CS105 Lab 10: Data Mining II

Brian Borucki – bborucki@bu.edu

R0

Say I'm trying to classify whether people are coffee drinkers or not

ID	State	Color	Coffee?
1	MA	Red	Yes
2	NH	Red	No
3	MA	Magenta	Yes

- R0 just ignores all input attributes and just looks at the output

R0

Say I'm trying to classify whether people are coffee drinkers or not

ID	State	Color	Coffee?
-	-	-	Yes
-	-	-	No
-	-	-	Yes

- R0 just ignores all input attributes and just looks at the output
- Which is the most common output?
- Always predict the most common output

R1

- We can do better than R0 … R1!

ID	State	Color	Coffee?
1	MA	Red	Yes
2	NH	Red	No
3	MA	Magenta	Yes

- For each value of each input attribute, find the most frequent class and create a rule
- Choose the rules with the higest accuracy

R1

- We can do better than R0 … R1!

ID	State	Color	Coffee?
1	MA	Red	Yes
2	NH	Red	No
3	MA	Magenta	Yes

- State: MA → Yes (2 / 2)
  NH → No (1/1)

  Overall Accuracy = 3/3

R1

- We can do better than R0 … R1!

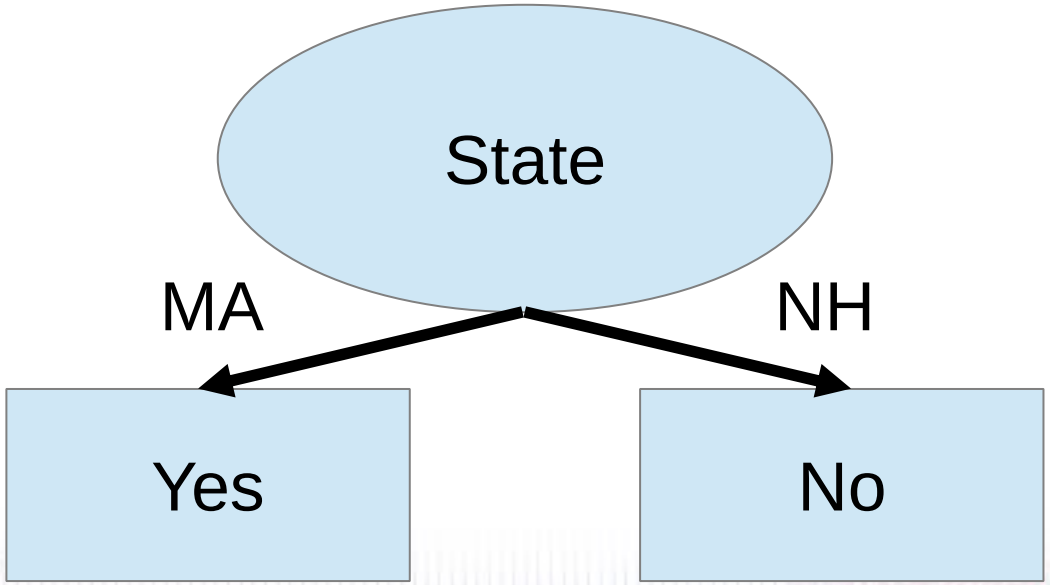
ID	State	Color	Coffee?
1	MA	Red	Yes
2	NH	Red	No
3	MA	Magenta	Yes

- Color: Magenta → Yes (1/1)
  Red → No (1/2)
  Overall Accuracy = 2/3

R1

- State had an overall accuracy of 100%
- Color had an overall accuracy of 66.66%

- So we create the following classifier:



Building a Tree

ID	State	Color	Coffee?
1	MA	Red	Yes
2	NH	Red	No
3	MA	Magenta	No
4	ME	Red	Yes
5	ME	Magenta	No

Color : Red → Yes (2/3)

Magenta → No (2/2)

Overall Acc = 4/5
Goodness = 4/5

State : MA → Yes (1/2)

NH → No (1/1)

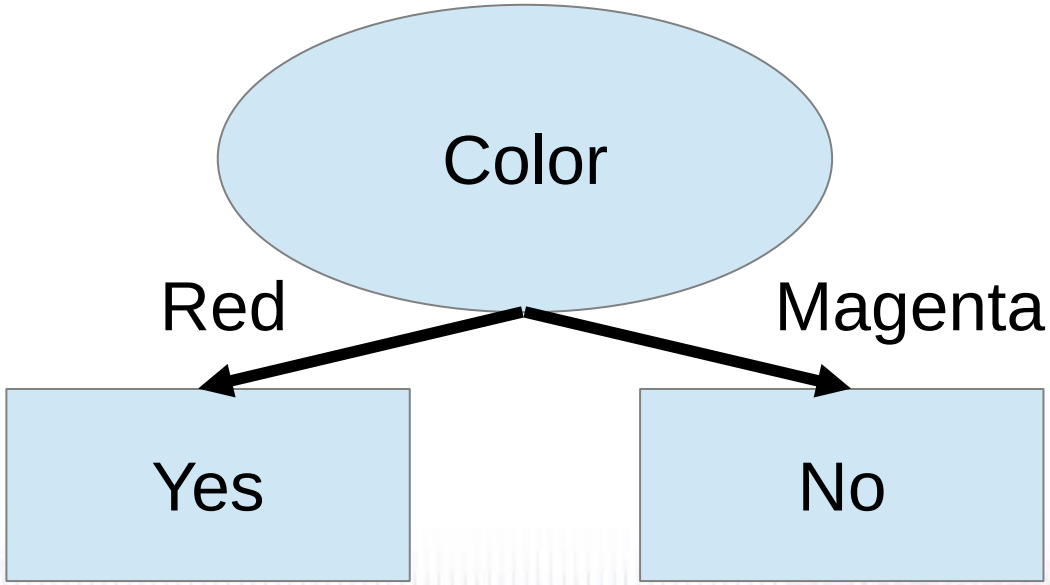
ME → Yes (1/2)

Overall Acc = 3/5
Goodness = 3/10

Building a Trees

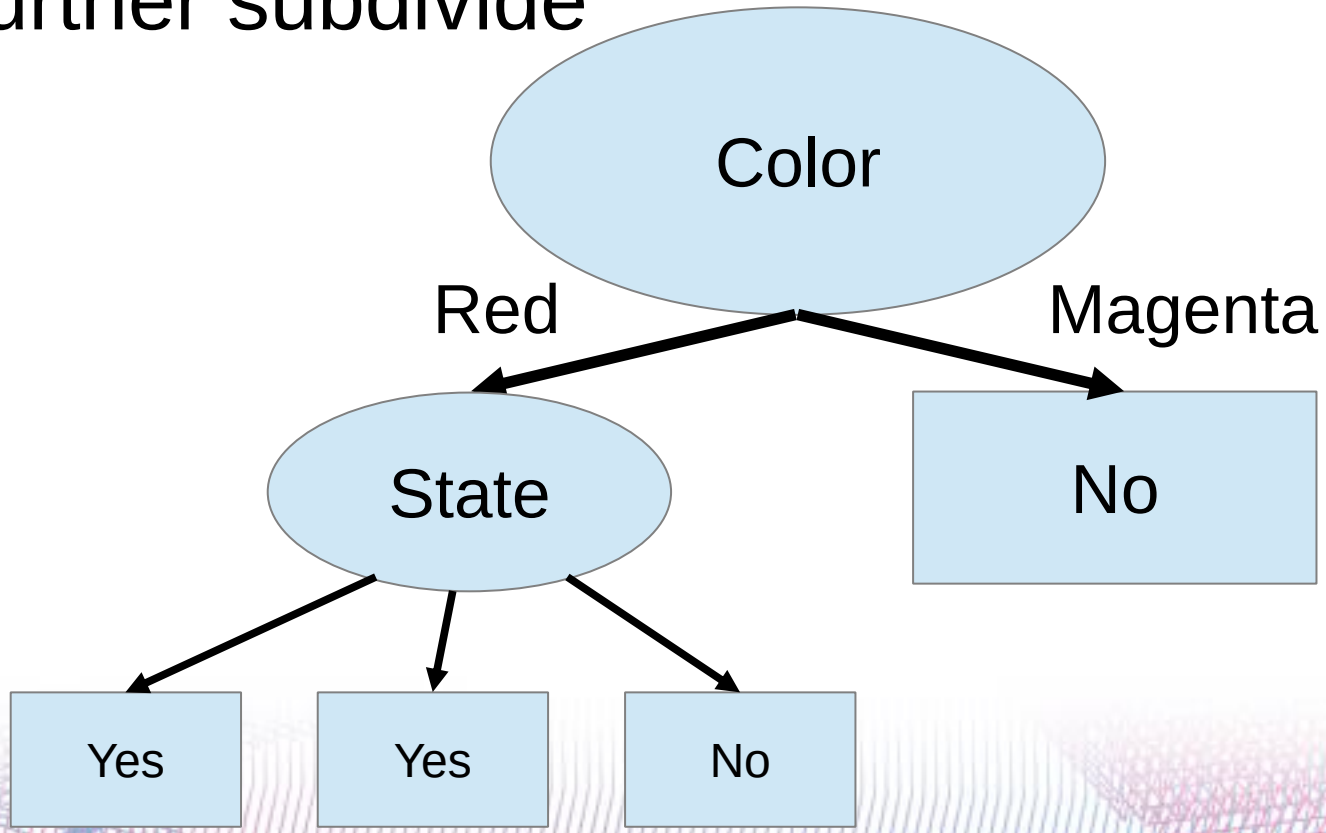
- State had an overall goodness of 30%
- Color had an overall goodness of 90%

- So we create the following classifier:



Building a Trees

- The “Red” subgroup still had some inaccuracies, can go back, recalculate goodness scores for remaining attributes, and further subdivide



Wait … what?

- We're just trained a model which now determines whether someone is a coffee drinker based on their favorite color....
- How does this make sense?

- There isn't enough training data and there aren't enough input attributes to 'filter' out the color attribute

Awesome Segue

- Choosing a data set is very important
- You need to have enough information to filter out noise (irrelevant information/anomalies)
- You need a problem that fits your dataset and a dataset that fits your problem
  - I can't model a generic person if I only have data on people from a specific place
  - If I only have data on people from a specific place, I shouldn't choose to try and model a generic person

Lab

- Nothing to submit, but I recommend you work through the questions and ask for help
- You may also use lab time to work on your proposals
- Questions?