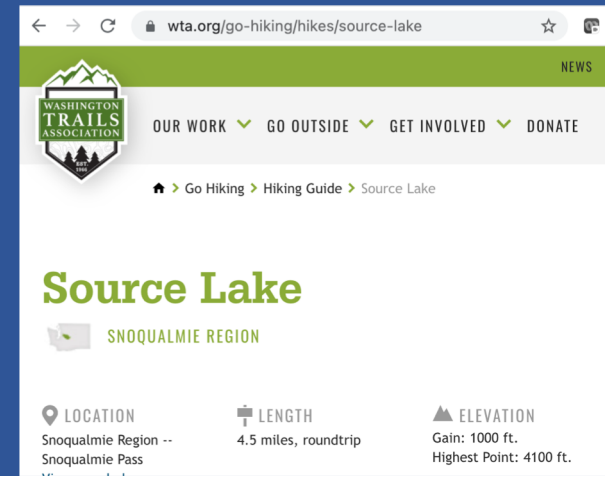
All about

PREDICTIVE MODELING

Oct 11th 2021



Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Canlis	American	$$$$	Y	Y	
Lark	Plant based	$$$$	N	Y	
Nue	Vegetarian	$$	N	N	
Kati Thai	Vegan	$	Y	Y	

MODEL

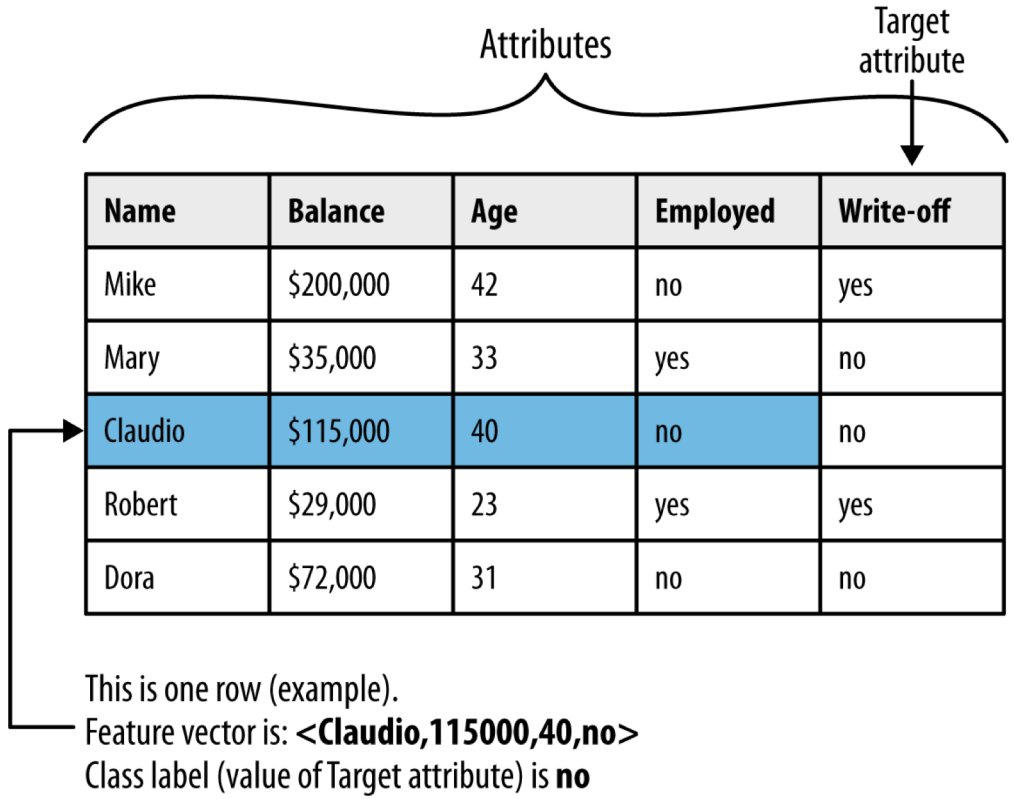
An abstraction of reality

Group Discussion

- 1. What according to you **is Data Driven Decision making (DDD)?**
- 2. What is the difference between **supervised and unsupervised learning**?
- 3. Briefly explain any 2 types of **supervised learning** algorithms
- 4. Briefly explain any 2 types of **unsupervised learning** algorithms
- 5. I have a dataset of candidates applying to graduate degree programs at various universities in the United States. It has details about their name, undergraduate GPA, test scores, nationality, gender, years of work experience, university they had applied to, and the result of their application (Admit or Reject). Based on the data and results that I already have; I want to predict whether a candidate will receive an Admit or Reject from a certain university. Select which type of machine learning problem it is:
  - Supervised Learning
  - Unsupervised Learning

Terminology

- Supervised learning
- Target
  - Dependent variable
- Attributes
  - Variables , features , predictors, explanatory variables
- Feature vector
- Data set



Models

- Prediction vs forecasting
  - Temporal
- Two types
  - Segmentation
  - Regression
- Descriptive vs predictive model
- Supervised learning
- The creation of models from data is known as model induction.
  - Induction vs deduction

Supervised Segmentation

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Canlis	American	$$$$	Y	Y	
Lark	Plant based	$$$$	N	Y	
Nue	Vegetarian	$$	N	N	
Kati Thai	Vegan	$	Y	Y	

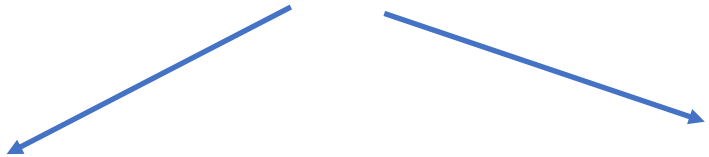
- Which variable ?
- Informative attribute
- Rarely split perfectly
- Type of attribute

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Lark	Plant based	$$$$	N	Y	
Nue	Vegetarian	$$	N	N	
Kati Thai	Vegan	$	Y	Y	

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Canlis	American	$$$$	Y	Y	
Kati Thai	Vegan	$	Y	Y	

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Kati Thai	Vegan	$	Y	Y	

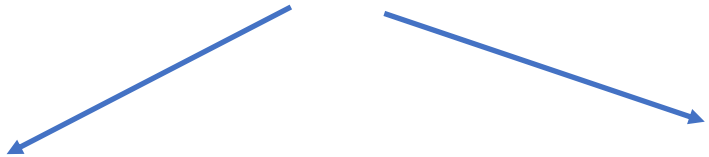
Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Canlis	American	$$$$	Y	Y	
Lark	Plant based	$$$$	N	Y	
Nue	Vegetarian	$$	N	N	
Kati Thai	Vegan	$	Y	Y	



Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Canlis	American	$$$$	Y	Y	
Kati Thai	Vegan	$	Y	Y	

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Lark	Plant based	$$$$	N	Y	
Nue	Vegetarian	$$	N	N	

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Canlis	American	$$$$	Y	Y	
Lark	Plant based	$$$$	N	Y	
Nue	Vegetarian	$$	N	N	
Kati Thai	Vegan	$	Y	Y	



Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	
Lark	Plant based	$$$$	N	Y	
Nue	Vegetarian	$$	N	N	
Kati Thai	Vegan	$	Y	Y	

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Likes
Canlis	American	$$$$	Y	Y	

How do I split

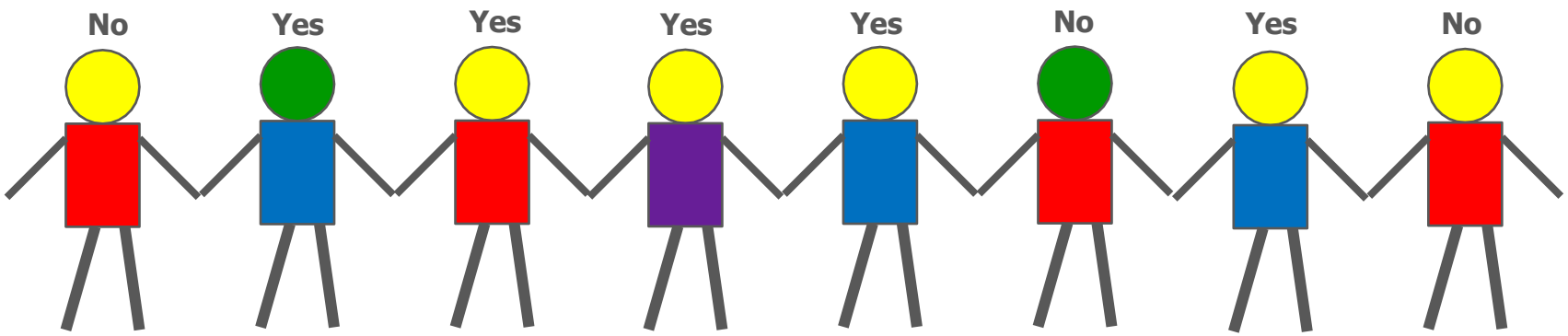
- Entropy
- Information Gain
- Entropy is a measure of disorder that can be applied to a set, such as one of our individual segments
- entropy = - p1 log (p1) - p2 log (p2) - ⋯

INFORMATIVE ATTRIBUTES

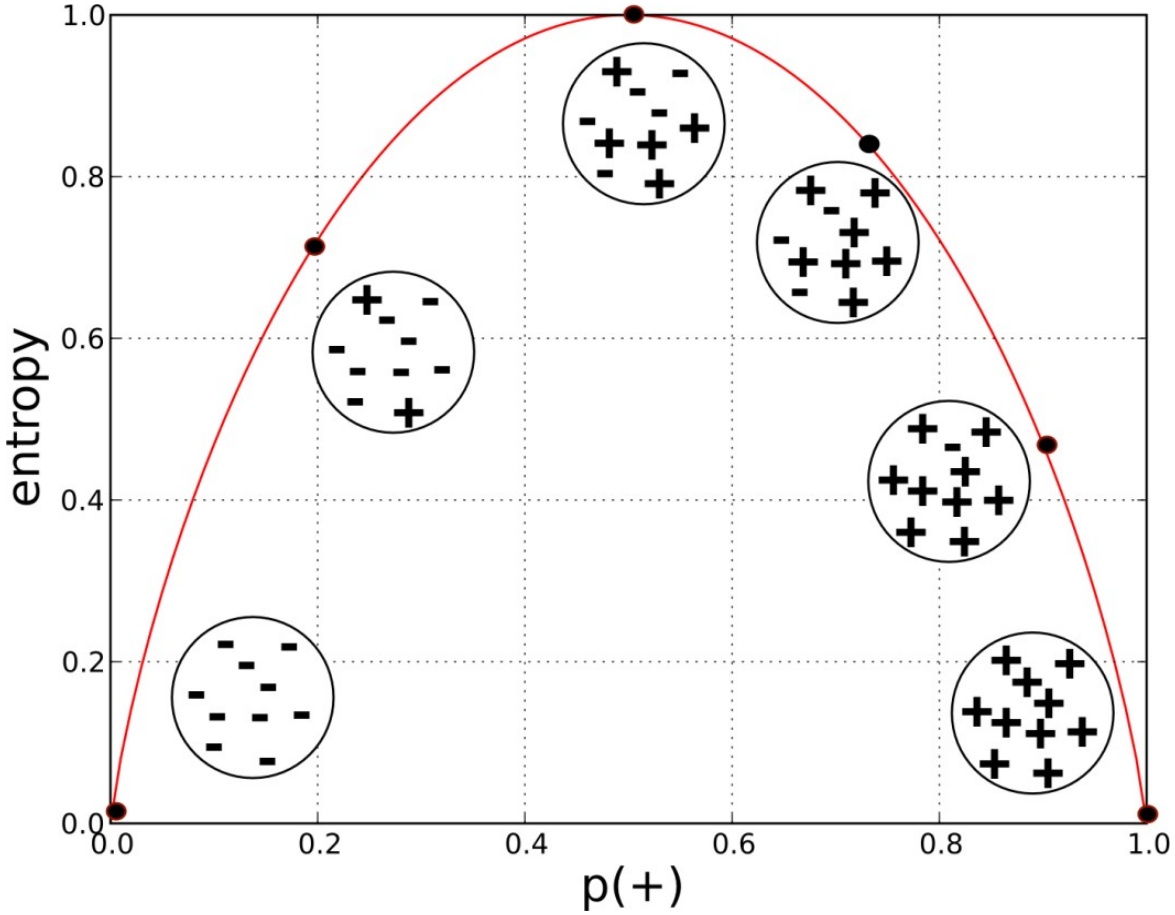
ENTROPY

Selecting Informative Attributes

Objective: Based on customer attributes, partition the customers into subgroups that are less impure – with respect to the class (i.e., such that in each group as many instances as possible belong to the same class)



Informative Attributes

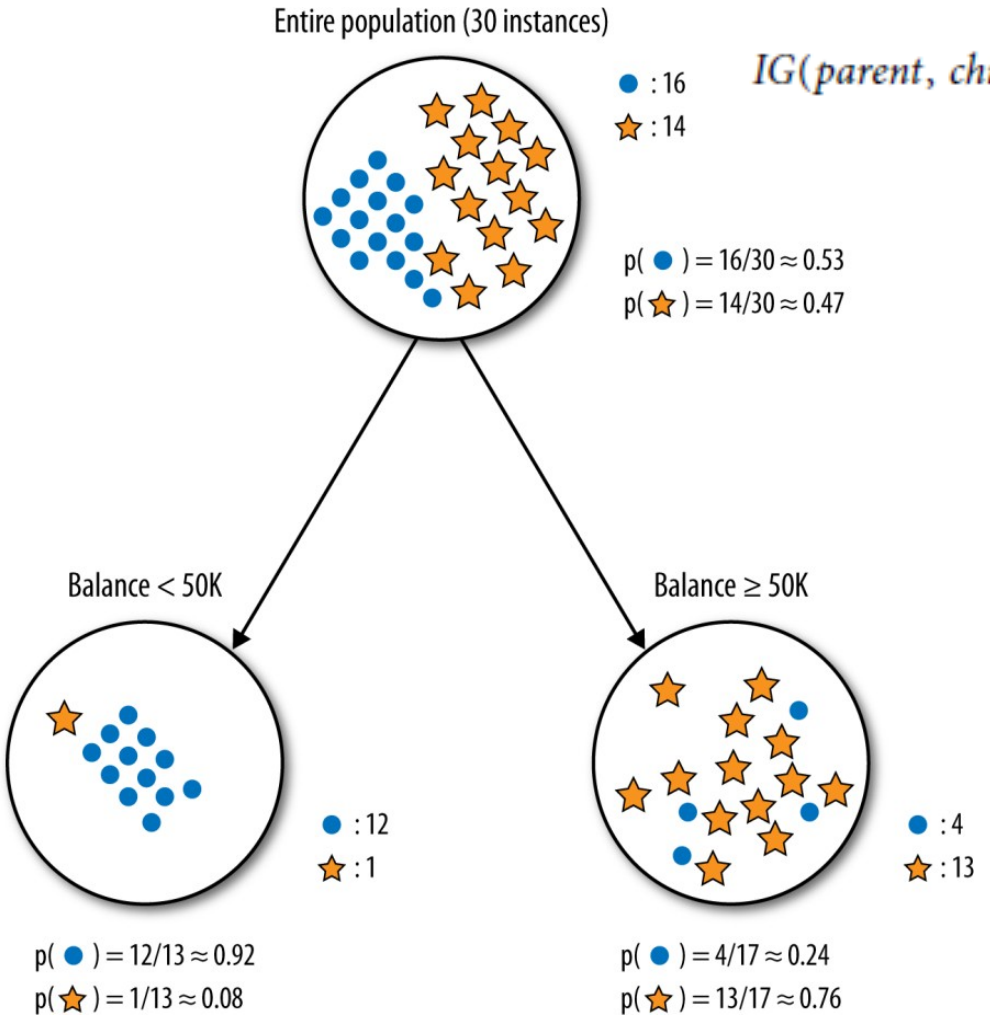


The most common splitting criterion is called **information gain** (IG)

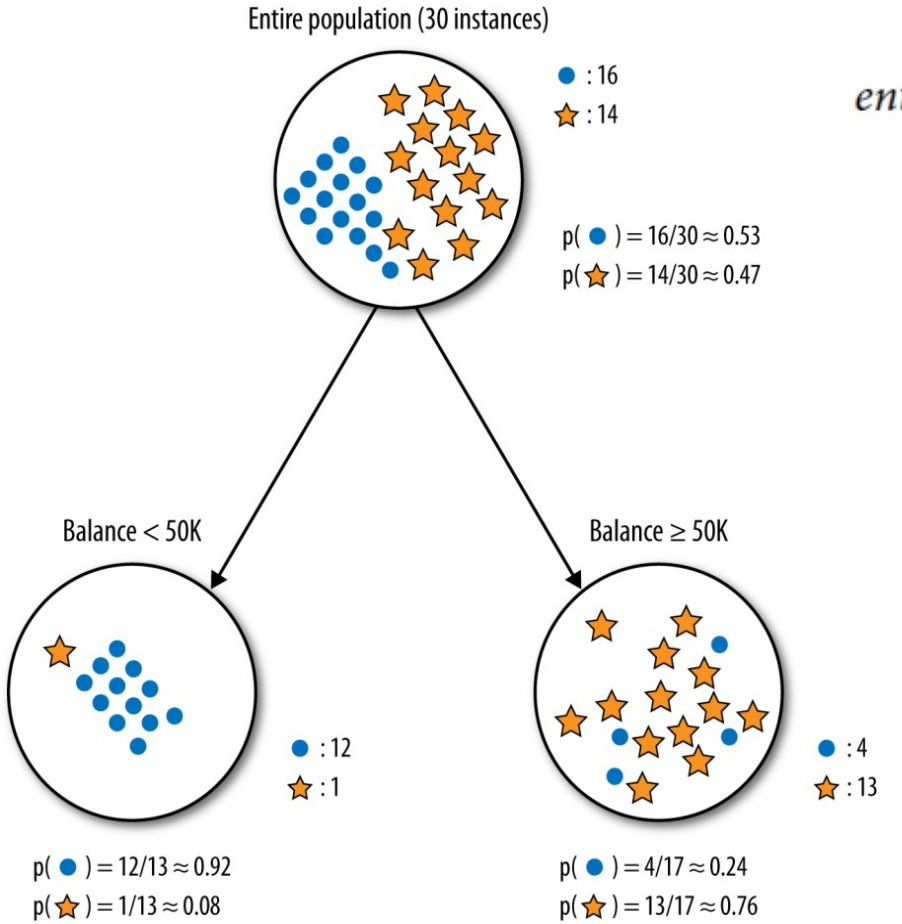
- It is based on a **purity measure** called **entropy**
  - *entropy* = $-p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots$
  - Measures the general disorder of a set

Informative GAIN

- Information gain measures the *change* in entropy due to any amount of new information being added



$$IG(parent, children) = entropy(parent) - [p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + \cdots]$$



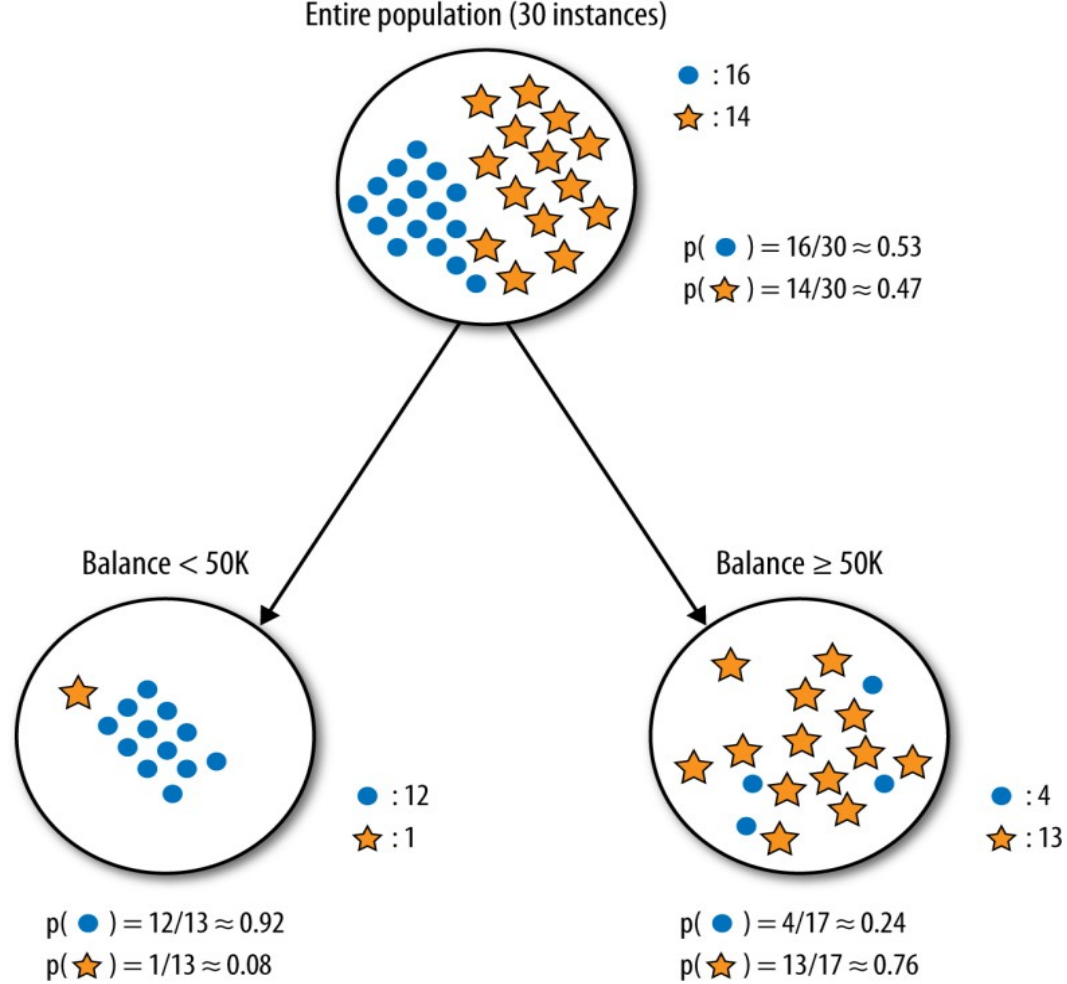
$$\begin{aligned} entropy(parent) &= -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)] \\ &\approx -[0.53 \times -0.9 + 0.47 \times -1.1] \\ &\approx 0.99 \quad \text{(very impure)} \end{aligned}$$

The entropy of the *left* child is:

$$\begin{aligned} entropy(Balance < 50K) &= -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)] \\ &\approx -[0.92 \times (-0.12) + 0.08 \times (-3.7)] \\ &\approx 0.39 \end{aligned}$$

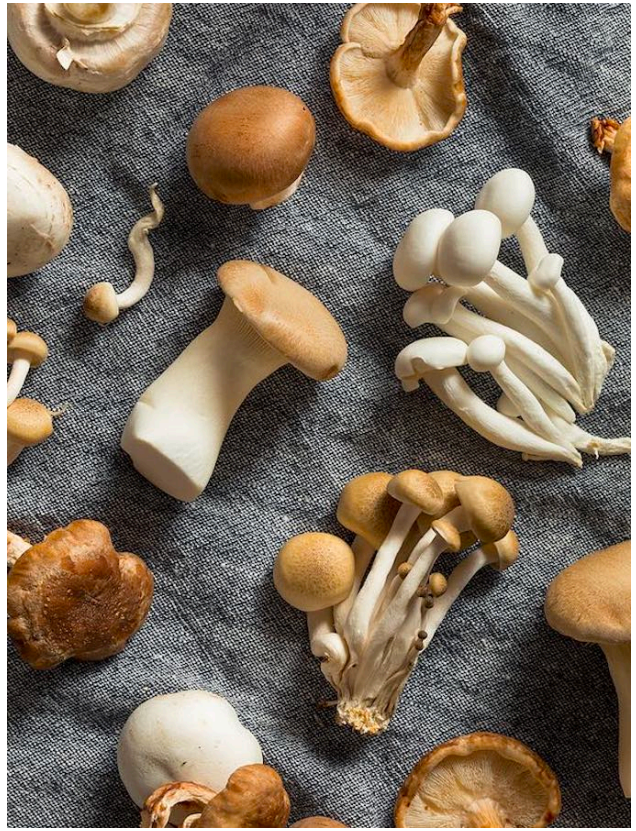
The entropy of the *right* child is:

$$\begin{aligned} entropy(Balance \geq 50K) &= -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)] \\ &\approx -[0.24 \times (-2.1) + 0.76 \times (-0.39)] \end{aligned}$$



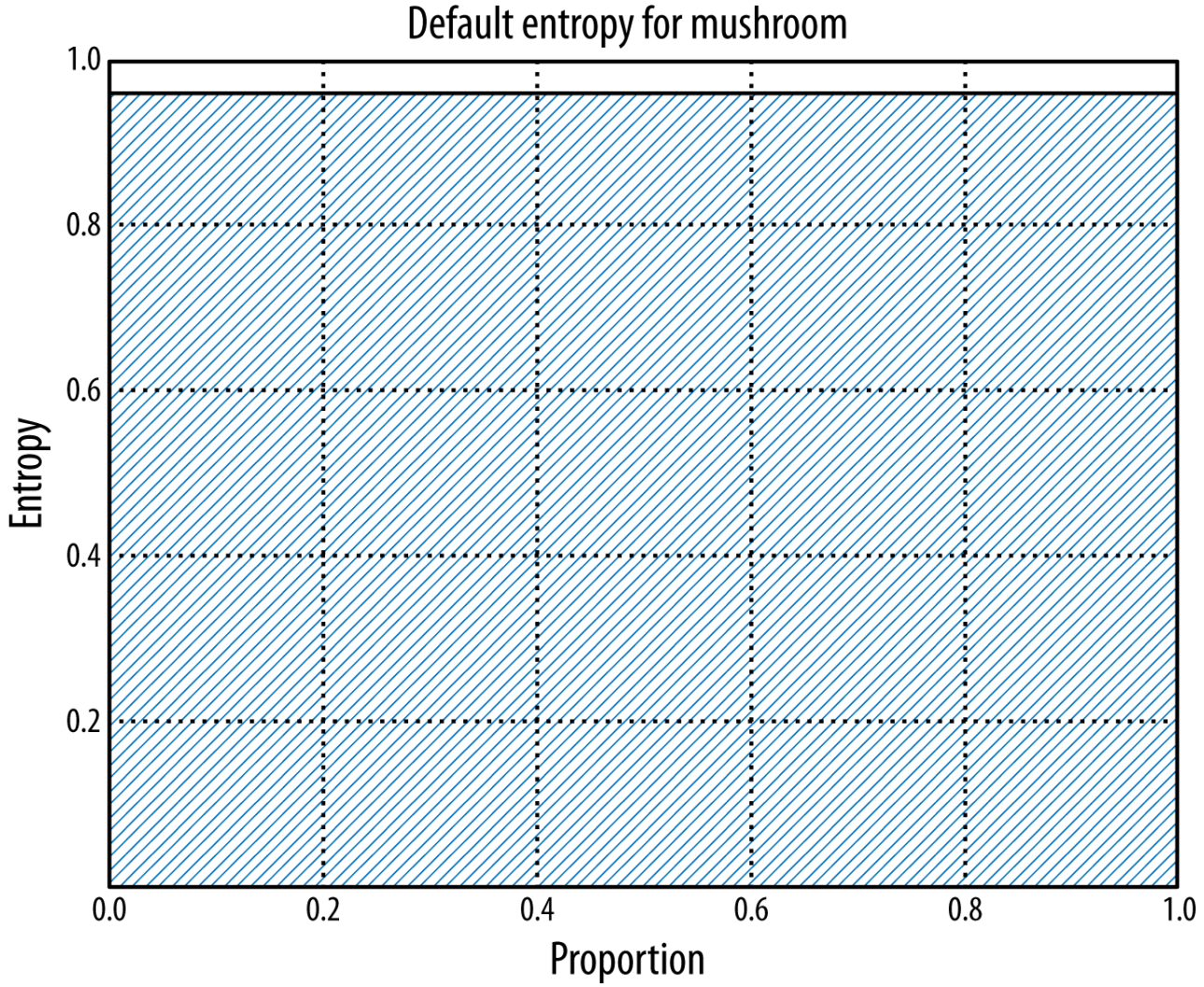
$$
\begin{aligned}
IG &= entropy(parent) - [p(\text{Balance} < 50\text{K}) \times entropy(\text{Balance} < 50\text{K}) \\
&\quad + p(\text{Balance} \geq 50\text{K}) \times entropy(\text{Balance} \geq 50\text{K})] \\
&\approx 0.99 - [0.43 \times 0.39 + 0.57 \times 0.79] \\
&\approx 0.37
\end{aligned}
$$

Example: Attribution Selection with Information Gain

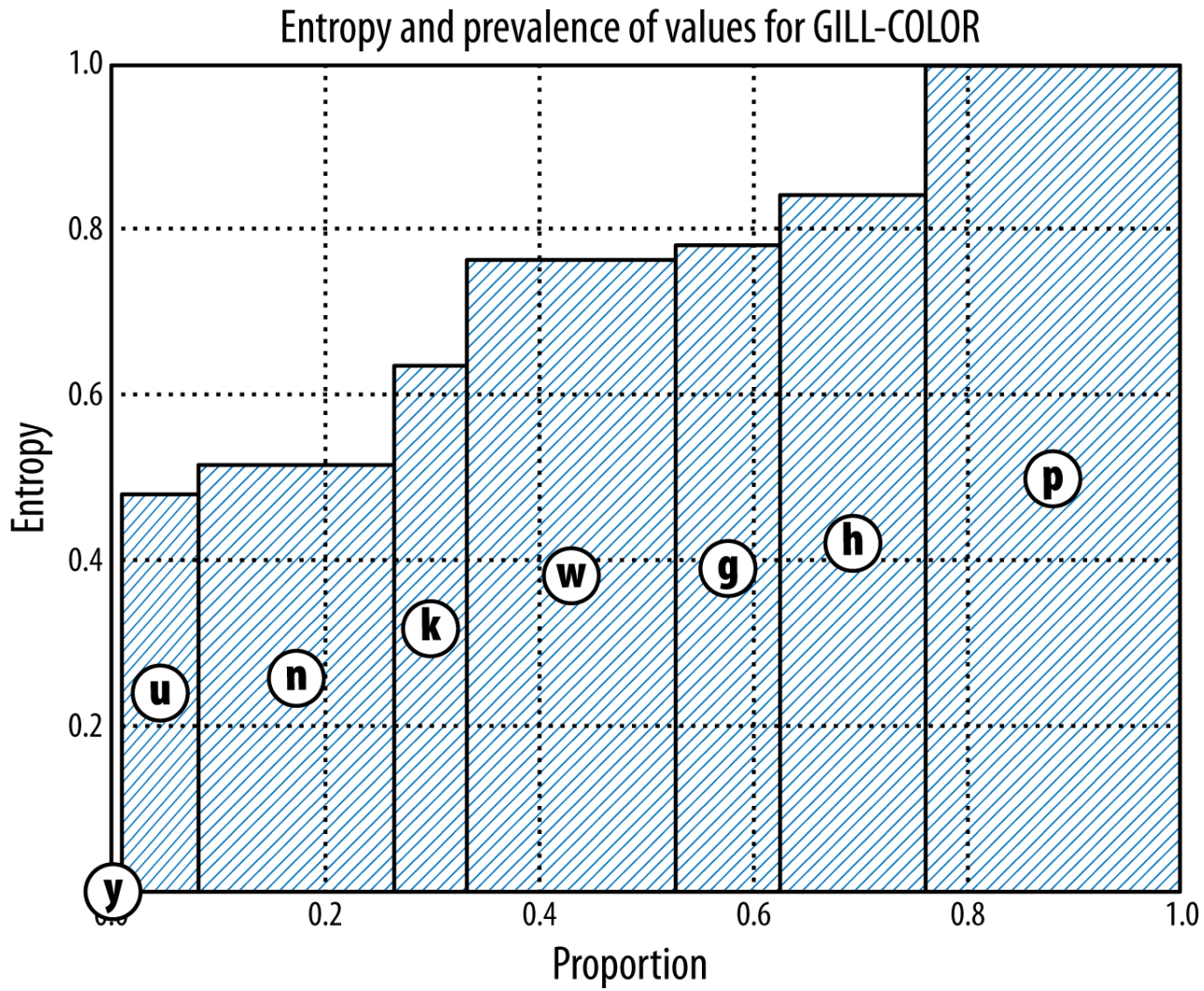


Attribute name	Possible values
CAP-SHAPE	bell, conical, convex, flat, knobbed, sunken
CAP-SURFACE	fibrous, grooves, scaly, smooth
CAP-COLOR	brown, buff, cinnamon, gray, green, pink, purple, red, white, yellow
BRUISES?	yes, no
ODOR	almond, anise, creosote, fishy, foul, musty, none, pungent, spicy
GILL-ATTACHMENT	attached, descending, free, notched
GILL-SPACING	close, crowded, distant
GILL-SIZE	broad, narrow
GILL-COLOR	black, brown, buff, chocolate, gray, green, orange, pink, purple, red, white, yellow
STALK-SHAPE	enlarging, tapering
STALK-ROOT	bulbous, club, cup, equal, rhizomorphs, rooted, missing
STALK-SURFACE-ABOVE-RING	fibrous, scaly, silky, smooth
STALK-SURFACE-BELOW-RING	fibrous, scaly, silky, smooth

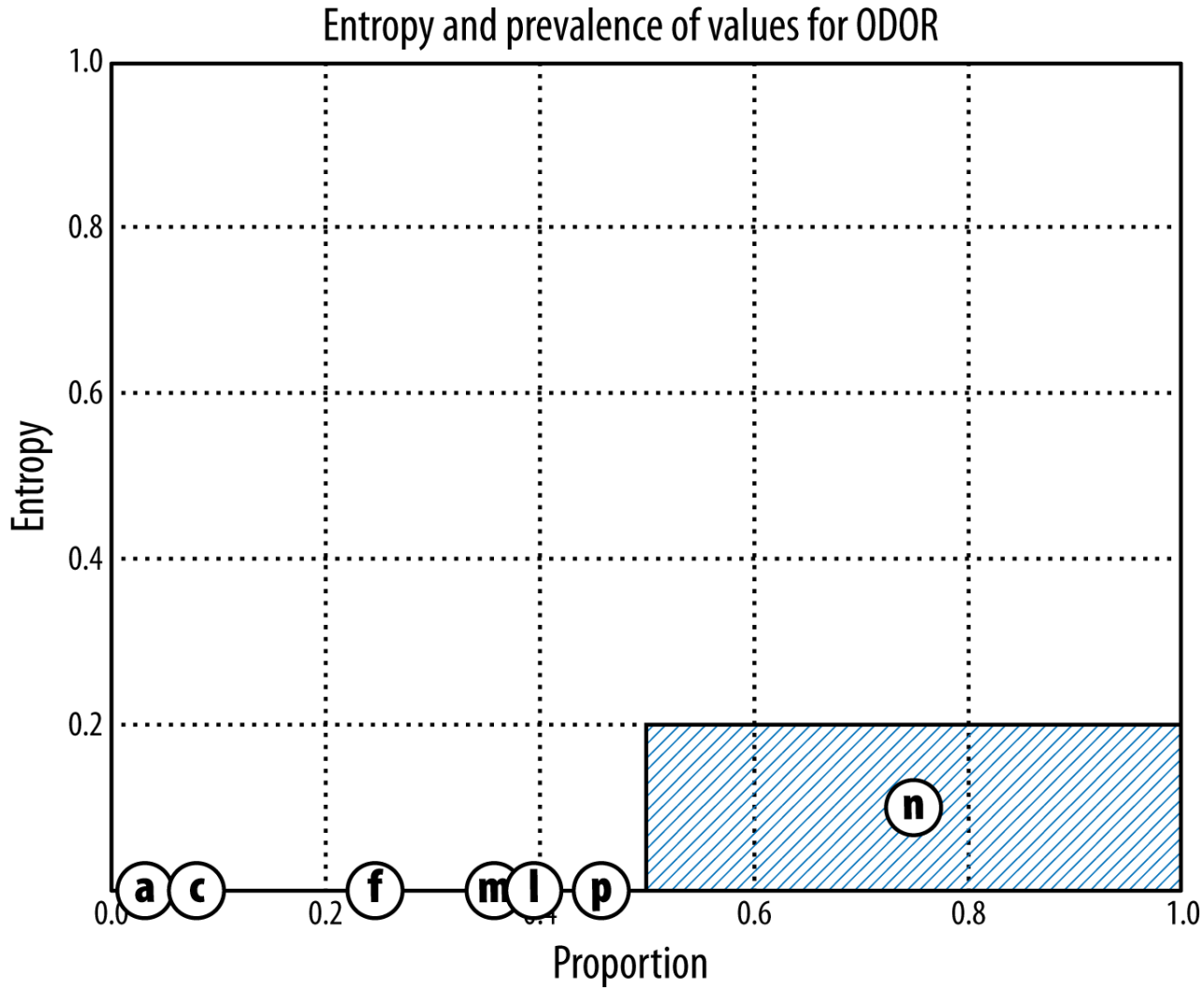
Example: Attribution Selection with Information Gain



Example: Attribution Selection with Information Gain



Example: Attribution Selection with Information Gain

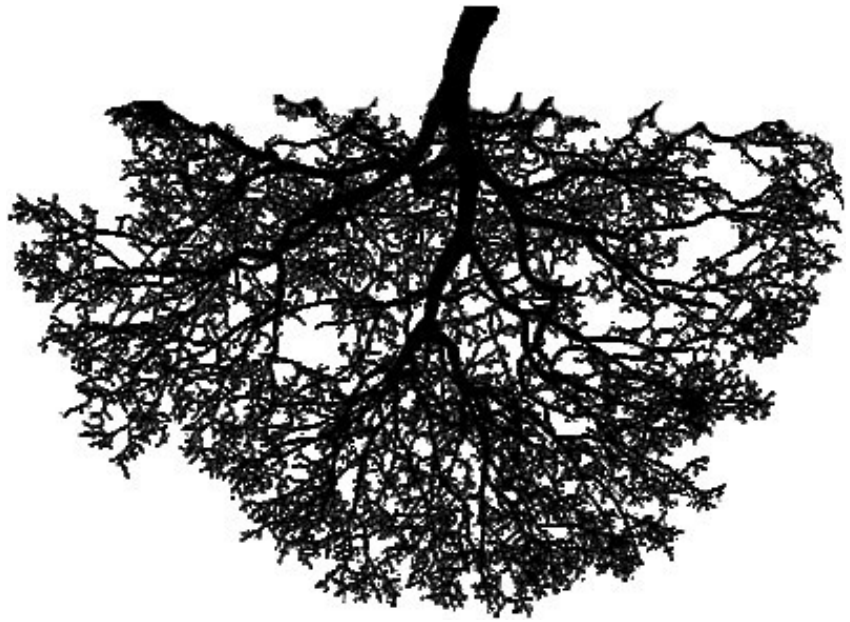


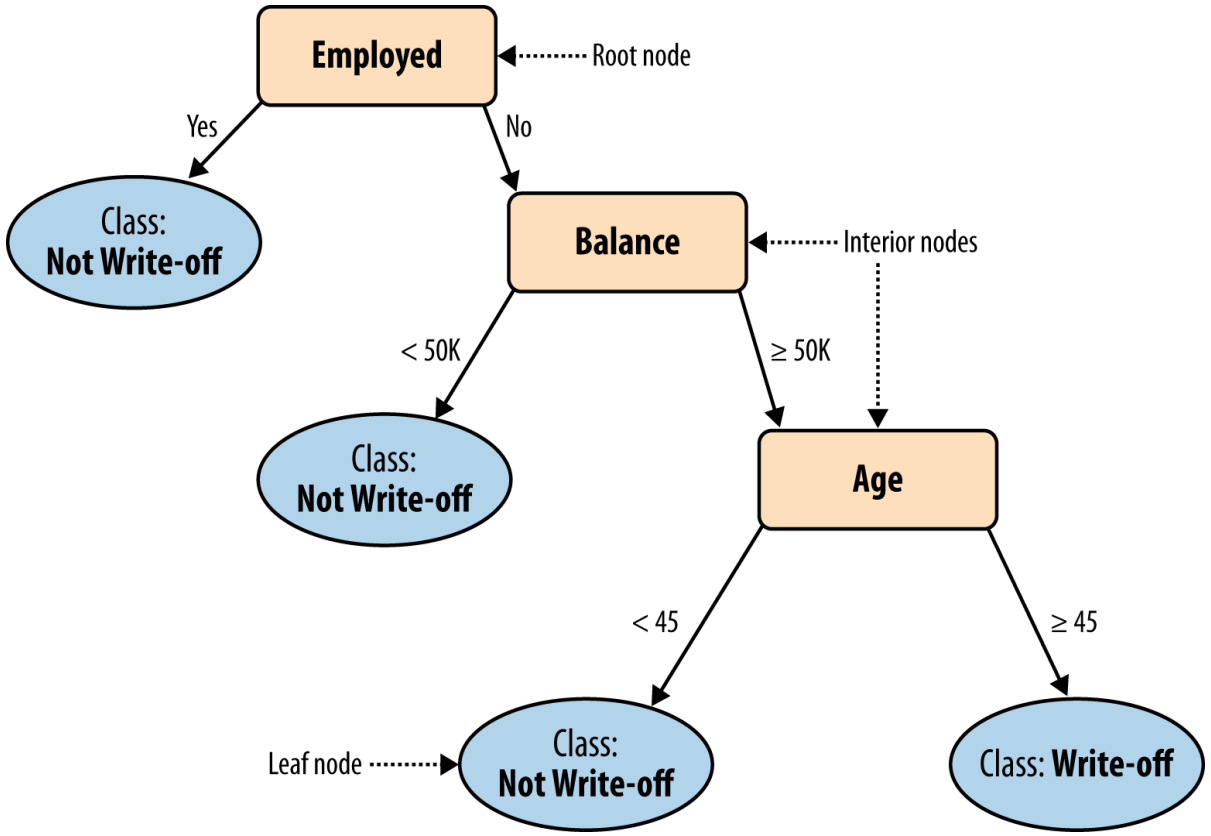
Multivariate Supervised Segmentation

- If we select the *single* variable that gives the most information gain, we create a very *simple* segmentation

- If we select multiple attributes each giving some information gain, how do we put them together?

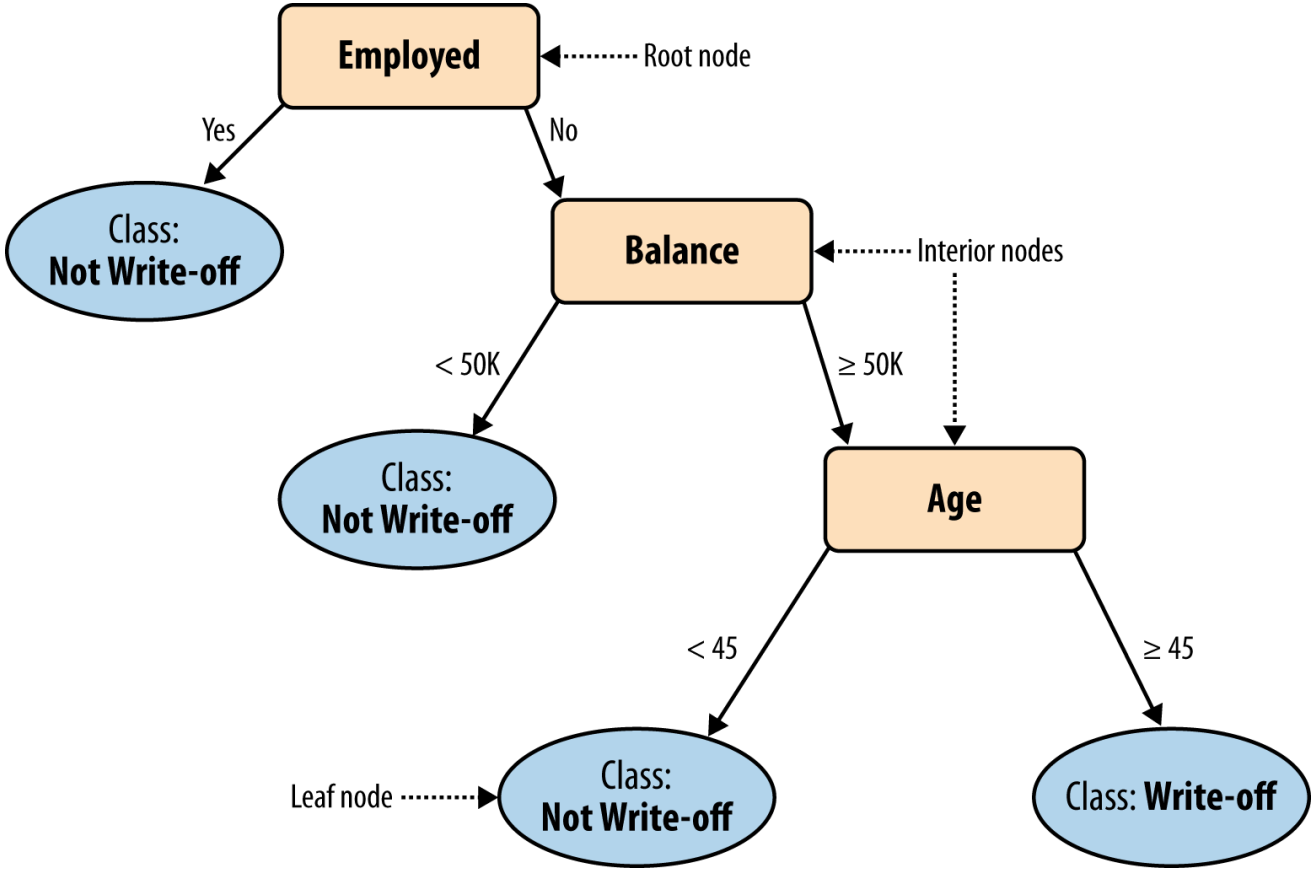
Tree-Structured Models



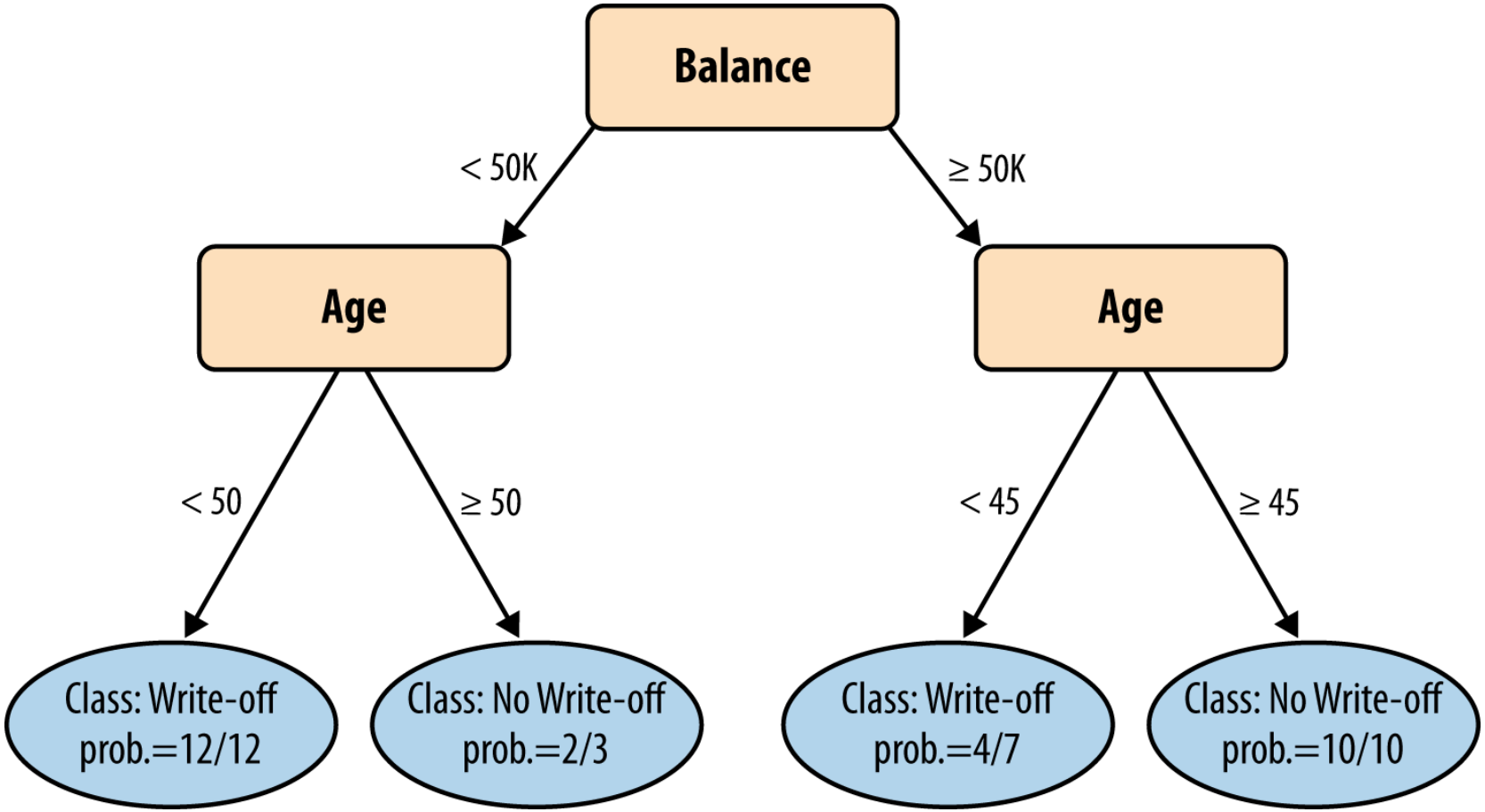


Tree-Structured Models

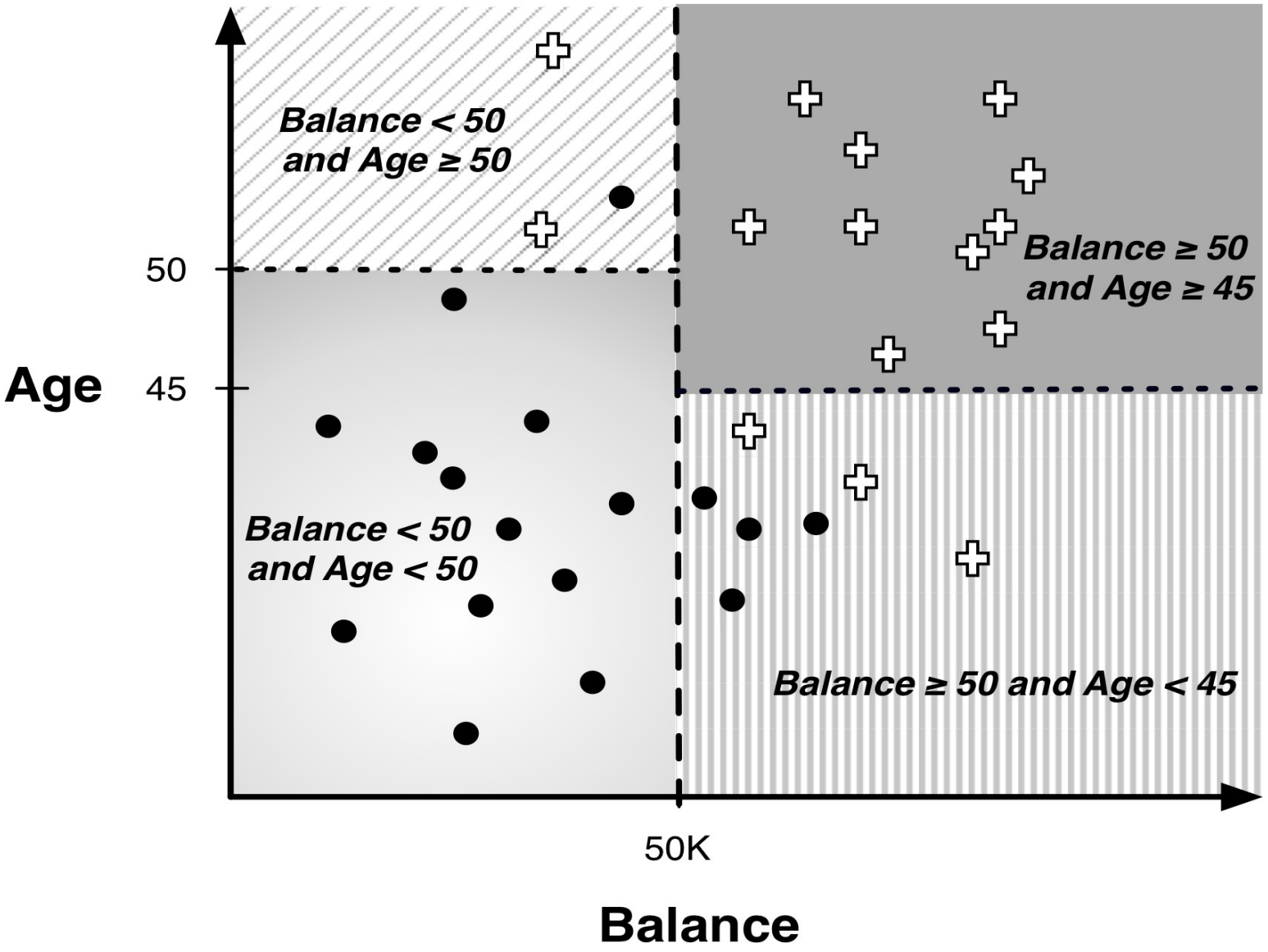
- Classify ‘John Doe’
  - Balance=115K, Employed=No, and Age=40

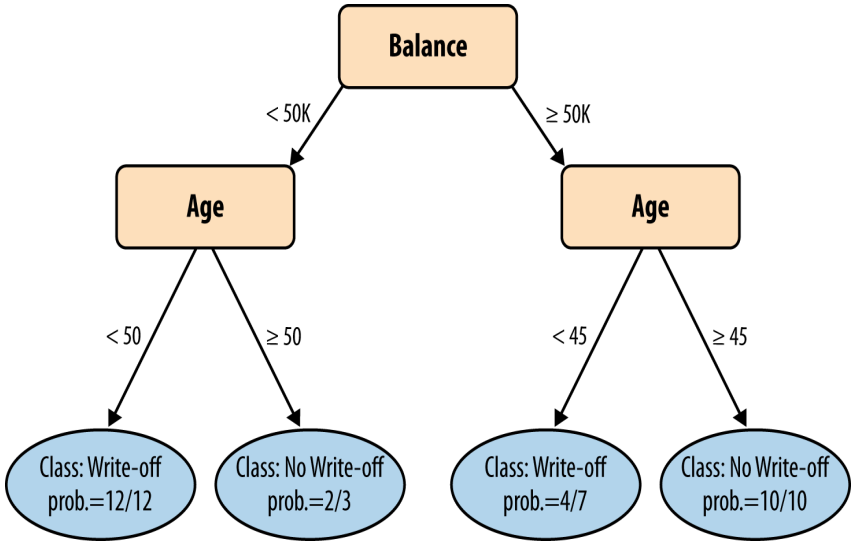


Visualizing Segmentations

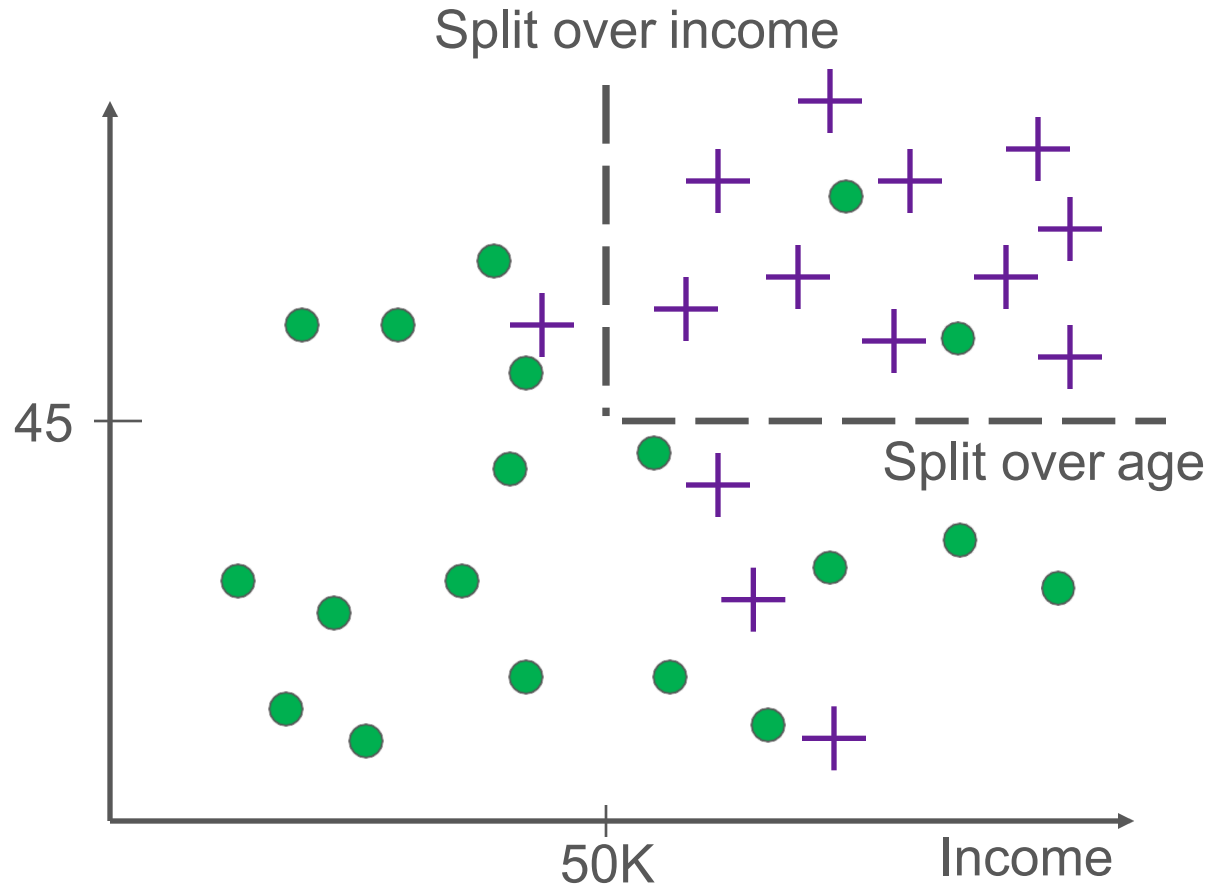


Visualizing Segmentations





Geometric interpretation of a model



Did not buy life insurance

Bought life insurance

Pattern:

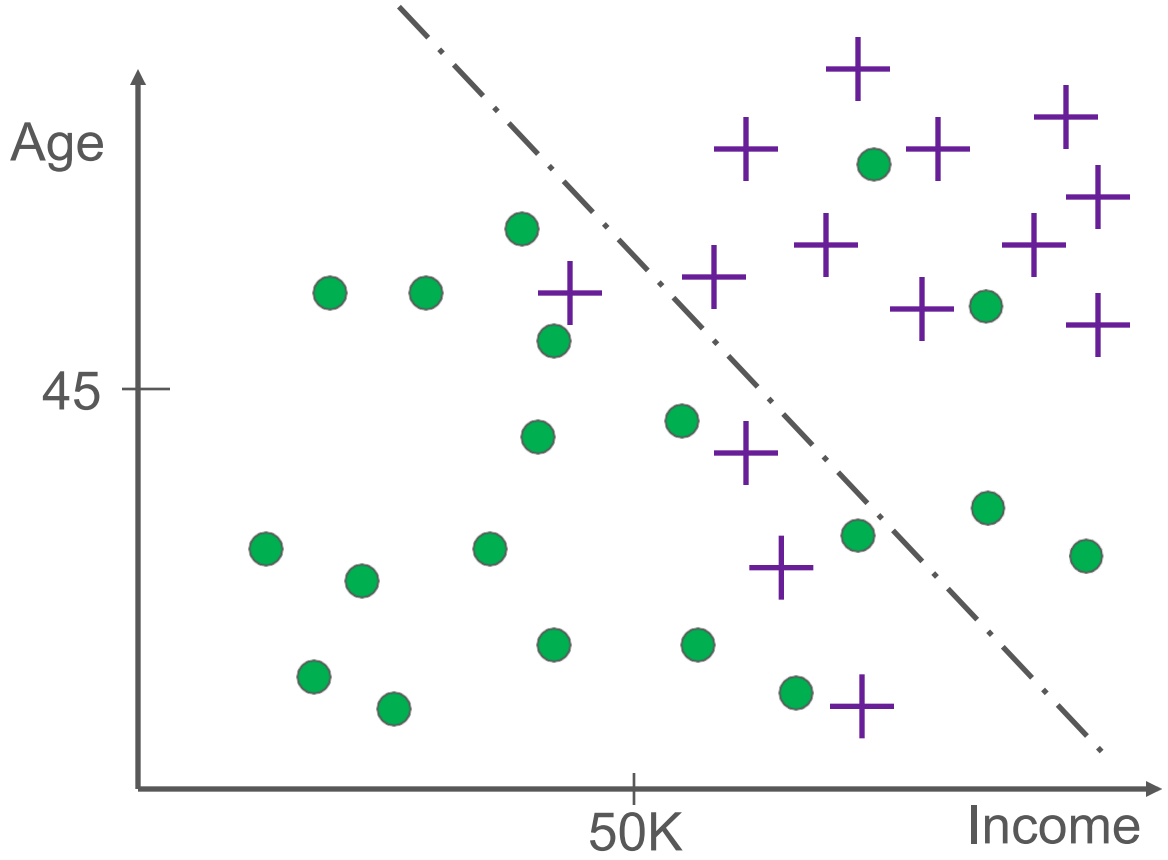
IF Balance >= 50K & Age > 45

THEN Default = ‘no’

ELSE Default = ‘yes’

Geometric interpretation of a model

What alternatives are there to partitioning this way?

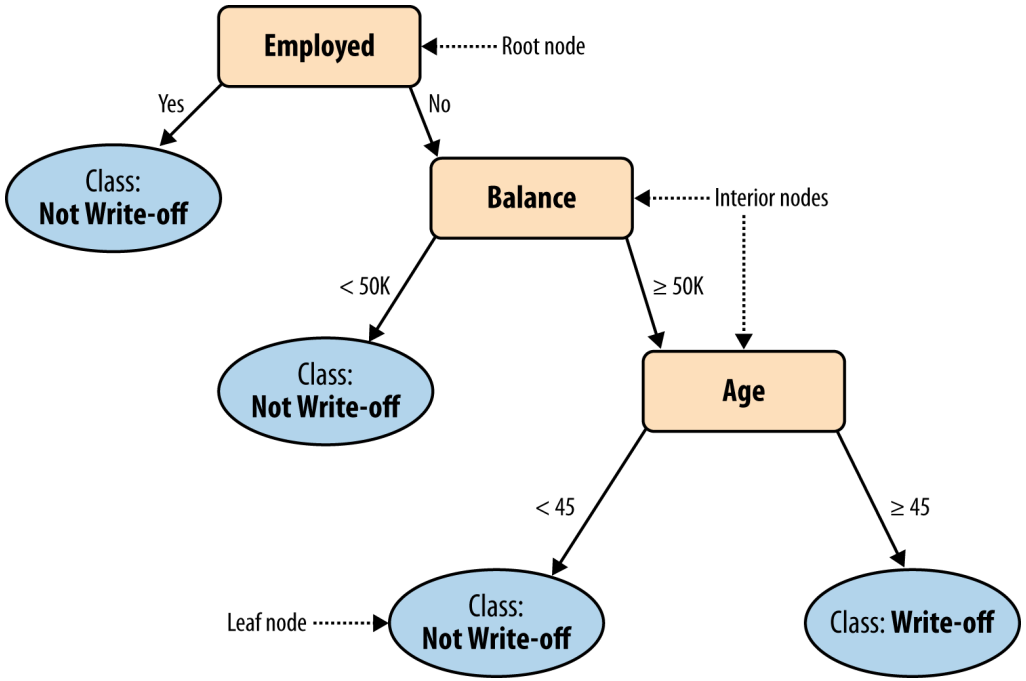


"True" boundary may not be closely approximated by a linear boundary!

Did not buy life insurance

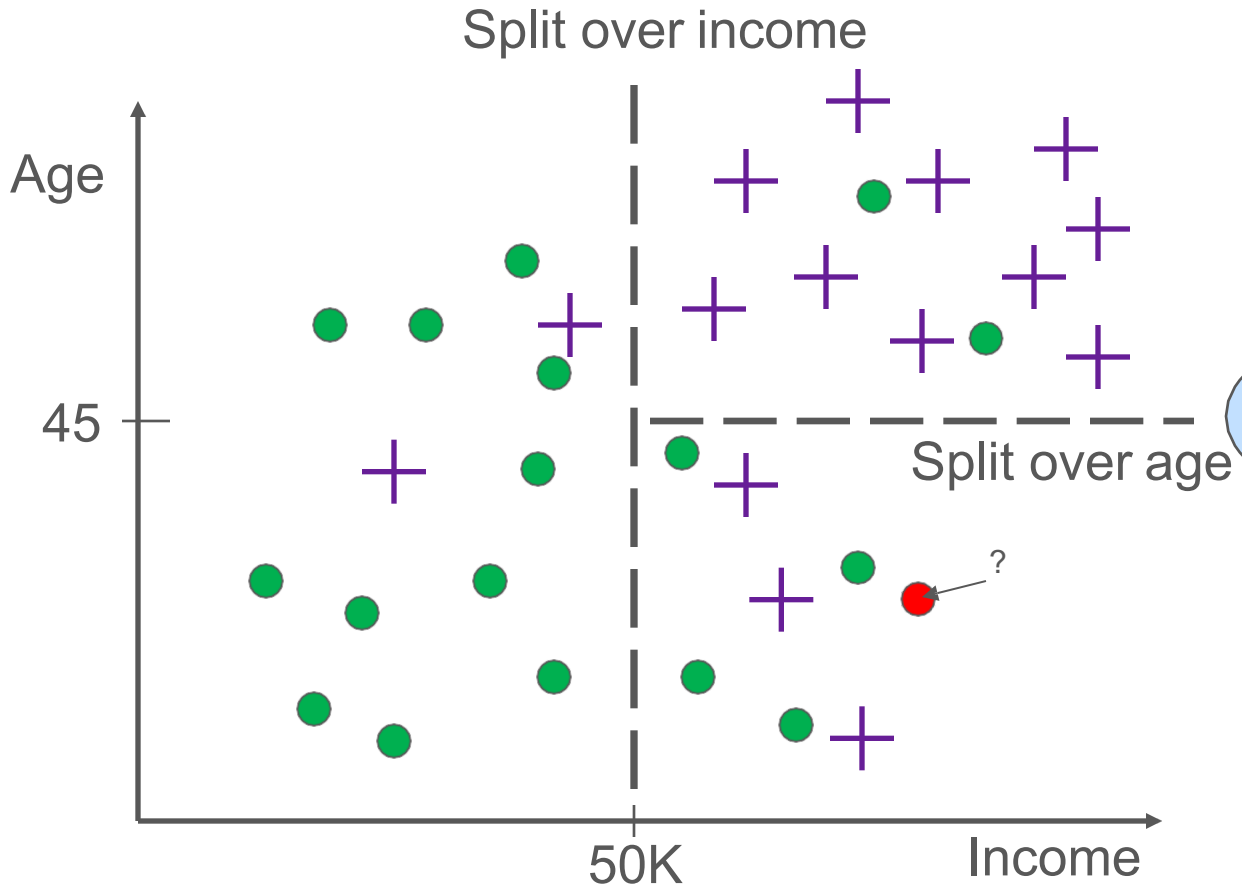
Bought life insurance

Trees as Sets of Rules

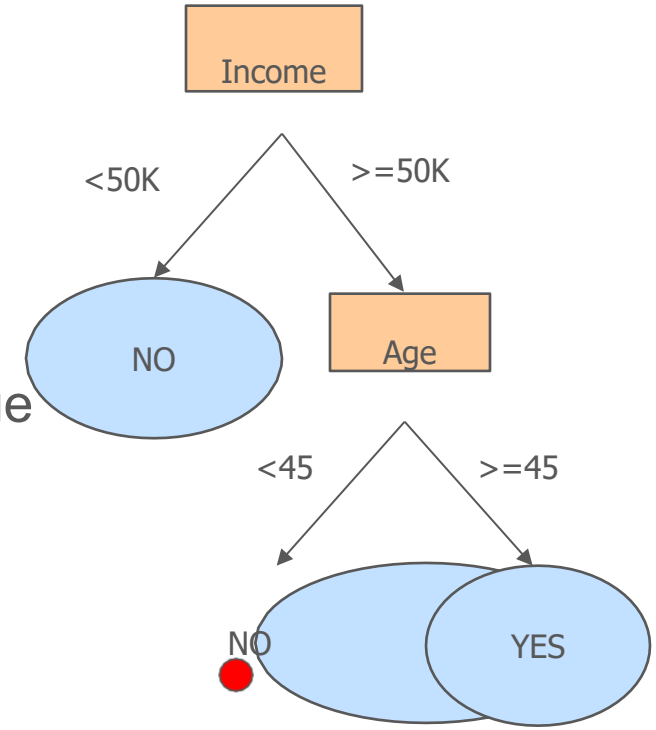


- IF (Employed = Yes) THEN Class=No Write-off
- IF (Employed = No) AND (Balance < 50k) THEN Class=No Write-off
- IF (Employed = No) AND (Balance ≥ 50k) AND (Age < 45) THEN Class=No Write-off
- IF (Employed = No) AND (Balance ≥ 50k) AND (Age ≥ 45) THEN Class=Write-off

What are we predicting?



Classification tree

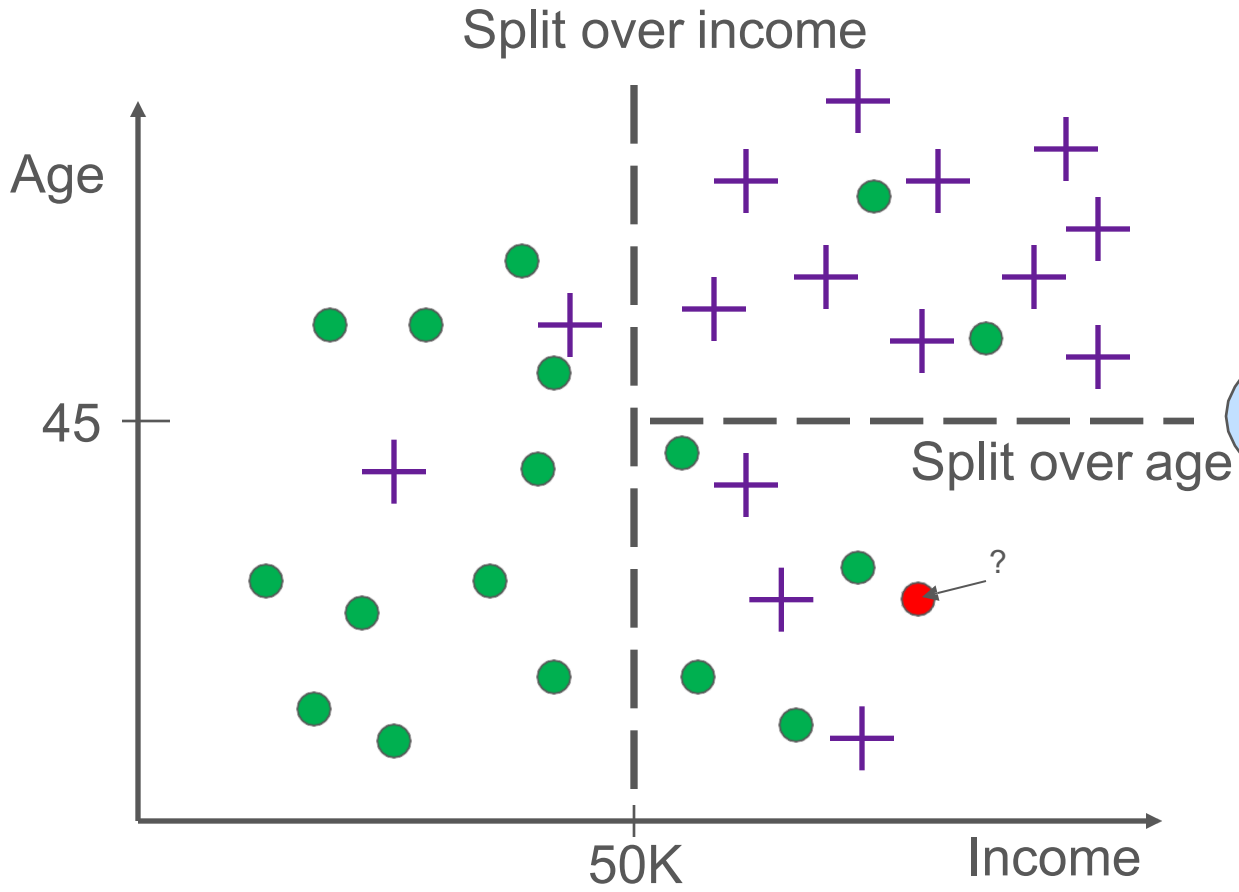


Did not buy life insurance

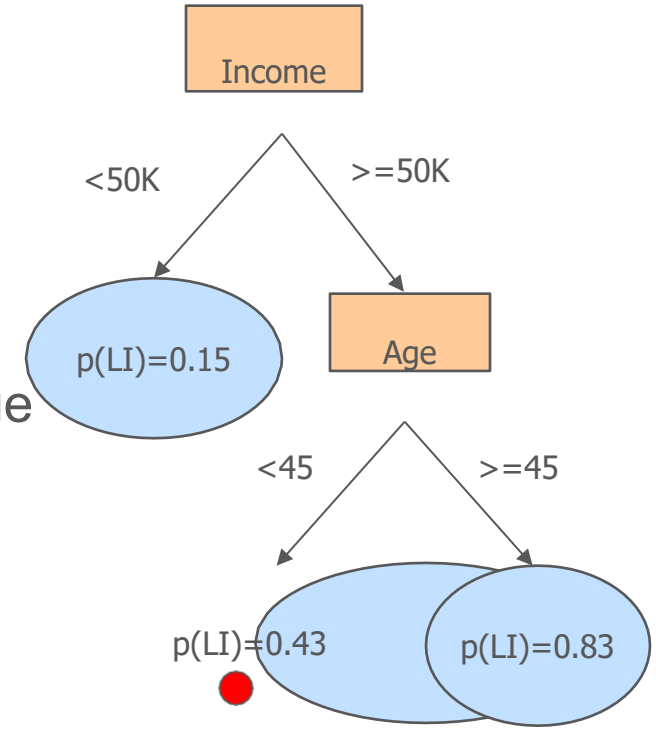
Bought life insurance

Interested in LI? = NO

What are we predicting?



Classification tree



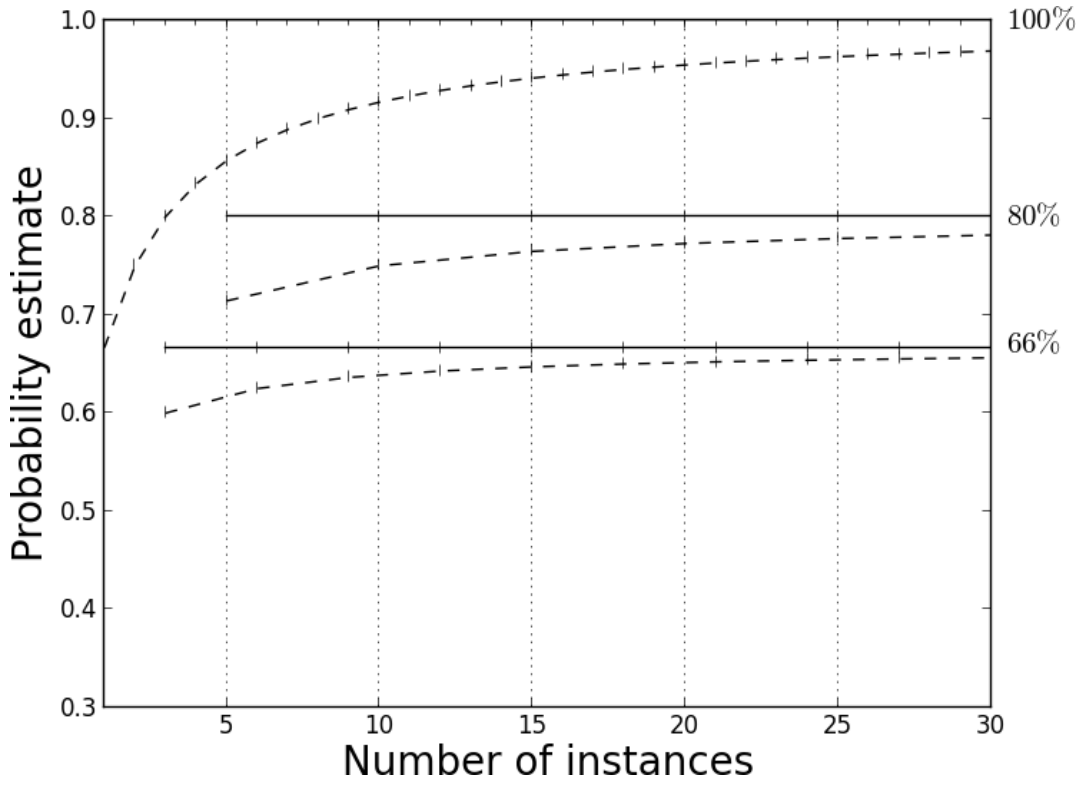
Did not buy life insurance

Bought life insurance

Interested in LI? = 3/7

Class probability

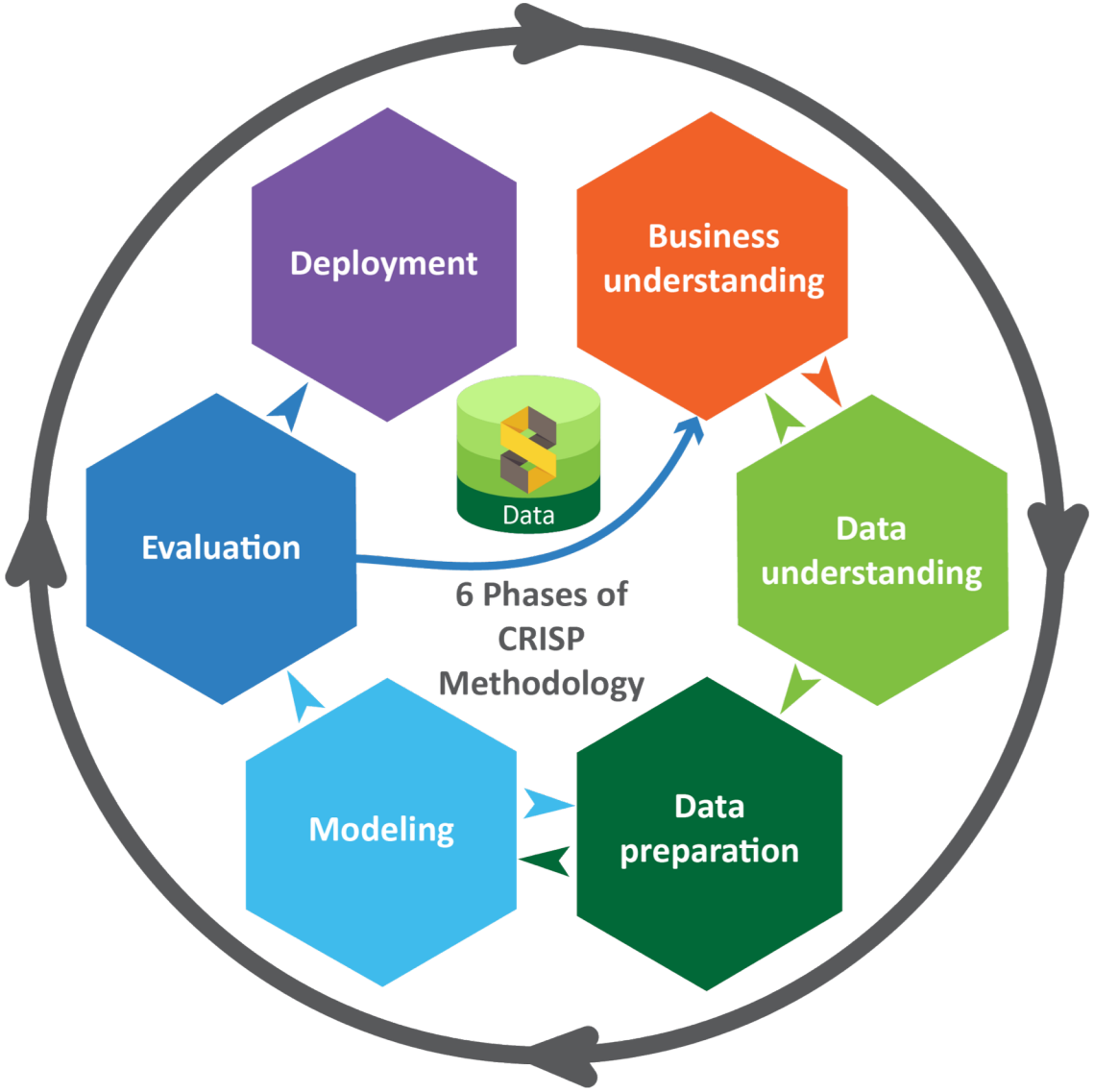
- $p(c) = \frac{n+1}{n+m+2}$,
  - where n is the number of examples in the leaf belonging to class c, and m is the number of examples not belonging to class c



Laplace Correction

MODULE 1 Revisit

Data Mining Process : CRISP-DM



Success in Data Science Projects

		Medical	Software Engineering	St Prediction
What do we want to achieve ?	What am I trying to achieve ?	Predict an onset | Raise a alarm | Screen | Treatment | Intervene | Help research (genetics, marrow match)		
	What do I know			
	What is my question of the data			
Data	What is my data assumption			
Approach	What is my approach selection			
Team	My team			
Quantify: risk, success	Risks			
	Success criteria			
	My metric ?			

OUTCOMES

DATA

APPROACH

TEAM