PLN para a Ciência Política e Políticas Públicas Públicas

Professora: Lorena Barberia

Semana 11

Universidade de São Paulo

Tópicos da Aula

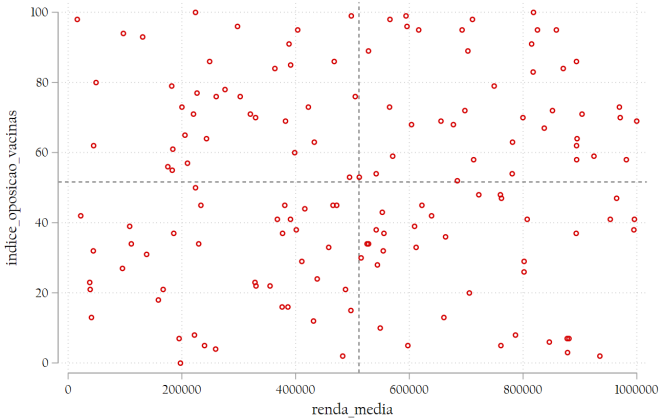
Universidade de São Paulo

Decision Trees

- Algoritmo não-linear.
- Regras para estratificar e segmentar.

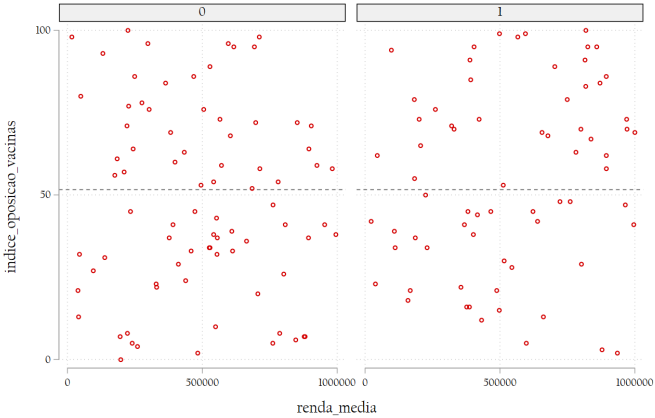
Universidade de São Paulo

Visão Geral de Decision Trees



Universidade de São Paulo

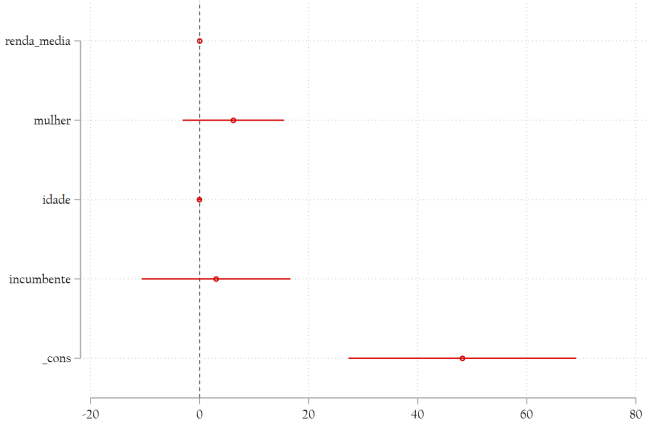
Decision Trees



Universidade de São Paulo

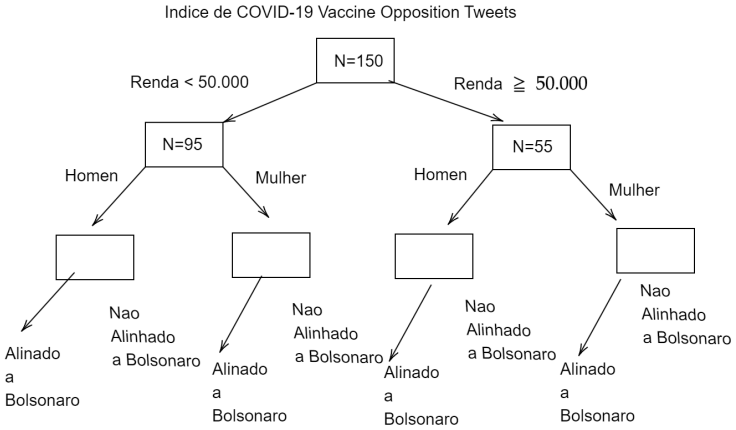
Os Limites do Modelo Linear

Índice = f(mulher, idade, incumbente)



Universidade de São Paulo

Decision Trees



Universidade de São Paulo

Entropy

- Como selecionamos o ponto em que segmentar/estratificar os dados?
- Nos modelos de classificação, e tal como o índice de Gini nos modelos quantitativos, a **entropia** assumirá um pequeno valor se o m-ésimo nó for puro.

$$E = -\sum \widehat{\rho_{mk}}\, log\widehat{\rho_{mk}} \tag{1}$$

ρ_{mk} represents the **proportion** of training observations in the **mth** region that are from the **kth** class.

Universidade de São Paulo

Decision Trees

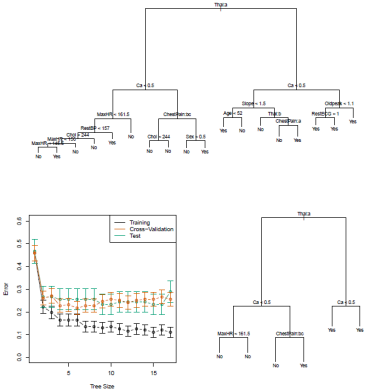


FIGURE 8.6. `Heart` *data.* Top: *The unpruned tree.* Bottom Left: *Cross-validation error, training, and test error, for different sizes of the pruned tree.* Bottom Right: *The pruned tree corresponding to the minimal cross-validation error.*

Bagging, Random Forest e Boosting

- Decision Trees sofrem de variância alta.
- Bootstrap aggregation, ou bagging, random forest, e boosting são técnicas para reduzir a variância.

Universidade de São Paulo

Bagging e Random Forest

- No caso de Bagging, construímos uma floresta de árvores de decisão baseadas em amostras de treinamento usando a técnica de *bootstrap*. As árvores são selecionadas independentemente em amostras aleatórias das observações.
- Em random forest, tambem construímos uma floresta de árvores de decisão baseadas em amostras de treinamento com bootstrap. Diferente de bagging, ao construir estes árvores de decisão, cada vez que uma divisão em uma árvore é considerada, uma amostra aleatória de m preditores são escolhidos como candidatos.

Universidade de São Paulo

Bagging and Random Forest

OOB= Out-of-bag observations

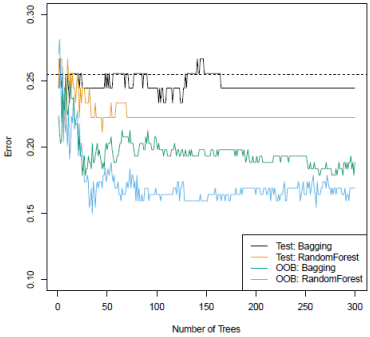


FIGURE 8.8. *Bagging and random forest results for the* `Heart` *data. The test error (black and orange) is shown as a function of B, the number of bootstrapped training sets used. Random forests were applied with* $m = \sqrt{p}$. *The dashed line indicates the test error resulting from a single classification tree. The green and blue traces show the OOB error, which in this case is — by chance — considerably*

Avaliação do Modelo

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Table 1. Confusion matrix with advanced classification metrics

Universidade de São Paulo

Laboratório 10

Universidade de São Paulo