Classical Machine Learning: Classification and Regression (I)

- Learn some techniques to understand your data and prepare your data for ML.
- Learn the concept, theory, toy example, and scikit-learn usage of a few interesting base classifiers.

Techniques to Understand Your Data

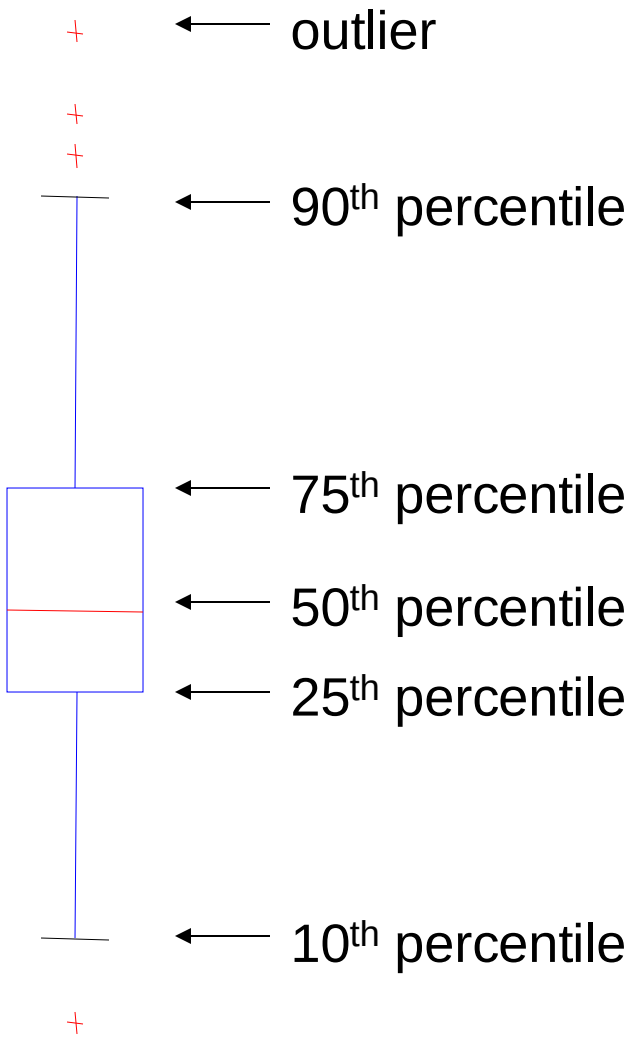
Understand Your Data with Descriptive Statistics and Visualization

Data_understand.ipynb

- Take a peek at your raw data.
- Review the dimensions of your dataset.
- Review the data types of attributes in your data.
- Summarize the distribution of instances across classes in your dataset.
- Summarize your data using descriptive statistics.
- Understand the relationships in your data using correlations.
- Review the skew of the distributions of each attribute.

Visualization Techniques: Box Plots

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data



Prepare your data for machine learning

C-S David Chen, Department of Civil Engineering, National Taiwan University

Data Preparation

Data_prepare.ipynb

- Rescale data.
- Standardize data.
- Normalize data.
- Binarize data.

Scikit-Learn Recipe

- Load the data.
- Split the dataset into the input feature matrix and output target vector for machine learning.
- Apply a pre-processing transform to the input variables.
- Summarize the data to show the change.

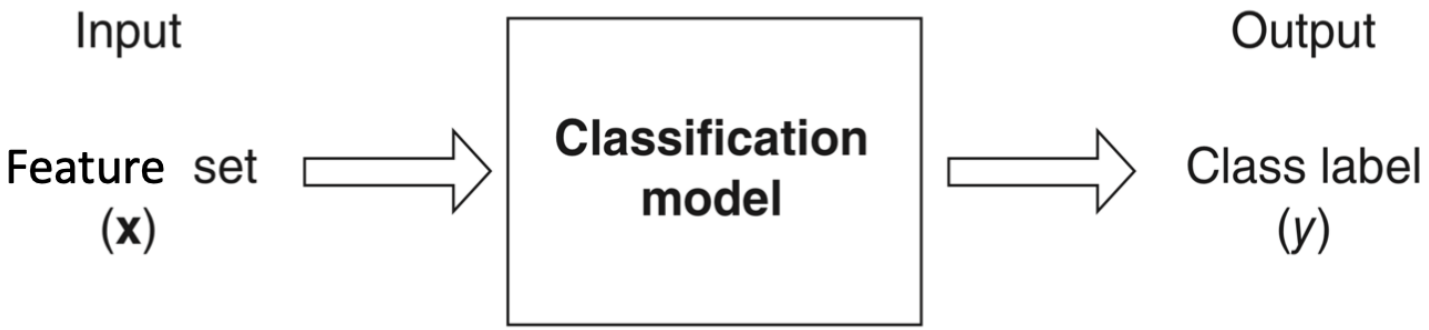
Classification algorithm walkthrough

C-S David Chen, Department of Civil Engineering, National Taiwan University

Classification

Classification uses models called classifiers to predict categorical (discrete, unordered) class labels.

Task	Feature set, **x** (or attribute set)	Class label, *y*
Spam filtering	Features extracted from email message header and content	spam or non-spam
Tumor identification	Features extracted from MRI scans	malignant or benign
Bridge warning	Features extracted from river velocity and depth	danger or safe



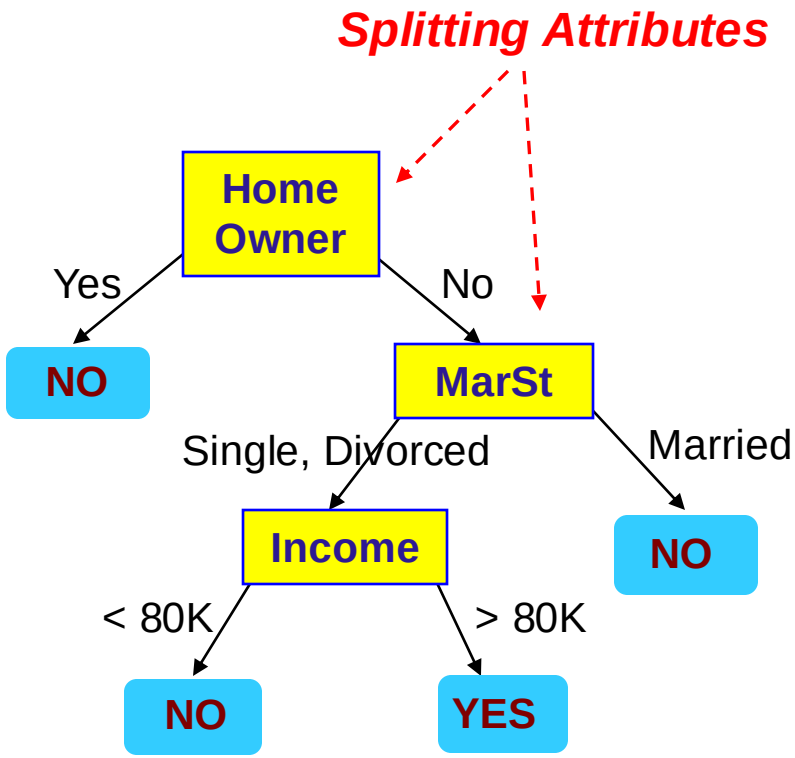
Base Classifier: Decision Tree

C-S David Chen, Department of Civil Engineering, National Taiwan University

Example of a Decision Tree

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	categorical	categorical	continuous	class
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

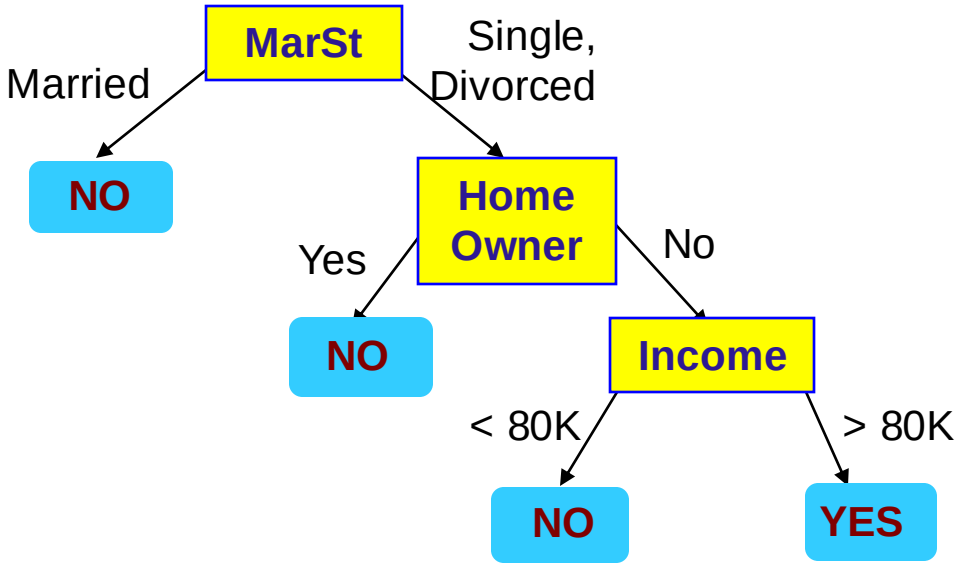
Training Data



Model: Decision Tree

Another Example of Decision Tree

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
		categorical	categorical	continuous	class
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



There could be more than one tree that fits the same data!

Decision Tree Induction

- Many Greedy Algorithms:
  - Hunt's Algorithm (one of the earliest)
  - CART
  - ID3, C4.5
  - SLIQ, SPRINT

A greedy algorithm is an approach for solving a problem by selecting the best option available at the moment. It doesn't worry whether the current best result will bring the overall optimal result.

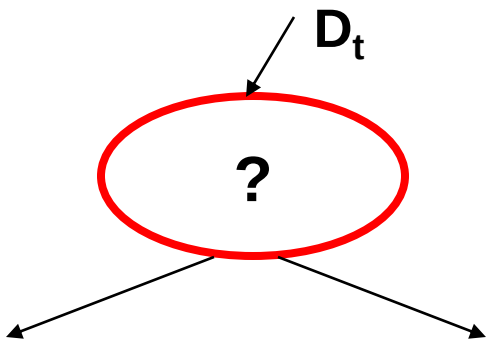
https://en.wikipedia.org/wiki/Greedy_algorithm

General Structure of Hunt's Algorithm

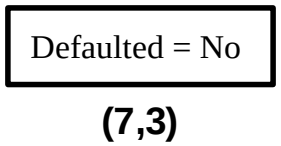
- Let D_t be the set of training records that reach a node t

- 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 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.

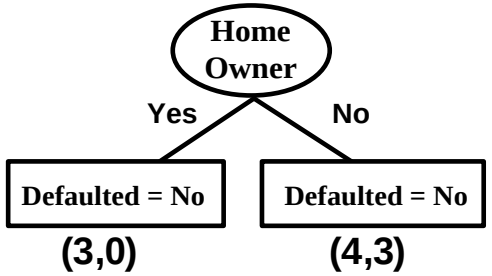
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



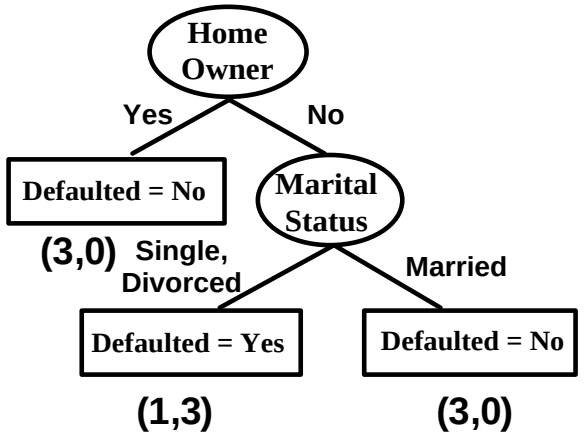
Hunt's Algorithm



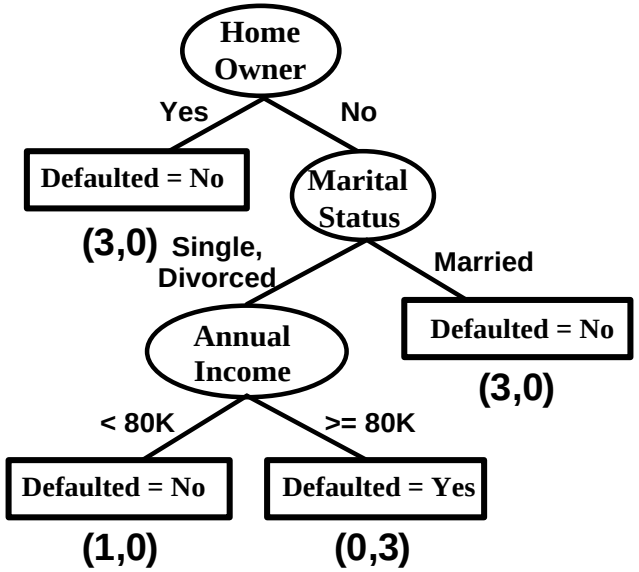
(a)



(b)



(c)



(d)

ID	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