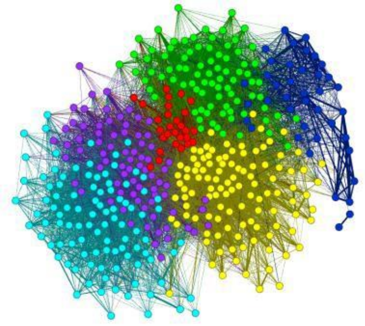
IST 736

Classification



What is a model?

- An attempt to **represent reality** through a particular lens.

- An **artificial construct** that does not contain unnecessary detail and makes a set of assumptions.

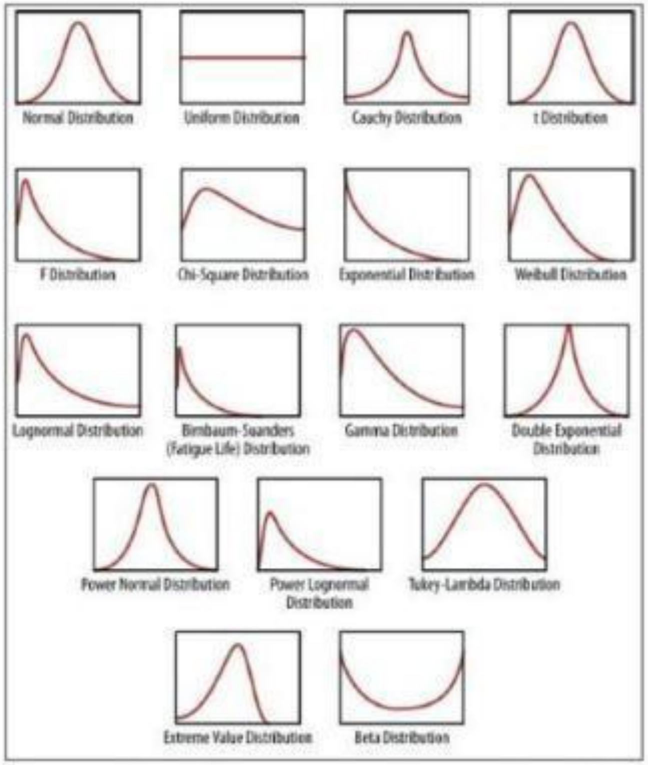
Statistical Modeling

- A way to express a **model using mathematics**
- The model designer makes an assumption about the **generative process** of the data
- The goal is to **estimate the parameters** of the model given a particular data set
- A **level of confidence** is always given for the model, e.g. confidence intervals

Statistical modeling questions

- What is the process that generated the data?
- What happened first?
- What influences what?
- What causes what?
- How canbI test these?

Common Distributions



- Basis for statistical models
- Natural processes generate **"shapes" of distributions** that can often be approximated by a mathematical function, given a few parameters that are estimated using the data.
- **Not all processes generate data that looks like a named distribution**.

From book: Doing Data Science

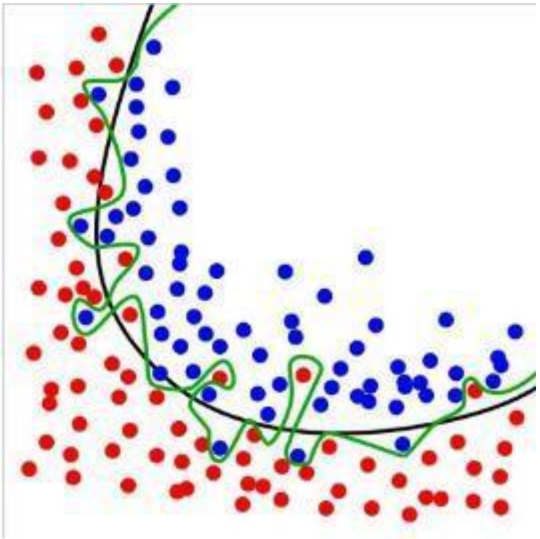
Deciding on a model to use

- Conduct **exploratory analysis**
- Develop a **hypothesis** to test
  - **Try a linear function first – why?**
    - Write down assumptions
    - Does this make sense?
  - If necessary, begin looking at more sophisticated models
    - Write assumptions
    - Does this make sense?

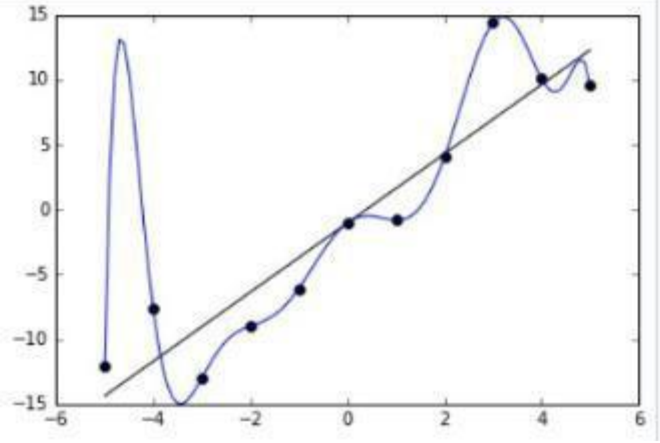
Fitting a Model

- When you **fit a model**, you **estimate its parameters** using real world collected data (samples).
- Fitting a model often requires **optimization techniques** and **algorithms**.
- **Over fitting** is a common problem that needs to be avoided.
  - Can end up describing random error or noise rather than the underlying distribution.
  - Can occur when a model is too complex.

Avoid testing and training using the same or overlapping data.



Visual Examples of overfitting



Images borrowed from: https://en.wikipedia.org/wiki/Overfitting

Learning Styles

Supervised Learning

- ***Labeled input*** data exist to train a model. The model is then used to predict the class on unseen data.

Unsupervised Learning

- Input data are ***not labeled*** and the result is not known.

Semi-supervised Learning

- Input data is a ***mix*** of labeled and unlabeled examples.

Reinforcement Learning

- A model that **interacts with and learns from its environment.**
- Feedback is provided as ***punishments and rewards in the environment.***

Supervised Examples:
Regression, Decision Tree, Random Forest, KNN, Logistic Regression, Naive Bayes, Support Vector Machines, Neural Networks

Unsupervised Examples:
kmeans clustering, Association Rules

Reinforcement Learning Examples:
Q-Learning, Temporal Difference (TD), Deep Adversarial Networks

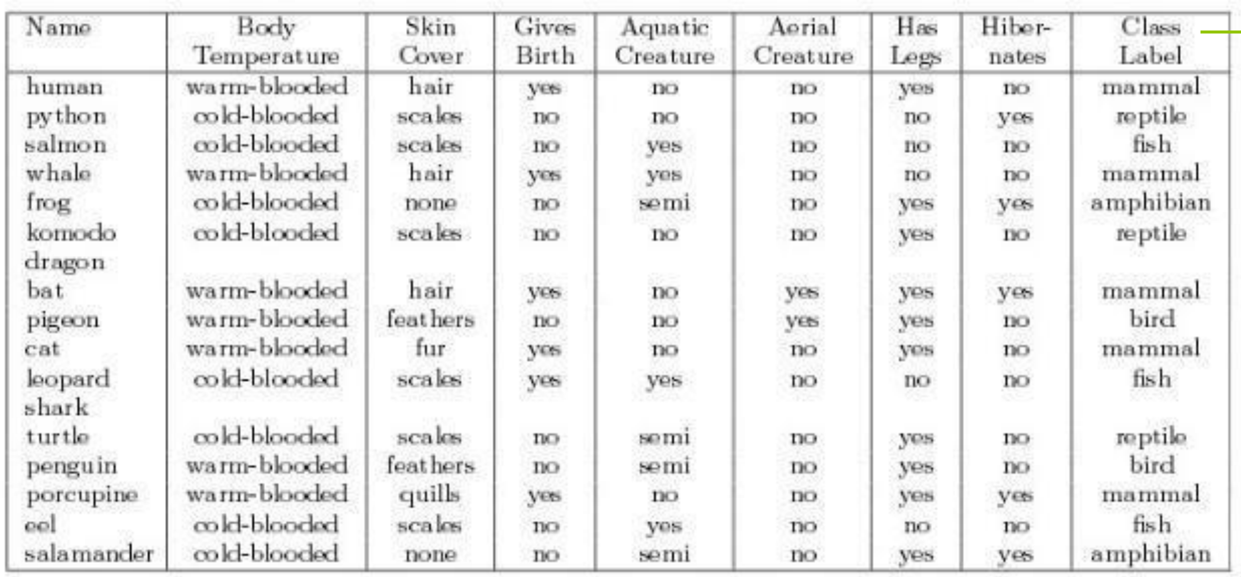
Interesting References:
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

What is Classification

Given: a collection of records/vectors (*training dataset)* Each record contains a set of *attributes (variable values)*, **one of the attributes must be the *class*.**

Goal: Find a *model* (some function of the variable values) to identify the **class of a new vector/record.**

Table 4.1. The vertebrate data set.



Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

CLASS

Cross-validation

- A *test set* is used to determine the accuracy of the model.
- Usually, the given data set is **divided into training and test sets**, with training sets used to build the model and the test set used to validate it.
- Training sets and testing sets should not overlap in values.
- **Cross-validation** (leave-one-out) is often used.

Concepts for ML Classification

Input data: collection of records (also called an instance or example).

- For example: tuple(**x**, *y*), where **x** is the set (vector) of known attributes (variable values) and *y* is the **class label (called the target)**.

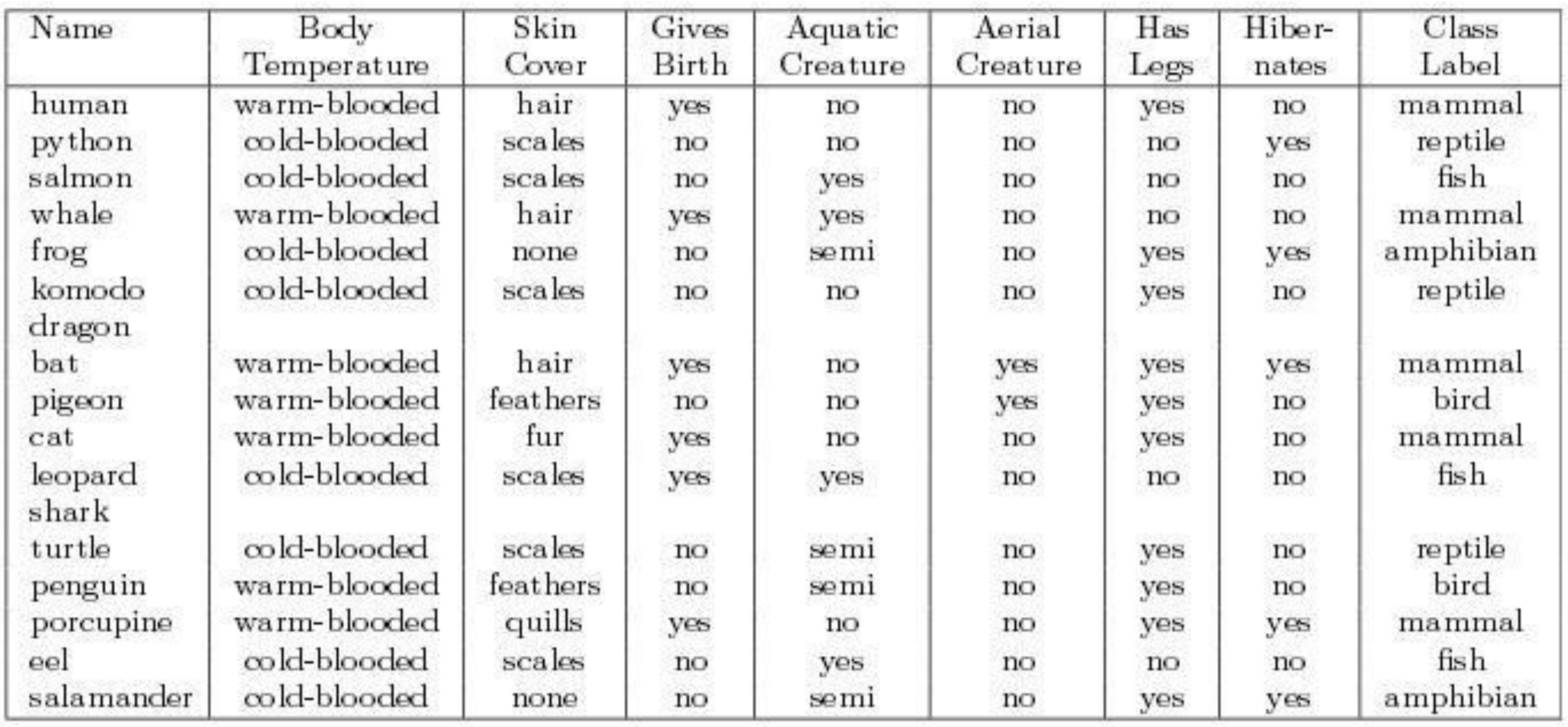
Classification: learning a target/class **function** f that maps any vector of attributes, **x** to a predefined class y.

f: **x** → y <u>f is a classification model</u>

- **Descriptive Modeling**: Classification model that can distinguish between objects of different classes.
- **Predictive Modeling**: Using a classification model to predict a label/class given a vector/record **x**

Example: Feature Table

Table 4.1. The vertebrate data set.



Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

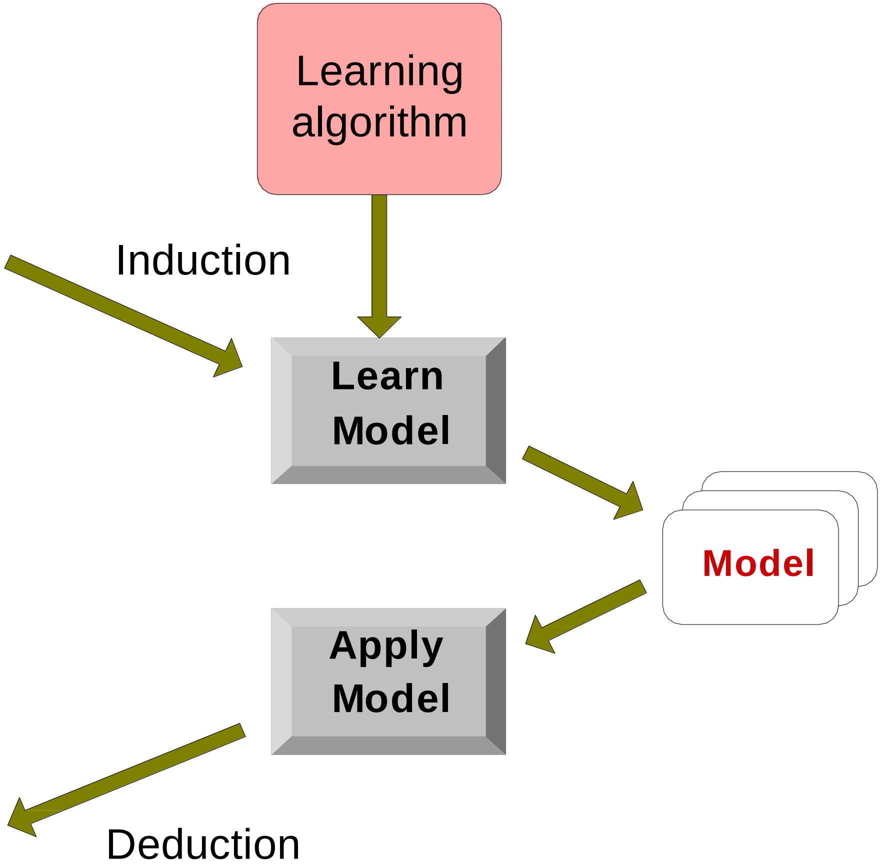
http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf, page 147

1. **What are the classes, y?**
2. **What are the records, x?**

Illustrating A 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



Learning algorithm

Induction

Learn Model

Model

Apply Model

Deduction

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

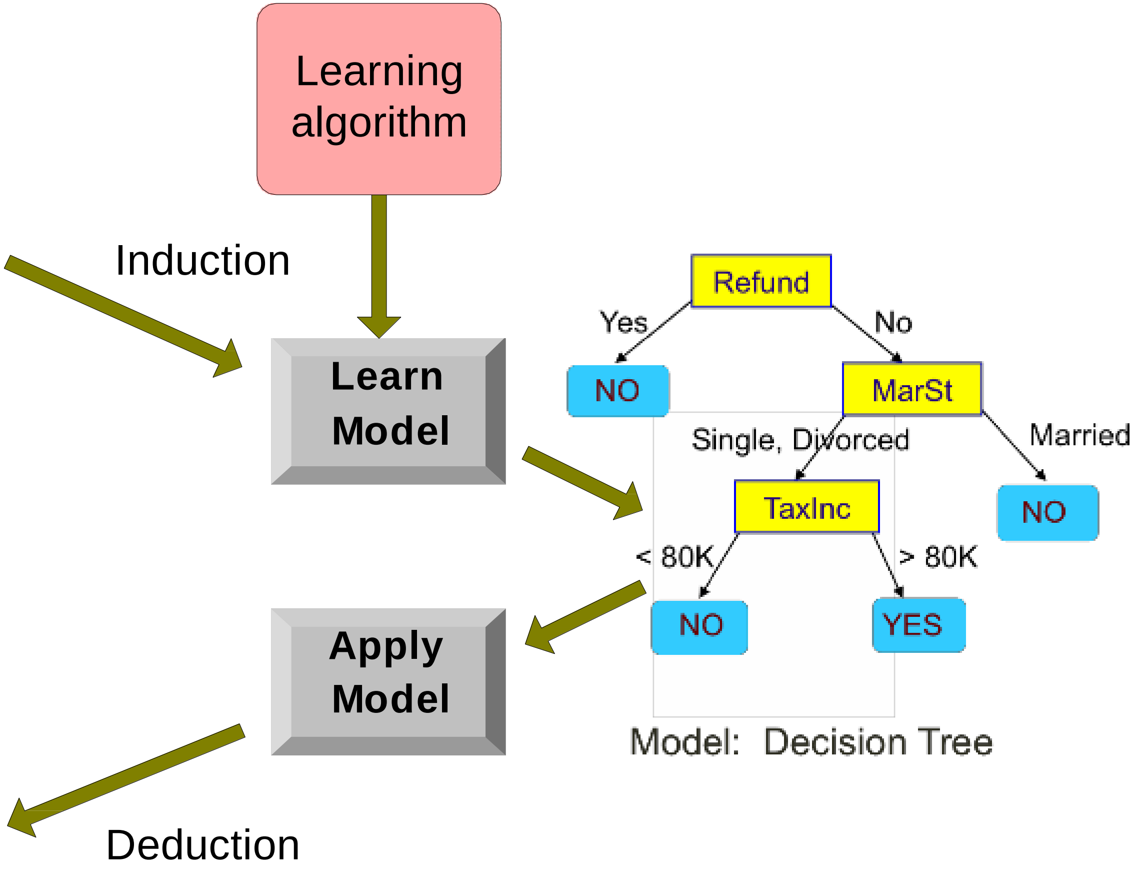
EXAMPLE: IsSomeone Cheating on Their Taxes?

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



Learning algorithm

Induction

Learn Model

Apply Model

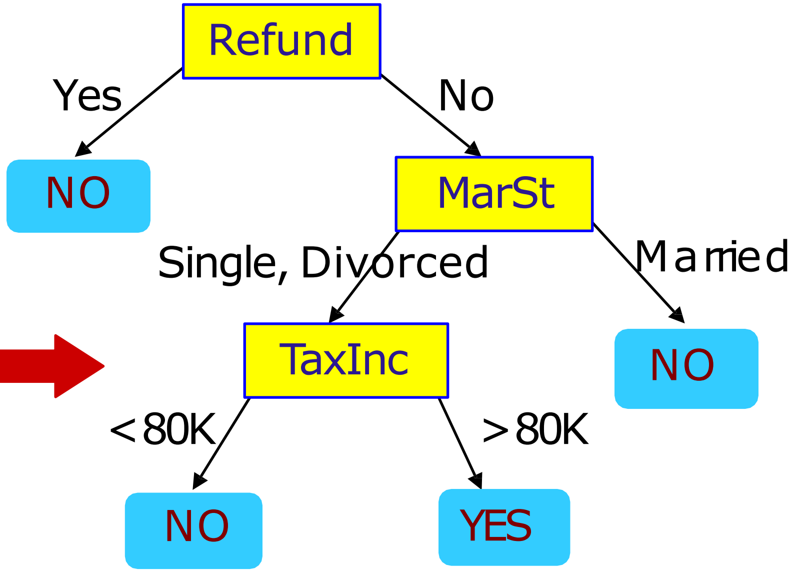
Deduction

Model: Decision Tree

Example of Training Data and Decision Tree Model

	categorical	categorical	continuous	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**

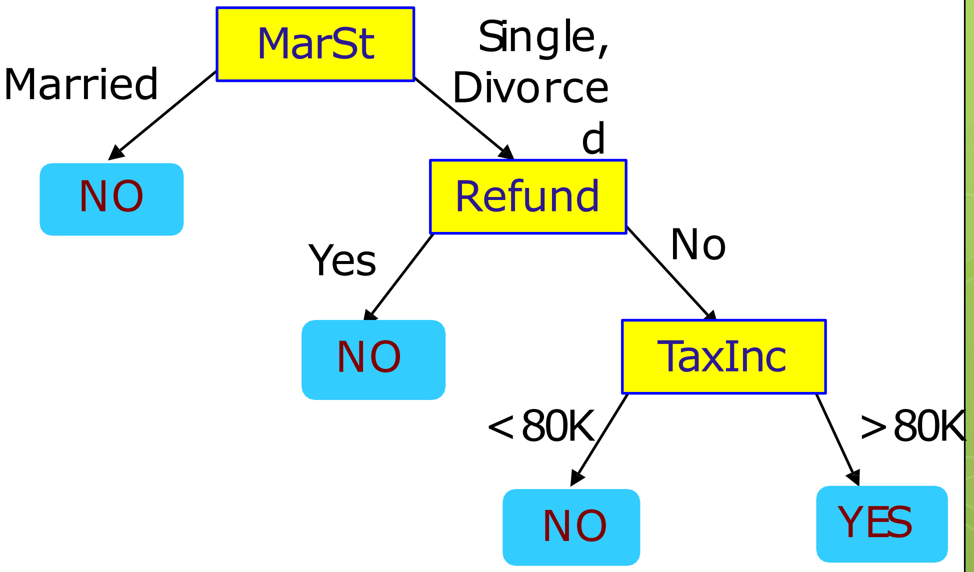
Training Data



Model: Decision Tree

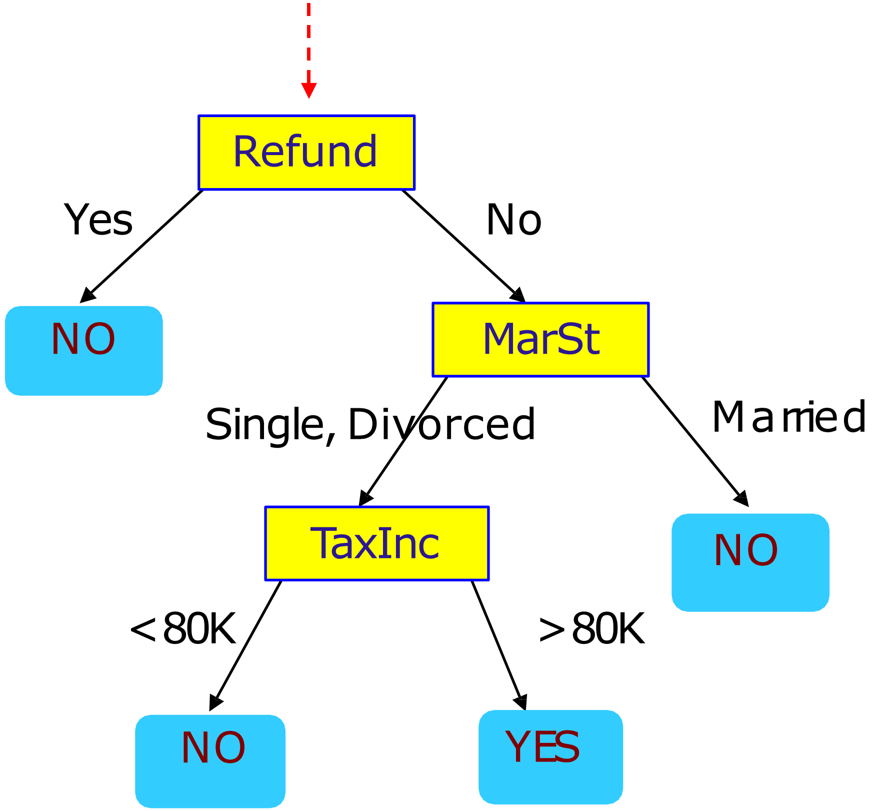
Another Example of Decision Tree – there are infinite tree options

	categorical	categorical	continuous	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**



Apply Model to Test Data

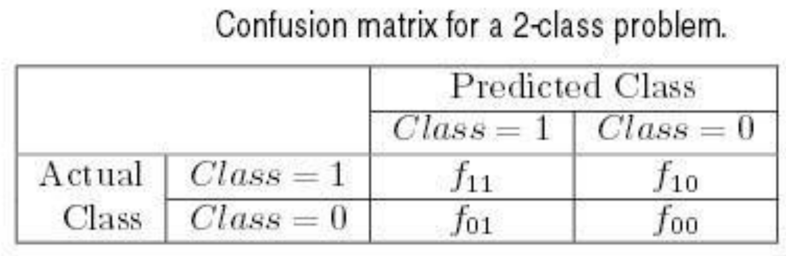
Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Performance Evaluation



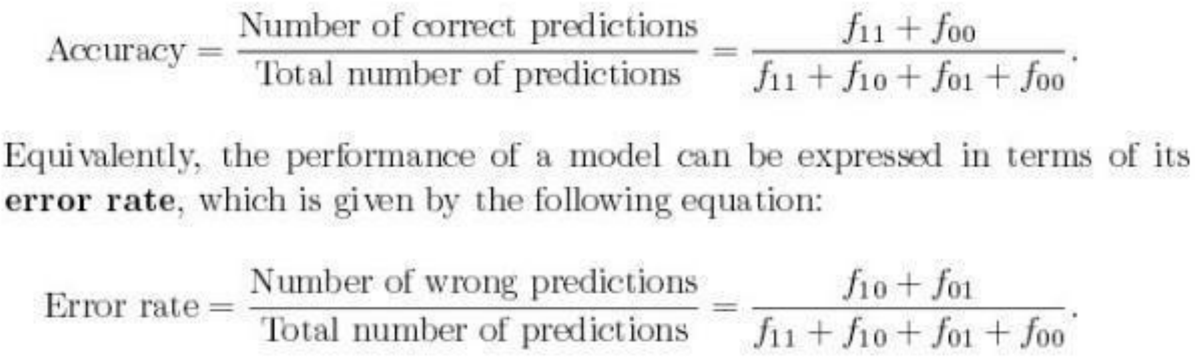
Confusion matrix for a 2-class problem.

		Predicted Class	
		$Class = 1$	$Class = 0$
Actual Class	$Class = 1$	f_{11}	f_{10}
	$Class = 0$	f_{01}	f_{00}

Total Num Correct = **f_{11} + f_{00}**

The performance of a classification model can be based on **counts** of test records **correctly** or **incorrectly** predicted.

f_{11}: Record was class 1 and was predicted as class 1 correctly
f_{01}: Record was class 0 and incorrectly predicted as Class 1



$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

Equivalently, the performance of a model can be expressed in terms of its **error rate**, which is given by the following equation:

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf, page 5

Metrics for Performance Evaluation: Confusion Matrix

Confusion Matrix:

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

		PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	True Positive	False Negative
	Class=No	False Positive	True Negative

Is accuracy always a good measure?

Can you think of an example when it is not?

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Example when Accuracy is not a good measure:

Consider a 2-class problem

- Number of Class 0 examples = 9990
- Number of Class 1 examples = 10

If the model predicts everything to be in class 0, accuracy is 9990/10000 = 99.9 %

- Accuracy is misleading because model does not detect any class 1 examples.

Using a Cost Matrix

	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of **misclassifying** class j, as class i

EXAMPLE: Computing Cost of Classification

This the actual prediction from the model

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

This is the Cost Matrix

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Accuracy

=(150+ 250)/
(150+40+60+250) = **80%**

Cost

=(150)(-1) +
(40)(100) +
(60)(1) +
(250)(0) = **3910**

Cost vs Accuracy

Count		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

Cost		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	**p.**	**q**
	Class=No	**q.**	**p**

Accuracy is proportional to cost if

Proof:

1. $C(Yes|No)=C(No|Yes) = q$
2. $C(Yes|Yes)=C(No|No) = p$

$N = a + b + c + d$

$\rightarrow b + c = N - a - d$

$\text{Accuracy} = (a + d)/N$

$$
\begin{aligned}
\text{Cost} &= p\,(a + d) + q\,(b + c) \\
&= p\,(a + d) + q\,(N - a - d) \\
&= p\,(a + d) + qN - q(a + d) \\
&= q\,N - (q - p)(a + d) \\
&= qN - N(q - p)\text{Accuracy} \\
&= N\,[q - (q\text{-}p)\,\text{Accuracy}]
\end{aligned}
$$

Classification Techniques

- **Decision Tree Methods**
- Instance-based Methods
- Bayesian algorithms (Naïve Bayes)
- Support Vector Machines
- Ensembles (Random Forest)