Fundamental Methods of Data Science

Class 5

Tomer Libal

Supervised segmentation

- In the exercise, you have used specific attributes to predict if a passenger survived or not
  - How did you choose these attributes?

Supervised segmentation

- In the exercise, you have used specific attributes to predict if a passenger survived or not
    - How did you choose these attributes?
- Some attributes (which ones?) offer more information than others
- How can we segment the population into groups that differ from each with respect to some quantity of interest?

Supervised segmentation

- In the exercise, you have used specific attributes to predict if a passenger survived or not
  - How did you choose these attributes?
- Some attributes (which ones?) offer more information than others
- How can we segment the population into groups that differ from each with respect to some quantity of interest?
- Informative attributes
  - Find knowable attributes that correlate with the target of interest

Supervised Segmentation

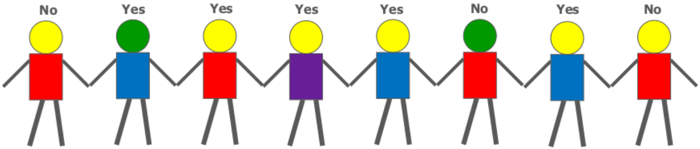
- How can we judge whether a variable contains important information about the target variable?
  - How much?

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)

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)

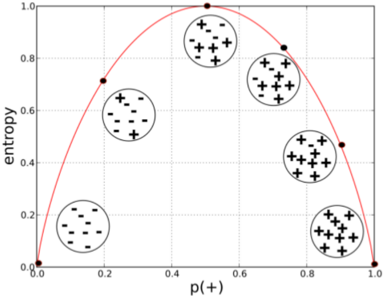


Selecting Informative Attributes

- 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

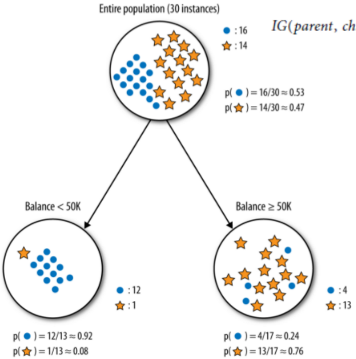
Selecting Informative Attributes

- 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



Information Gain

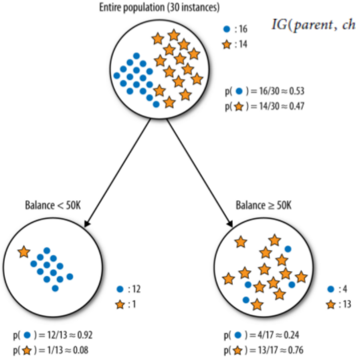
- **Information gain** measures the change in entropy due to any amount of new information being added



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

Information Gain

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

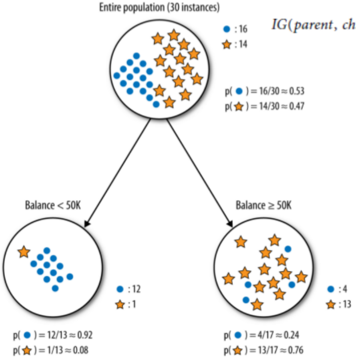


$$IG(parent, children) = entropy(parent) - [p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + \cdots]$$

- What is the entropy of the left child? And the right?

Information Gain

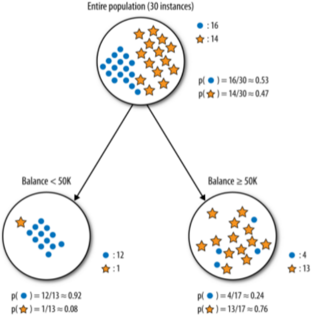
- **Information gain** measures the change in entropy due to any amount of new information being added



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

- What is the entropy of the left child? And the right?
- What is the IG?

Information Gain



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

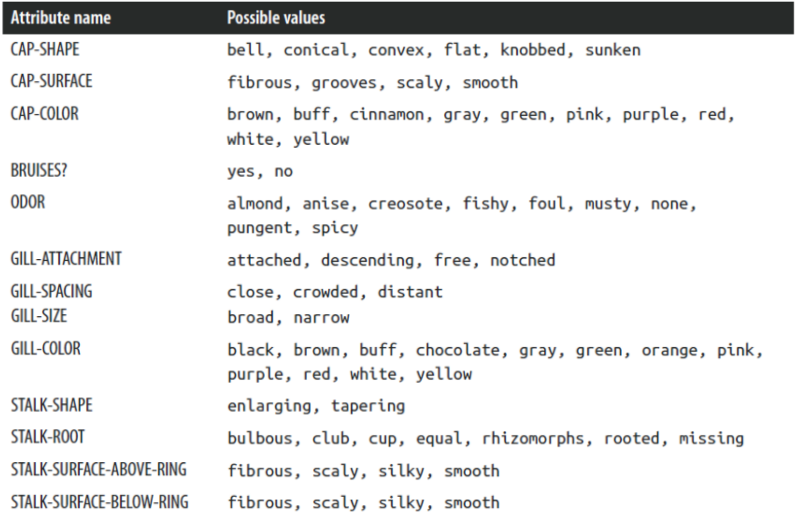
Attribute Selection

- Reasons for selecting only a subset of attributes:
  - Better insights and business understanding
  - Better explanations and more tractable models
  - Reduced cost
  - Faster predictions
  - Better predictions!
    - Over-fitting (to be continued . . .)

Example: Attribution Selection with Information Gain

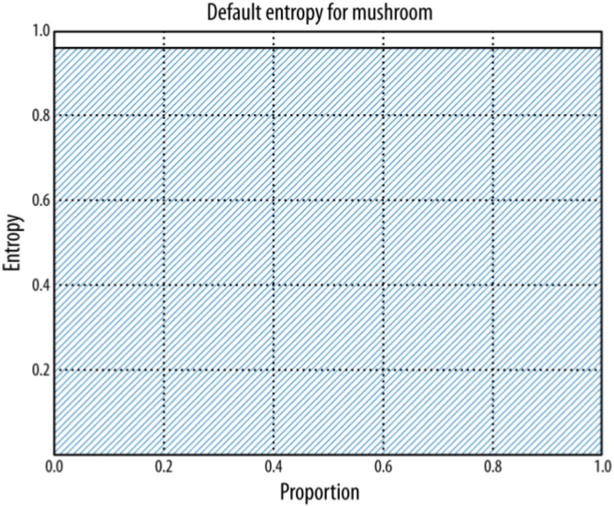
- This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family
- Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended
  - This latter class was combined with the poisonous one
- The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy

Example: Attribution Selection with Information Gain

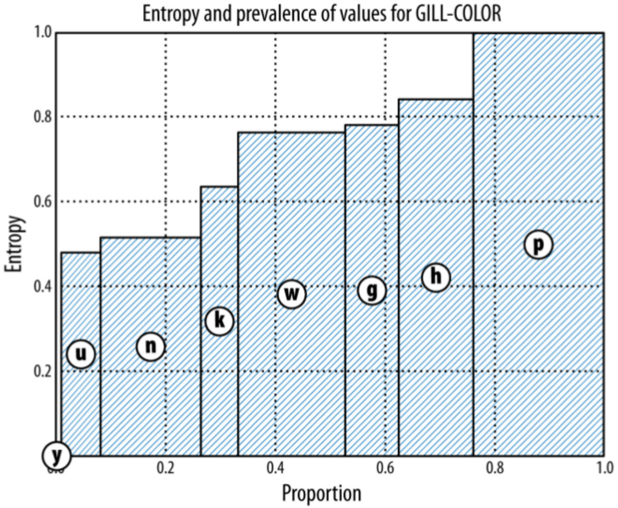


Attribute name	Possible values
CAP-SHAPE	bell, conical, convex, flat, knobbed, sunken
CAP-SURFACE	fibrous, grooves, scaly, smooth
CAP-COLOR	brown, buff, cinnamon, gray, green, pink, purple, red, white, yellow
BRUISES?	yes, no
ODOR	almond, anise, creosote, fishy, foul, musty, none, pungent, spicy
GILL-ATTACHMENT	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

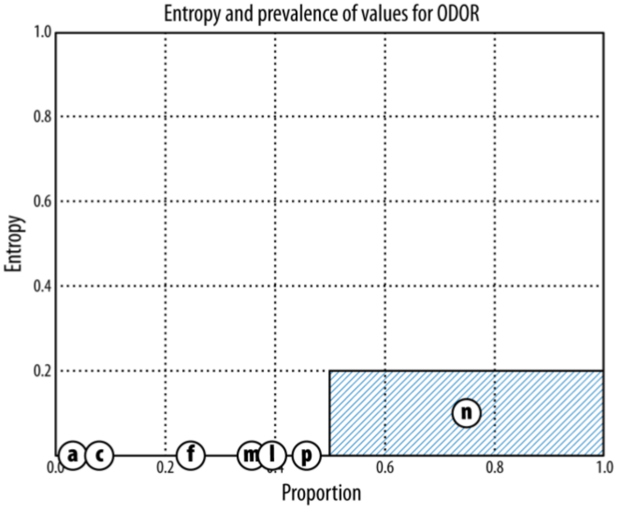
Example: Attribution Selection with Information Gain



Example: Attribution Selection with Information Gain



Example: Attribution Selection with Information Gain


Entropy and prevalence of values for ODOR
Entropy
Proportion
a
c
f
m
l
p
n