Classification

UE 141 Spring 2013

Jing Gao
SUNY Buffalo

Classification

features **class labels**

patient	temp.	blood pres.	heart rate	Sick?
	99	110	90	Yes
	100	120	100	Yes
	96	130	65	No

labeled

training

a model: f(x)=y: features → class labels

patient	temp.	blood pres.	heart rate	Sick?
	98	130	80	
	115	110	95	

test

unlabeled

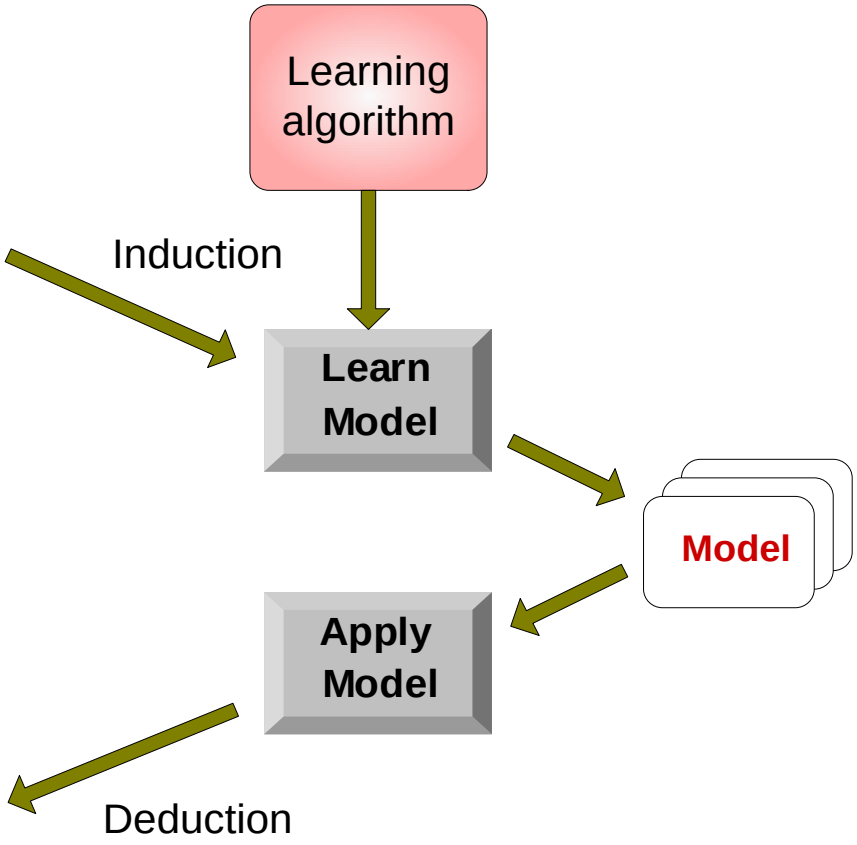
Illustrating Classification Task

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



Classification Techniques

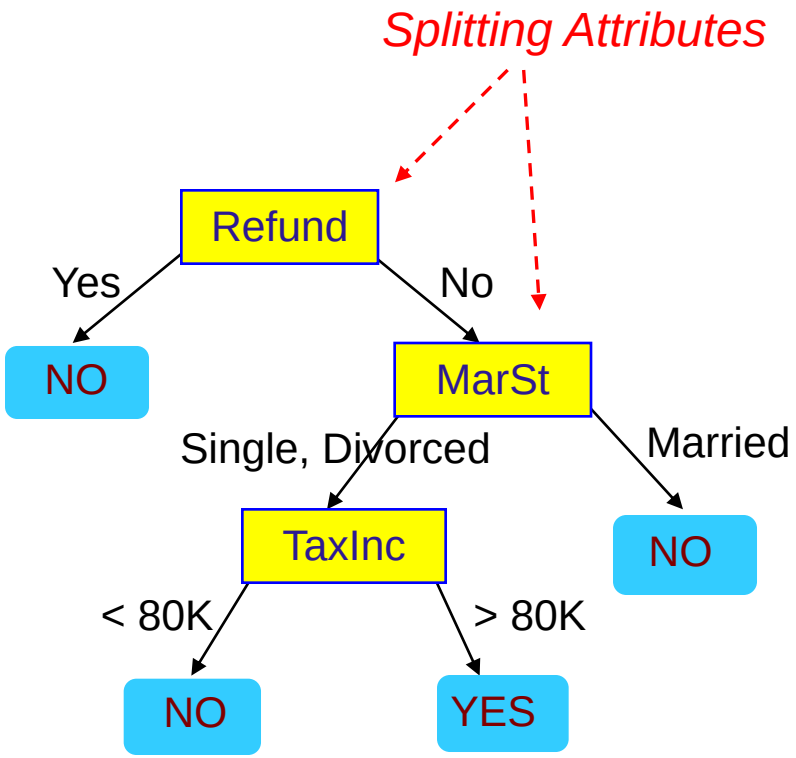
- Decision Tree
- Naïve Bayes
- Logistic Regression
- Support Vector Machines
- K nearest neighbor
- Ensemble learning
- ……

Example of a Decision Tree

class

Tid	Refund	Marital Status	Taxable Income	Cheat
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

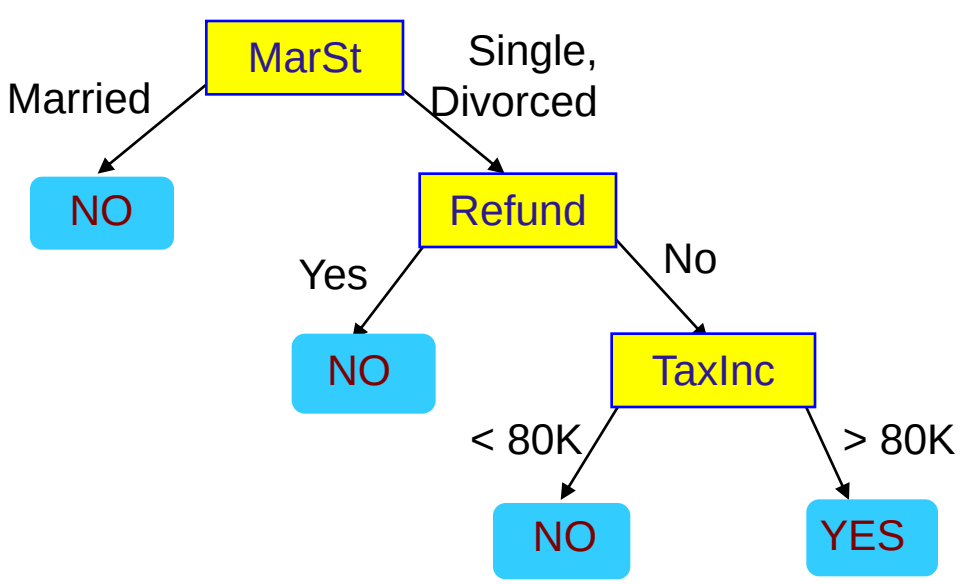
Training Data



Model: Decision Tree

Another Example of Decision Tree

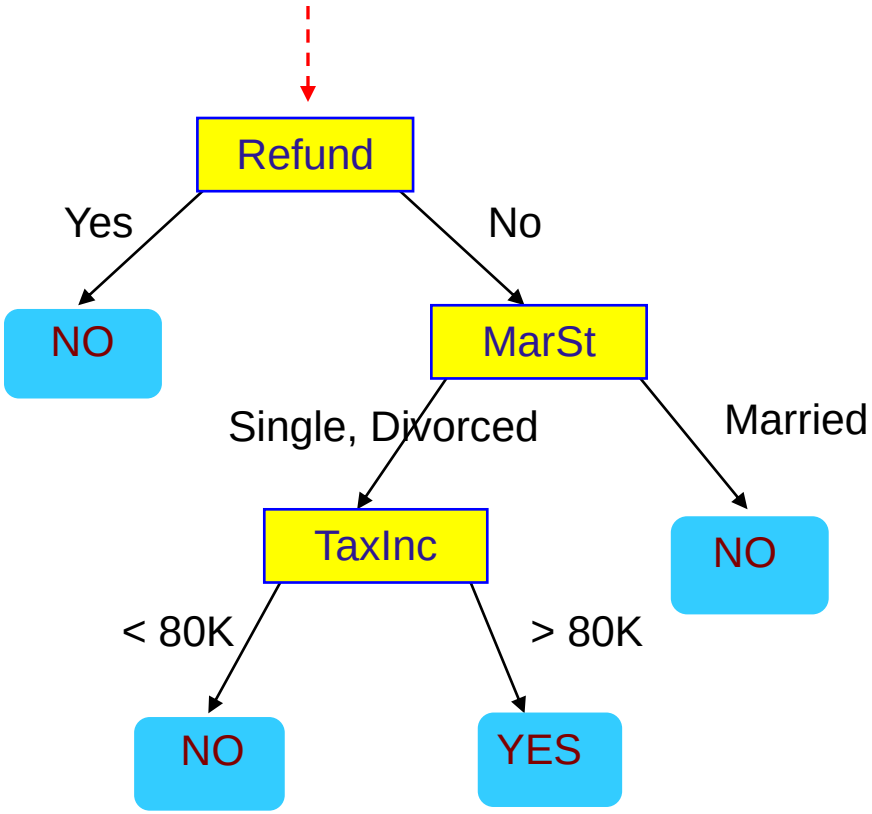
Tid	Refund	Marital Status	Taxable Income	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



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

Apply Model to Test Data

Start from the root of tree.



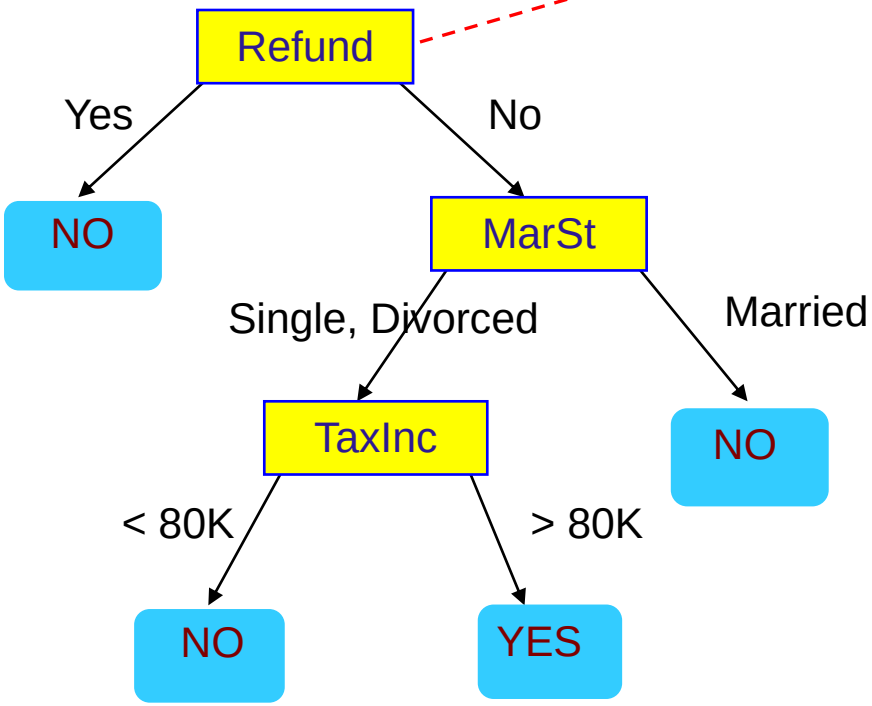
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

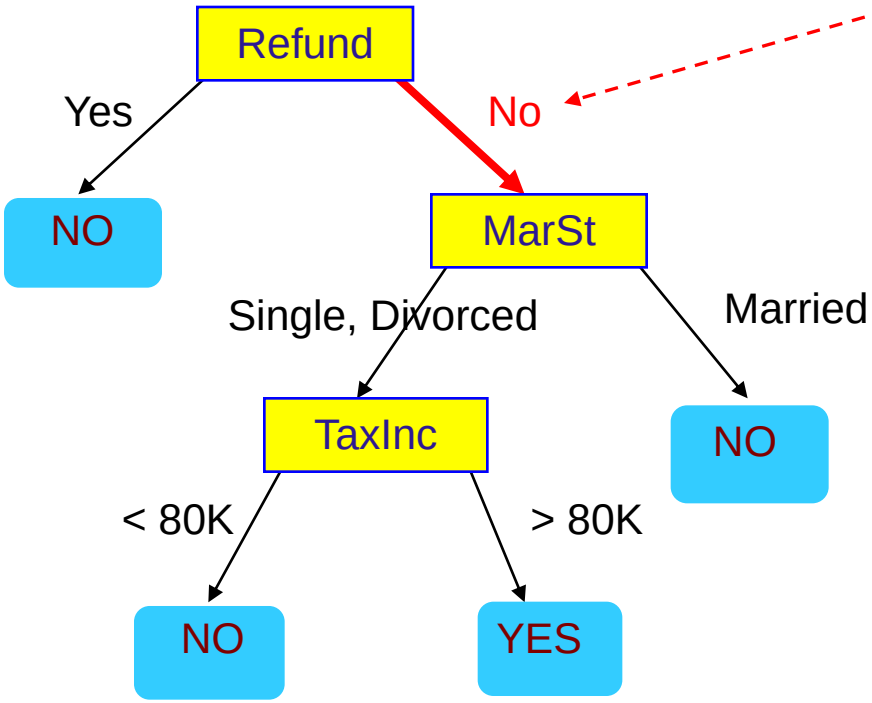
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

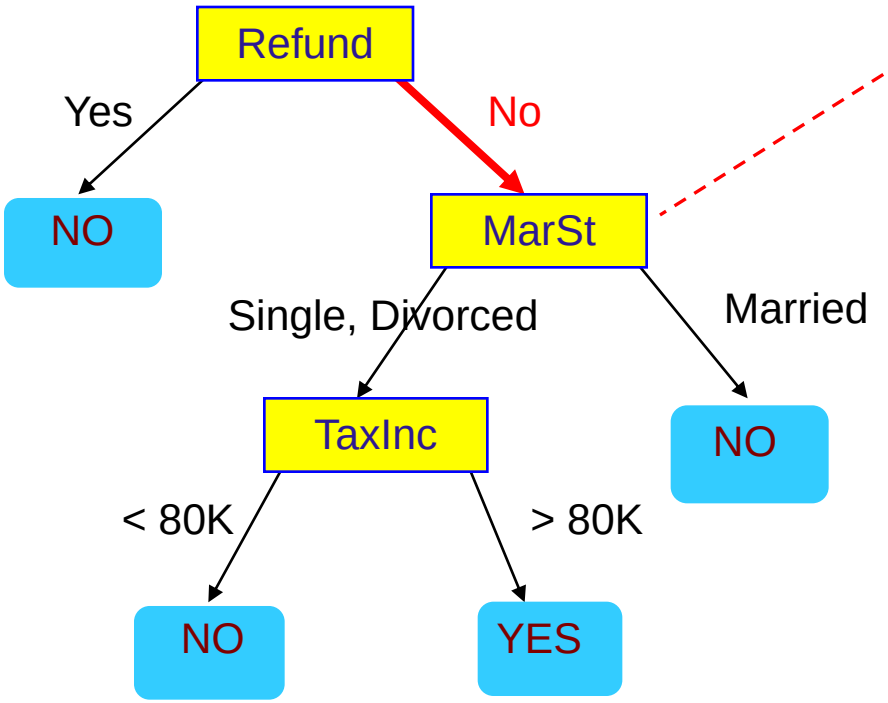
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

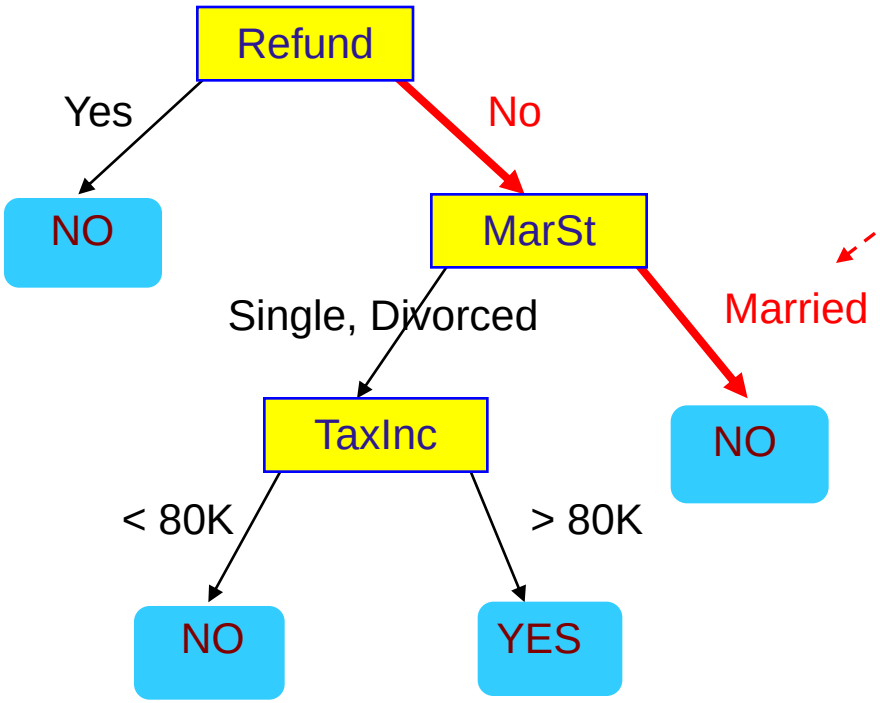
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

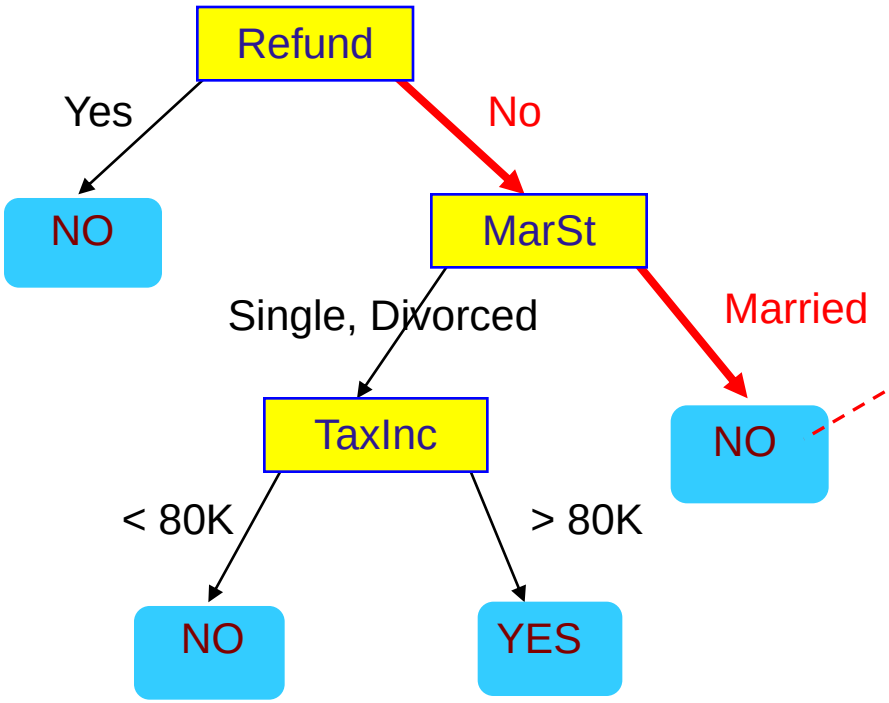
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

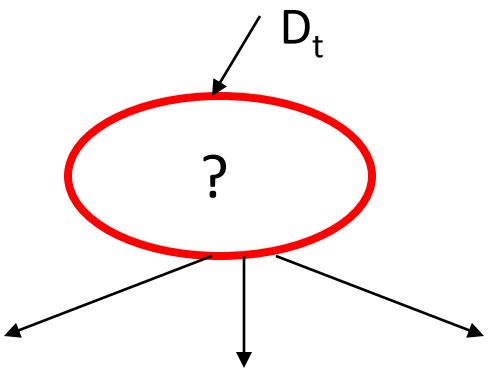


Assign Cheat to “No”

Build a Decision Tree

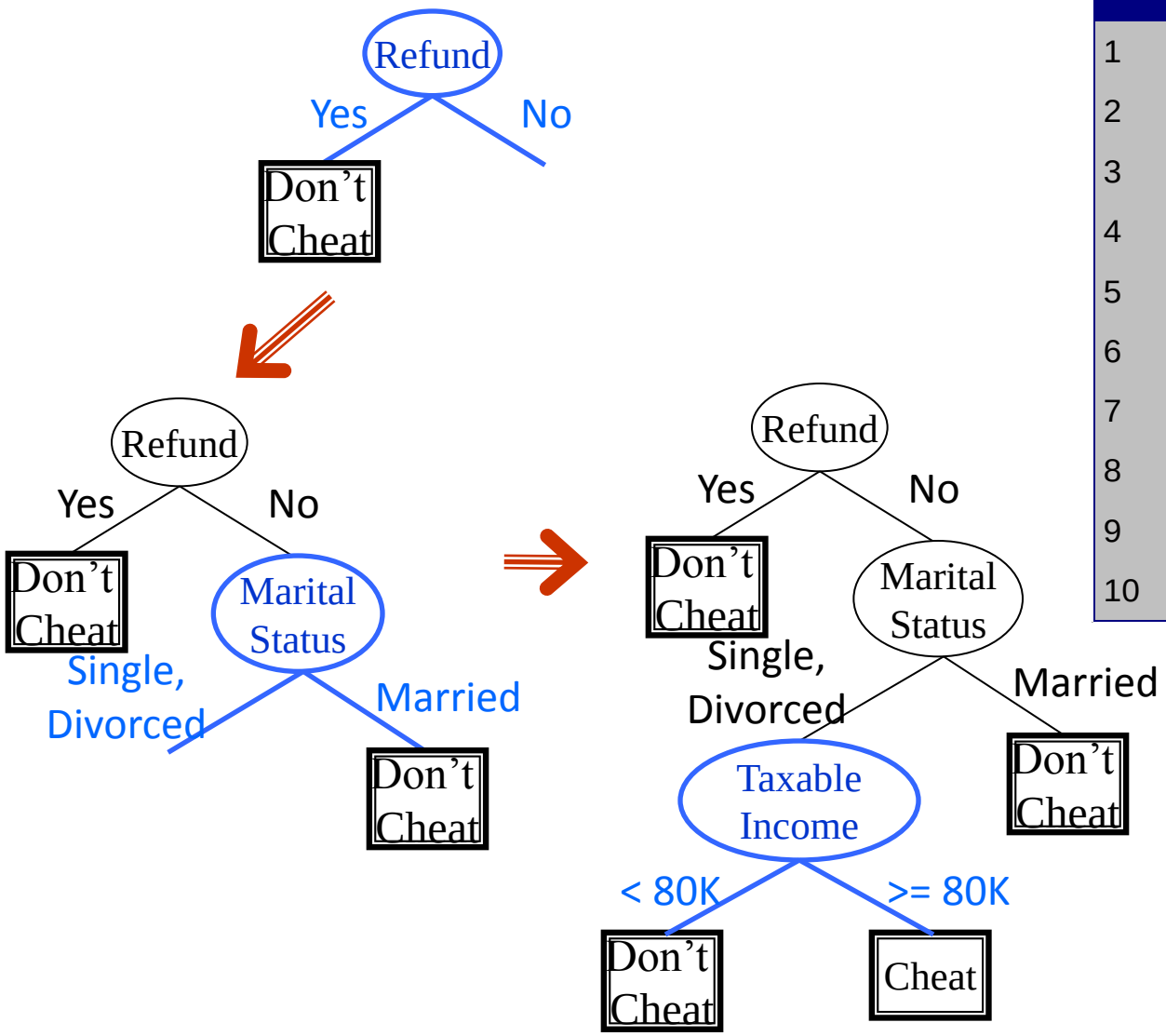
- Let D_t be the set of training records that reach a node t
- **General Procedure:**
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t contains records that belong to more than one class, use an attribute to split the data into smaller subsets.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

Tid	Refund	Marital Status	Taxable Income	Cheat
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



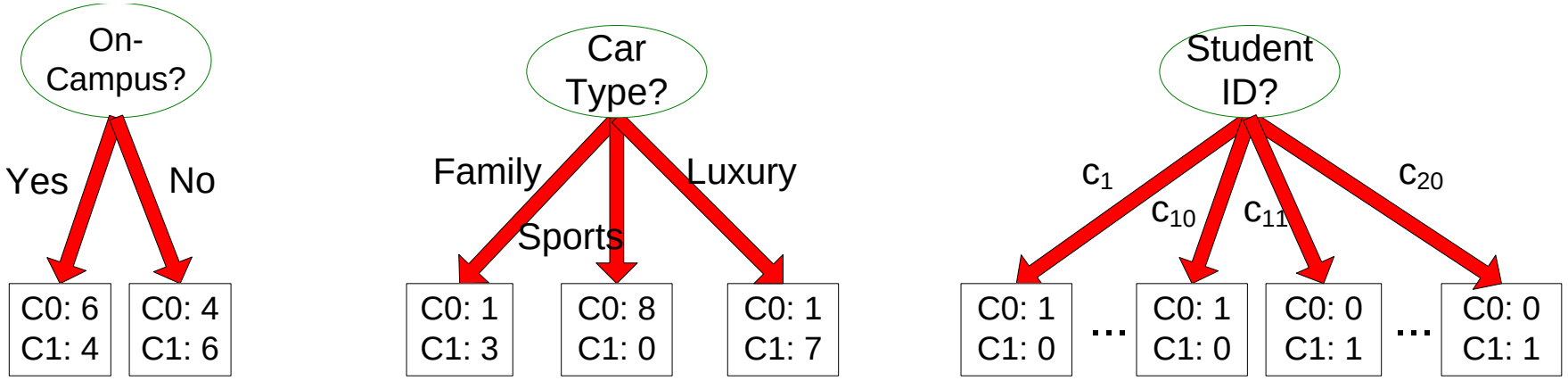
Example

Tid	Refund	Marital Status	Taxable Income	Cheat
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



How to determine the Best Split

Before Splitting: 10 records of class 0,
10 records of class 1



Which test condition is the best?

Classification in Experiments

training

patient	temp.	blood pres.	heart rate	Sick?
	99	110	90	Yes
	100	120	100	Yes
	96	130	65	No

test

patient	temp.	blood pres.	heart rate	Sick?
	98	130	80	Yes(predicted) No (true)
	115	110	95	Yes (predicted) Yes (true)

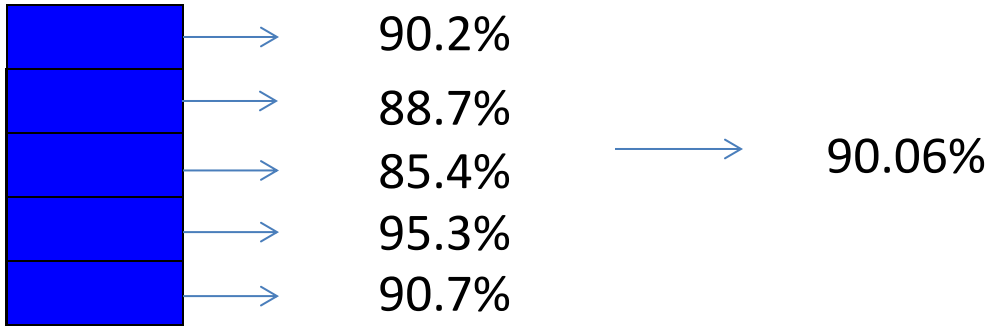
Metrics for Performance Evaluation

$$\text{Accuracy} = \frac{\#\,correct\ predictions}{\#\,total\ predictions}$$

Example. The test set includes 100 examples. When we use the model to predict the class labels of 100 examples, 90 of them are the same as the true labels, and 10 make mistakes. Then the accuracy of this model is 90%.

k-fold Cross Validation

Partition the data set into *k* disjoint subsets. Each time use *k*-1 subsets as training, and the remaining 1 as test. Repeat this process *k* times and calculate the average classification accuracy.



Typically k=10—10-fold cross validation

Lazy vs. Eager Learning

- **Lazy vs. eager learning**
  - **Lazy learning**: Simply stores training data and waits until it is given a test record (one-step)
  - **Eager learning**: Given a set of training records, constructs a classification model before predicting on test data (two-step)
- **Comparison**
  - The prediction phase of eager learning is usually short and most of the time is spent on training, instead, lazy learning does not have a training phase, but prediction time is long
  - Lazy learning focuses on “local” behavior while eager learning builds a “global” model

Nearest Neighbor Classifiers

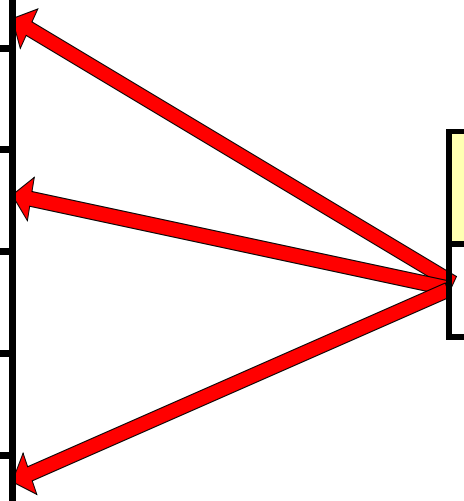
- Store the training records
- Use training records to predict the class label of unseen cases

Set of Stored Cases

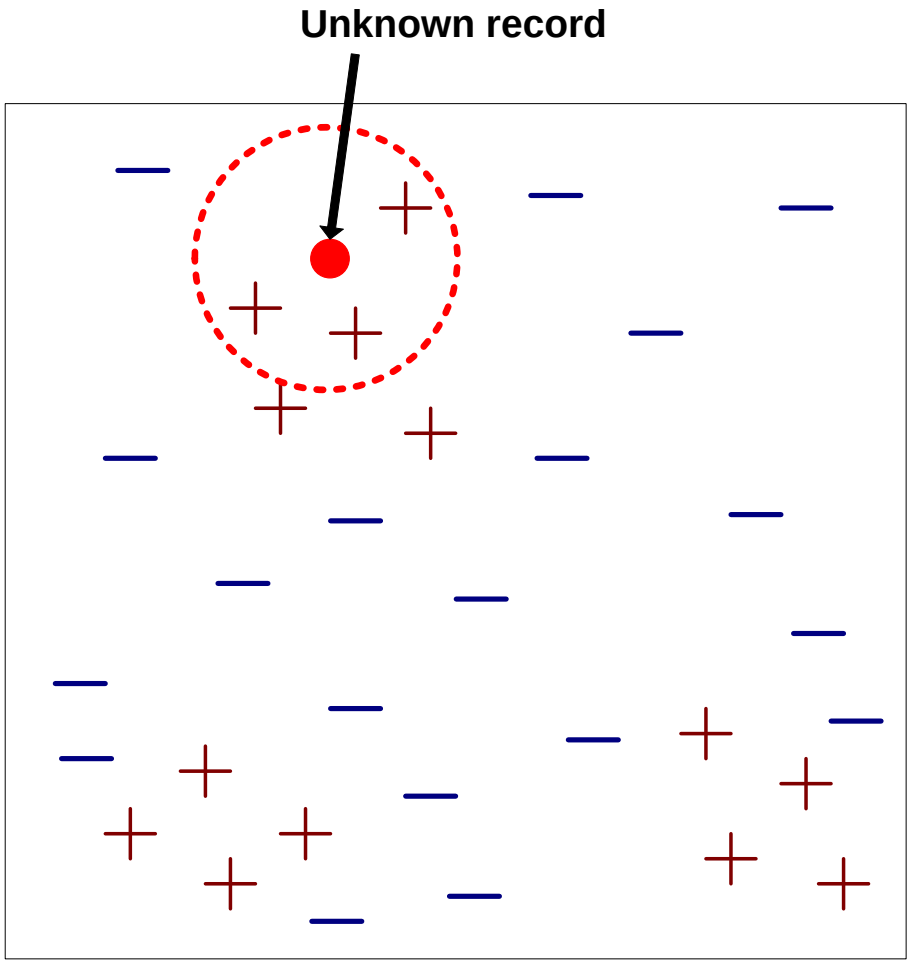
Atr1	……	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

Unseen Case

Atr1	……	AtrN

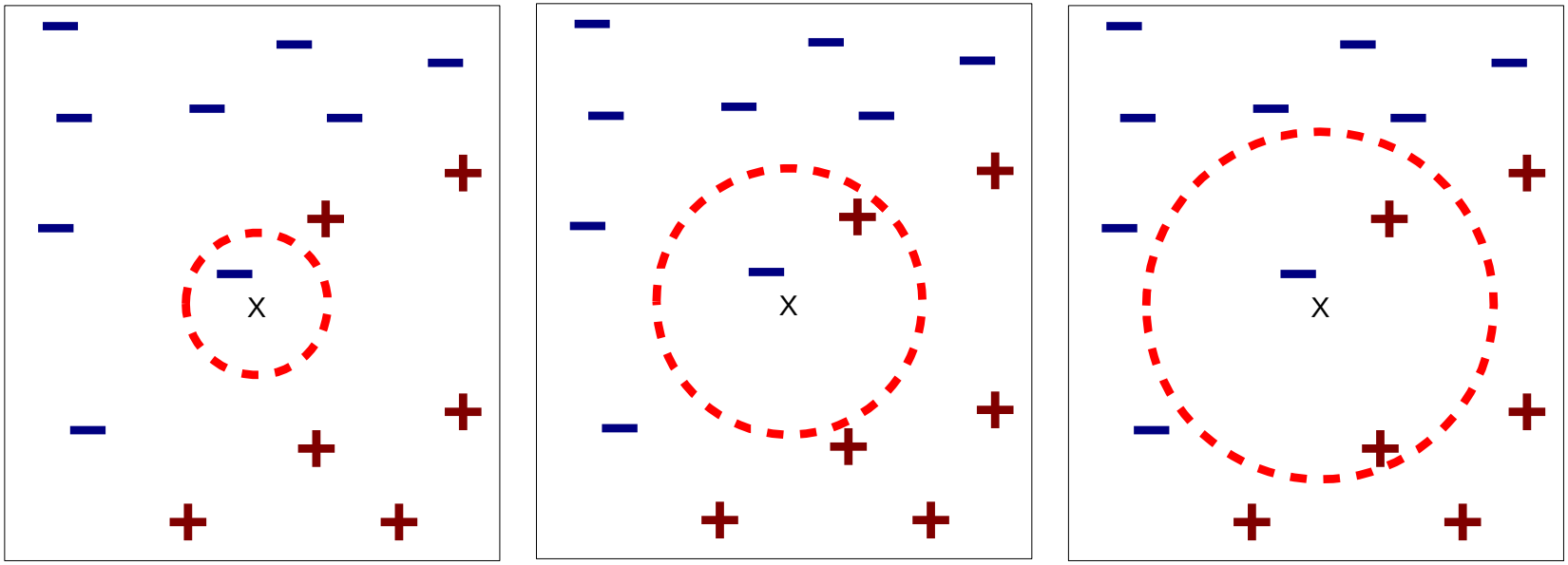


Nearest-Neighbor Classifiers



- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve
- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

Definition of Nearest Neighbor

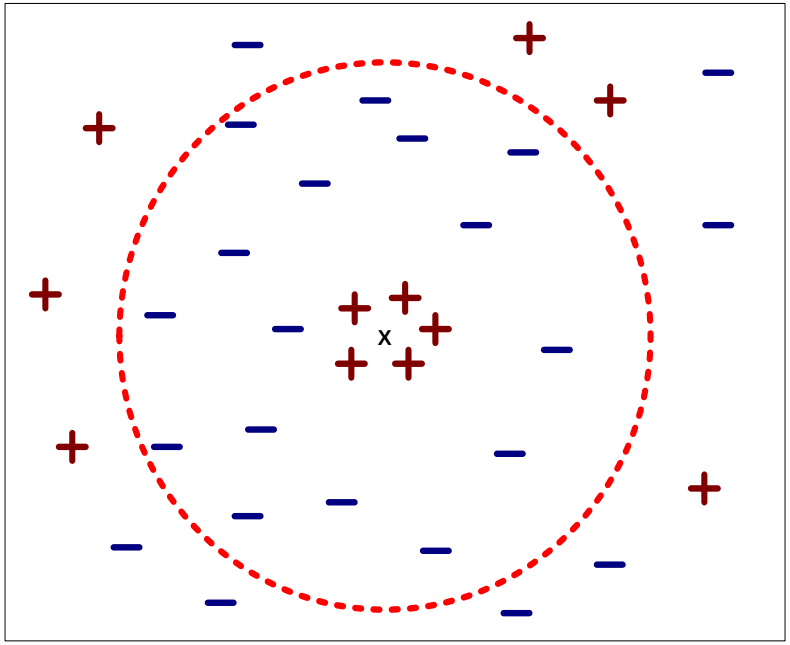


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

Nearest Neighbor Classification

- **Choosing the value of k:**
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes



Question

- **Data mining tasks are not independent. For example, we can use association analysis to conduct classification.**
  - The question is: How can we do that?
  - The topic is called rule-based classification.