Decision trees and random forest

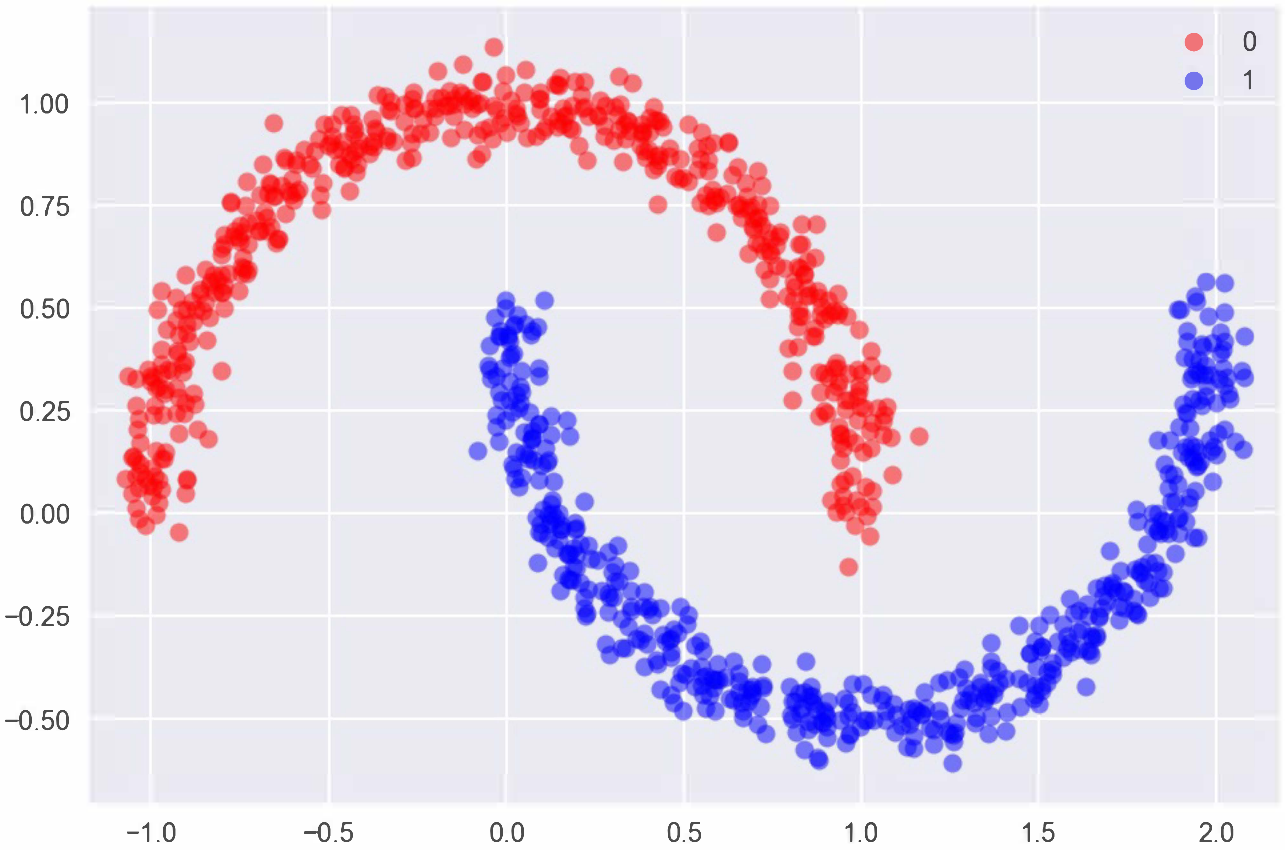
Last time: generalized linear model (GLM)

$$\theta = a_0 + a_1 x_1 + ... + a_n x_n$$

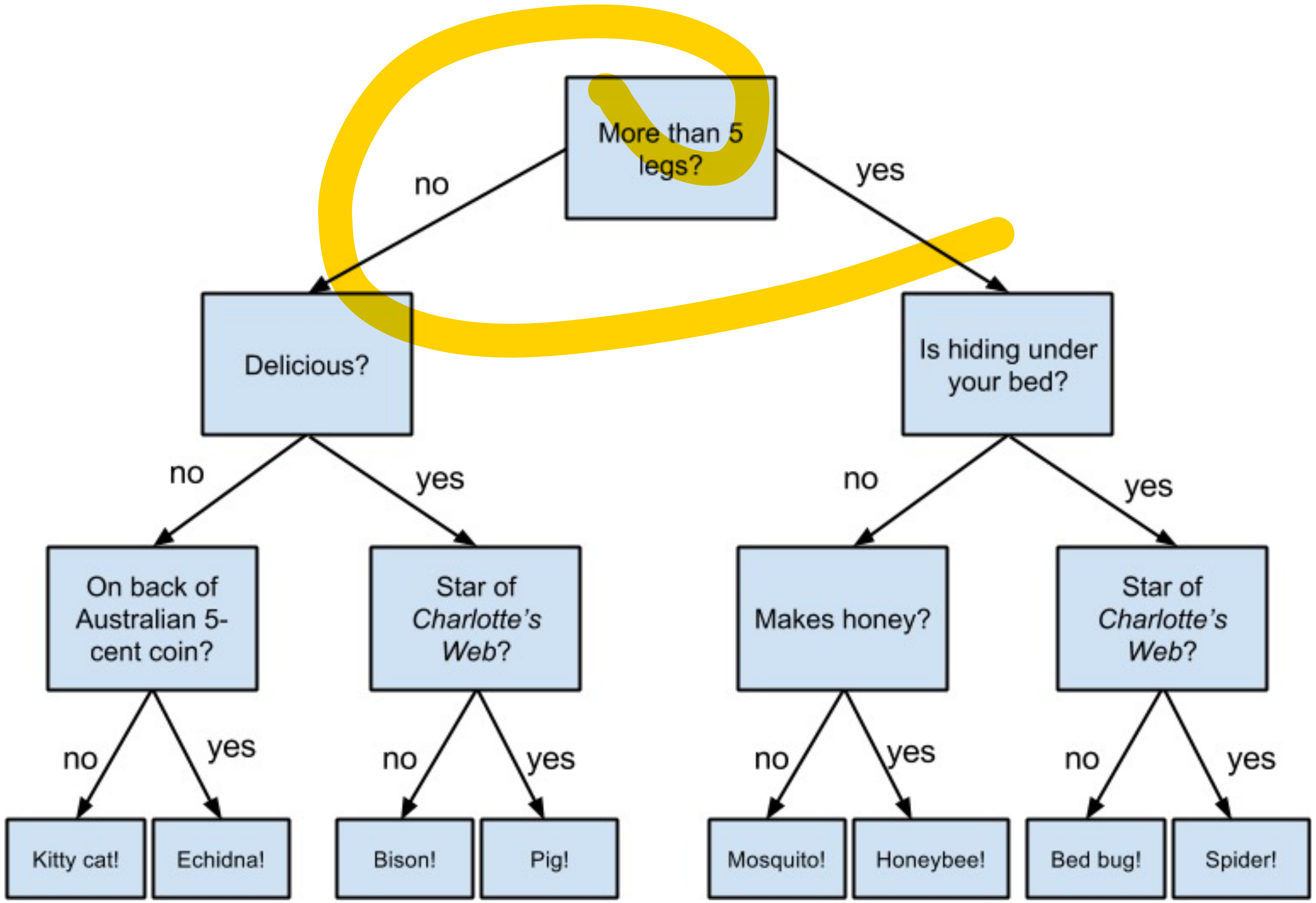
$$y \sim g(\theta)$$

- x are independent variables or “predictors”
- a are model coefficients (linear)
- theta is a parameter of the GLM (for regular linear regression g is the identity)

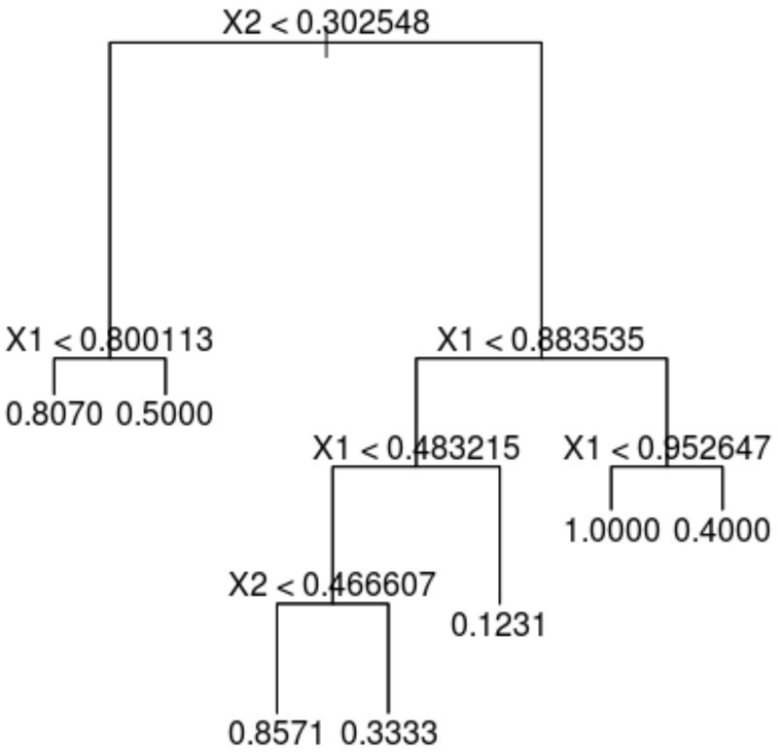
But your data might be non–linear. We need a decision boundary to separate this data.

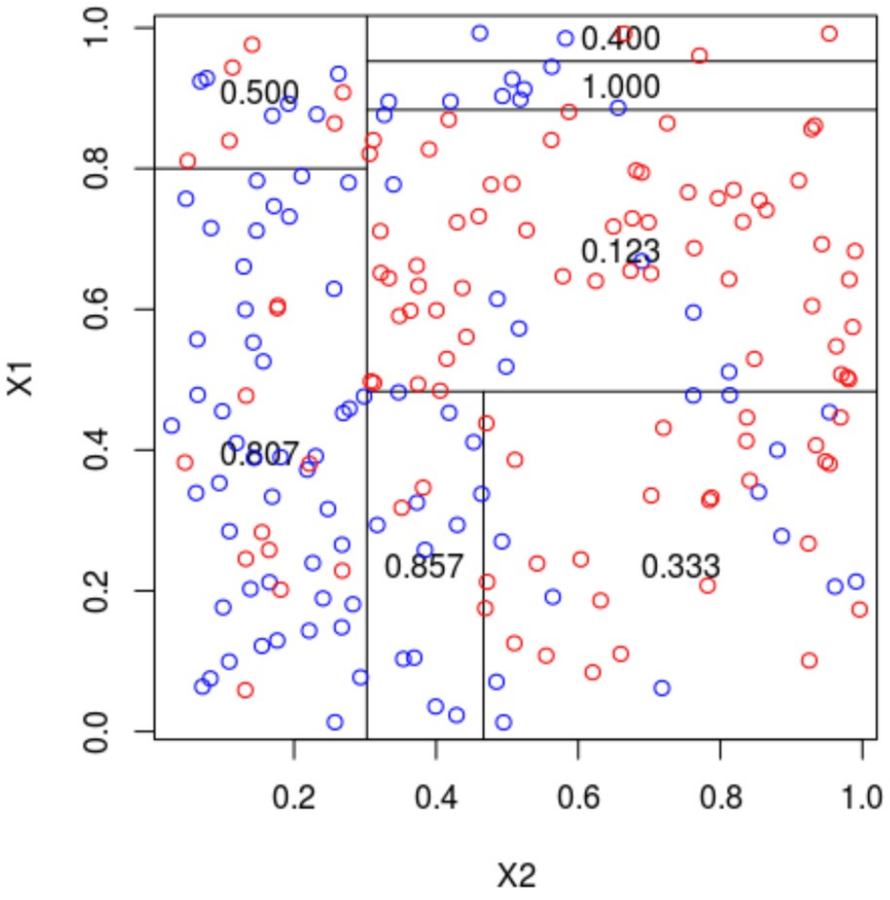


decision tree



decision tree





how to grow a decision tree

- pick the split (variable and cutoff) that best separates the data at current node
- for classification, best is typically determined by gini impurity
- stop growing tree when all nodes are “pure” (all one label) or node contains a low number of data points

gini impurity

- suppose a particular split separates the data into two groups:
- group A: 1 (12) and 0 (5)   group B: 1 (3) and 0 (11)
- GI_A=1-(12/17)^2-(5/17)^2
- GI_B=1-(3/14)^2-(11/14)^2
- GI=GI_A+GI_B   want to be as small as possible

bootstrap

- resample (with replacement) a dataset of the same size as original data
- repeat, compute a statistic on each sampled dataset
- analyze how the statistic changes over different sampled datasets to quantify uncertainty

bootstrap aggregation (bagging)

- resample (with replacement) a dataset of the same size as original data
- repeat, compute a *model* on each sampled dataset
- *aggregate* the model over the sampled datasets to reduce overfitting or quantify uncertainty

random forests:
bagging decision trees

- a single decision tree fully grown overfits the data (the prediction will be perfect on training data)
- random forests grow many trees over randomly sampled subsets of the data and then aggregate (average or majority vote)
- to reduce variance further, choose splits at each node from a random subset of predictor variables

R packages

- decision tree: **rpart, tree, party**
- random forests: **randomForest**
- cross validation and parameter tuning: **caret**

Review for exam

Exam will cover

- sections 1-3 in r4ds book
  - Explore
  - Wrangle
  - Program
- basic statistics

Explore

- data visualization in ggplot
    - Basic plots like scatter plot, histogram, time series, bar chart
    - aesthetics
        - Color plot by data group
        - Change size, shape of markers
        - facet_wrap to have multiple plots split out by a variable
        - Editing x, y axes, title

Wrangle (data manipulation)

- Tidy
  - pipes
  - filter
  - group_by
  - summarise
  - merging/joining data together (join)

Program

- Data structures in R (list, array, tibble)
- Functions in R
- Loops, logical statements in R

Statistics

- central limit theorem and application to quantifying uncertainty (CI, p-value)
- Bootstrap
- Continuous vs discrete random variable
- Linear and logistic regression