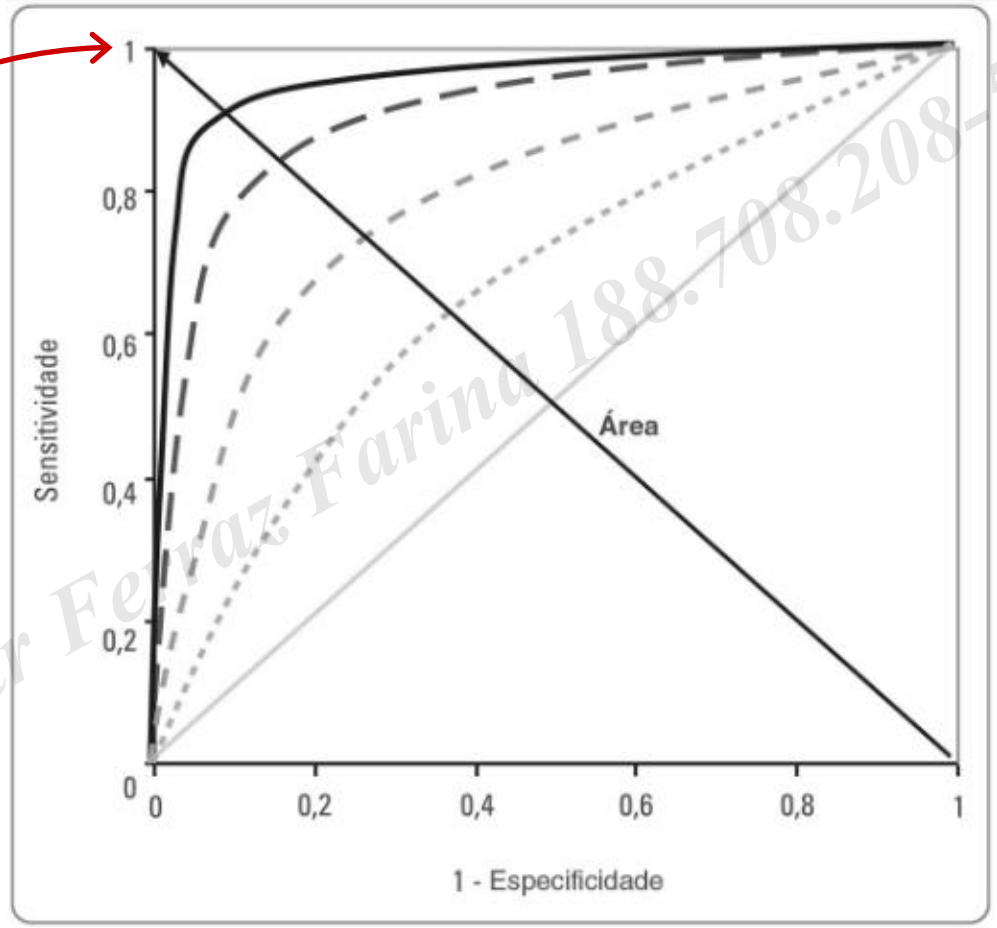
- É um gráfico que mostra a variação da sensibilidade em função de 1 – especificidade
- Independe da escolha de um *cutoff* para a classificação
- Pode-se avaliar um modelo com base na área abaixo da curva ROC (AUC-ROC):
  - Quanto maior a AUC, melhor a capacidade preditiva do modelo na amostra em análise
  - Maiores AUC indicam maiores sensibilidade e especificidade

Critérios de avaliação dos modelos: classificação

Curva ROC

Classificador Ideal:

Sensibilidade = 100%
1 – Especificidade = 0%



Fonte: Fávero e Belfiore (2024, Capítulo 13)

Critérios de avaliação dos modelos: regressão

Como identificar a qualidade das estimativas?

- Calcula-se um **erro de previsão** fundamentado na diferença entre os valores observados e previstos para as observações. Quanto menor o erro, melhor é o ajuste do modelo estimado

- Erro Quadrático Médio: $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

- Erro Absoluto Médio: $MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$

- Coeficiente R^2: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$ Obs.: quanto mais próximo de 1 estiver o R^2, melhor!

Alguns procedimentos para estimação

1. Selecionar o modelo de interesse
2. Separação aleatória da amostra completa entre dados de treino e teste
3. Estimação do modelo
   - Aqui é feita a seleção dos hiperparâmetros (pode ser por meio de um Grid Search e CV)
4. Avaliação dos resultados na base de treinamento
5. Avalição dos resultados na base de testes
   - Houve bom ajuste? *Overfitting*, *Underfitting*?

Referências

- Fávero, Luiz Paulo; Belfiore, Patrícia. (2024). Manual de análise de dados: estatística e machine learning com Excel®, SPSS®, Stata®, R® e Python®. 2 ed. Rio de Janeiro: LTC.

- Géron, Aurélien. (2021). Mão à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow. 2ª ed. Rio de Janeiro: Alta Books.

- Therneau, Terry M.; Atkinson, Elizabeth J. An Introduction to Recursive Partitioning Using the RPART Routines. Mayo Foundation. October 21, 2022.

MBA USP ESALQ

Obrigado!

Wilson Tarantin Junior | linkedin.com/in/wilson-tarantin-junior-359476190