STA 529 2.0 Data Mining

Dr Thiyanga S. Talagala

Random Forests

Lecture 4

Random Forests vs AdaBoost

Category	Random Forest	AdaBoost
Tree size	Fully grown	Decision stump
Tree construction	Independent	Each stump is made by taking the previous stump's mistakes into account
Say for each tree	Equal	Some stumps get more say
	parallel	sequential

Data

	smoking	family_history	height	cancer
1	Yes	Yes	5.3	Yes
2	No	Yes	6.0	Yes
3	Yes	No	5.2	Yes
4	Yes	Yes	5.0	Yes
5	No	Yes	5.9	No
6	No	Yes	4.7	No
7	Yes	No	5.2	No
8	Yes	Yes	5.4	No

Creating the first decision stump

1) Initial weights

	smoking	family_history	height	cancer	Ini_weights
1	Yes	Yes	5.3	Yes	0.125
2	No	Yes	6.0	Yes	0.125
3	Yes	No	5.2	Yes	0.125
4	Yes	Yes	5.0	Yes	0.125
5	No	Yes	5.9	No	0.125
6	No	Yes	4.7	No	0.125
7	Yes	No	5.2	No	0.125
8	Yes	Yes	5.4	No	0.125

Decision stump (In-class)

Smoking ?

Height? Splitting point?

Family_history?

Decision stump

Your turn

Compute node Gini coefficient for Smoking and Family History and Height > 5.1

Frist decision stump

In-class

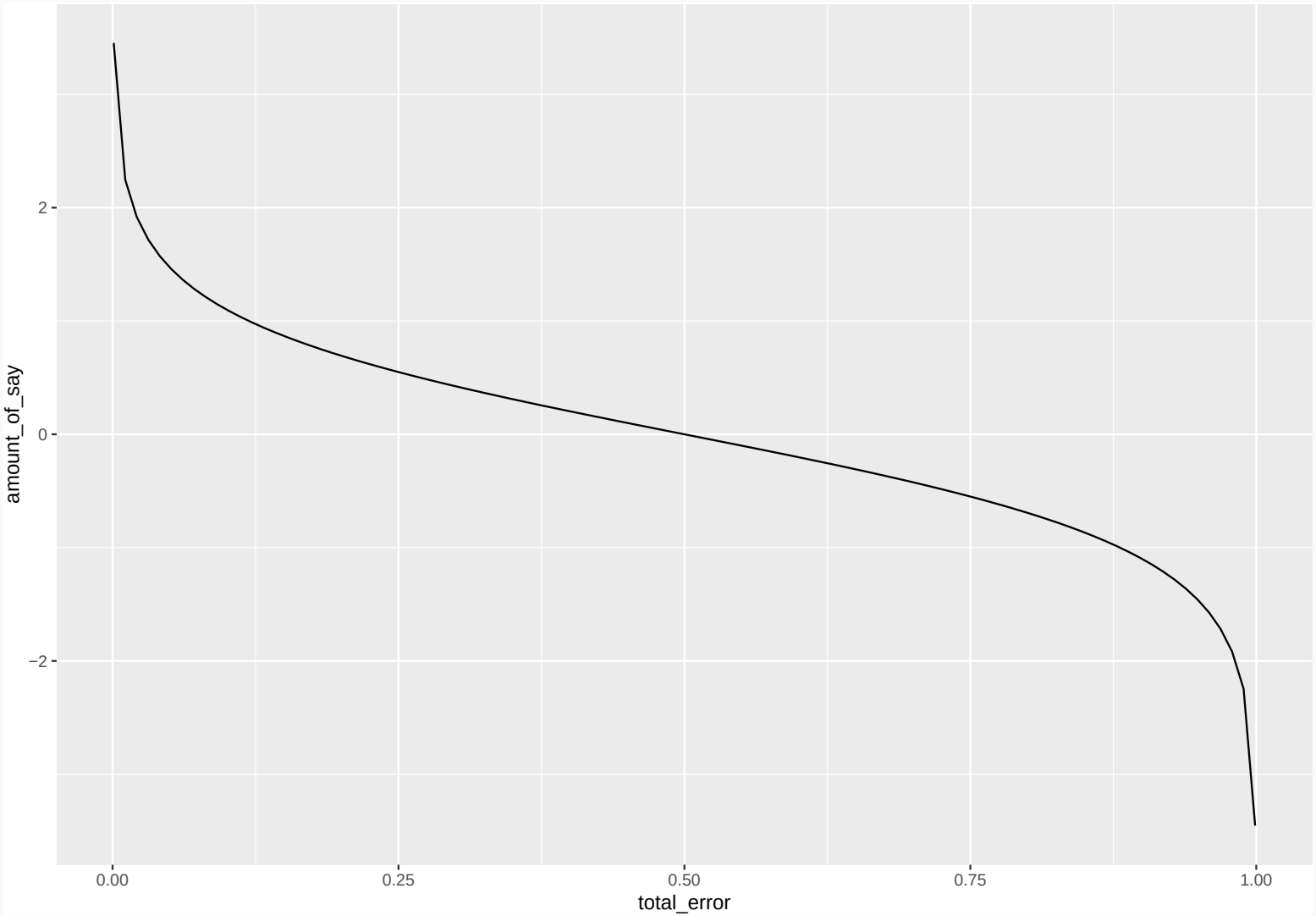
Compute total error.

Amount of say

$$\text{Amount of say} = \frac{1}{2} log\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$

Compute the amount of say for the first decision stump.

Amount of say vs Total error

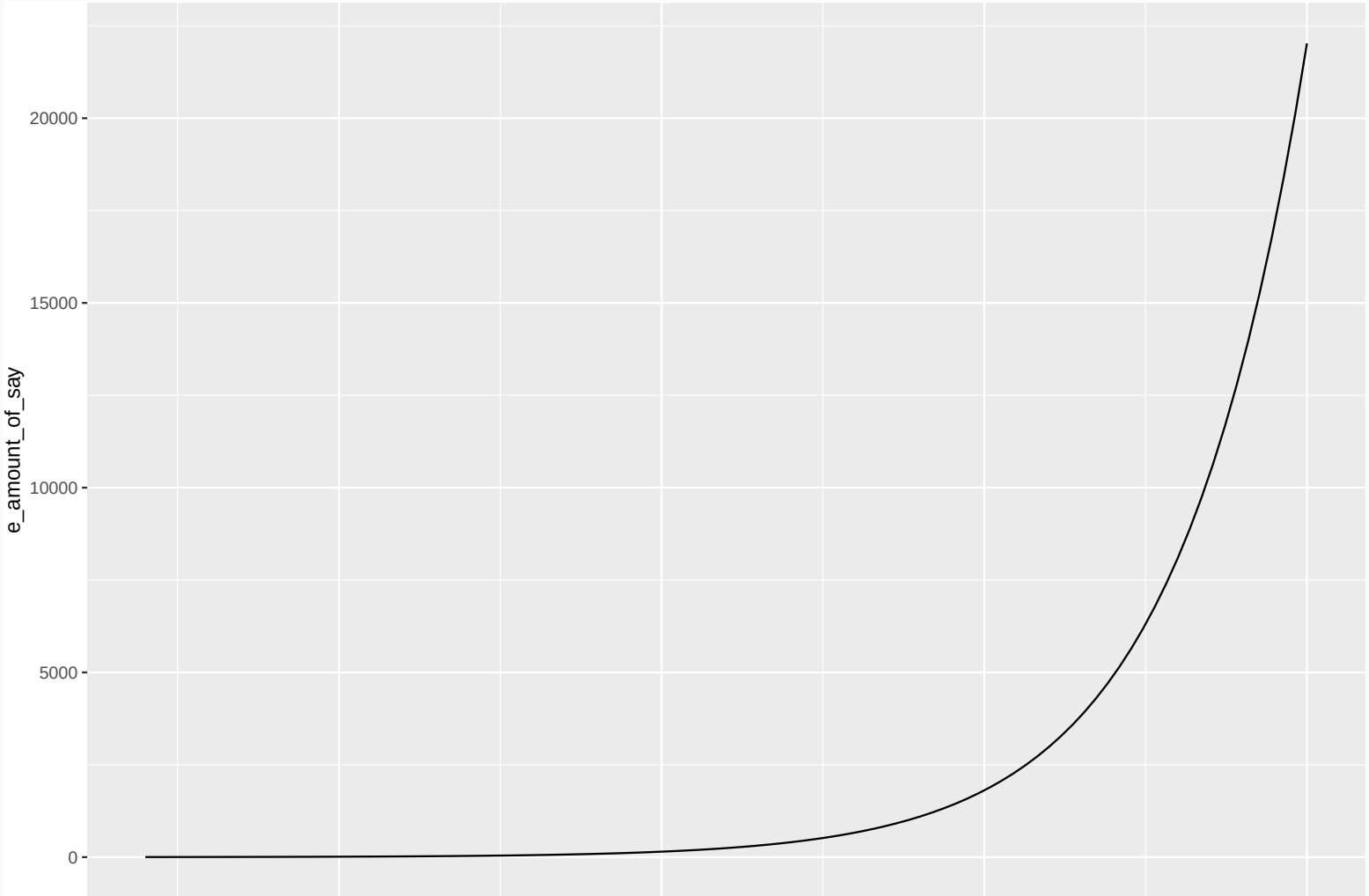


Building the next decision stump

	smoking	family_history	height	cancer	Ini_weights
1	Yes	Yes	5.3	Yes	0.125
2	No	Yes	6.0	Yes	0.125
3	Yes	No	5.2	Yes	0.125
4	Yes	Yes	5.0	Yes	0.125
5	No	Yes	5.9	No	0.125
6	No	Yes	4.7	No	0.125
7	Yes	No	5.2	No	0.125
8	Yes	Yes	5.4	No	0.125

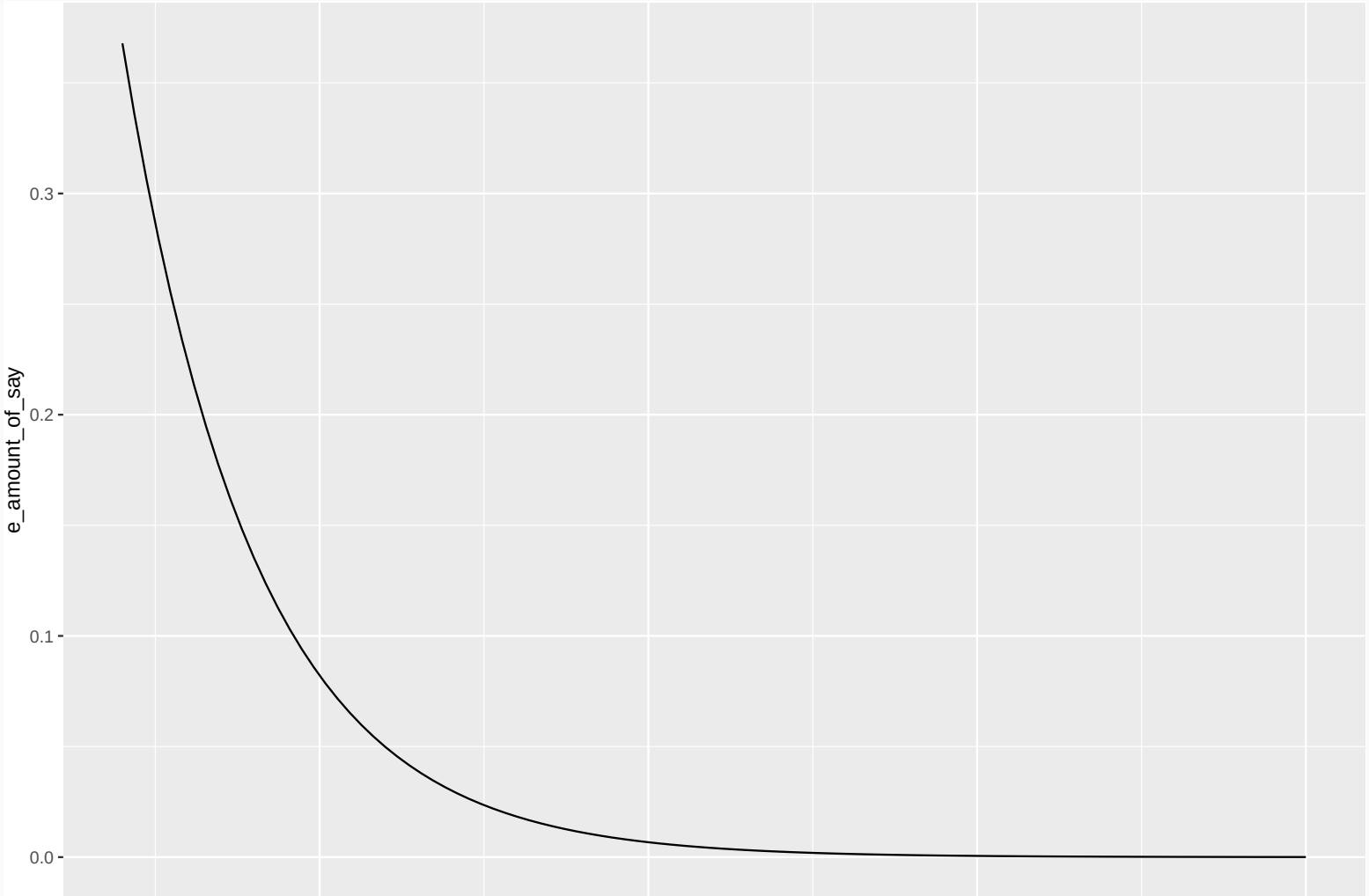
New observation weights for incorrect classifications

$$\text{New observation weight} = \text{initial weight} \times e^{\text{Amount of say}}$$



New observation weights for correct classifications

$$\text{New observation weight} = \text{initial weight} \times e^{\text{-Amount of say}}$$



Initial weights + new weights

	ini_weights	new_weights
1	0.125	0.05
2	0.125	0.05
3	0.125	0.05
4	0.125	0.33
5	0.125	0.05
6	0.125	0.05
7	0.125	0.05
8	0.125	0.05

Initial weights + new weights + normalized weights

	ini_weights	new_weights	normalized_weights
1	0.125	0.05	0.07352941
2	0.125	0.05	0.07352941
3	0.125	0.05	0.07352941
4	0.125	0.33	0.48529412
5	0.125	0.05	0.07352941
6	0.125	0.05	0.07352941
7	0.125	0.05	0.07352941
8	0.125	0.05	0.07352941

Create the next decision stump - inclass

Adaboost final decision - inclass

Adaboost vs Gradient Boosting Algorithm

Category	AdaBoost	Gradient Boosting
Tree size	Decision stump	Statrts by making a single leaf and then grow trees
Tree construction	Up weights the observations misclassified before	identify observations by large residuals computed in the previous iteration
Say for each tree	Not Equal	Equal
Tree construction	sequential	sequential

Gradient Boosting - Regression

Dataset

	colour	gender	tusk_length	bmi
1	blue	M	1.6	2.3
2	brown	F	1.6	2.4
3	blue	F	1.5	3.2
4	black	M	1.8	3.4
5	brown	M	1.5	3.5
6	blue	F	1.4	2.1

Initial guess

Compute residuals

Predict BMI

Continue building trees

Final prediction