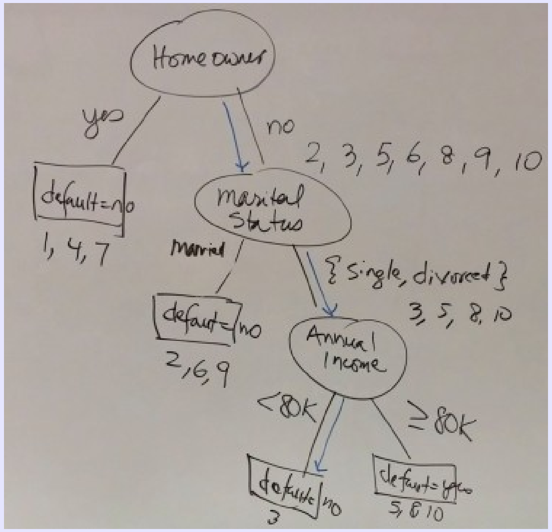
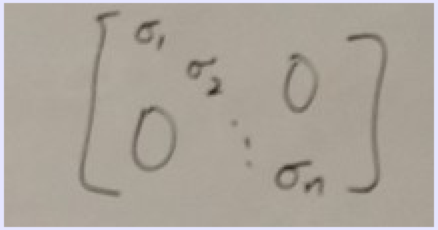
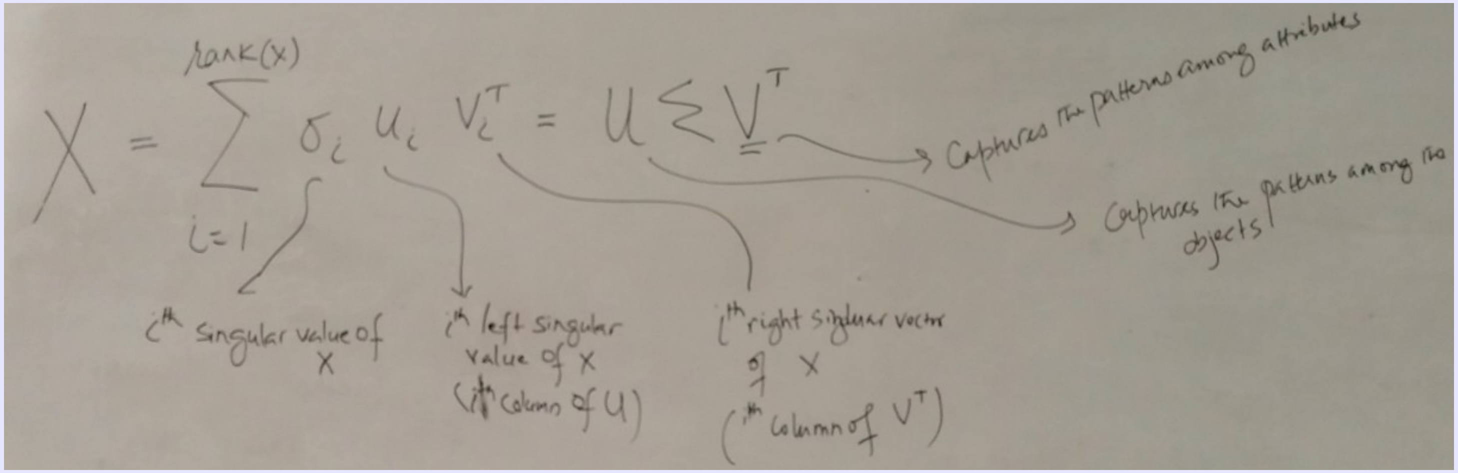
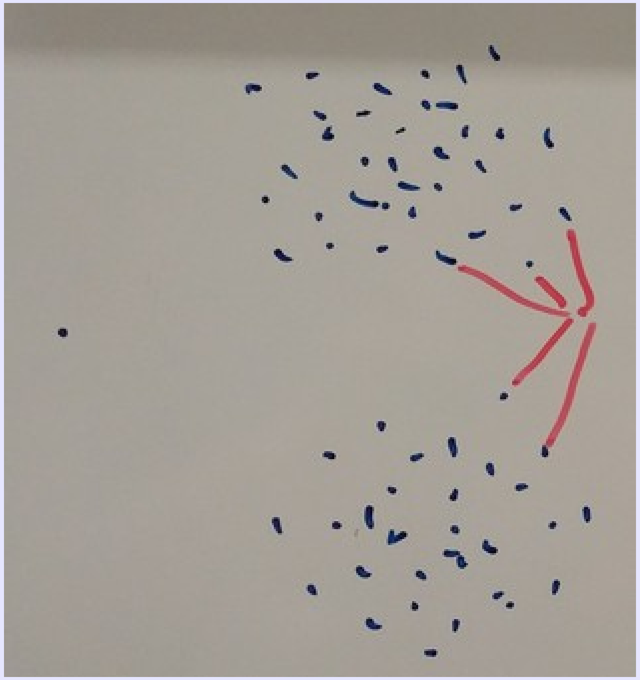








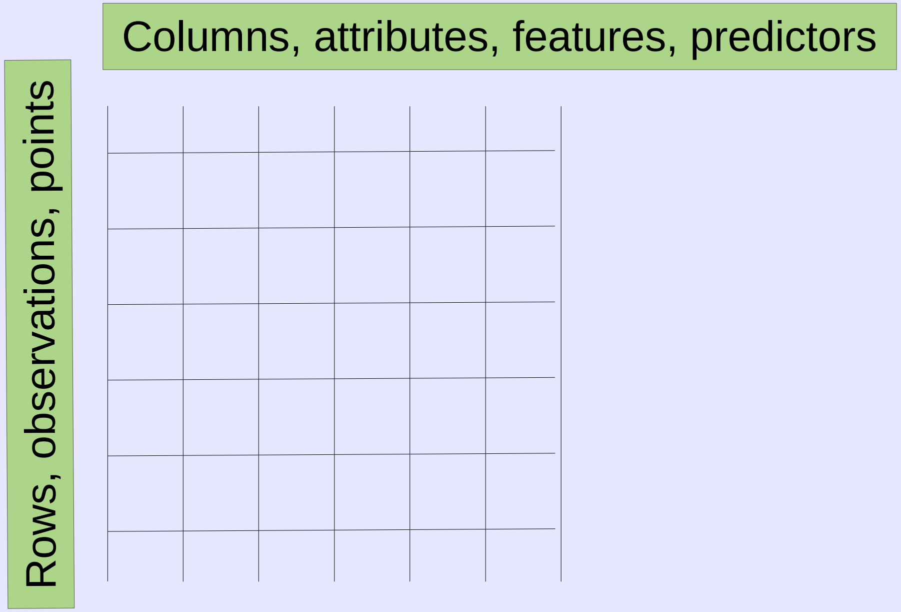
CS 422-04: Data Mining

Vijay K. Gurbani, Ph.D., Illinois Institute of Technology

Lecture 4: **Components of Learning**
Decision Trees

Components of learning

- Recall, most data mining / machine learning algorithms operate on matrices.
- The canonical picture to keep in mind is this:



Components of learning

- Example of a *matrix* data layout.

Projection of x Load	**Projection of y load**	**Distance**	**Load**	**Thickness**
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Components of learning

- Example of a *document* data layout.

Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

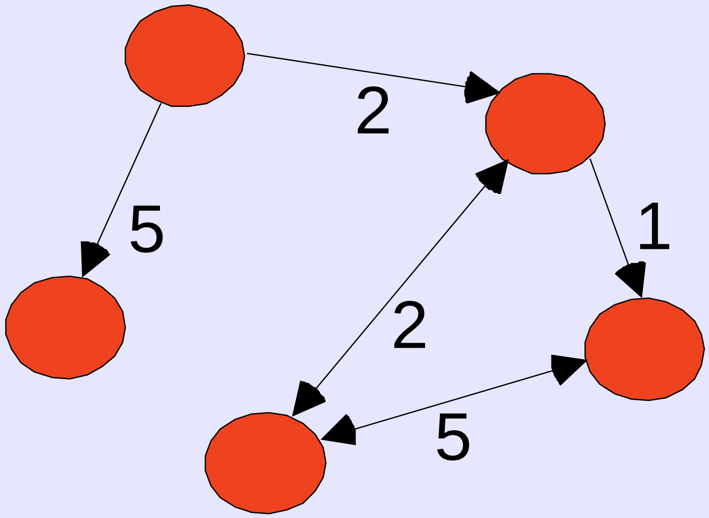
Components of learning

- Example of a *transaction* data layout.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk

Components of learning

- Example of a *graph* data layout.

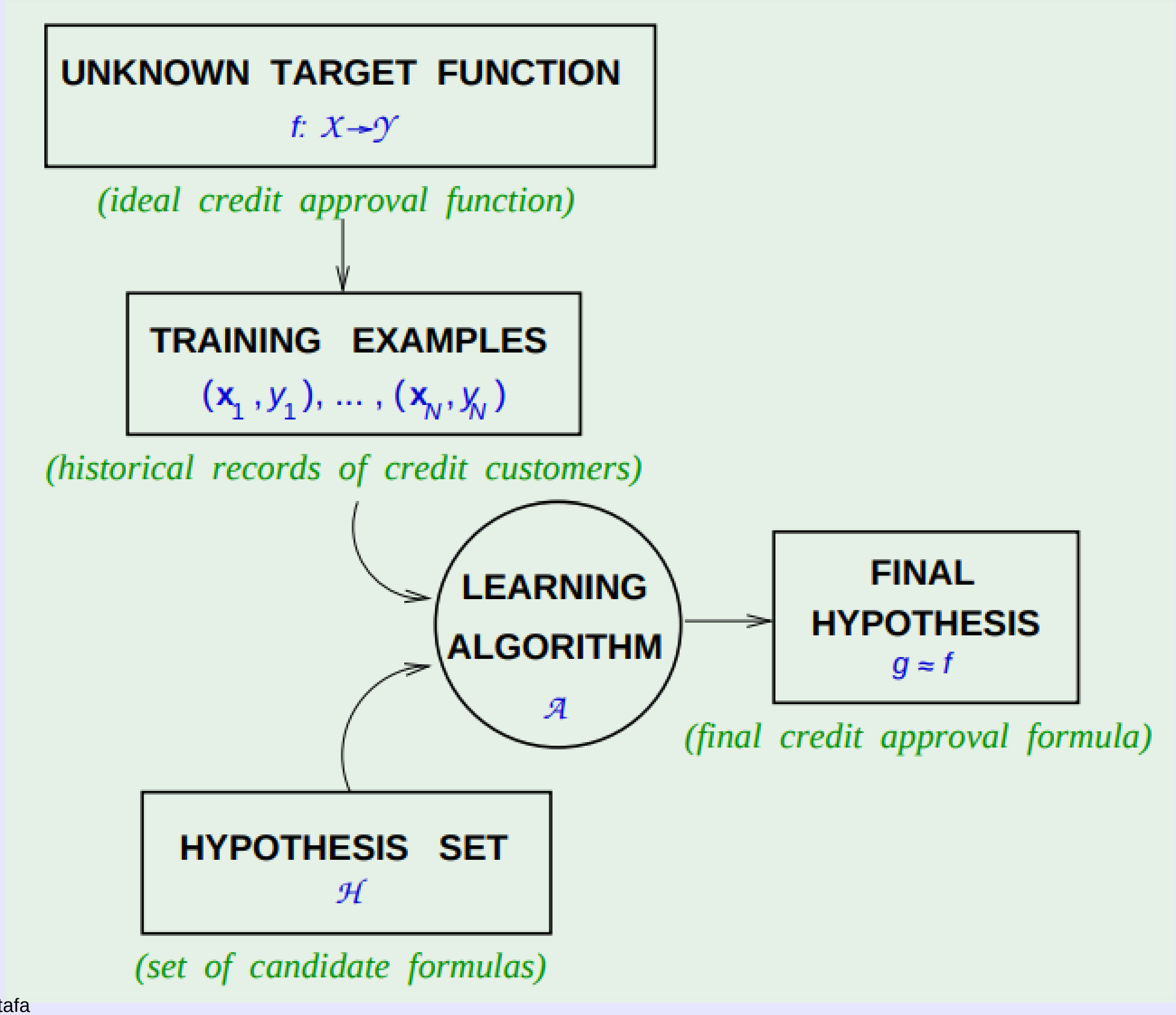


- As it turns out, graphs can be represented as matrices.

Components of learning

- Formalism:
  - Input: $\mathcal{X}$, A matrix (n-dimension, n >= 1) of attributes
  - Output: $\vec{\mathcal{Y}}$, the response vector
  - Target function: $f : \mathcal{X} \to \mathcal{Y}$
  - Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
  - Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$
  - Hope: $g \approx f$

Components of learning



Slide source: Prof. Yaser S. Abu-Mostafa
Learning from Data, 2012.

Components of learning

Machine learning

- **Supervised**
  - Decision Tree
  - Rule-based
  - Neural Networks
  - Naive Bayes
  - SVM
  - Regression
- **Unsupervised**
  - Clustering
  - Mixture models
  - EM algorithms
- **Reinforcement**
  - Actions to maximize reward (or minimize punishment)
  - Ex. Go, chess playing

Components of learning

- Machine learning: *Generalizing to cases we have not seen before!*
- But wait … can’t we simply see all or most of the data?

Components of learning

- Machine learning: *Generalizing to cases we have not seen before!*
- But wait … can’t we simply see all or most of the data?
- Suppose: You have data that consists of 1,000 Boolean fields, and you have 1,000,000,000,000 records in a database.
- How much insight do these 1 trillion records represent?

Components of learning

- Theoretically, you will need 2^{1000} records to represent all of your data.
- The 1 trillions records are one *gazillionth** of 1 percent of 2^{1000}!
  - * Gazillionth = 10E-285!!

Components of learning

- Theoretically, you will need 2^{1000} records to represent all of your data.
- The 1 trillions records are one *gazillionth** of 1 percent of 2^{1000}!
  - * Gazillionth = 10E-285!!

Morals:
- Curse of dimensionality is real
- Generalization is how we deal with combinatorial explosion!

Data Types

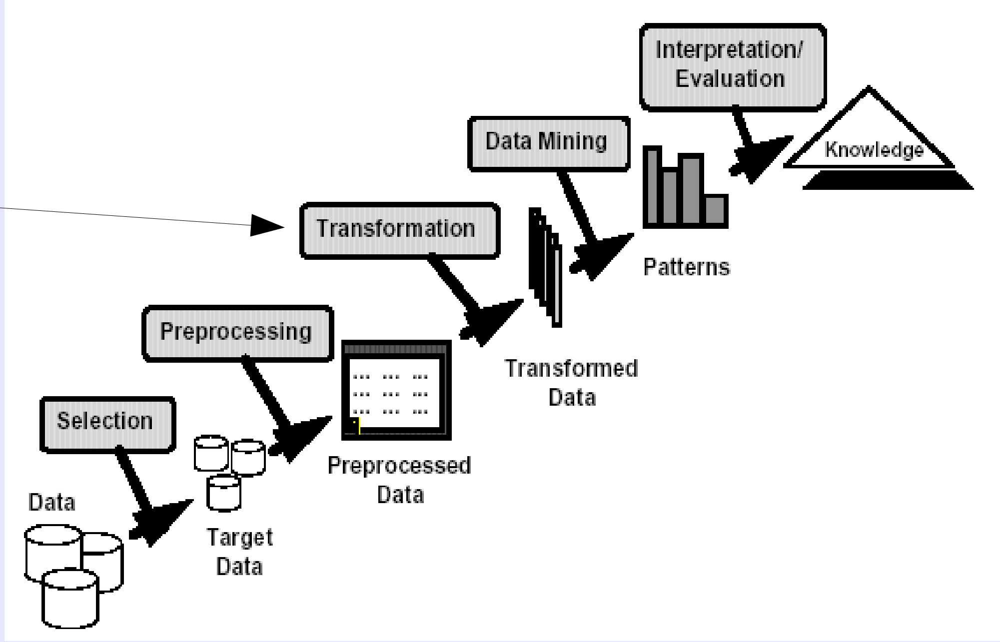
- R has the following data types to represent attributes:
  - Numeric
  - Integer
  - Factor
  - Character

Data Types

- R has the following data types to represent attributes:
  - Numeric: Can take “float” or “double” values.
  - Integer: Cannot take decimal or fraction values.
  - Factor: An enumeration data type that takes only certain values: {“blue”, “green”, “red”}; or {0, 1, 2}.
    - Values of a factor can be *ordinal*, i.e., order of values matter. Example: {“small”, “medium”, “large”} is different than {“small”, “large”, “medium”}.
    - Factors are also referred to as *categorical* variables.
  - Character: Single character or character strings.

Data Types

- Certain algorithms have an affinity for certain data types:
  - Certain classification requires that numeric (or continuous) data be represented as categorical (factor) attributes.
  - Association algorithms prefer a binary attribute (a factor of 0 and 1).
- One of the important step during the transformation phase is to ensure that algorithms get the attribute in the form they can operate on it.



Data Types

- Example: Binarization (Tan, Ch. 2)

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Data Types

- Example: Discretization

  {10, 20, 30, 1, 22, 25, 2, 18, 15}

- Step 1: Sort:

  {1, 2, 10, 15, 18, 20, 22, 25, 30}

- Step 2: Create split points

  {1, 2, 10, 15, 18, 20, 22, 25, 30}

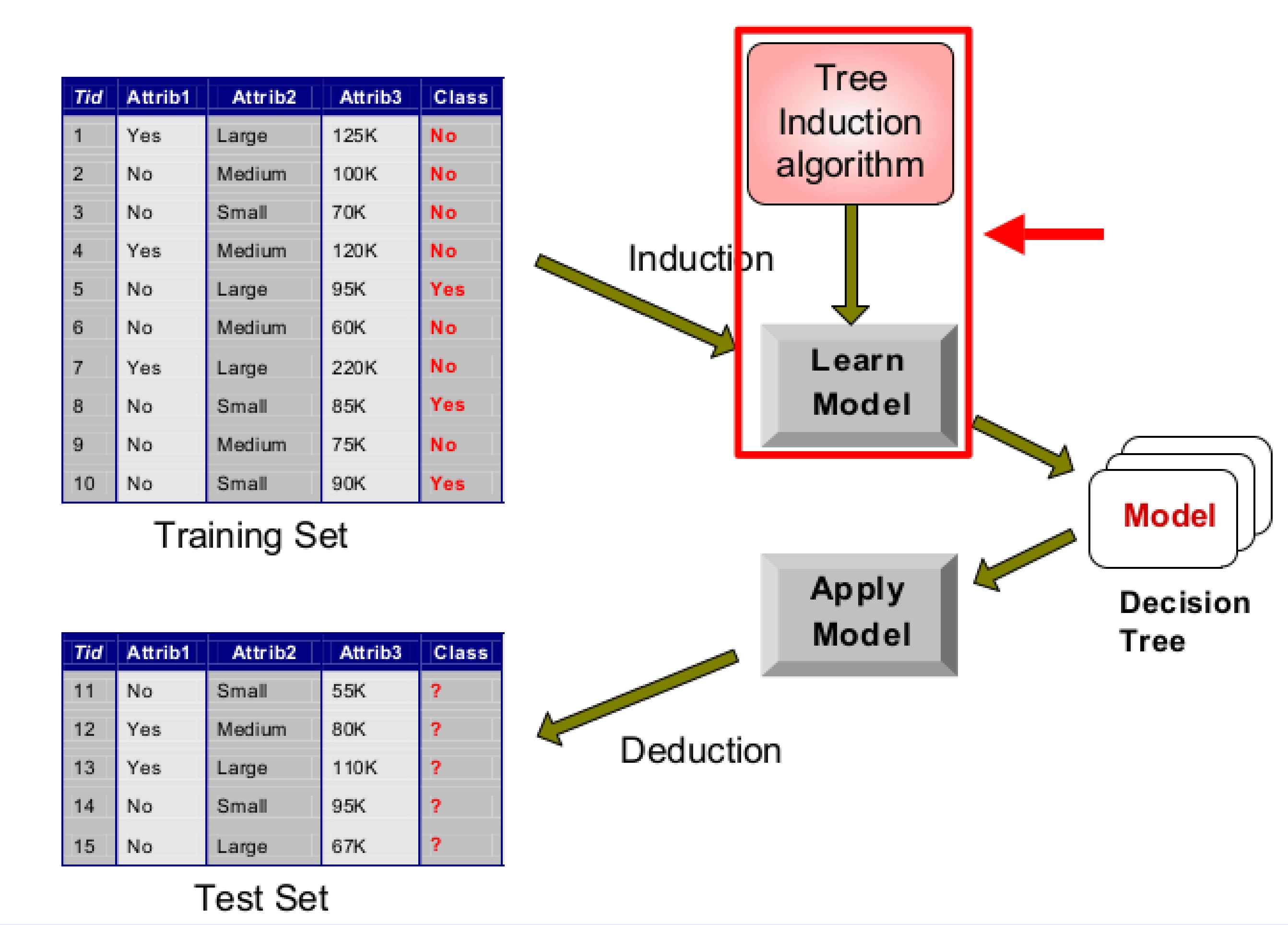
- Step 3: Map split values to discrete categorical variables; e.g.: {1, 2, 10} → "Small", ...

Decision tree

- Our first classification algorithm.
- **Classification**: The task of learning a target function, *g*, that maps each attribute set x to one of the predefined class labels, $\vec{y}$.
- Let's play a game of 20 questions.
  - Category: Movie.
  - Ask me questions!

Decision tree

- A bird's eye view.



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

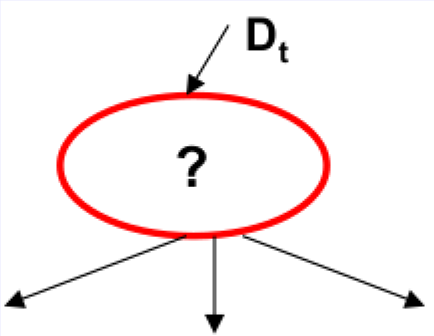
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Decision tree: Hunt's algorithm

- Let D_t be the set of training records that reach a node t
- General Procedure:
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
  - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Decision tree: Hunt's algorithm

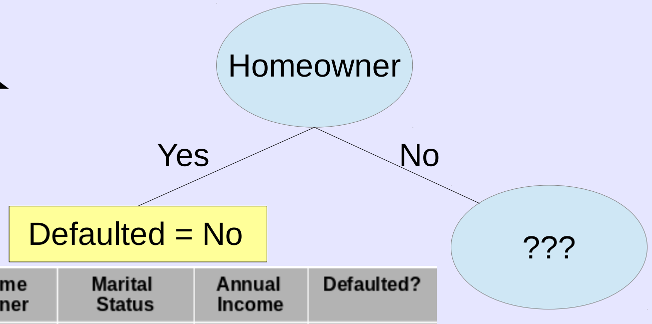
Default class = No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Decision tree: Hunt’s algorithm

Default class = No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Tid	Home owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125K	No
4	Yes	Married	120K	No
7	Yes	Divorced	220K	No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
2	No	Married	100K	No
3	No	Single	70K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Decision tree: Hunt's algorithm

Default class = No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Homeowner

Yes

No

Defaulted = No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125K	No
4	Yes	Married	120K	No
7	Yes	Divorced	220K	No

Marital Status

Tid	Home owner	Marital Status	Annual Income	Defaulted?
2	No	Married	100K	No
3	No	Single	70K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Married

Defaulted = No

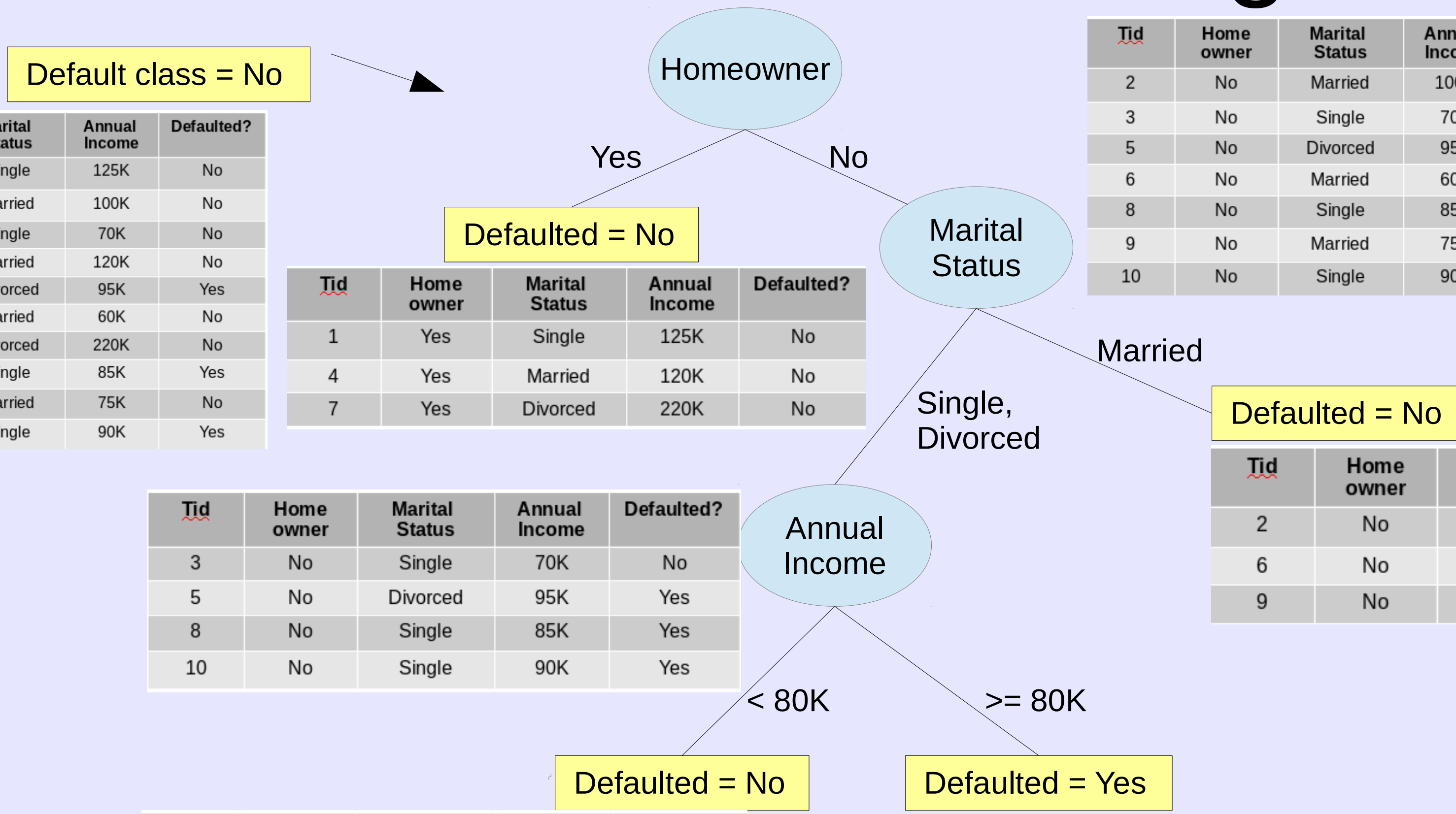
Tid	Home owner	Marital Status	Annual Income	Defaulted?
2	No	Married	100K	No
6	No	Married	60K	No
9	No	Married	75K	No

Single, Divorced

Tid	Home owner	Marital Status	Annual Income	Defaulted?
3	No	Single	70K	No
5	No	Divorced	95K	Yes
8	No	Single	85K	Yes
10	No	Single	90K	Yes

???

Decision tree: Hunt's algorithm



Tid	Home owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tid	Home owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125K	No
4	Yes	Married	120K	No
7	Yes	Divorced	220K	No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
2	No	Married	100K	No
3	No	Single	70K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tid	Home owner	Marital Status	Annual Income	Defaulted?
2	No	Married	100K	No
6	No	Married	60K	No
9	No	Married	75K	No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
3	No	Single	70K	No
5	No	Divorced	95K	Yes
8	No	Single	85K	Yes
10	No	Single	90K	Yes

Tid	Home owner	Marital Status	Annual Income	Defaulted?
3	No	Single	70K	No

Tid	Home owner	Marital Status	Annual Income	Defaulted?
5	No	Divorced	95K	Yes
8	No	Single	85K	Yes
10	No	Single	90K	Yes