Árvores de Decisão: Bagging, Random Forest e Boosting

André Menezes & Daniel Oliveira

Organização

- Árvores de Decisão
- Bagging e Random Forest
- Boosting
- Aplicação
- Considerações

Árvore de Decisão

- Metodologia não paramétrica apropriada para descrever a relação entre uma variável resposta y_i e um conjunto de covariáveis $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$.
- Consiste em particionar recursivamente o espaço das covariáveis conforme algum critério ótimo.
- Os resultados são **compreensíveis**, porém **não robustos** e com **baixa acurácia** preditiva.
- Seja $R_1, \ldots, R_M$ partições do espaço das covariáveis. O modelo é especificado por

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m \, I(\mathbf{x} \in R_m)$$

em que c_m é um "modelo local" para cada partição.

Árvore de Classificação

- Para uma variável resposta com K classes temos que

$$c_m = \arg\left[\max_k \hat{p}_{mk}\right]$$

em que

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} I(y_i = k)$$

para $k = 1, \ldots, K$.

Bagging e Random Forest

- Gerar B amostras Bootstrap com reposição, ajustar as árvores de decisão e combinar o conjunto de predições, isto é,

$$\widehat{f}_{\text{bag}}(\mathbf{x}) = \frac{1}{B}\sum_{j=1}^{B}\widehat{f}_{(j)}(\mathbf{x})$$

em que $\widehat{f}_{(j)}(\mathbf{x})$ é modelo ajustado na j-ésima amostra Bootstrap.

- No Random Forest para cada amostra Bootstrap escolhe-se aleatoriamente um subconjunto $k < p$ dos preditores.
- Em problemas de classificação $k \approx \sqrt{p}$.
- Os modelos $\widehat{f}_{(j)}$ são treinados de forma independente.

Boosting

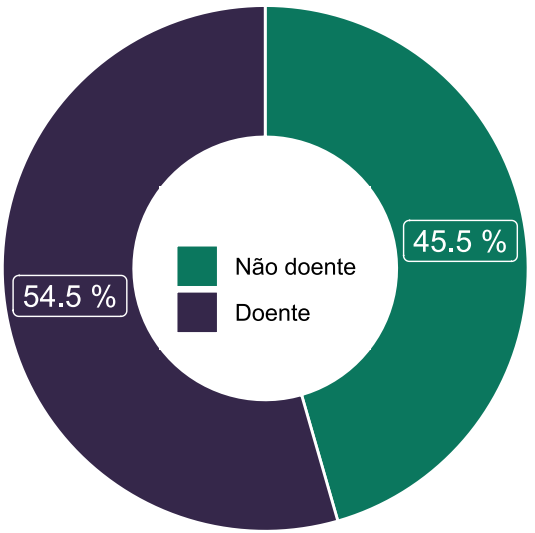
- Seja $\mathbf{T} = \{(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)\}$ a amostra de treinamento, em que $y_i \in \{-1, 1\}$.
- Algoritmo AdaBoost.M1:

  1. Atribuir os pesos $w_i = 1/N, i = 1, \ldots, N$.
  2. Para $b = 1, \ldots, B$:
     - Ajuste um classificador $f_b(\mathbf{x})$ usando os pesos w_i na amostra $\mathbf{T}$.
     - Calcule $\text{err}_b = \dfrac{\sum_{i=1}^{N} w_i I(y_i \neq f_b(\mathbf{x}_i))}{\sum_{i=1}^{N} w_i}$ e $\alpha_b = \log((1 - \text{err}_b)/\text{err}_b)$.
     - Atualize $w_i \leftarrow w_i \cdot \exp[\alpha_b I(y_i \neq f_b(\mathbf{x}_i))], i = 1, \ldots, N$.
  3. Retorne o modelo $f(\mathbf{x}) = \text{sinal}\left[\sum_{b=1}^{B} \alpha_b \, f_b(\mathbf{x})\right]$.

- Os modelos são treinados sequencialmente focando onde o classificador anterior performou mal.

Conjunto de Dados

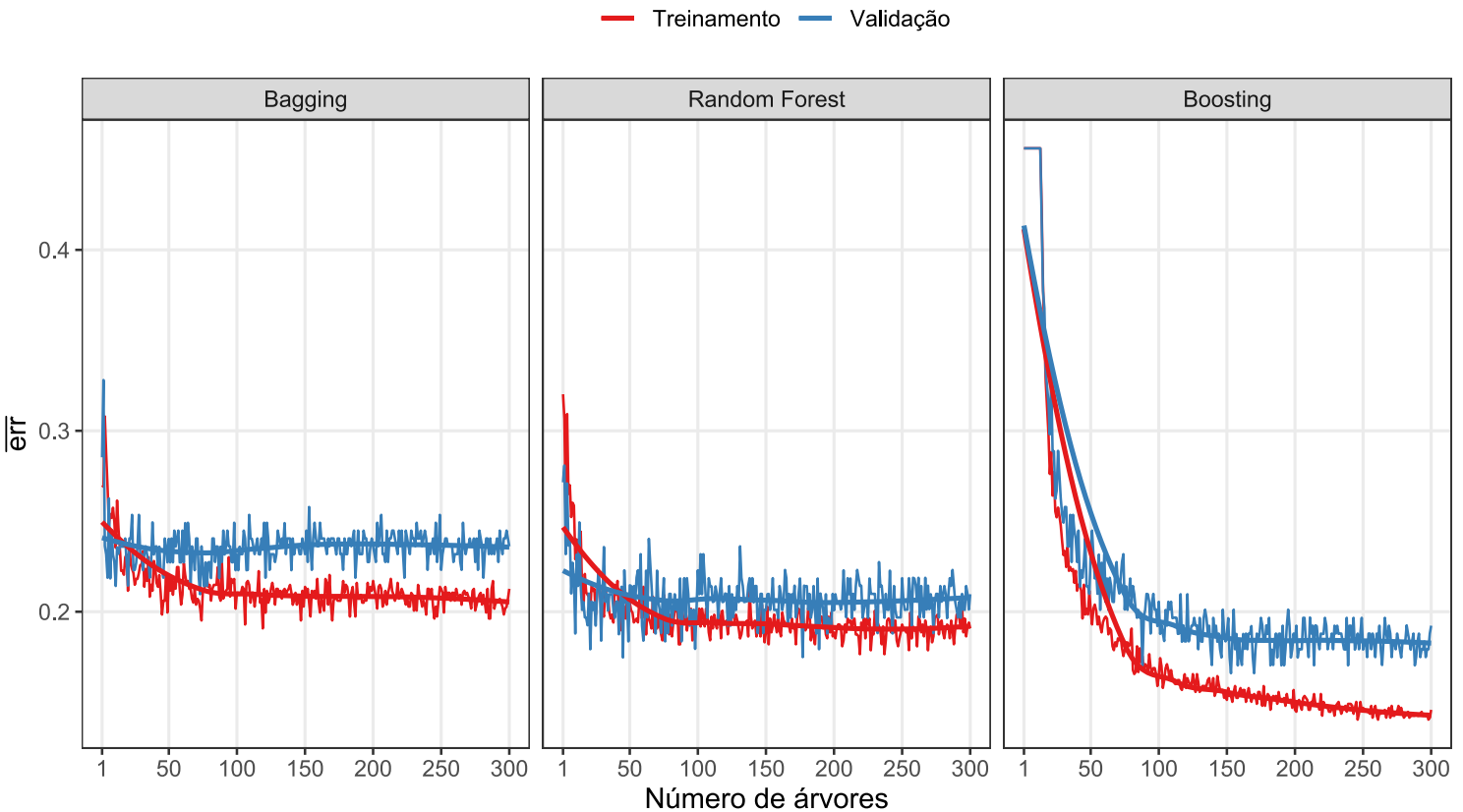
- Fonte: kaggle
- Contexto: informações hospitalares e características pessoais de 303 pacientes.
- Objetivo: classificar se determinar indivíduo tem doença no coração.
- 14 preditores: 6 contínuos e 8 categóricos.



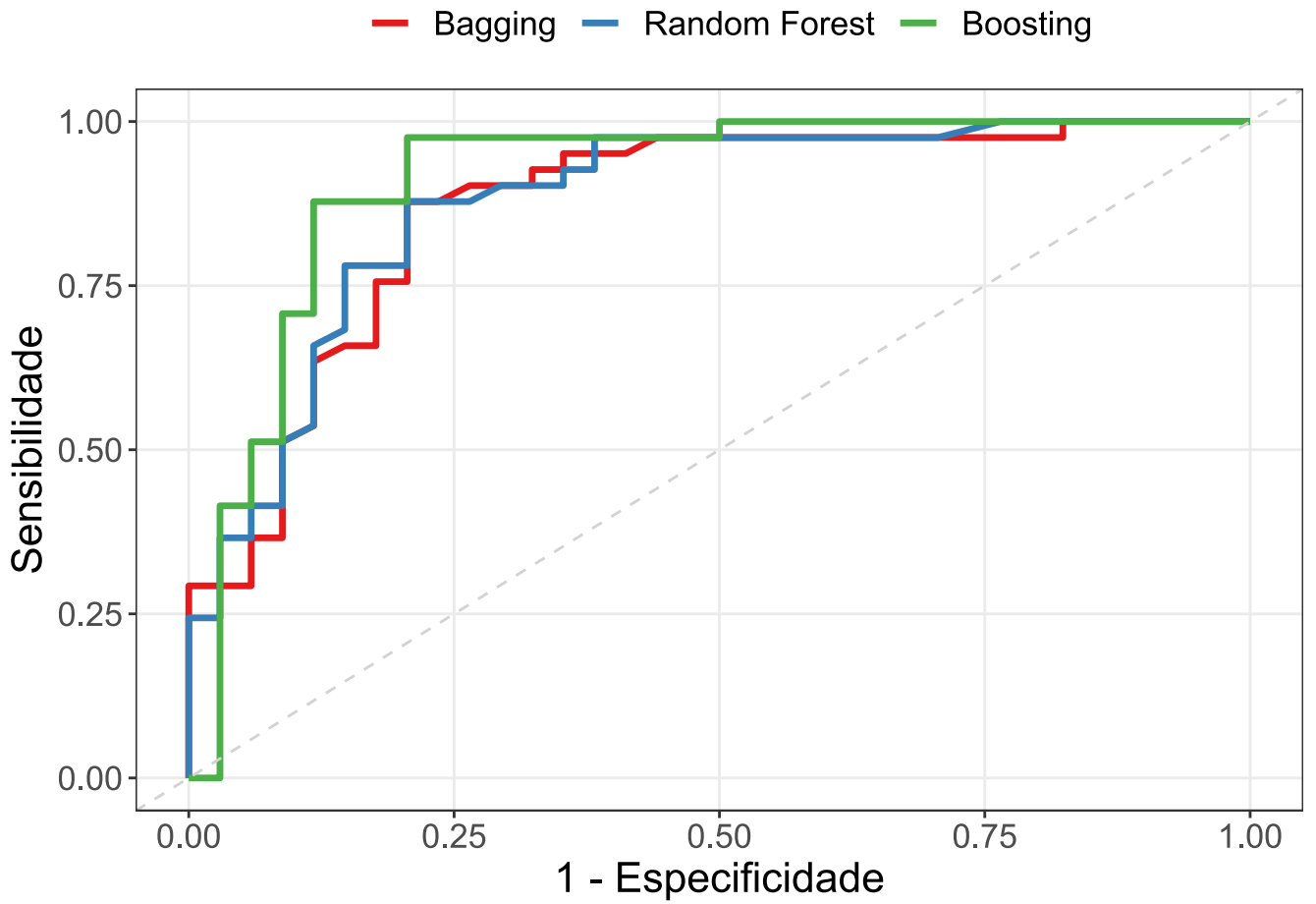
Recursos Computacionais

- Toda análise foi conduzida no software R, versão 3.6.1.
- `randomForest`: métodos Bagging e Random Forest.
- `gbm`: algoritmo AdaBoost.M1.
- `rsample`: validação cruzada.
- `ROCR`: curva ROC.

Número de Árvores



Desempenho na Amostra Teste



Matrizes de Confusão

Bagging

	Predito		
Observado	**Não doente**	**Doente**	**Total**
Não doente	27 (36.00%)	7 (9.33%)	34 (45.33%)
Doente	8 (10.67%)	33 (44.00%)	41 (54.67%)
Total	35 (46.67%)	40 (53.33%)	75 (100.00%)

Random Forest

	Predito		
Observado	**Não doente**	**Doente**	**Total**
Não doente	27 (36.00%)	7 (9.33%)	34 (45.33%)
Doente	7 (9.33%)	34 (45.33%)	41 (54.67%)
Total	34 (45.33%)	41 (54.67%)	75 (100.00%)

Boosting

	Predito		
Observado	**Não doente**	**Doente**	**Total**
Não doente	27 (36.00%)	7 (9.33%)	34 (45.33%)
Doente	4 (5.33%)	37 (49.33%)	41 (54.67%)
Total	31 (41.33%)	44 (58.67%)	75 (100.00%)

Comparação dos Classificadores

Modelo	`$\mathrm{Err}_{\mathcal{T}}$`	**AUC**	**Sens.**	**Espec.**
Logística	0.1333	0.9067	0.8529	0.8780
LDA	0.3200	0.7120	0.5294	0.8049
QDA	0.3333	0.7109	0.6176	0.7073
Bagging	0.2000	0.8702	0.7941	0.8049
Random Forest	0.1867	0.8784	0.7941	0.8293
Boosting	0.1467	0.9125	0.7941	0.9024

Considerações

- Regressão logística e Boosting apresentaram boa performance preditiva.
- Até que ponto vale a pena perder a interpretabilidade da regressão logística?
- O quão viável é utilizar Boosting ou Random Forest como métodos para tomada decisão?

Referências

- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining Inference and Prediction.** 2nd. ed. Springer, 2009.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning.** Springer, 2013.
- MURPHY, K. P. **Machine Learning: A Probabilistic Perspective.** MIT Press, 2012.