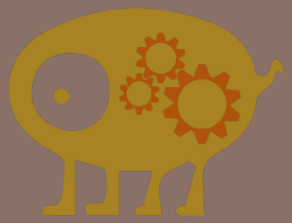


Data Mining
資料探勘

Classification

Hung-Yu Kao, Fall 2019

Big “machine learning” Data

- In Youtube
  - Query “machine learning” → 10M results
  - Query “Introduction to machine learning” → 3.3M results
- Huge amount of online courses on “Machine Learning”
  - In Coursera: 1,252 courses are related to “Machine Learning”

Machine learning

- Machine learning develops algorithms for making predictions from data
- Machine learning is no Voodoo
  - The word “predictions” can be misleading

運動大數據分析 v.s. 報明牌

Machine Learning

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

-- Tom M. Mitchell, 1997

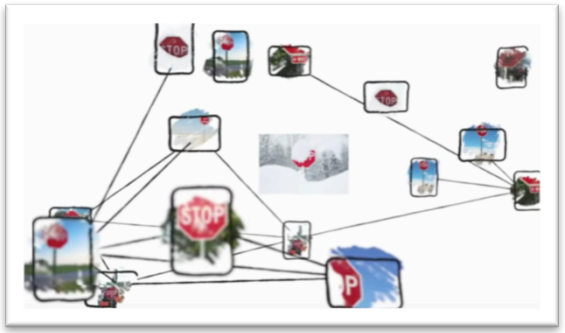
Machine Learning

- Model + Parameter + Learner

Model: makes the predictions and identifications

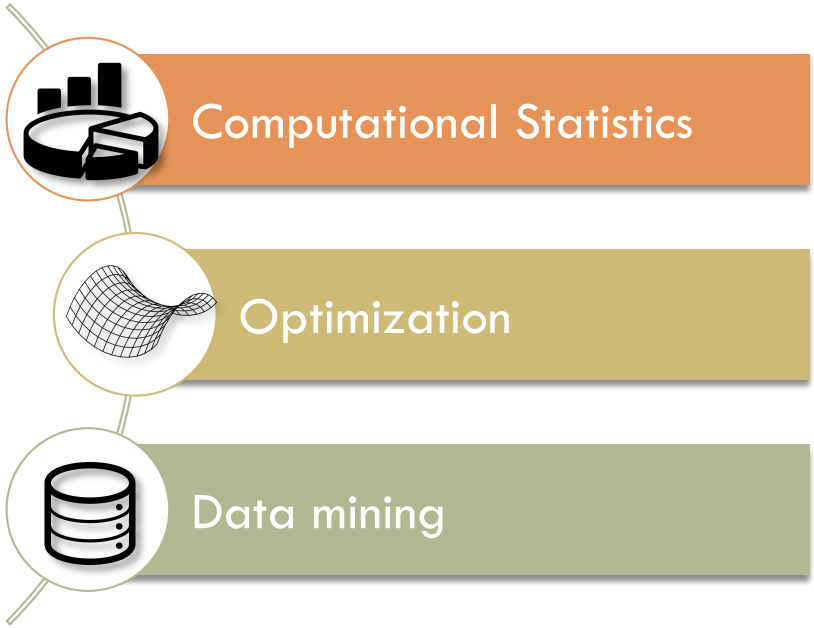
Parameter: Signals and factors that models need to make its decision

Learner: A system that adjusts the parameters by looking at differences between predictions and the actual outcomes



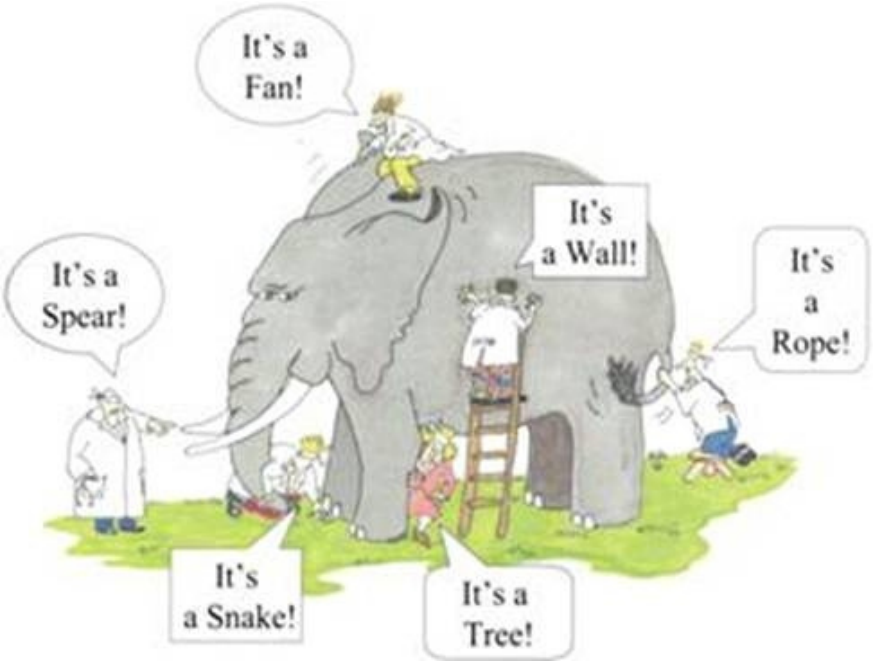
Machine Learning

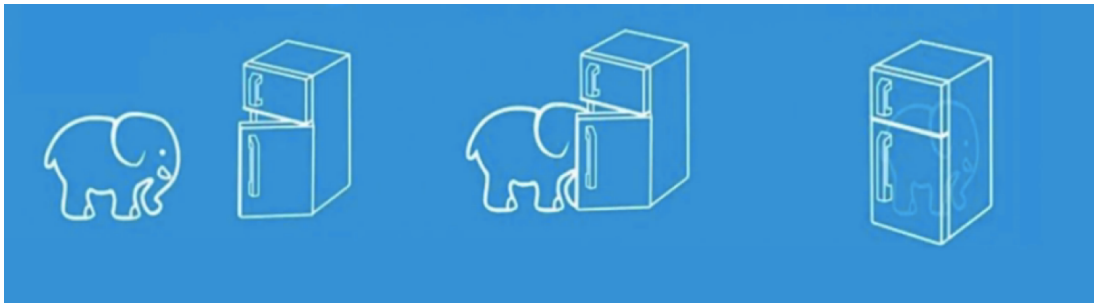
- Closely related to (overlaps with) Computational Statistics, Probability, Optimization, and Data Mining
- Type of Tasks
  - Supervised Learning
    - classification
  - Unsupervised Learning
  - Reinforcement Learning



Classification example

Memorize / Learn / Understand





Supervised vs. Unsupervised Learning

- Supervised learning (classification) 鑑往知來
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering) 看圖說故事
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - 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 set used to build/train the model and test set used to validate it.

Issues Regarding Classification and Prediction (1): Data Preparation

- Data cleaning
  - Preprocess data in order to reduce noise and handle missing values
- Relevance analysis (feature selection)
  - Remove the irrelevant or redundant attributes
- Data transformation
  - Generalize and/or normalize data

Issues regarding classification and prediction (2): Evaluating Classification Methods

- Predictive accuracy
- Speed and scalability
  - time to construct the model
  - time to use the model
- Robustness
  - handling noise and missing values
- Scalability
  - efficiency in disk-resident databases
- Interpretability:
  - understanding and insight provided by the model
- Goodness of rules
  - decision tree size
  - compactness of classification rules

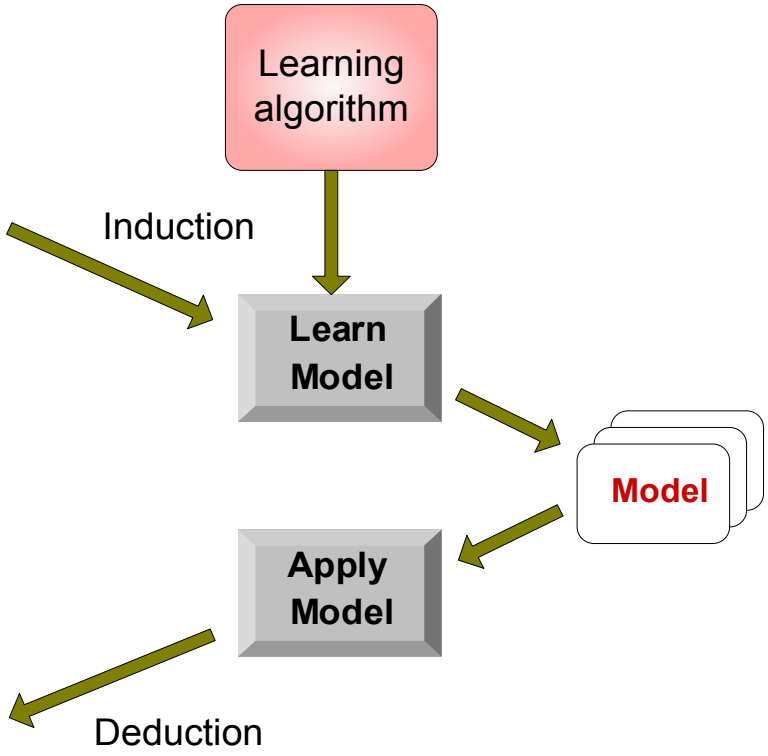
Illustrating Classification Task

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	?

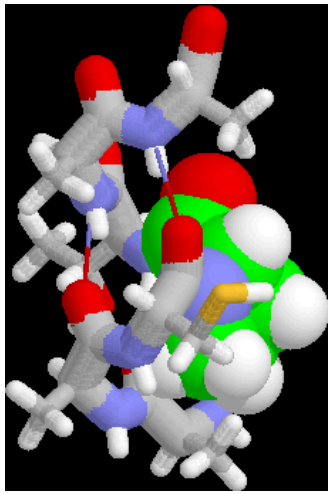
Test Set



Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent

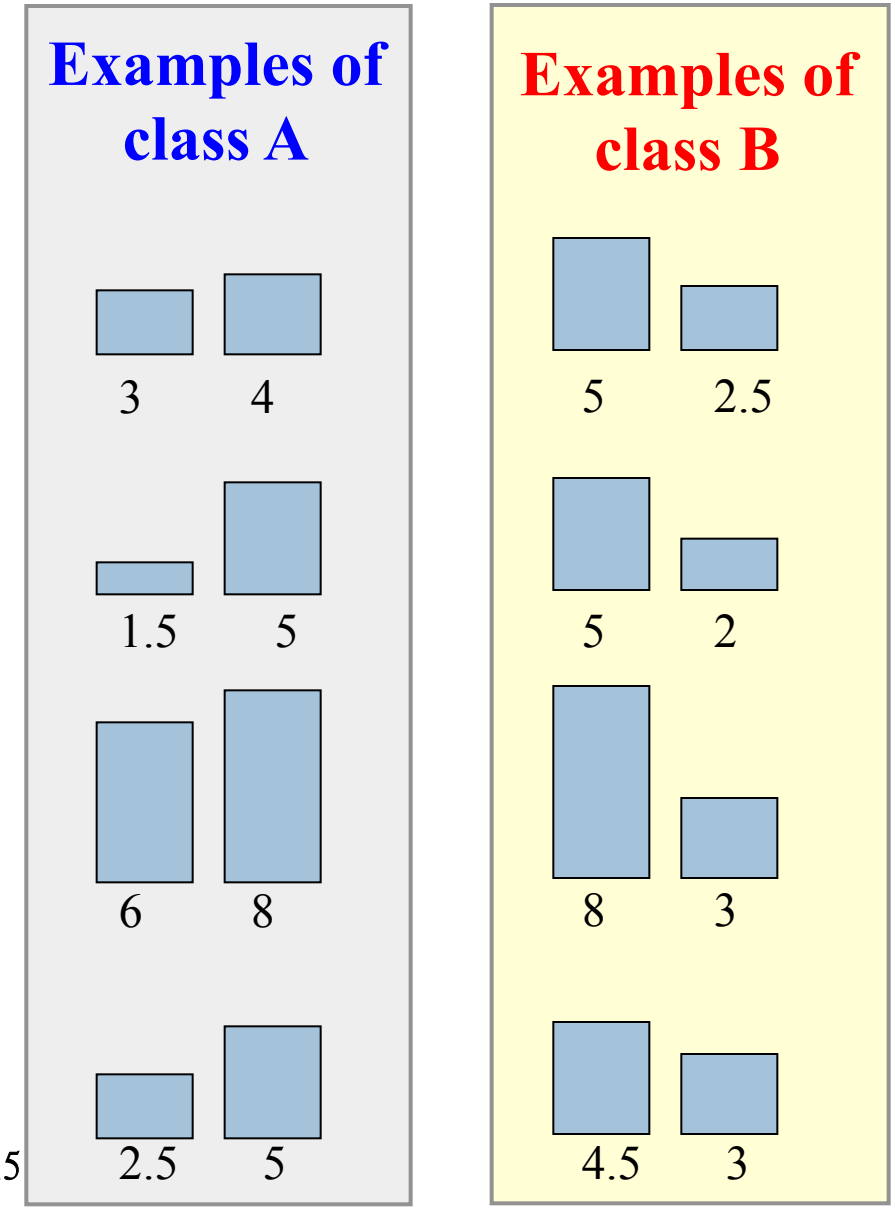
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



- Categorizing news stories as finance, weather, entertainment, sports, etc

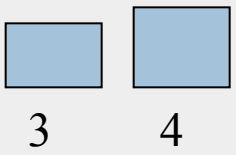
Kaggle: https://www.kaggle.com/competitions

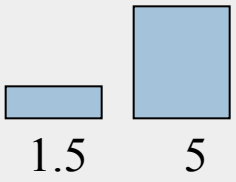
Simple Question 1

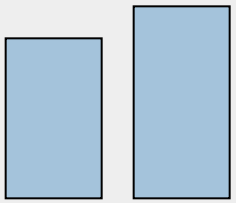


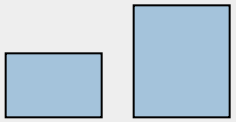
Simple Question 1

Examples of class A

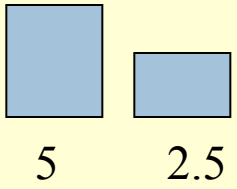


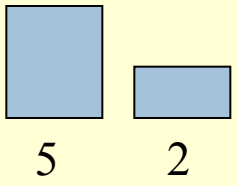


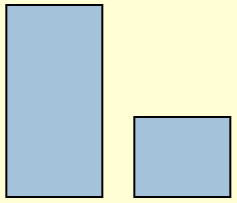


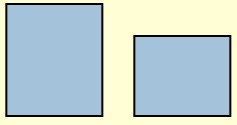


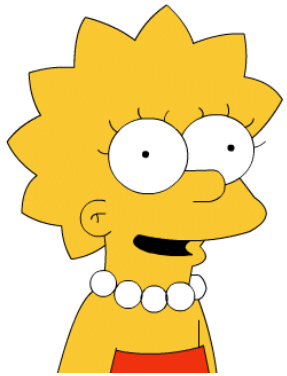
Examples of class B



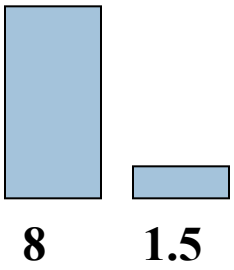




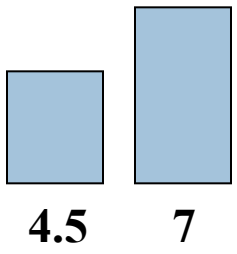




What class is this object?

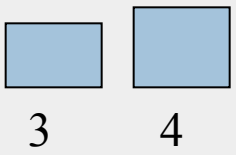


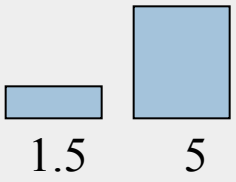
What about this one, A or B?

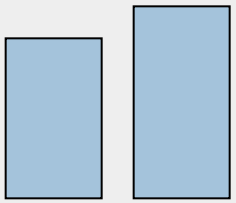


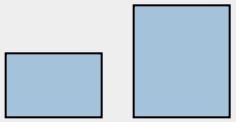
Simple Question 1

Examples of class A

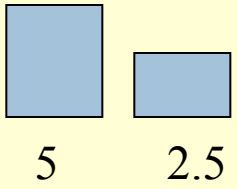


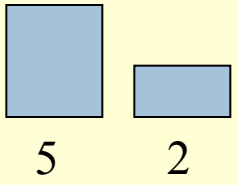


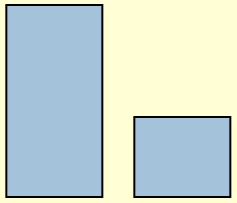


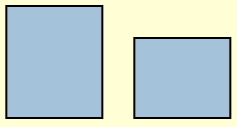


Examples of class B

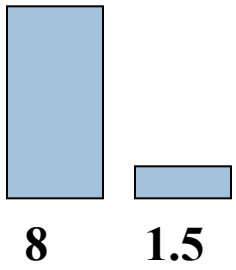








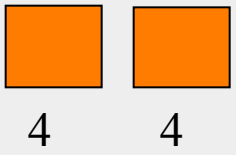
This is a B!

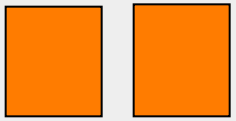


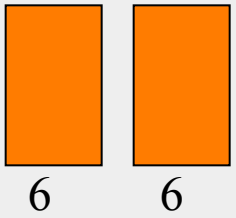
Here is the rule.
If the left bar is smaller than the right bar, it is an A, otherwise it is a B.

Simple Question 2

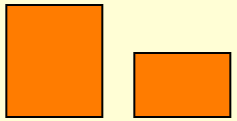
Examples of class A

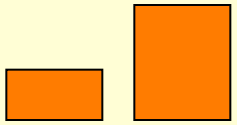


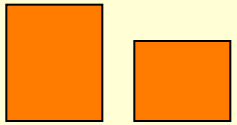


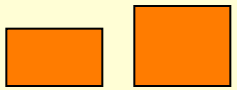


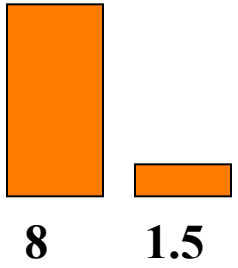
Examples of class B

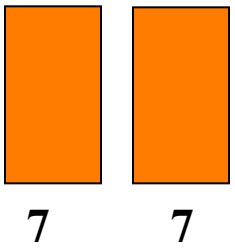






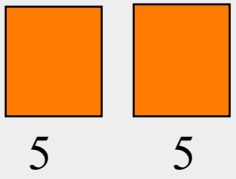


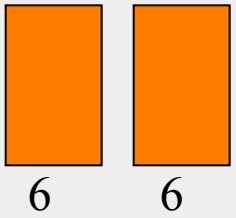


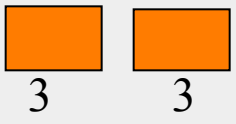


Simple Question 2

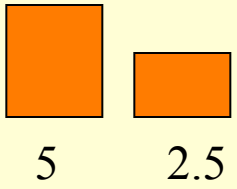
Examples of class A

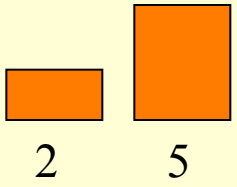


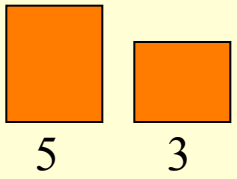


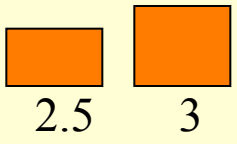


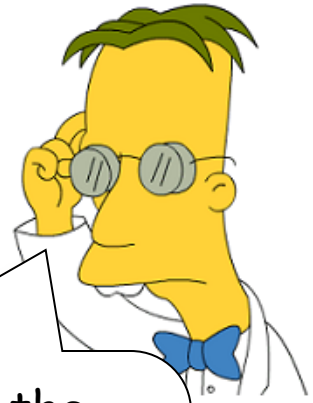
Examples of class B





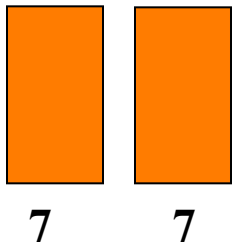






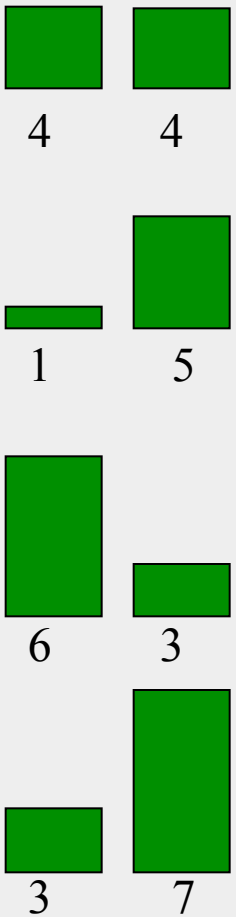
The rule is as follows, if the two bars are equal sizes, it is an A. Otherwise it is a B.

So this one is an A.

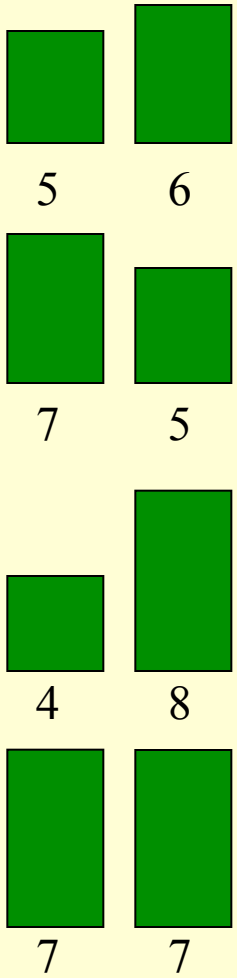


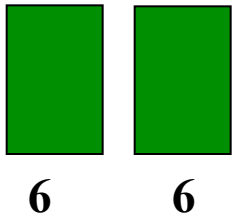
Simple Question 3

Examples of class A



Examples of class B

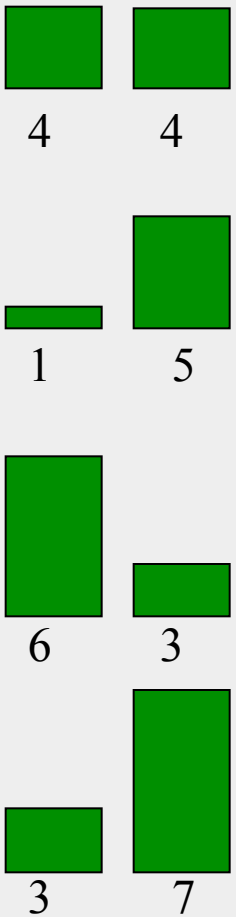




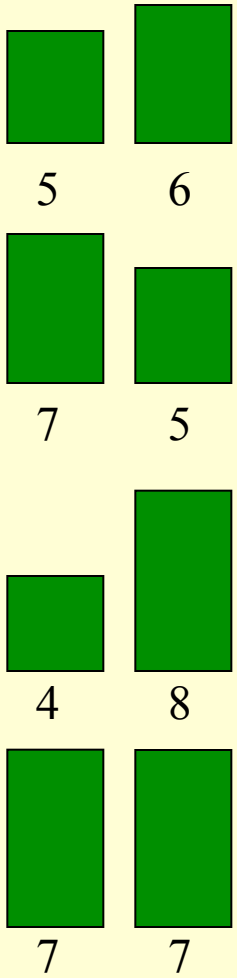
This one is really hard!
What is this, A or B?

Simple Question 3

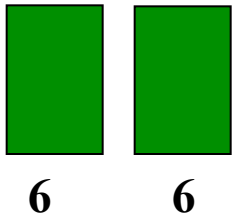
Examples of class A



Examples of class B



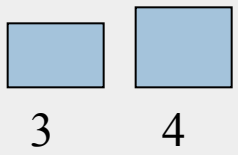
It is a B!

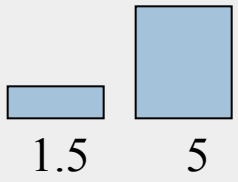


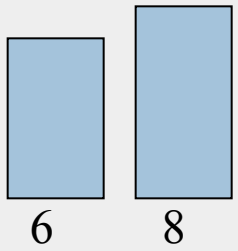
The rule is as follows, if the square of the sum of the two bars is less than or equal to 10, it is an A. Otherwise it is a B.

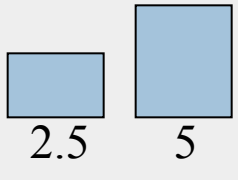
Simple Question 1

Examples of class A

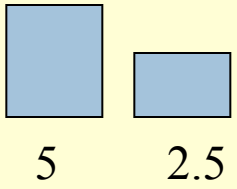


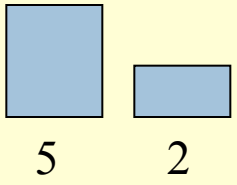


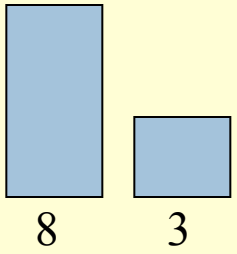


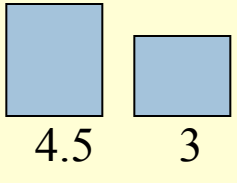


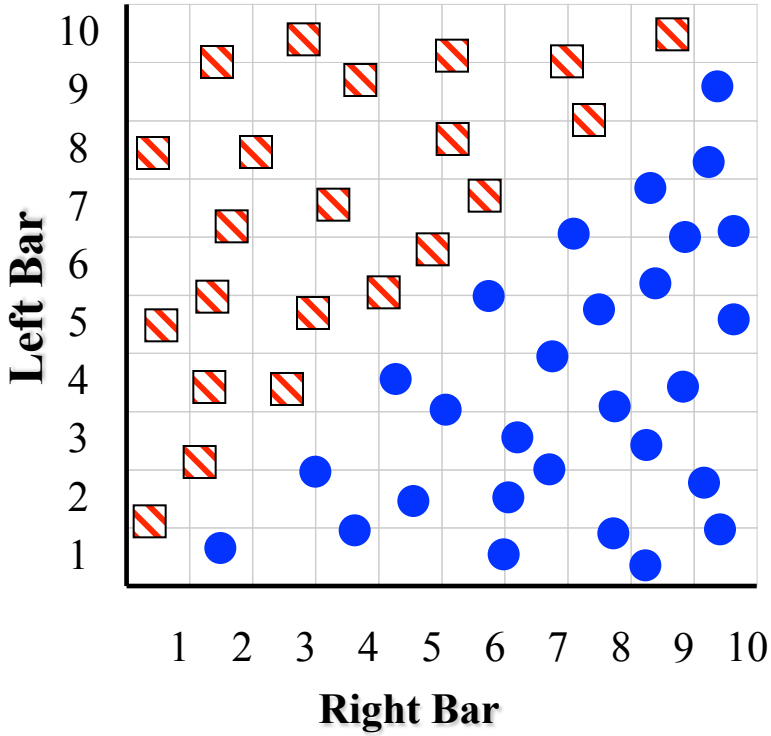
Examples of class B







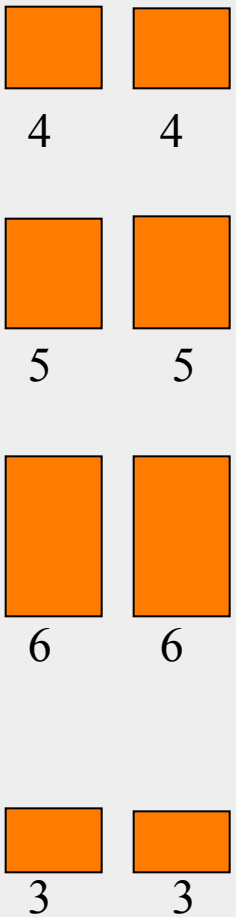




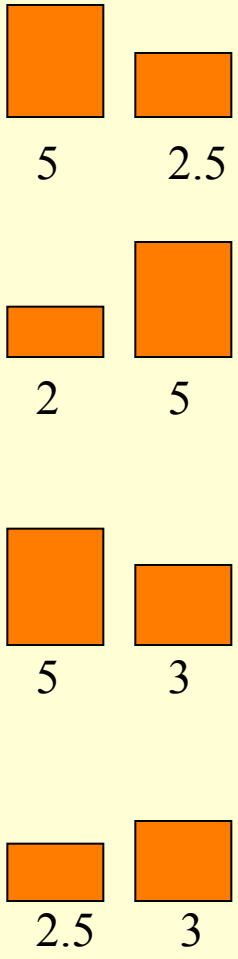
Here is the rule again. If the left bar is smaller than the right bar, it is an A, otherwise it is a B.

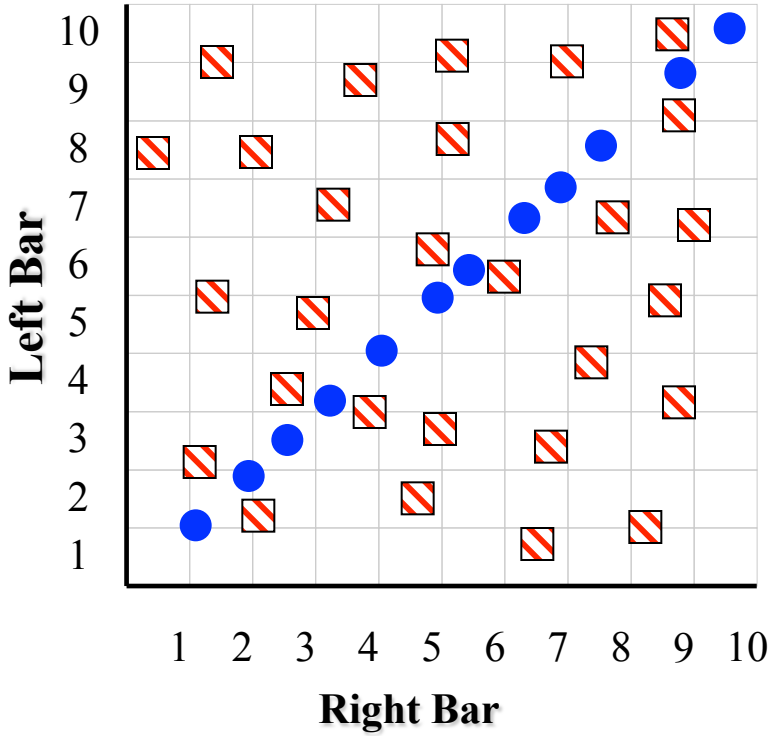
Simple Question 2

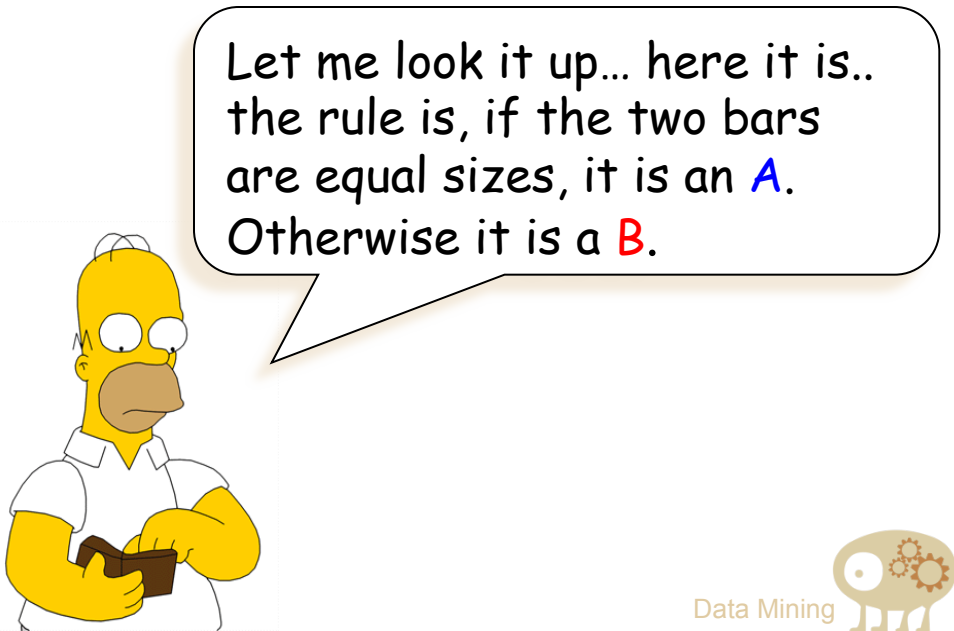
Examples of class A



Examples of class B

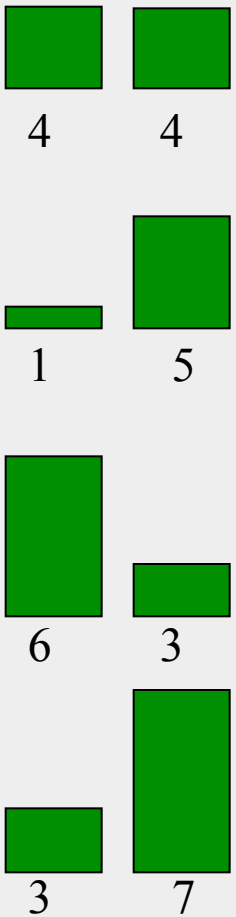




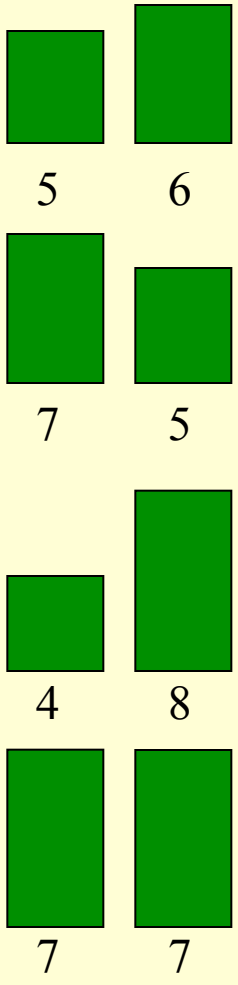


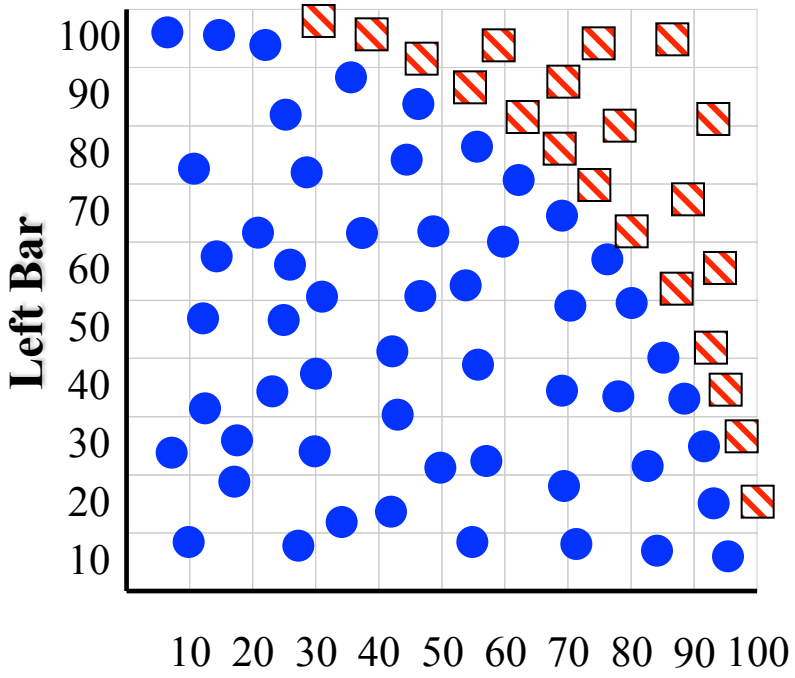
Simple Question 3

Examples of class A



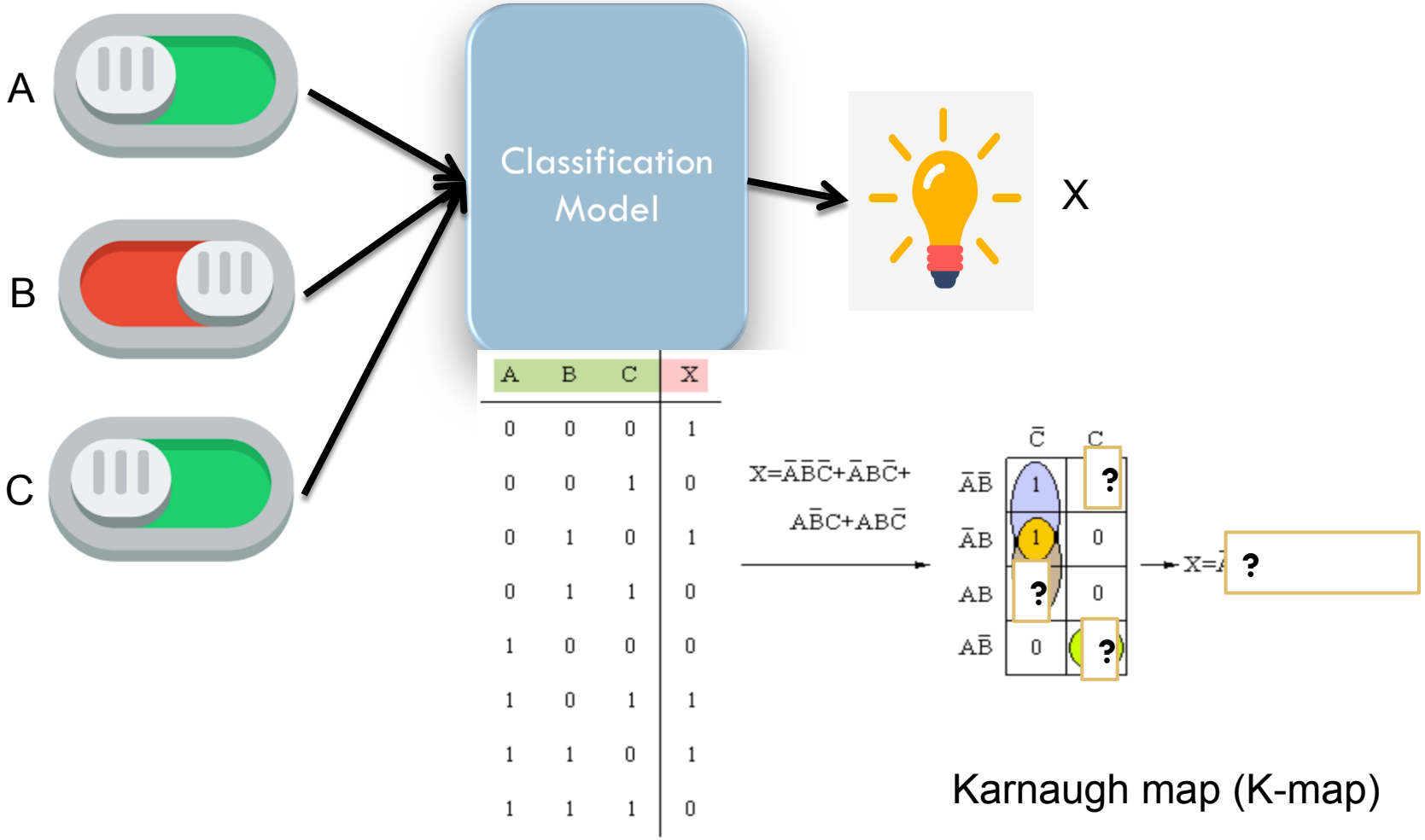
Examples of class B





The rule again:
if the square of the sum of the two bars is less than or equal to 10, it is an A. Otherwise it is a B.

Simple Data-Oriented Classification



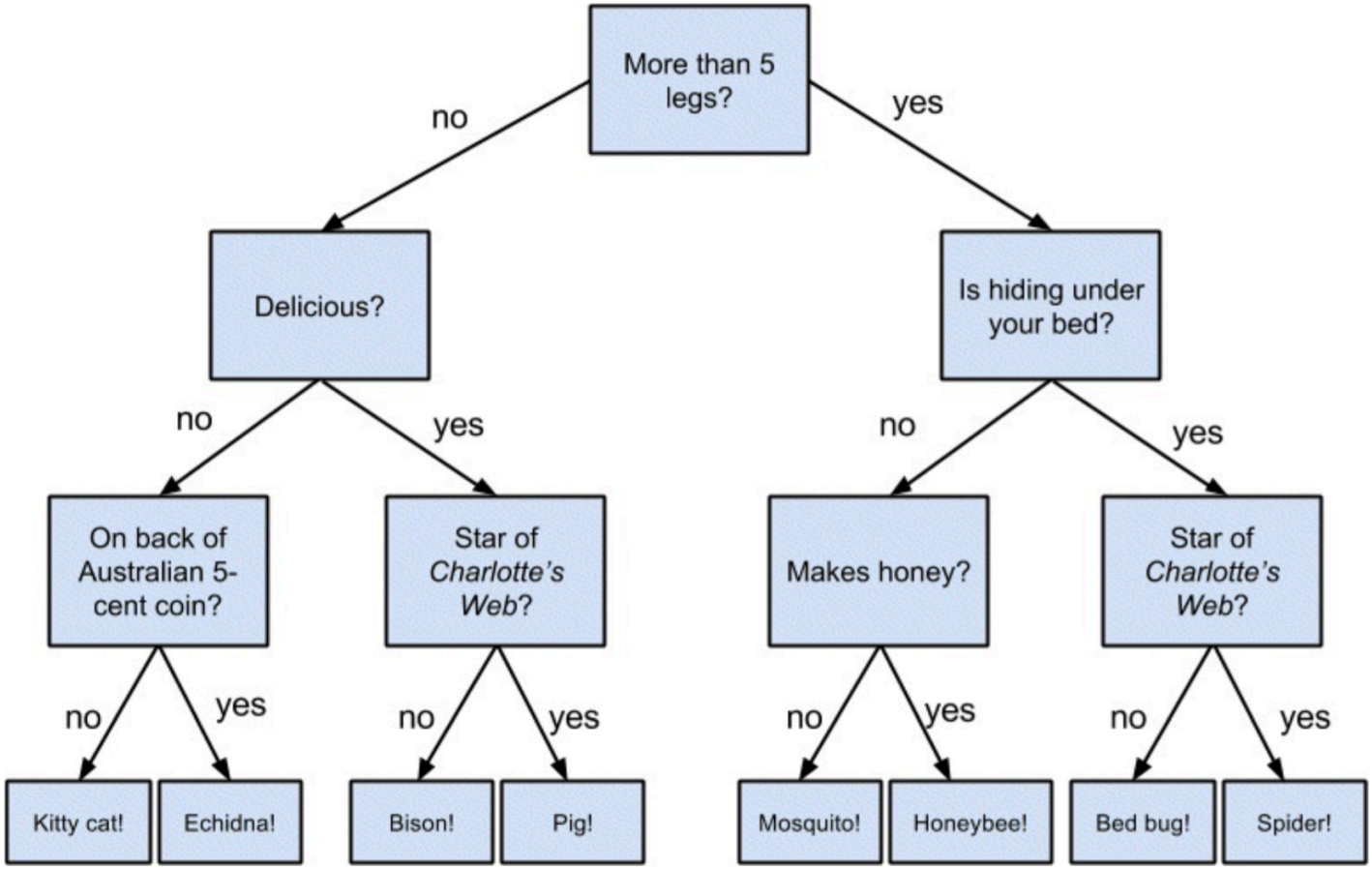
Karnaugh map (K-map)

Classification Techniques

- Decision Tree based Methods
- Memory based reasoning (K-NN)
- Regression
- Neural Networks
  - DNN
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines
- Random Forest

Decision Tree

20 Questions



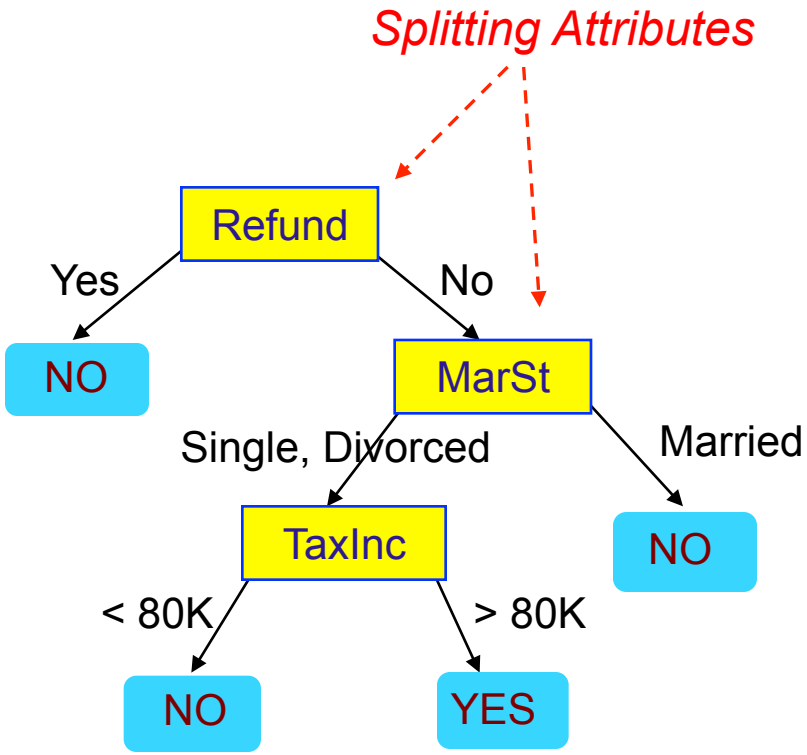
"*Guess the animal*" decision tree

Example of a Decision Tree

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

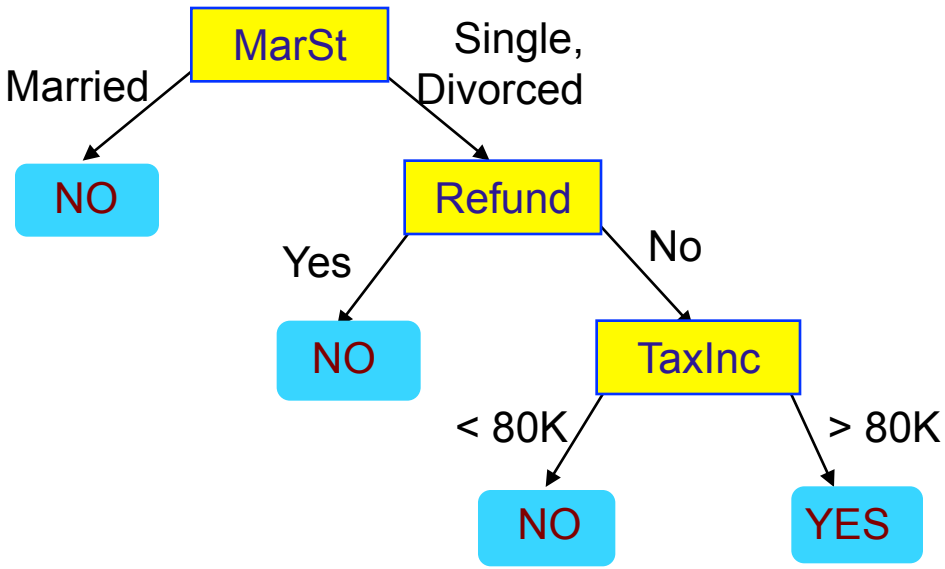
Training Data



Model: Decision Tree

Another Example of Decision Tree

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	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!

Test data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

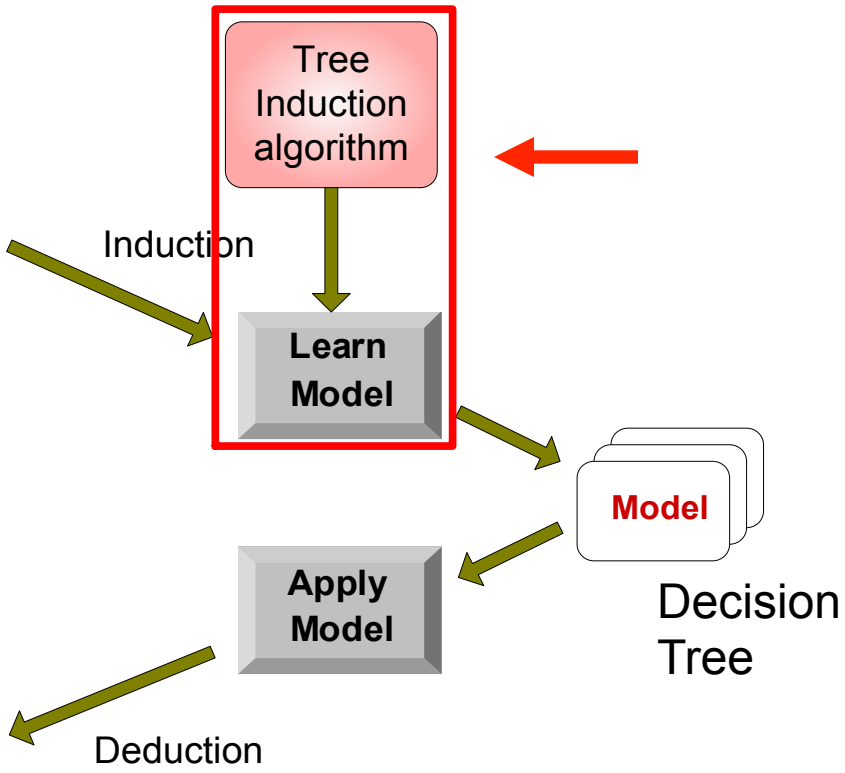
Decision Tree Classification Task

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



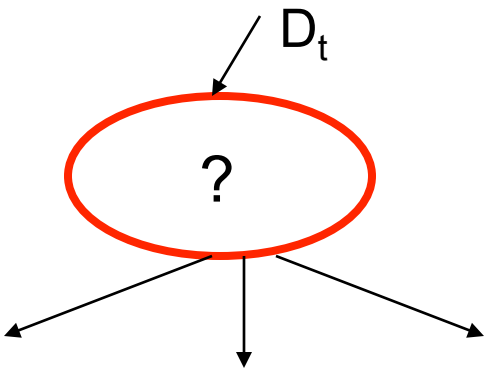
Decision Tree Induction

- Many Algorithms:
  - Hunt's Algorithm (one of the earliest)
  - CART
  - ID3, C4.5, C5.0, See5
  - SLIQ,SPRINT

General Structure of Hunt's Algorithm

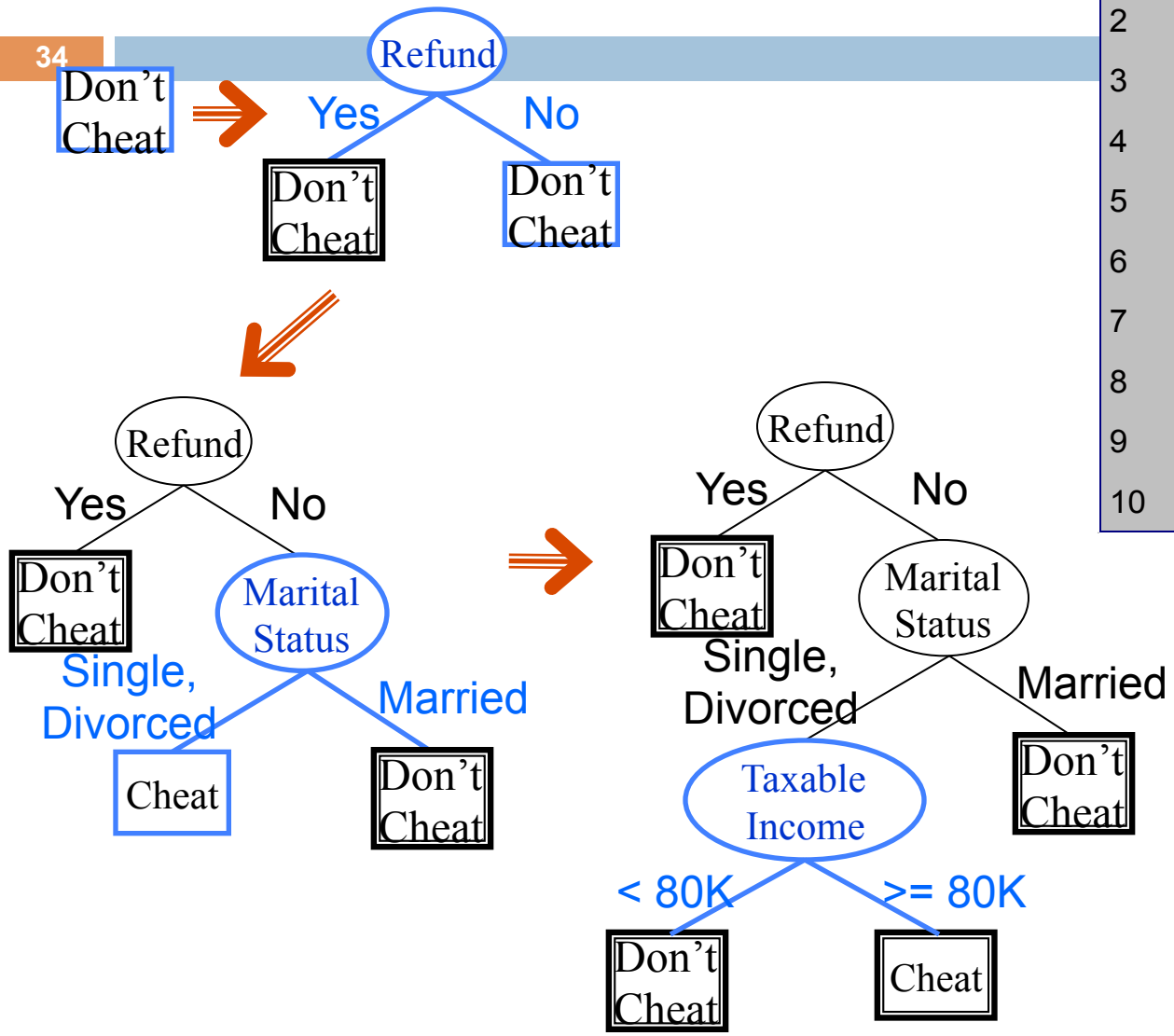
- Let D_t be the set of training records that reach a node t
- General Procedure:
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
  - If D_t contains records that belong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

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



Hunt's Algorithm

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



Tree Induction

- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion.

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting

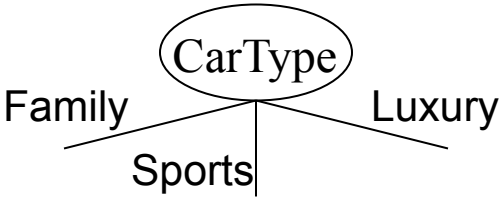
How to Specify Test Condition?

- Depends on attribute types
  - Nominal: {hot, cool, cold}
  - Ordinal: {A, B, C}, {small, medium, large}
  - Continuous: {1.25, 2, -3}

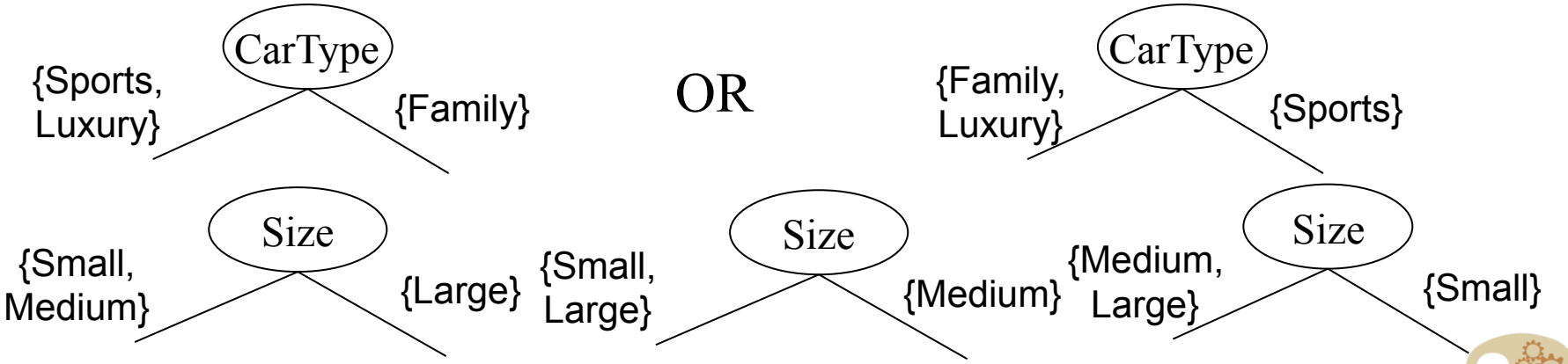
- Depends on number of ways to split
  - 2-way split
  - Multi-way split

Splitting Based on Nominal Attributes

- Multi-way split: Use as many partitions as *distinct values*.



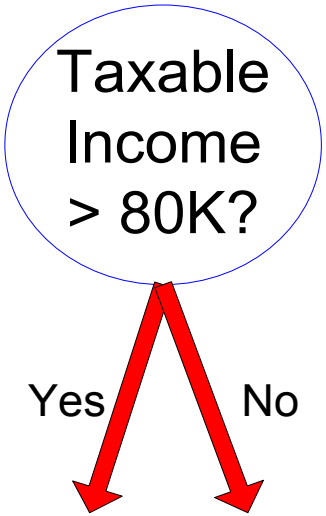
- Binary split: Divides values into two subsets.
  Need to find optimal partitioning.



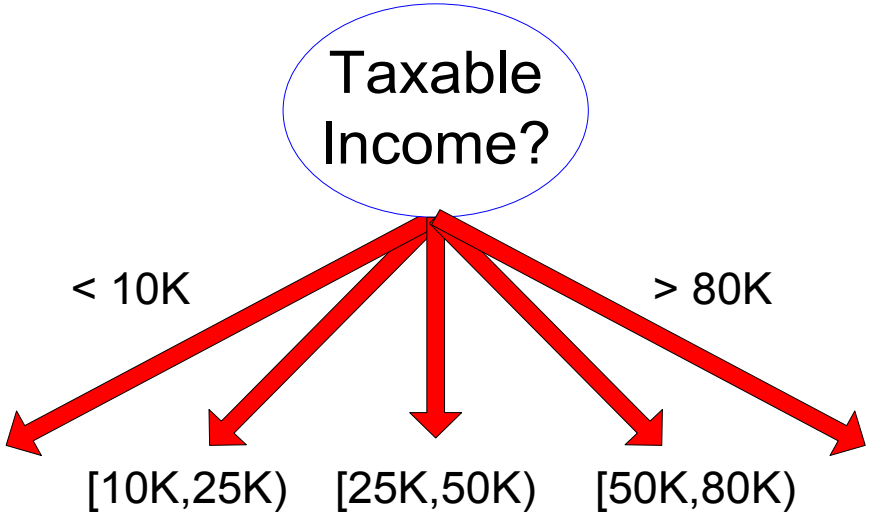
Splitting Based on Continuous Attributes

- Different ways of handling
  - Discretization to form an ordinal categorical attribute
    - Static – discretize once at the beginning
    - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.
  - Binary Decision: $(A < v)$ or $(A \geq v)$
    - consider all possible splits and finds the best cut
    - can be more compute intensive

Splitting Based on Continuous Attributes



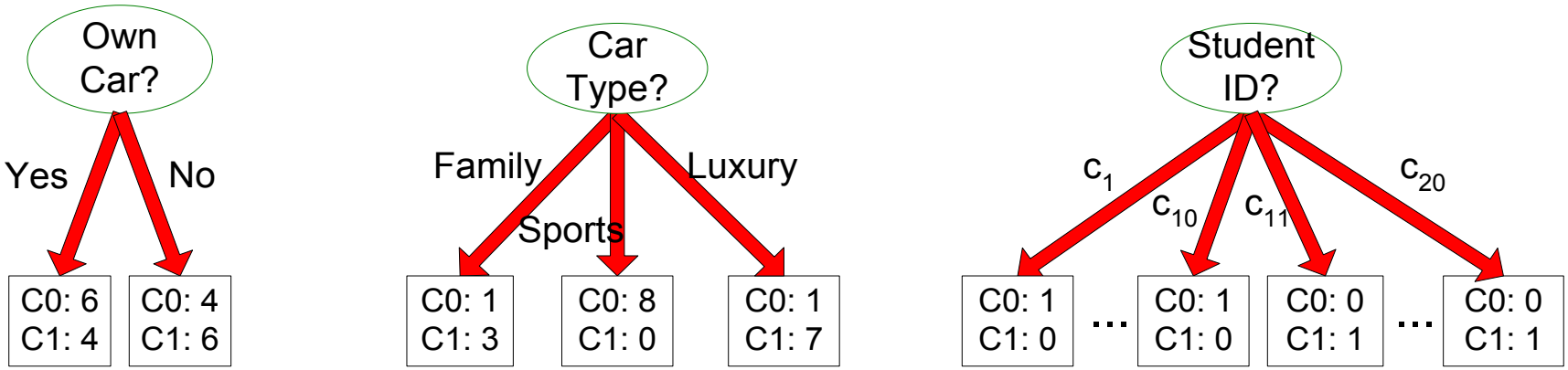
(i) Binary split



(ii) Multi-way split

How to determine the Best Split

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



Which test condition is the best?

How to determine the Best Split

- Greedy approach:
  - Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity:

C0: 5 C1: 5	C0: 9 C1: 1
Non-homogeneous, High degree of impurity	Homogeneous, Low degree of impurity

Measures of Node Impurity

- Gini Index

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

- Entropy

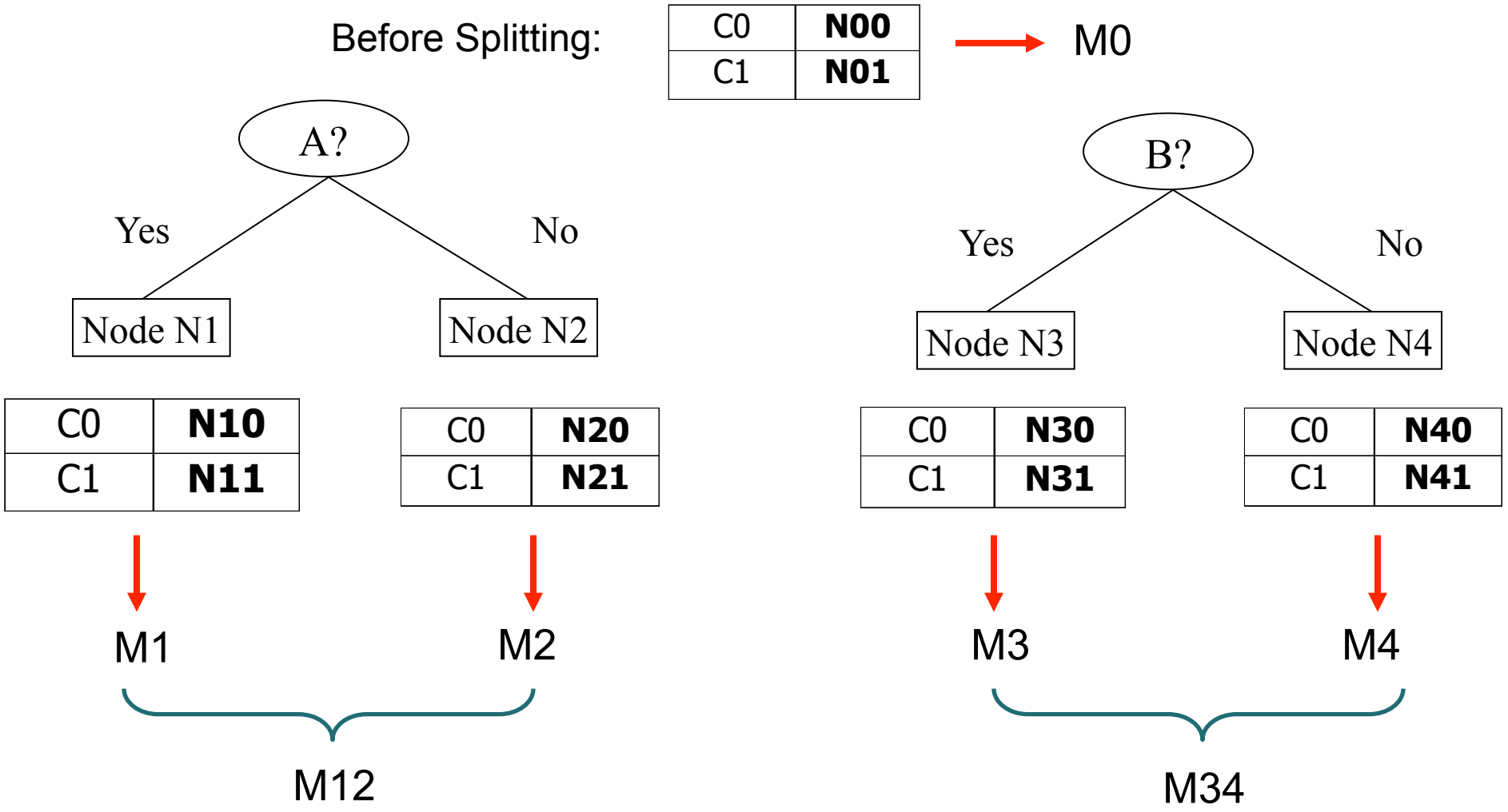
$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

- Misclassification error

$$Error(t) = 1 - \max_j p(j \mid t)$$

$p(j \mid t)$ is the relative frequency of training instances that belong to class j at note t

How to Find the Best Split



Gain = M0 – M12 vs M0 – M34

Measure of Impurity: GINI

- Gini Index for a given node t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- Maximum (1 - $1/n_c$) when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

Examples for computing GINI

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

Splitting Based on GINI

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,

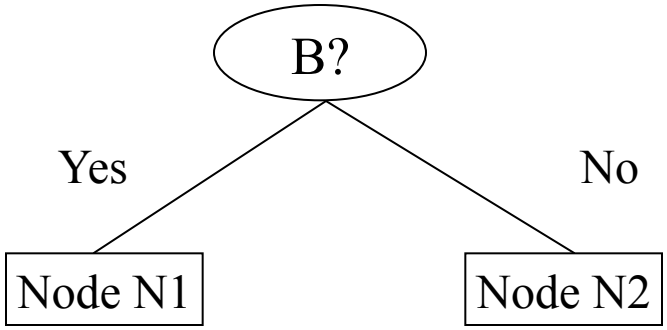
$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where, n_i = number of records at child i,

n = number of records at node p.

Binary Attributes: Computing GINI Index

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.



	Parent
C1	**6**
C2	**6**
Gini = 0.500	

Gini(N1)
= 1 – $(5/7)^2$ – $(2/7)^2$
= 0.408

Gini(N2)
= 1 – $(1/5)^2$ – $(4/5)^2$
= 0.32

	N1	N2
C1	**5**	**1**
C2	**2**	**4**
Gini=0.333		

Gini(Children)
= 7/12 * 0.408 +
5/12 * 0.320=
0.371

Categorical Attributes: Computing Gini Index

- For each distinct value, gather counts for each class in the dataset
- Use the count matrix to make decisions

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

Two-way split
(find best partition of values)

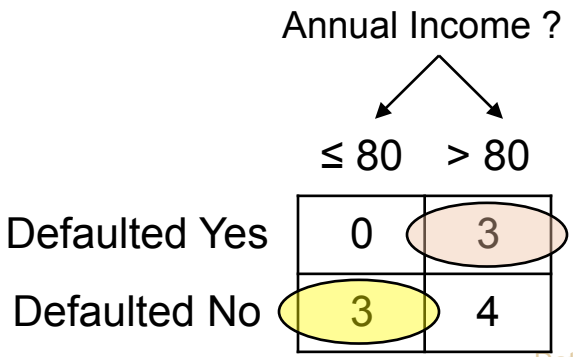
	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Continuous Attributes: Computing Gini Index

- Use Binary Decisions based on one value
- Several Choices for the splitting value
  - Number of possible splitting values = Number of distinct values
- Each splitting value has a count matrix associated with it
  - Class counts in each of the partitions, A < v and A ≥ v
- Simple method to choose best v
  - For each v, scan the database to gather count matrix and compute its Gini index
  - Computationally Inefficient! Repetition of work.

ID	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



Continuous Attributes: Computing Gini Index...

- For efficient computation: for each attribute,
  - Sort the attribute on values
  - Linearly scan these values, each time updating the count matrix and computing gini index
  - Choose the split position that has the least gini index

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income										
Sorted Values →	60	70	75	85	90	95	100	120	125	220
Split Positions →	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	*0.300*	0.343	0.375	0.400	0.420

Alternative Splitting Criteria based on INFO

- Entropy at a given node t:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

  - Measures homogeneity of a node.
    - Maximum (log n_c) when records are equally distributed among all classes implying least information
    - Minimum (0.0) when all records belong to one class, implying most information
  - Entropy based computations are similar to the GINI index computations

Example of Entropy

Entropy can be measured for a set, e.g.:

$$S = \{a, a, a, a, a, a, a, a, b, b, b, b, b\}$$

$$H(p_1, p_2, ..., p_s) = \sum_{i=1}^{s} (p_i log(1/p_i))$$

(8 a's and 5 b's, 13 total)

$$entropy(S) = -\left[\left(\frac{8}{13}(\log_2 \frac{8}{13})\right) + \left(\frac{5}{13}(\log_2 \frac{5}{13})\right)\right] = 0.96124$$

Remember negative! *for the a's* *for the b's*

Examples for computing Entropy

$$Entropy(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Splitting Based on INFO...

- Information Gain:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Parent Node, p is split into k partitions;

n_i is number of records in partition i

- Measures Reduction in Entropy achieved because of the split. Choose the split that achieves most reduction (maximizes GAIN)
- Used in ID3 and C4.5
- Disadvantage: Tends to prefer splits that result in large number of partitions, each being small but pure.

Splitting Based on INFO...

- Gain Ratio:

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO} \qquad SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

Parent Node, p is split into k partitions

n_i is the number of records in partition i

- Adjusts Information Gain by the entropy of the partitioning (SplitINFO). Higher entropy partitioning (large number of small partitions) is penalized!
- Used in C4.5
- Designed to overcome the disadvantage of Information Gain

Splitting Criteria based on Classification Error

- Classification error at a node t :

$$Error(t) = 1 - \max_i P(i \mid t)$$

- Measures misclassification error made by a node.
  - Maximum (1 - $1/n_c$) when records are equally distributed among all classes, implying least interesting information
  - Minimum (0.0) when all records belong to one class, implying most interesting information

Examples for Computing Error

$$Error(t) = 1 - \max_i P(i \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Error = 1 – max (0, 1) = 1 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

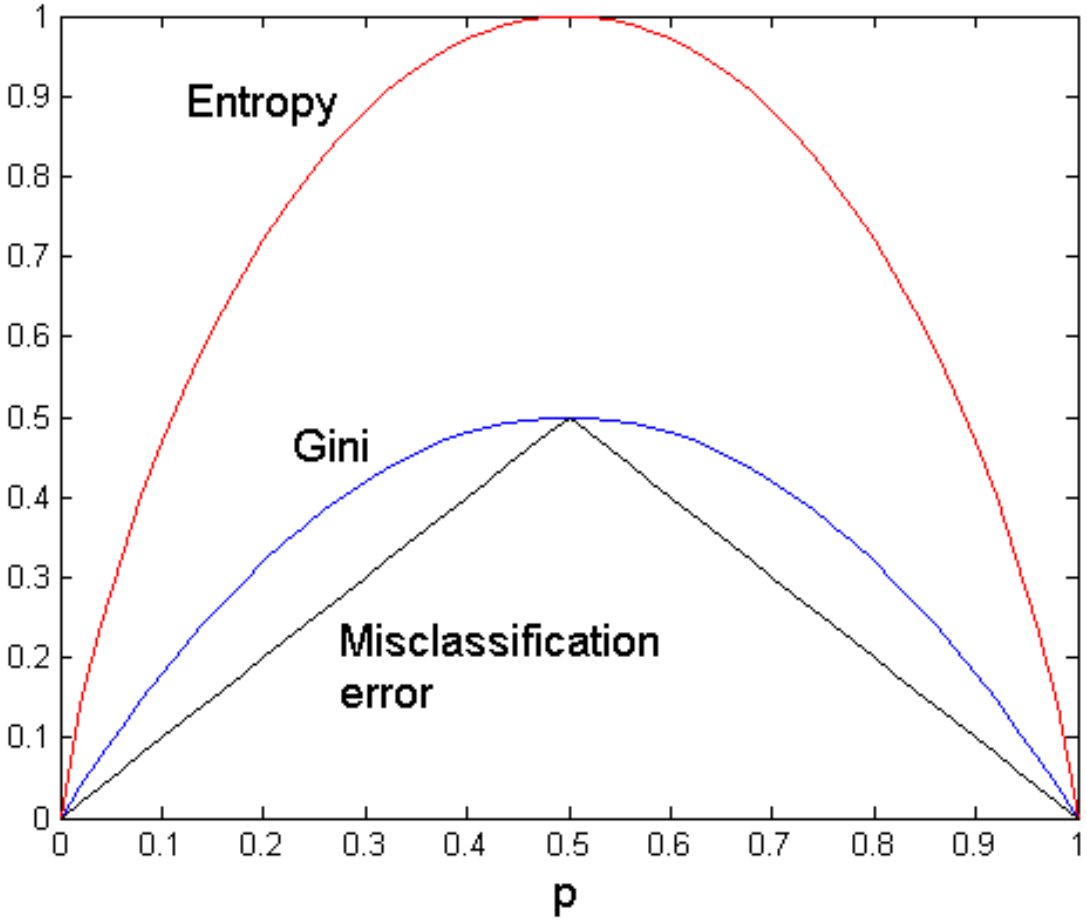
C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

Comparison among Splitting Criteria

For a 2-class problem:



Stopping Criteria for Tree Induction

- Stop expanding a node when all the records belong to the same class
- Stop expanding a node when all the records have similar attribute values
- Early termination (why?)

Decision Tree Based Classification

- Advantages:
  - Inexpensive to construct
  - Extremely fast at classifying unknown records
  - Easy to interpret for small-sized trees
  - Accuracy is comparable to other classification techniques for many simple data sets

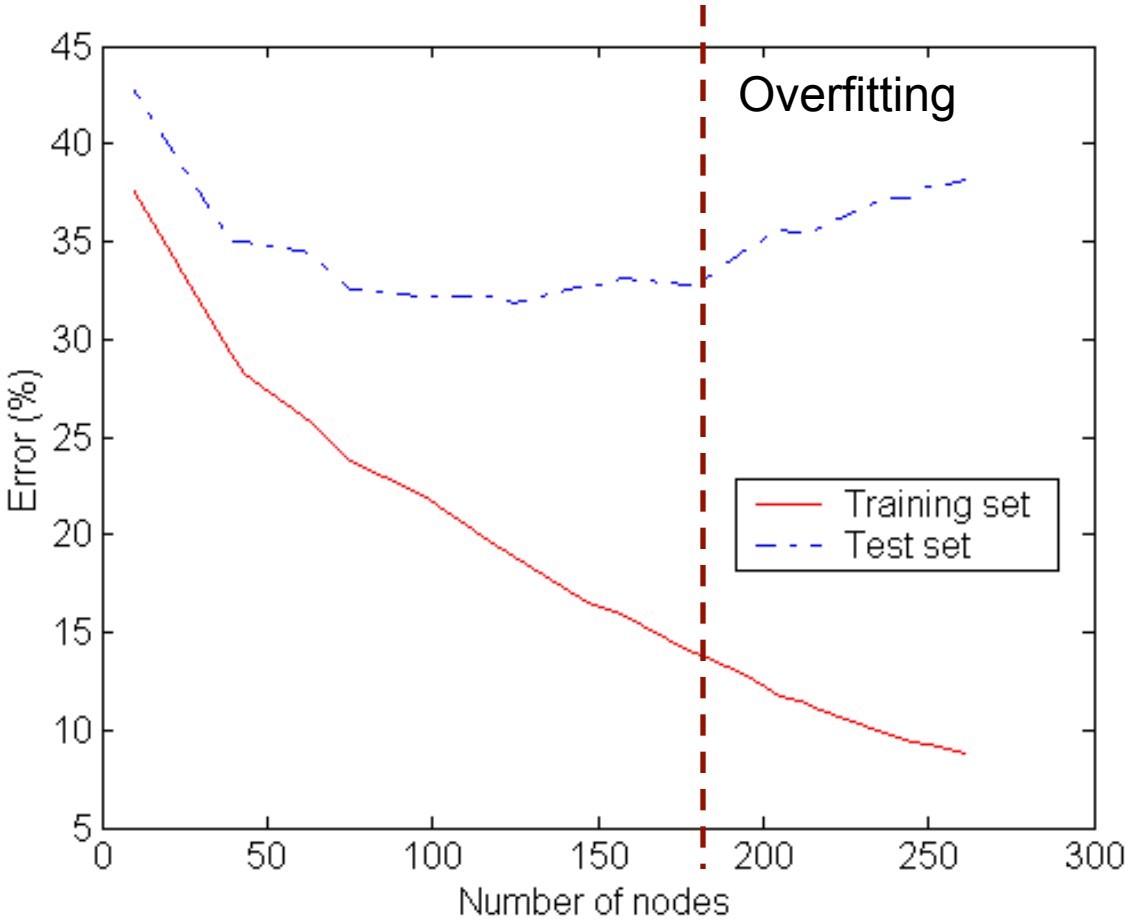
Decision Tree Learning

- Extremely popular method
  - Credit risk assessment
  - Medical diagnosis
  - Market analysis
- Good at dealing with symbolic feature
- Easy to comprehend
  - Compared to logistic regression model and support vector machine

Practical Issues of Classification

- Underfitting and Overfitting
- Missing Values
- Costs of Classification

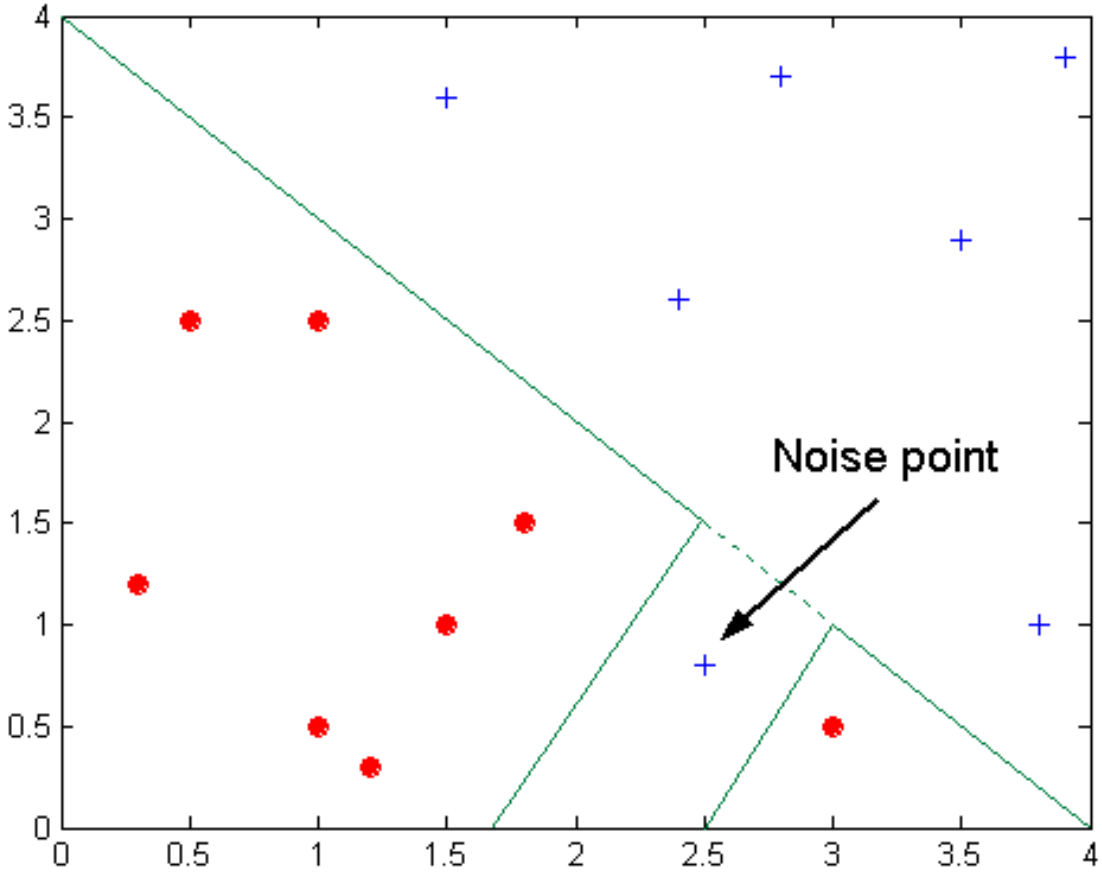
Underfitting and Overfitting



Overfitting with
(1) Data
(2) Feature

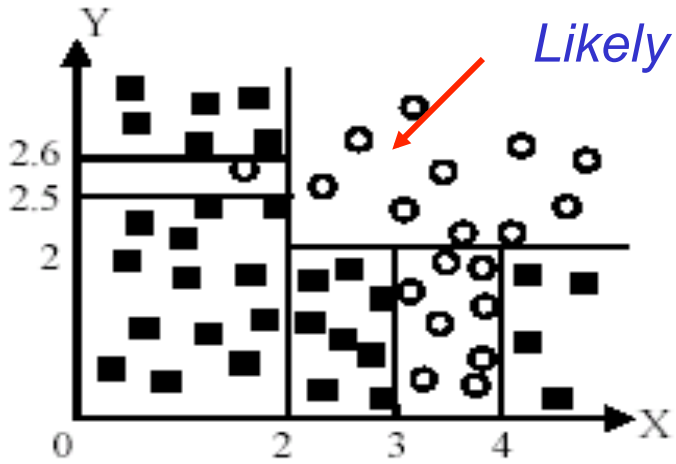
Underfitting: when model is too simple, both training and test errors are large

Overfitting due to Noise



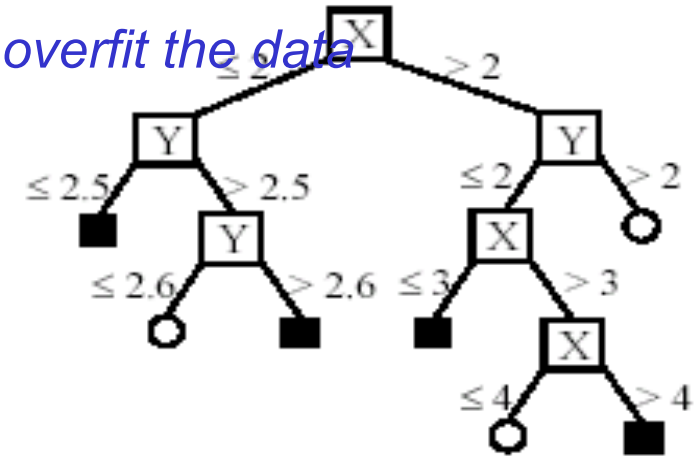
Decision boundary is distorted by noise point

An overfitting example

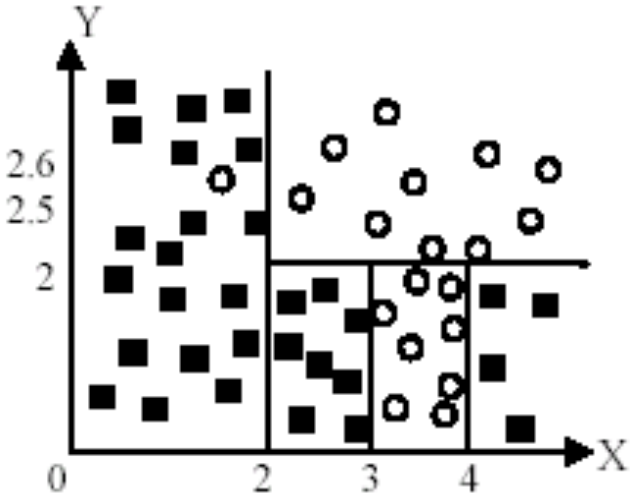


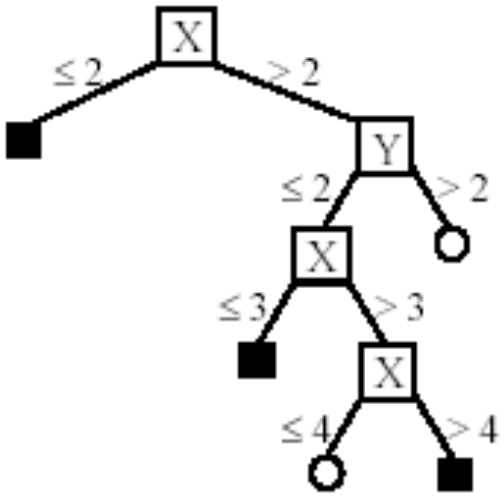
(A) A partition of the data space

Likely to overfit the data

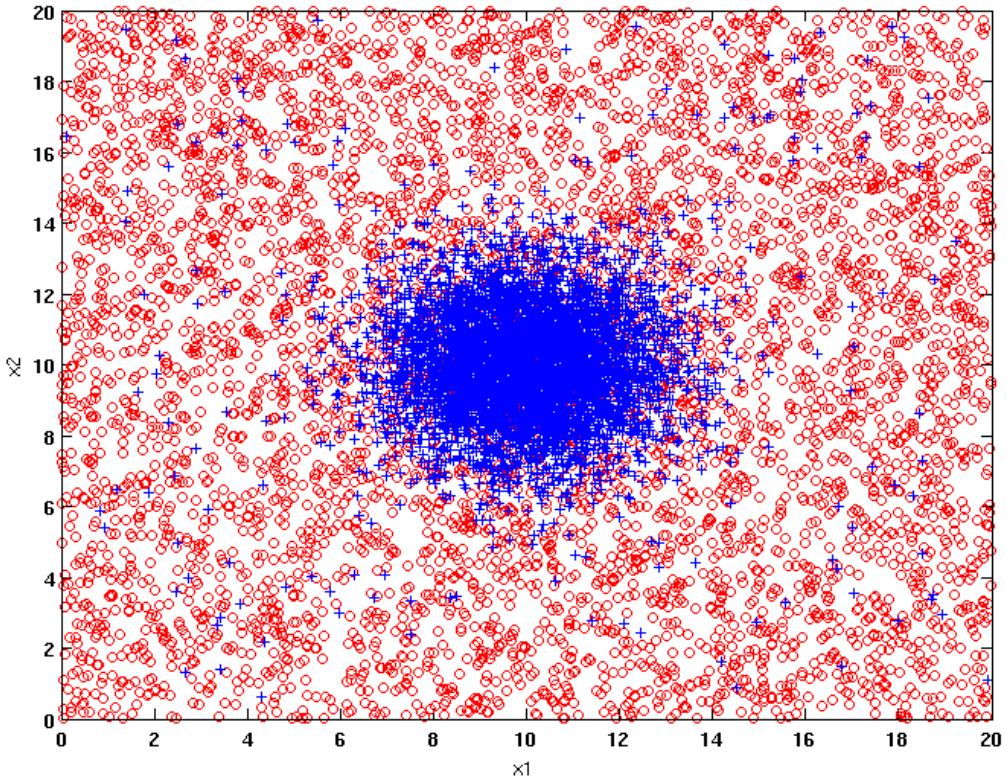


(B). The decision tree





Example Data Set



Two class problem:

+ : 5200 instances

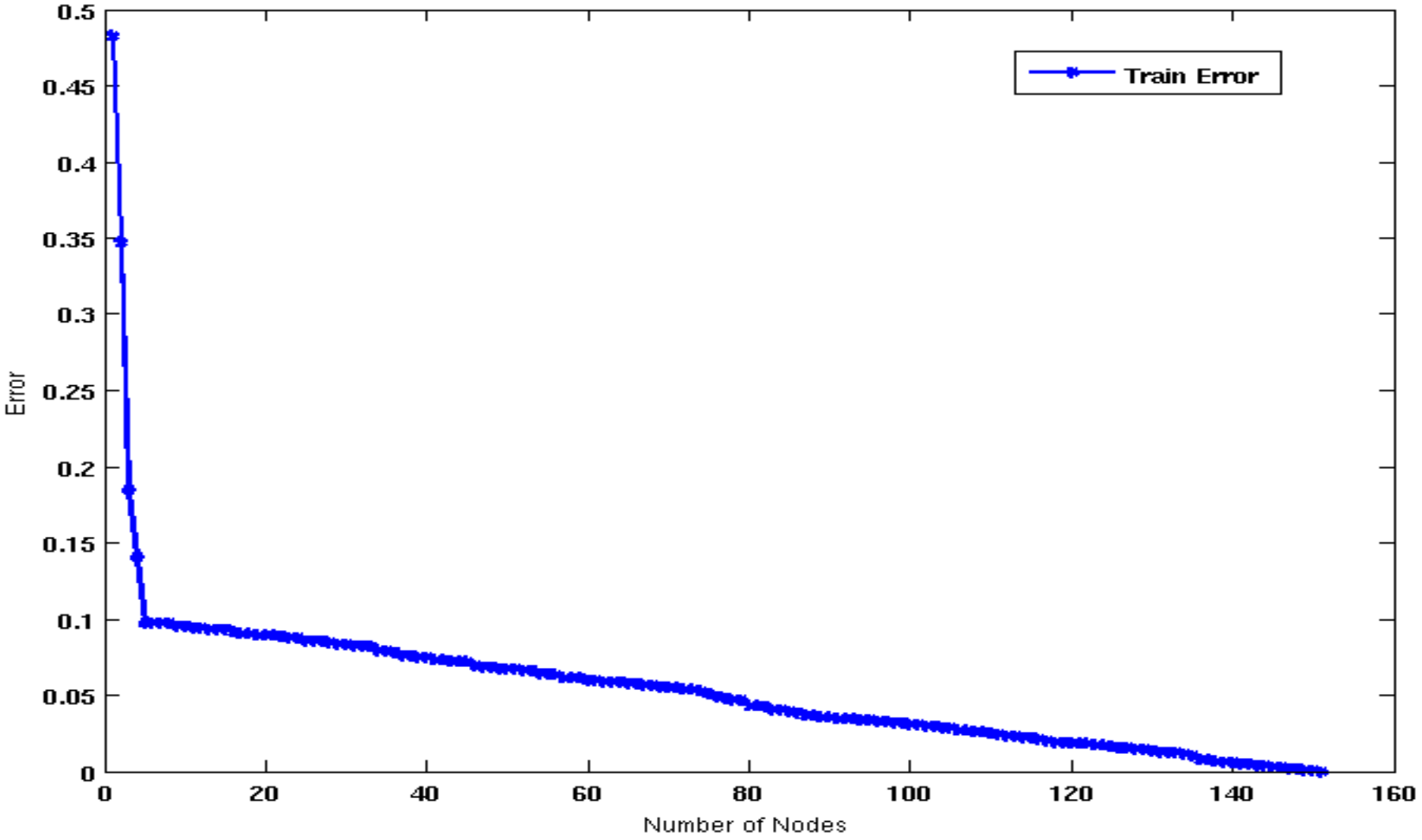
- **5000 instances generated from a Gaussian centered at (10,10)**
- **200 noisy instances added**

o : 5200 instances

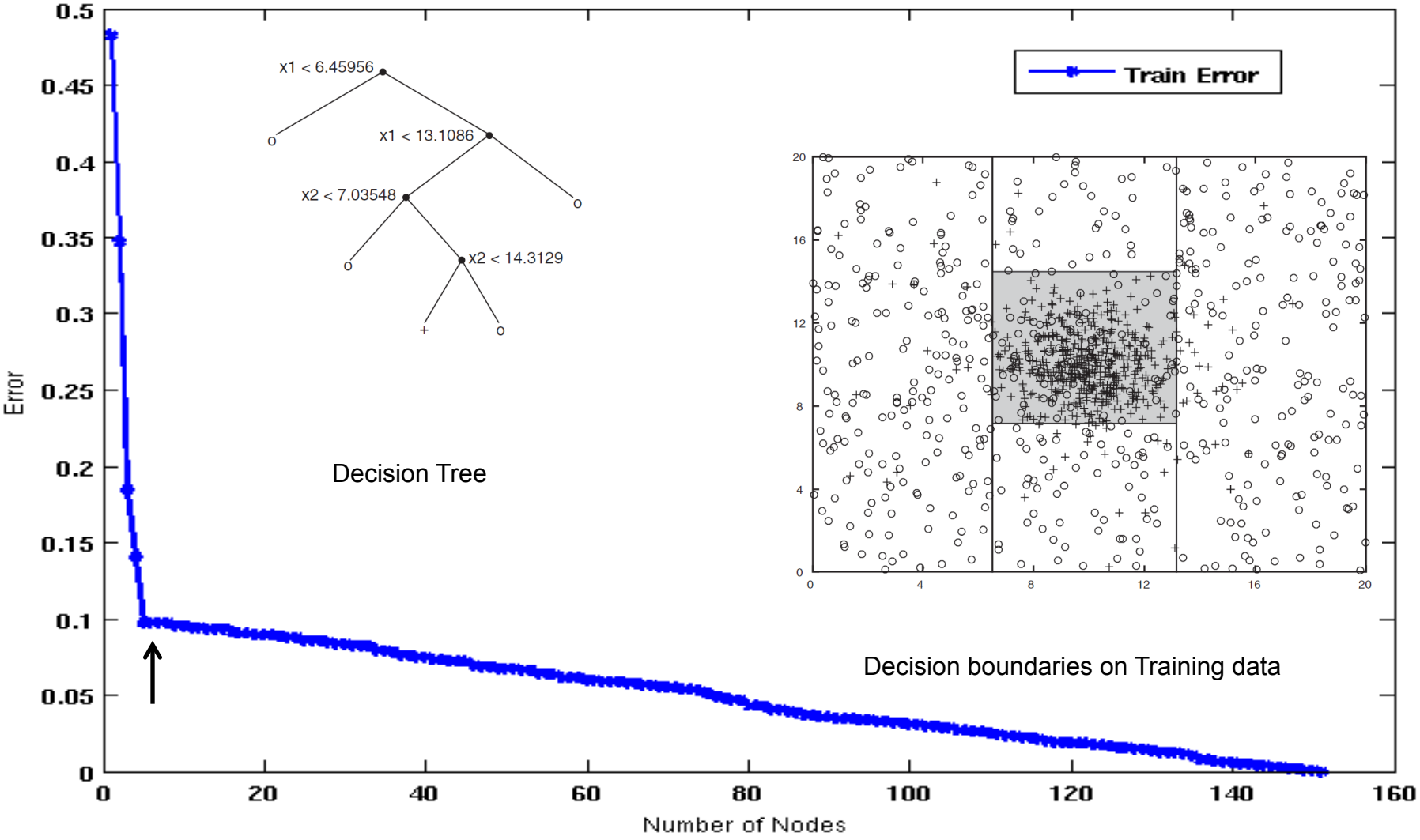
- **Generated from a uniform distribution**

10 % of the data used for training and 90% of the data used for testing

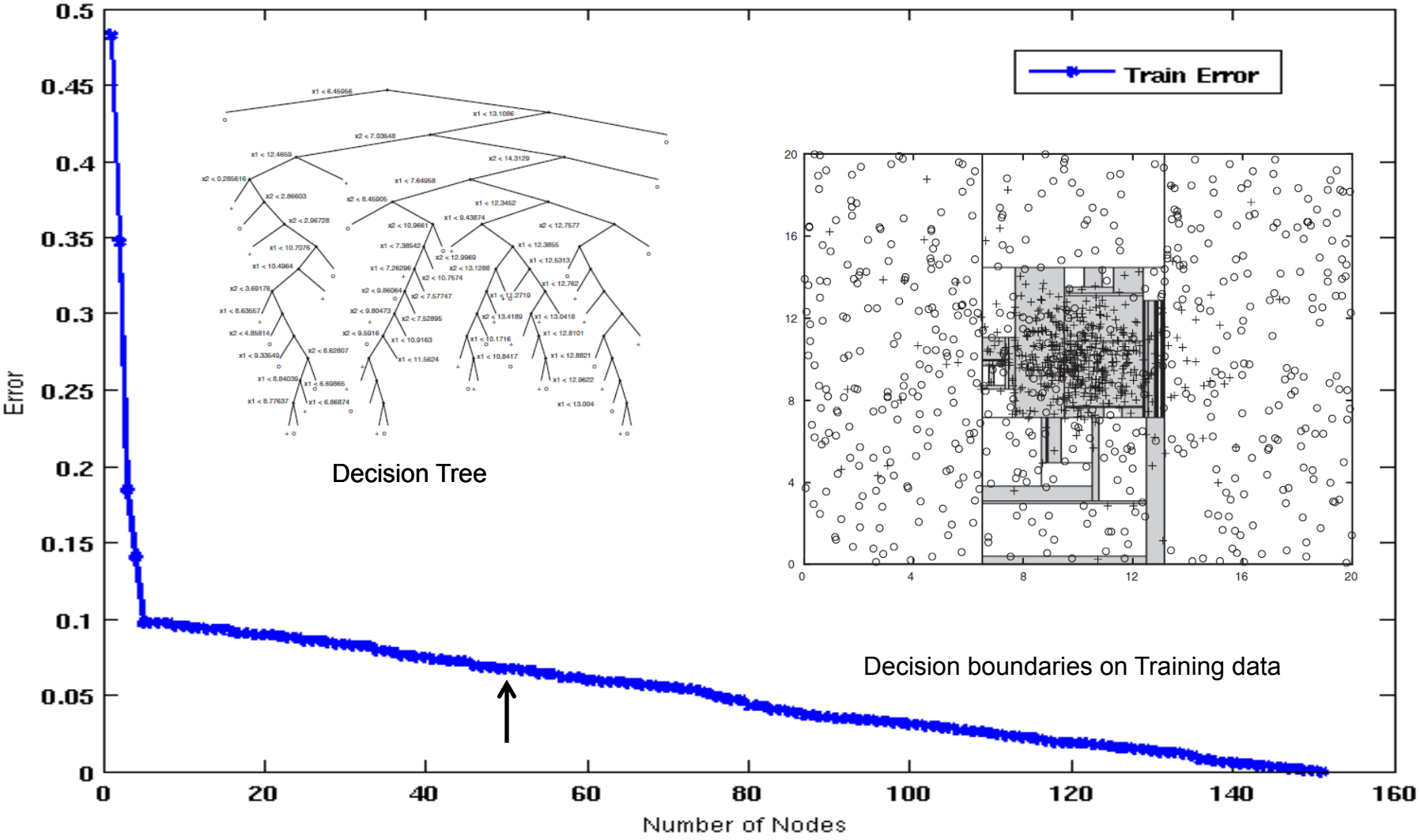
Increasing number of nodes in Decision Trees



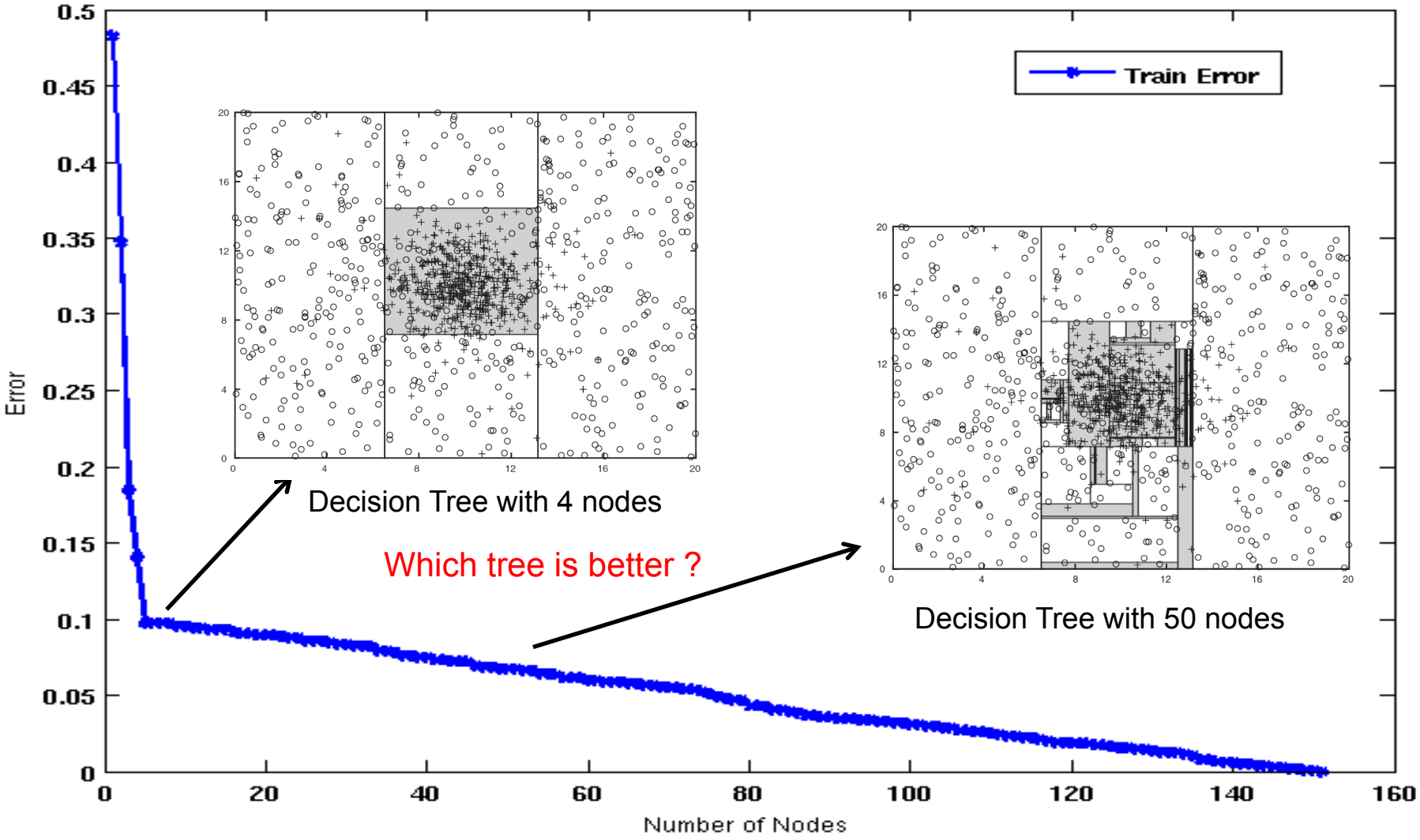
Decision Tree with 4 nodes



Decision Tree with 50 nodes



Which tree is better?



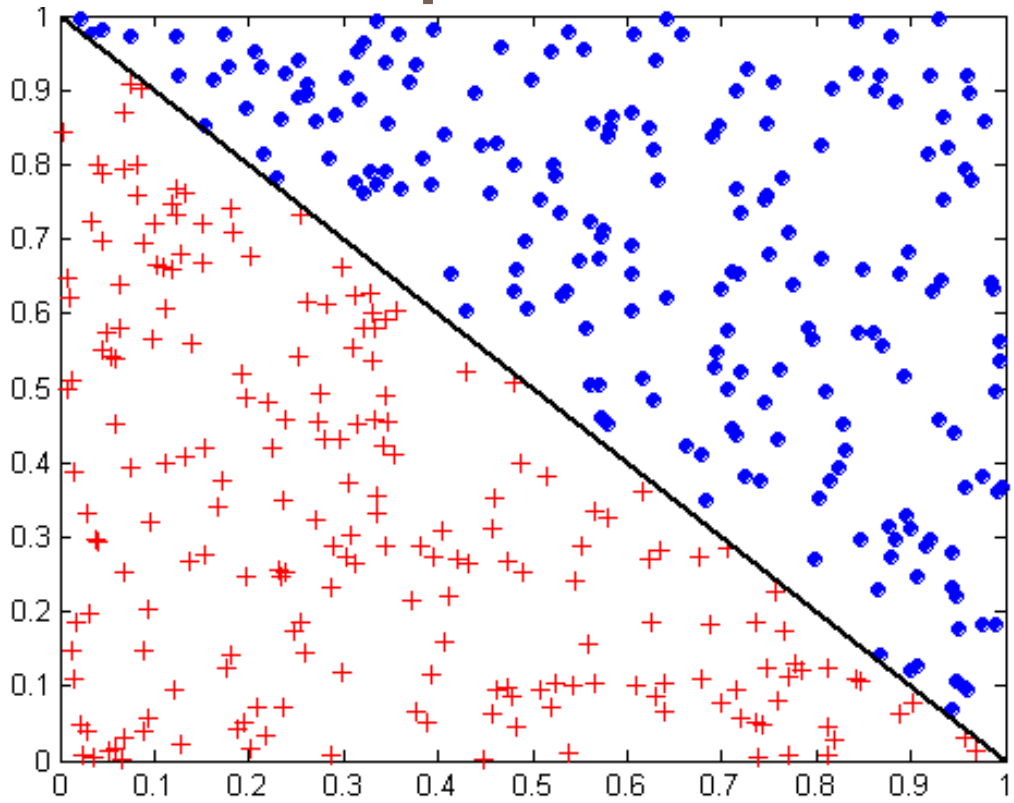
How to Address Overfitting

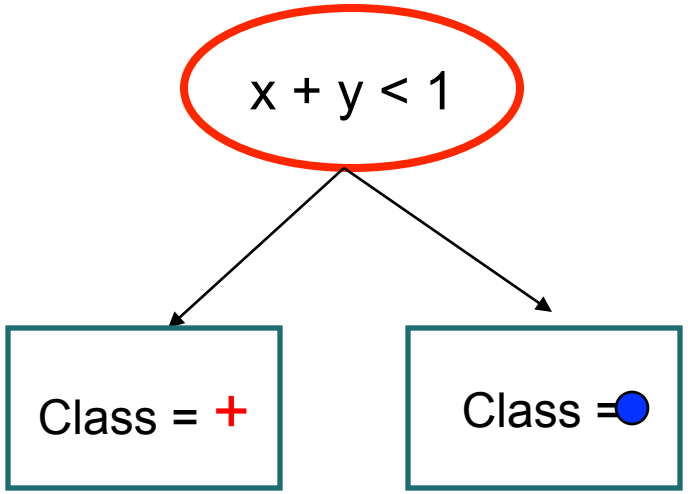
- Pre-Pruning (Early Stopping Rule)
  - Stop the algorithm before it becomes a fully-grown tree
  - Typical stopping conditions for a node:
    - Stop if all instances belong to the same class
    - Stop if all the attribute values are the same
  - More restrictive conditions:
    - Stop if number of instances is less than some user-specified threshold
    - Stop if class distribution of instances are independent of the available features (e.g., using χ^2 test)
    - Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain).

How to Address Overfitting…

- Post-pruning
  - Grow decision tree to its entirety
  - Trim the nodes of the decision tree in a bottom-up fashion
  - If generalization error improves after trimming, replace sub-tree by a leaf node.
  - Class label of leaf node is determined from majority class of instances in the sub-tree
  - Can use MDL for post-pruning

Oblique Decision Trees

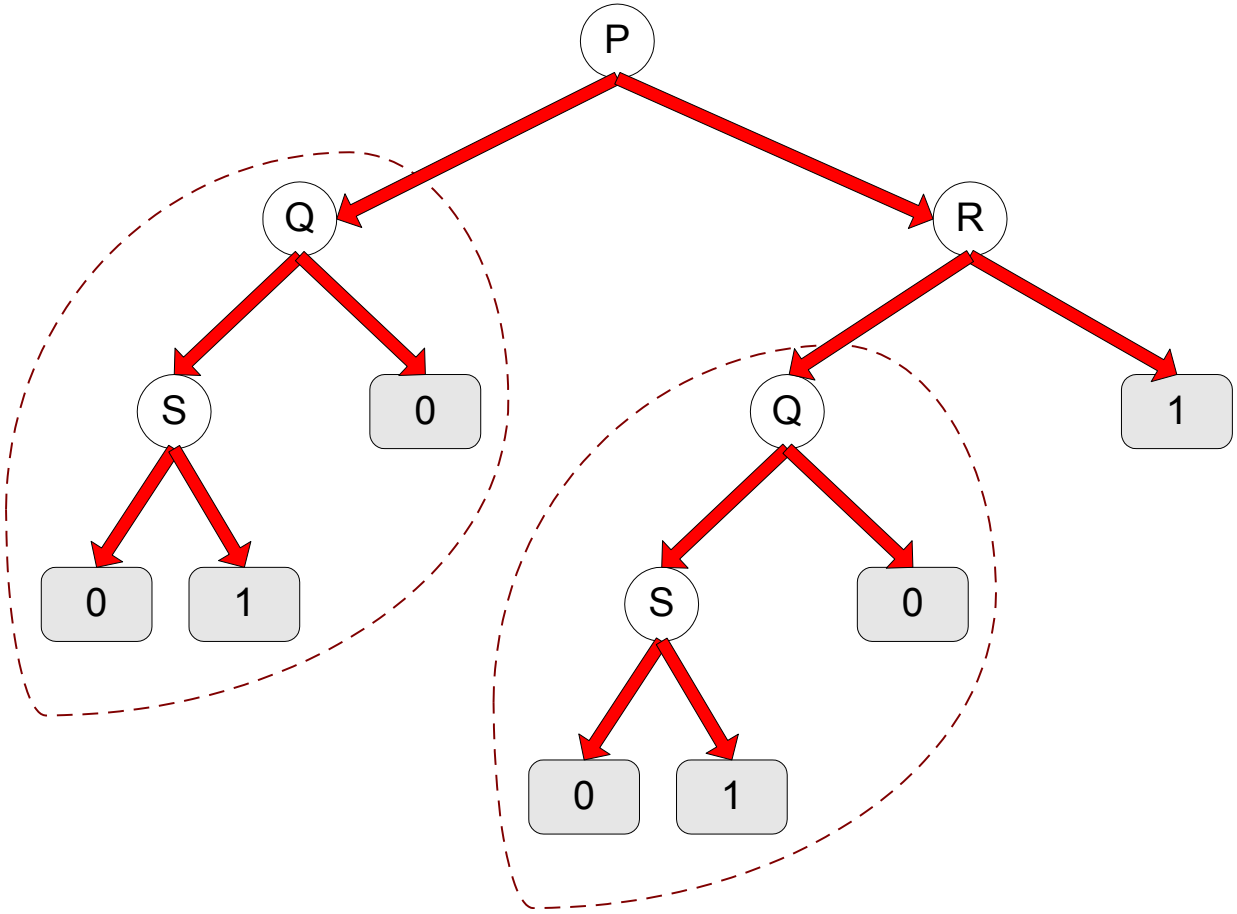




F1	F2	C
+	+	1
-	+	0
+	-	0
-	-	1

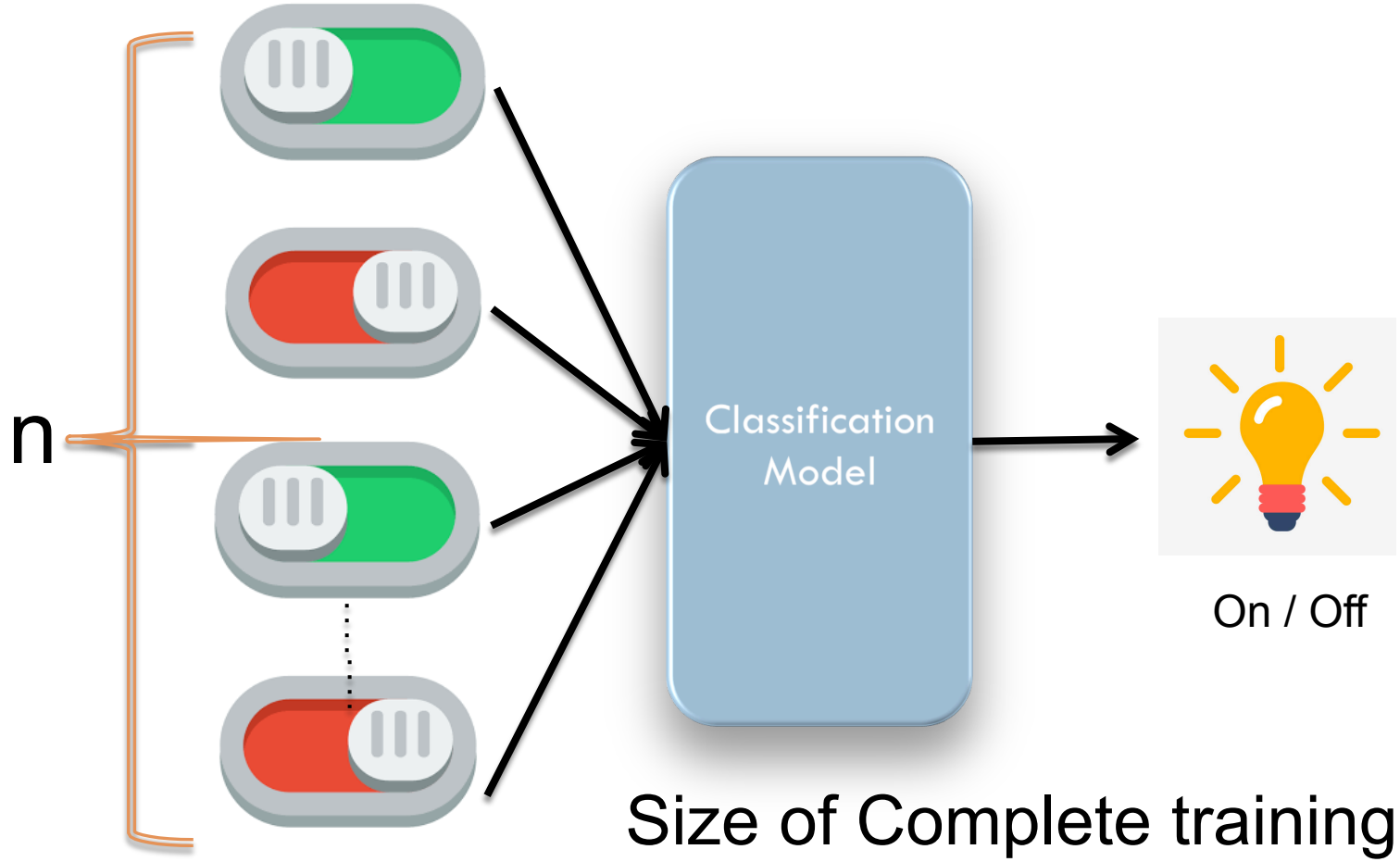
- Test condition may involve multiple attributes
- More expressive representation
- Finding optimal test condition is computationally expensive

Tree Replication



- Same subtree appears in multiple branches

Curse of Dimensionality



Size of Complete training data=2^n

Decision tree Example

A	B	C	D	class
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1

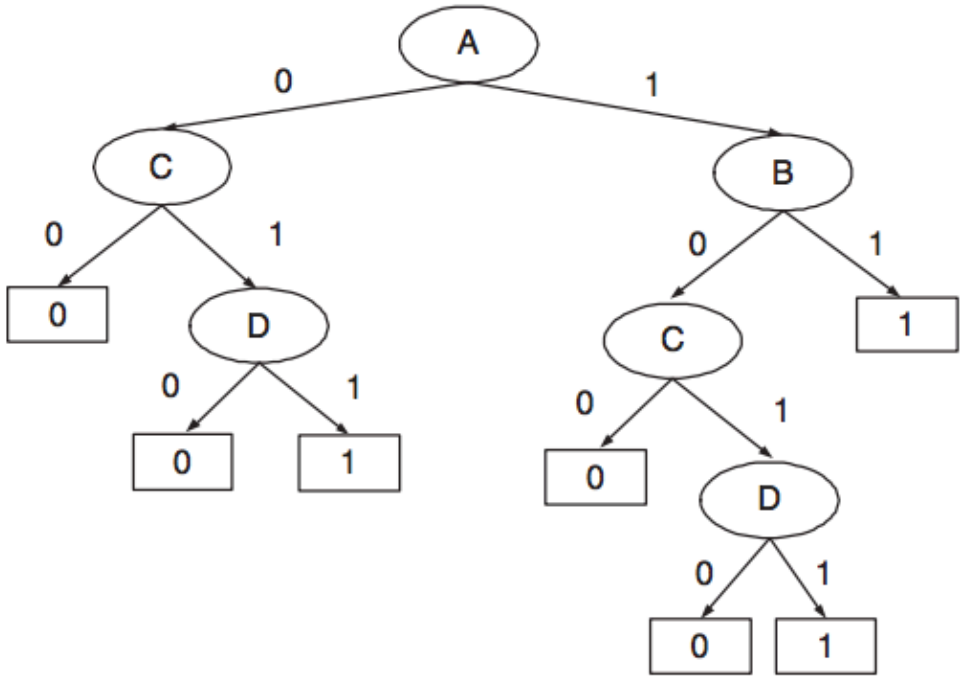


Figure 3.17. Decision tree for the Boolean function $(A \wedge B) \vee (C \wedge D)$.

Complex Decision Tree

- Regression Tree, Random Forest

`sklearn.ensemble`: Ensemble Methods

The `sklearn.ensemble` module includes ensemble-based methods for classification, regression and anomaly detection.

User guide: See the Ensemble methods section for further details.

`ensemble.AdaBoostClassifier` ([…])	An AdaBoost classifier.
`ensemble.AdaBoostRegressor` ([base_estimator, …])	An AdaBoost regressor.
`ensemble.BaggingClassifier` ([base_estimator, …])	A Bagging classifier.
`ensemble.BaggingRegressor` ([base_estimator, …])	A Bagging regressor.
`ensemble.ExtraTreesClassifier` ([…])	An extra-trees classifier.
`ensemble.ExtraTreesRegressor` ([n_estimators, …])	An extra-trees regressor.
`ensemble.GradientBoostingClassifier` ([loss, …])	Gradient Boosting for classification.
`ensemble.GradientBoostingRegressor` ([loss, …])	Gradient Boosting for regression.
`ensemble.IsolationForest` ([n_estimators, …])	Isolation Forest Algorithm
`ensemble.RandomForestClassifier` ([…])	A random forest classifier.
`ensemble.RandomForestRegressor` ([…])	A random forest regressor.
`ensemble.RandomTreesEmbedding` ([…])	An ensemble of totally random trees.
`ensemble.VotingClassifier` (estimators[, …])	Soft Voting/Majority Rule classifier for unfitted estimators.

Instance-Based Classifiers

Set of Stored Cases

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

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

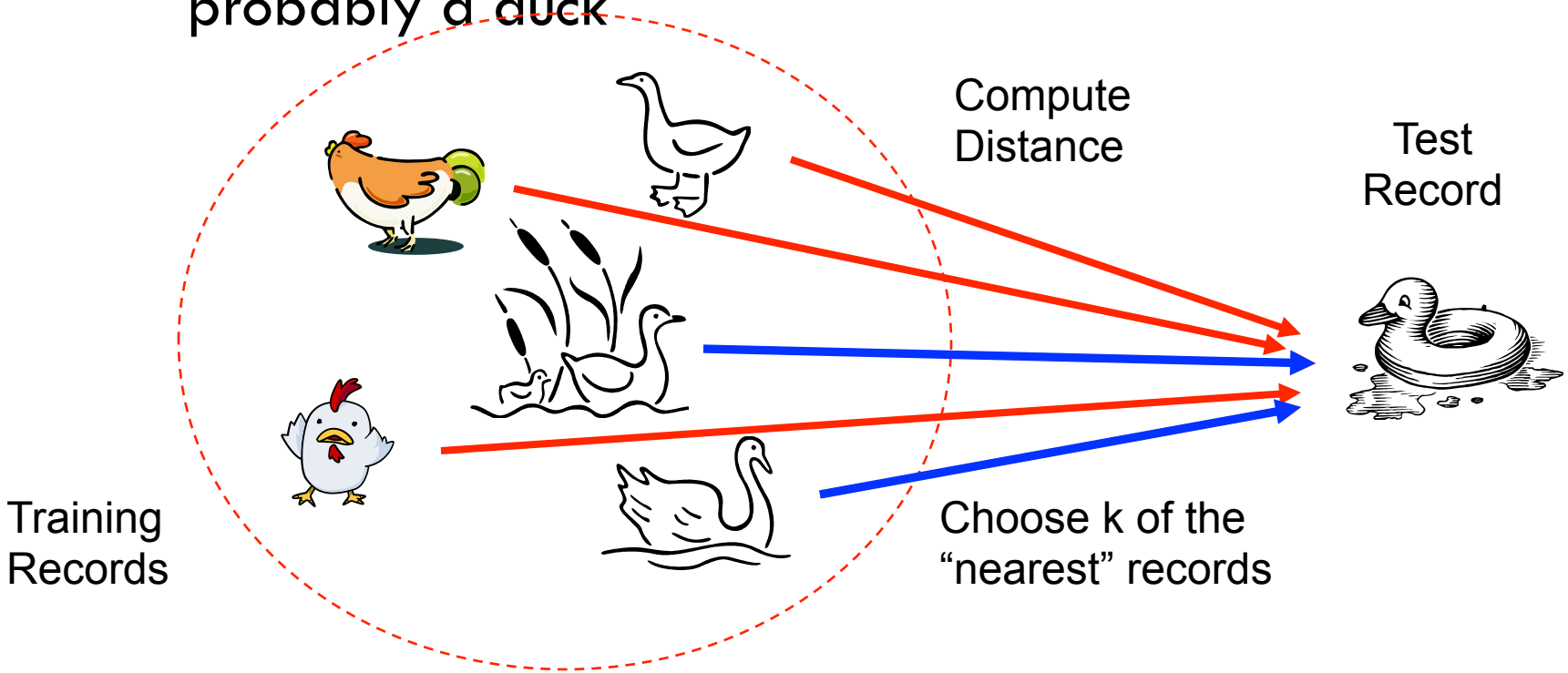
Unseen Case

Atr1	……...	AtrN

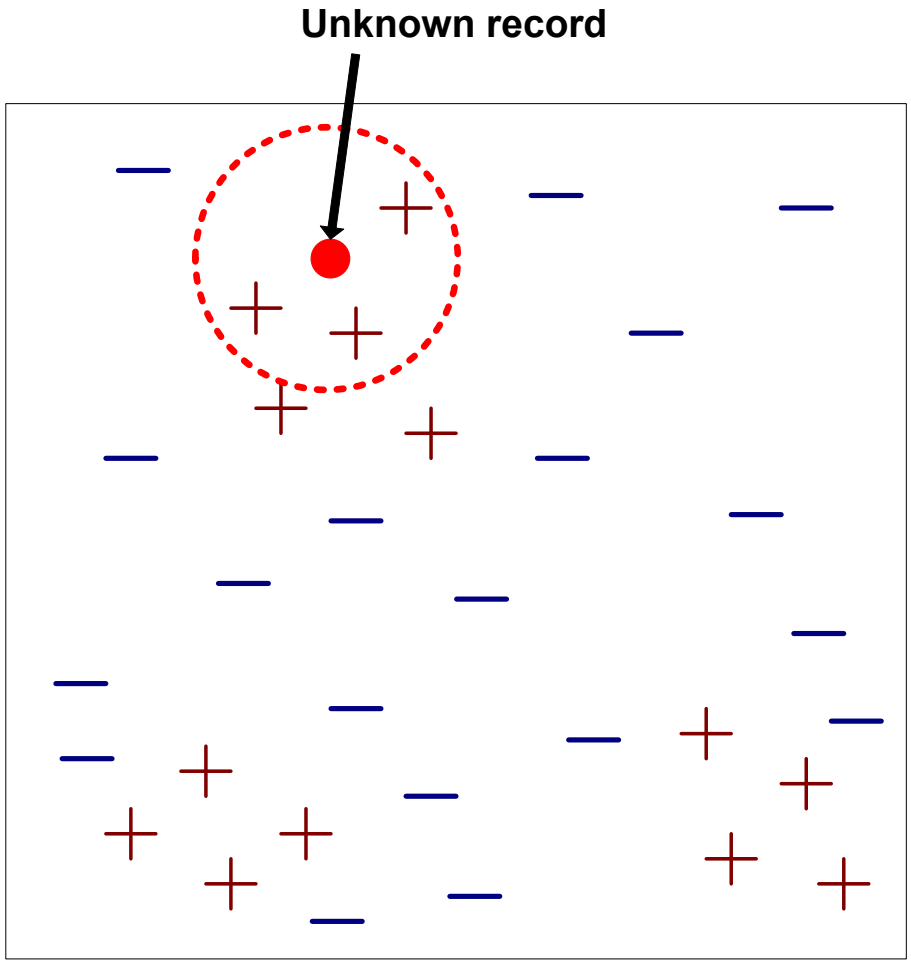
- Nearest neighbor: Uses k “closest” points (nearest neighbors) for performing classification

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



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)

Nearest Neighbor Classification

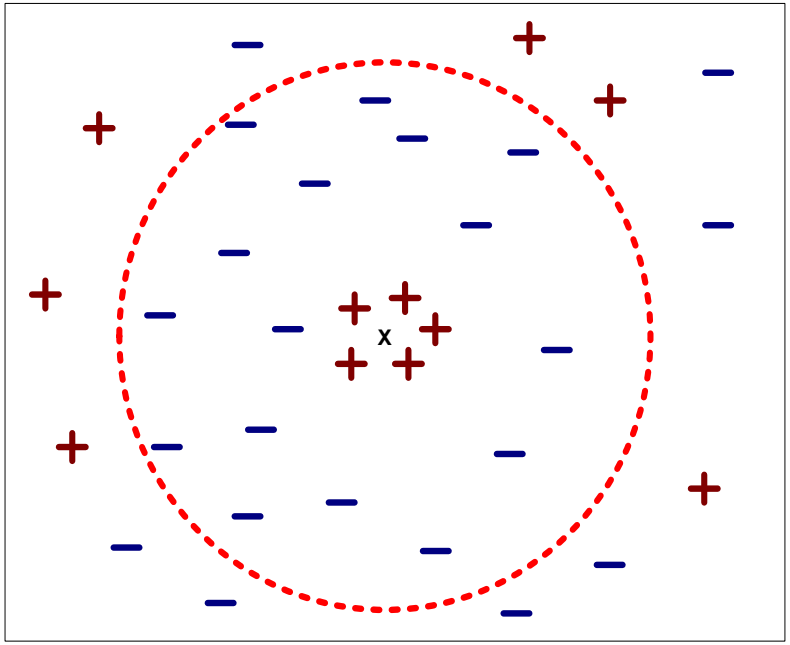
- k-NN classifiers are lazy learners
  - It does not build models explicitly
- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, $w = 1/d^2$

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



Nearest Neighbor Classification…

- Scaling issues
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M

Nearest Neighbor Classification…

- Problem with Euclidean measure:
  - High dimensional data
    - curse of dimensionality
  - Can produce counter-intuitive results

1 1 1 1 1 1 1 1 1 1 1 0		1 0 0 0 0 0 0 0 0 0 0 0	
0 1 1 1 1 1 1 1 1 1 1 1	vs	0 0 0 0 0 0 0 0 0 0 0 1	
d = 1.4142		d = 1.4142	

- Solution: Normalize the vectors to unit length

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

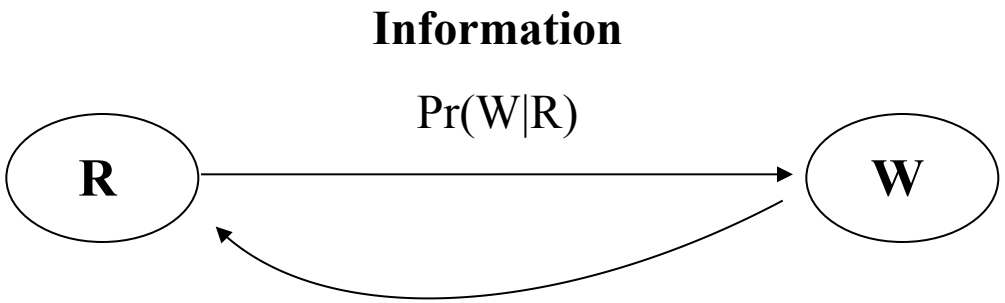
Bayes' Rule

If R= the set of favor news and W="*open source*"?

	R	¬R
W	0.7	0.4
¬W	0.3	0.6

R: It rains

W: The grass is wet



100d	80d	20d
	R	¬R
W	56	8
¬W	24	12

$$\frac{\Pr(W \mid R)\Pr(R)}{\Pr(W)} = \frac{0.7*0.8}{56+8/100} = \frac{0.7*0.8}{0.64}$$

$$\frac{\Pr(W \mid R)\Pr(R)}{\Pr(W)} = \frac{\Pr(W \mid R)\Pr(R)}{\Pr(W \mid R)\Pr(R)+\Pr(W \mid \neg R)\Pr(\neg R)} = \frac{0.7*0.8}{0.7*0.8+0.4*0.2} = \frac{7}{8}$$

Bayesian Classifiers

- Consider each attribute and class label as random variables

- Given a record with attributes (A_1, A_2,…,A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)

- Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability P(C | A_1, A_2, …, A_n) for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes
    P(C | A_1, A_2, …, A_n)
  - Equivalent to choosing value of C that maximizes
    P(A_1, A_2, …, A_n|C) P(C)

- How to estimate P(A_1, A_2, …, A_n | C)?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_i)\, P(A_2 \mid C_i) \ldots P(A_n \mid C_i)$
  - Can estimate $P(A_i \mid C_i)$ for all A_i and C_i.
  - New point is classified to C_i if $P(C_i) \prod P(A_i \mid C_i)$ is maximal.

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

- Class: $P(C) = N_c/N$
  - e.g., P(No) = 7/10,
    P(Yes) = 3/10
- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

P(Status=Married|No) = 4/7
P(Refund=Yes|Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption
  - Two-way split: $(A < v)$ or $(A > v)$
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i | c)$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (A_i, c_i) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

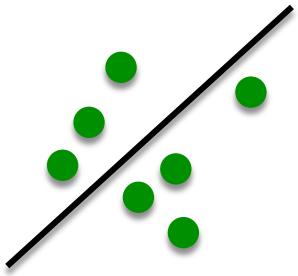
P(A|M)P(M) > P(A|N)P(N)

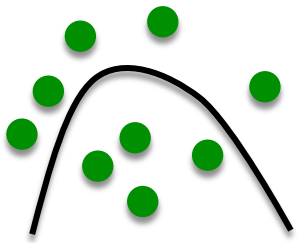
=> Mammals

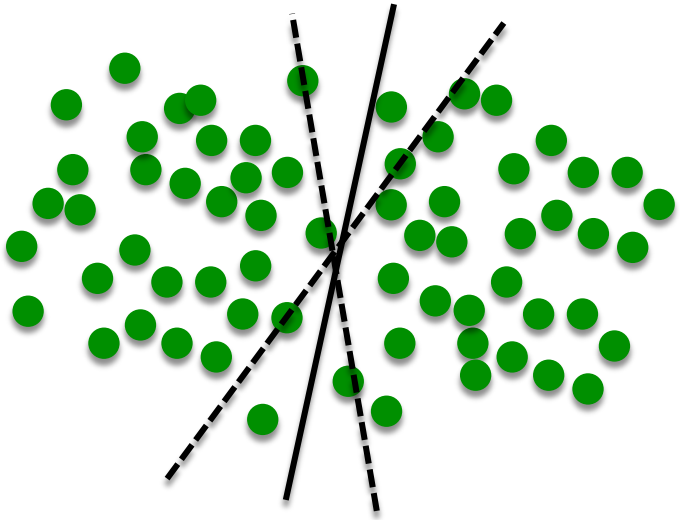
Naïve Bayes (Summary)

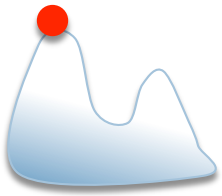
- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as *Bayesian Belief Networks* (BBN)

Regression





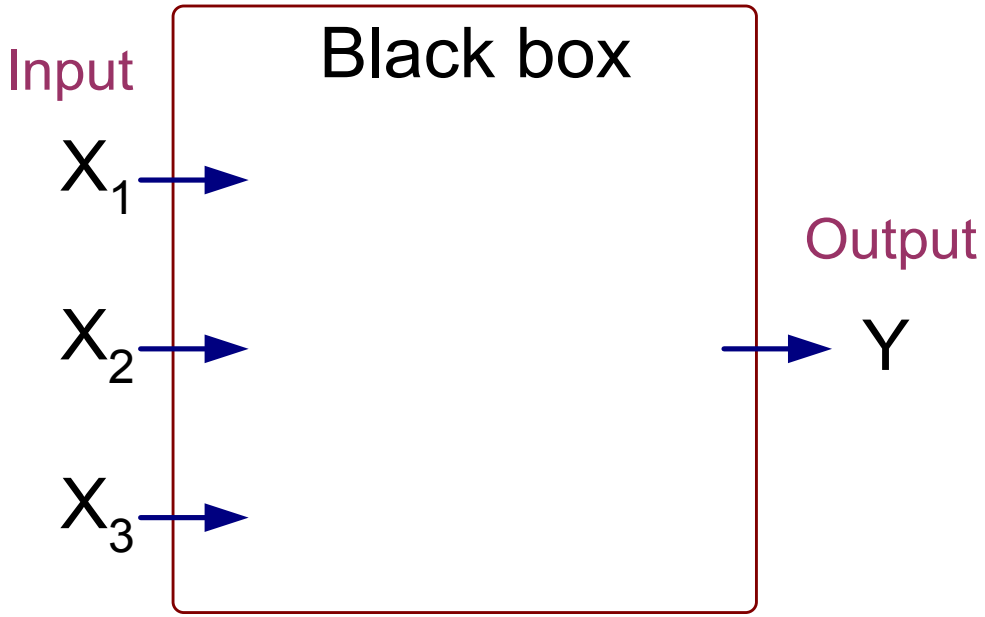




Gradient Descent

Artificial Neural Networks (ANN)

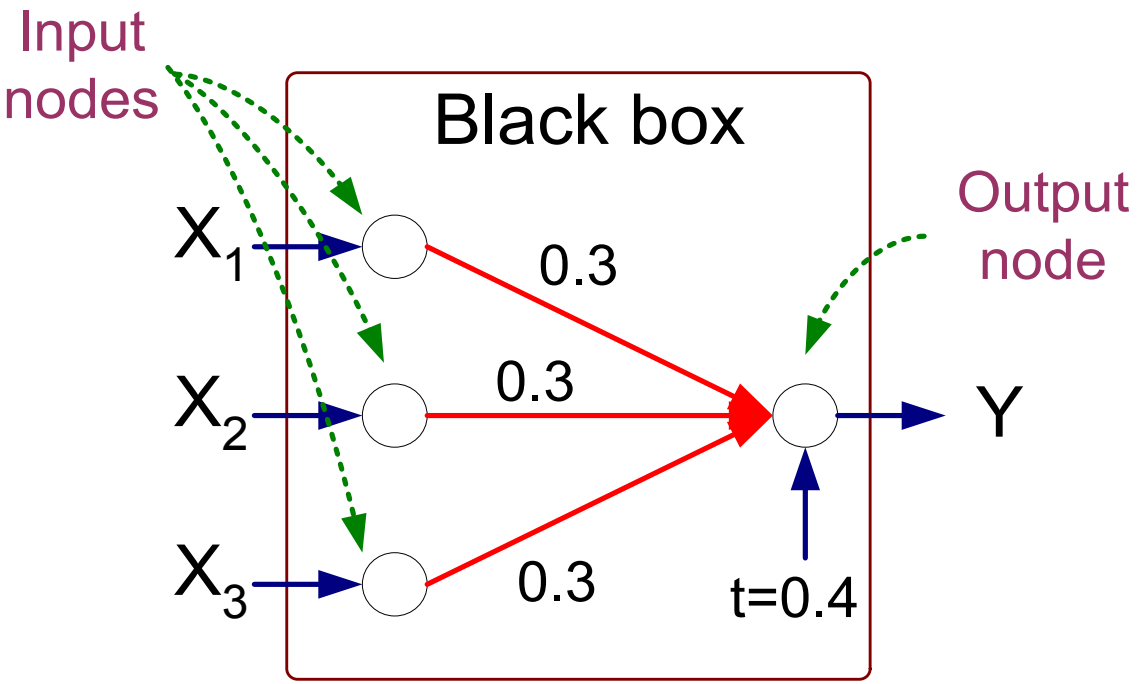
X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0



Output Y is 1 if at least two of the three inputs are equal to 1.

Artificial Neural Networks (ANN)

X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0

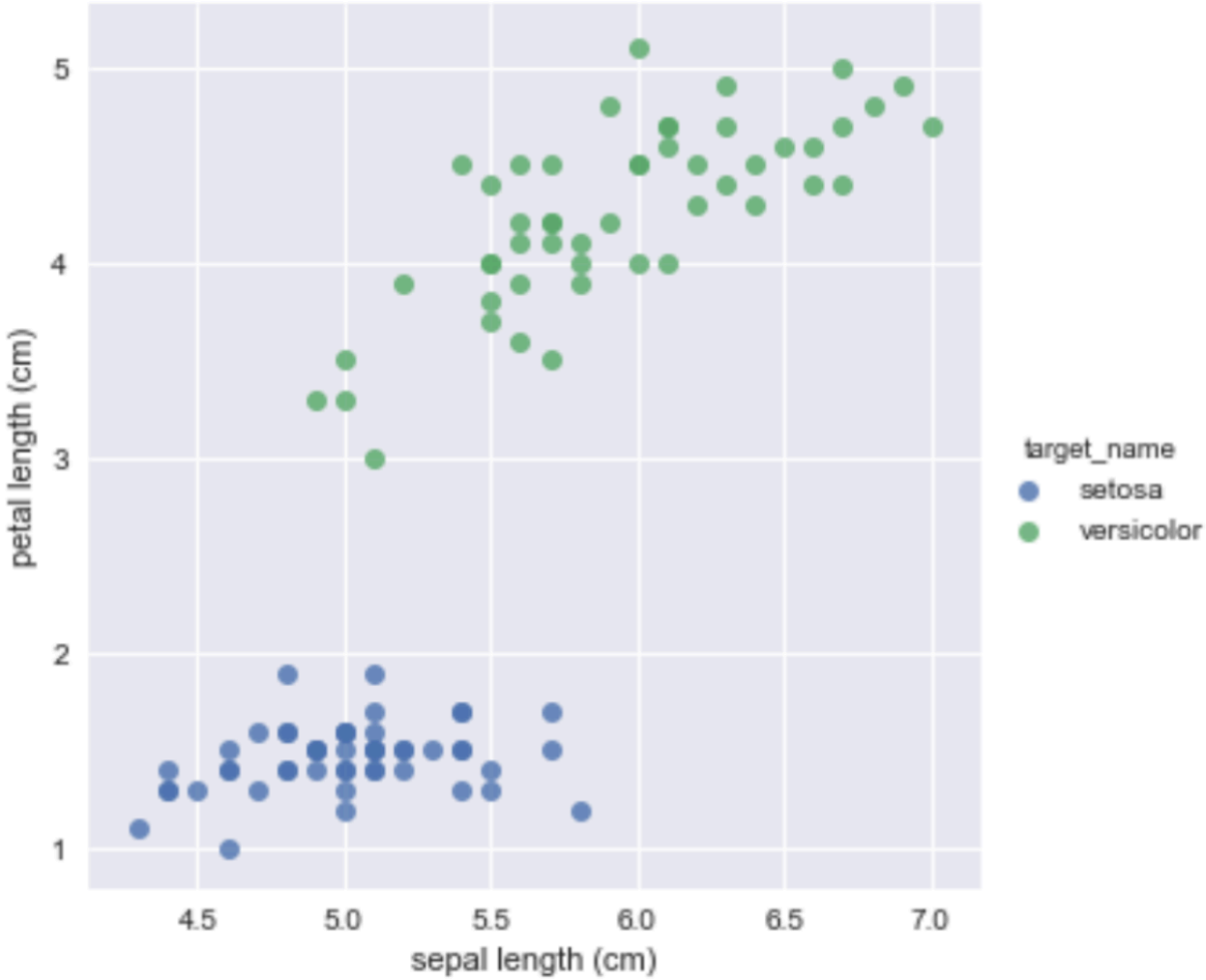


Perceptron (Perceptron Learning Algorithm, PLA)

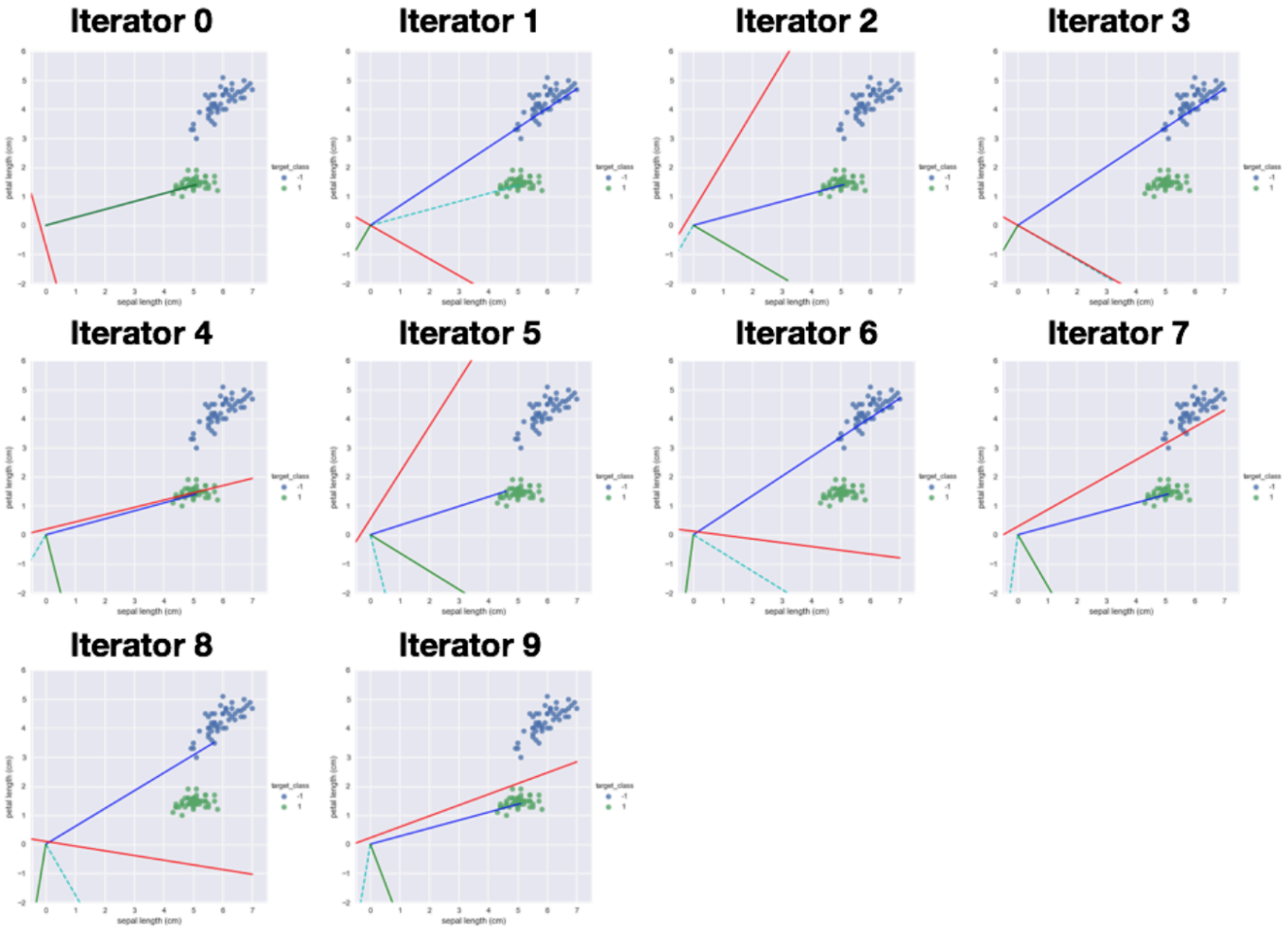
$$Y = I(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4 > 0)$$

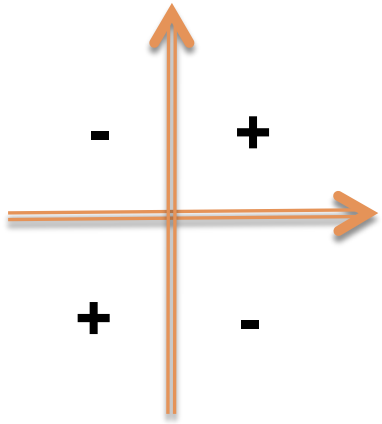
$$\text{where } I(z) = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Perceptron Example



Perceptron Example

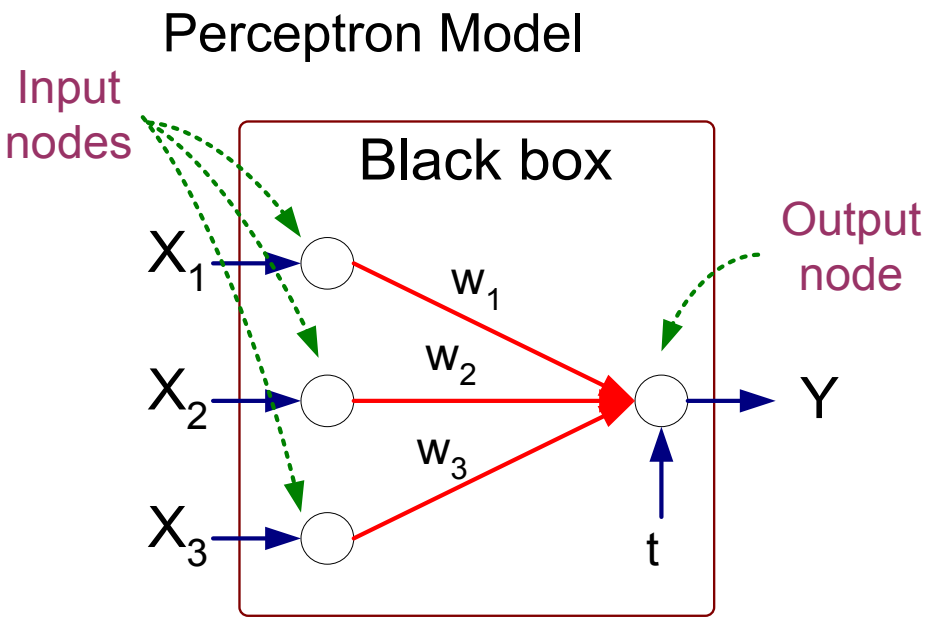




X1	X2	C
T	T	+
F	T	-
T	F	-
F	F	+

Artificial Neural Networks (ANN)

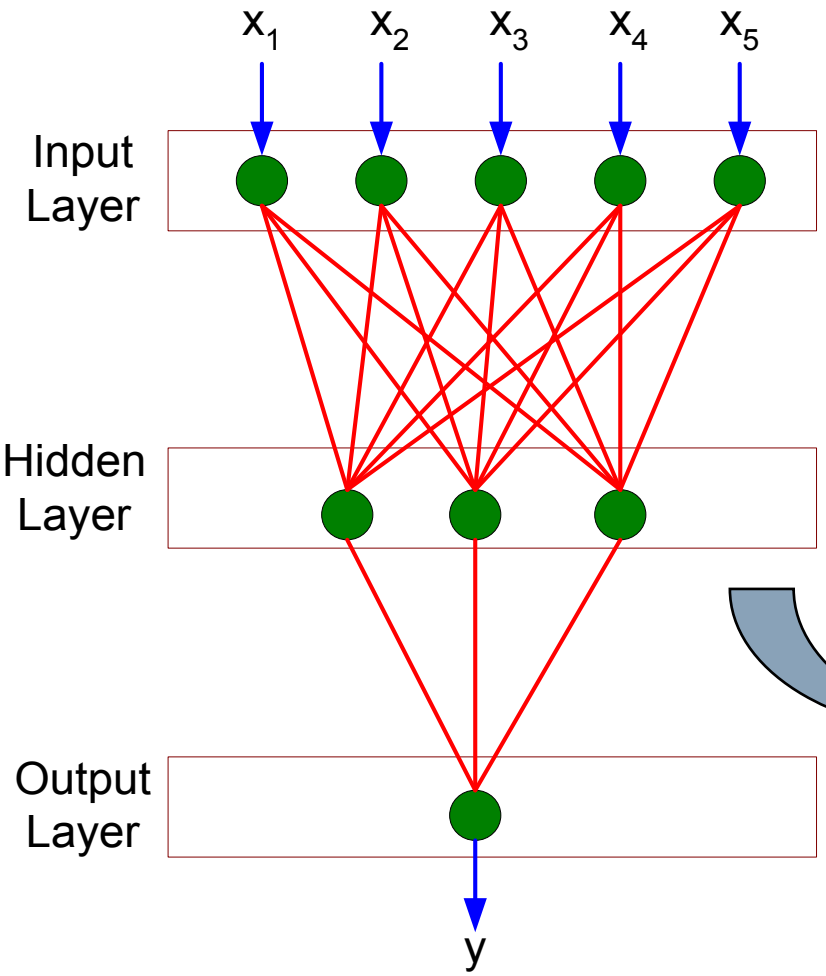
- Model is an assembly of inter-connected nodes and weighted links
- Output node sums up each of its input value according to the weights of its links
- Compare output node against some threshold t

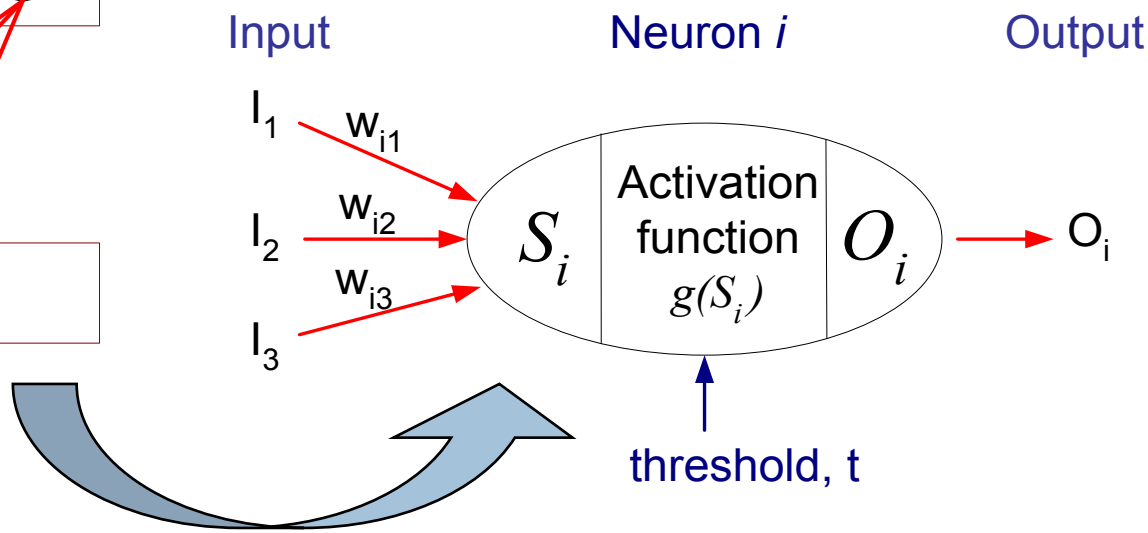


$$Y = I(\sum_i w_i X_i - t) \quad \text{or}$$

$$Y = sign(\sum_i w_i X_i - t)$$

General Structure of ANN





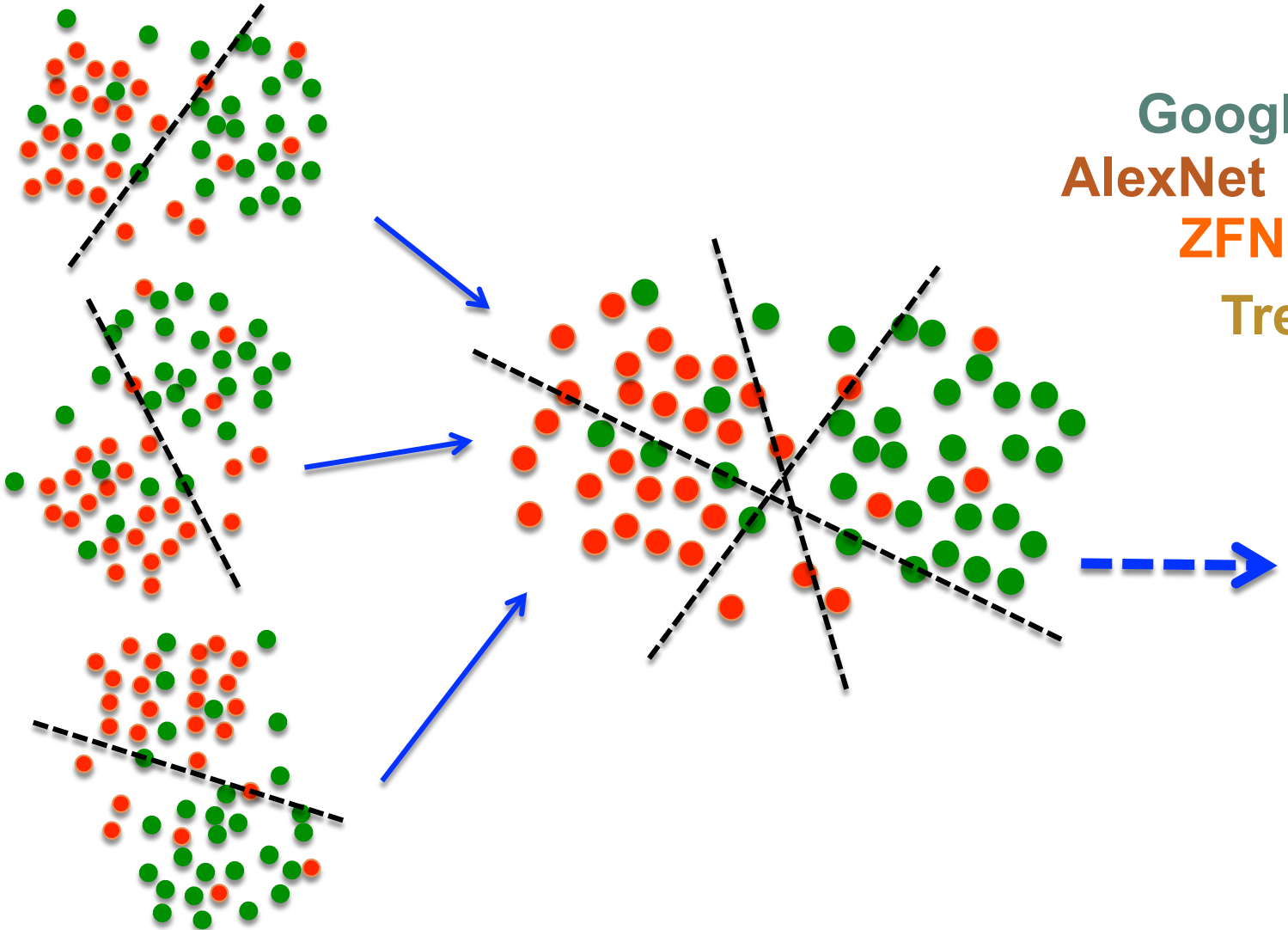
Training ANN means learning the weights of the neurons

Algorithm for learning ANN

- Initialize the weights (w_0, w_1, …, w_k)

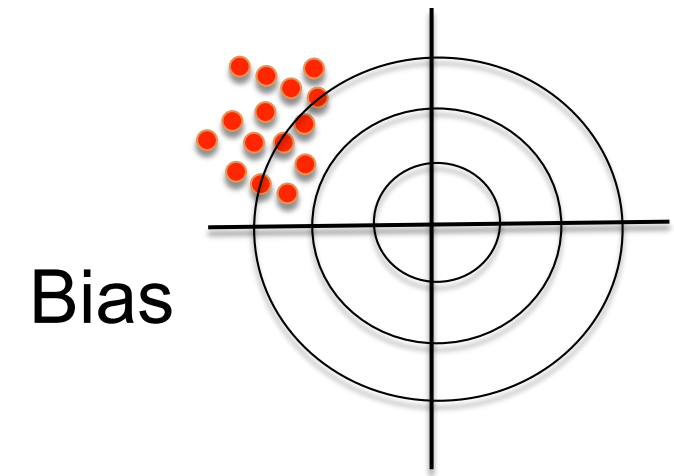
- Adjust the weights in such a way that the output of ANN is consistent with class labels of training examples
  - Objective function: $E = \sum_i [Y_i - f(w_i, X_i)]^2$
  - Find the weights w_i's that minimize the above objective function
    - e.g., backpropagation algorithm

ANN

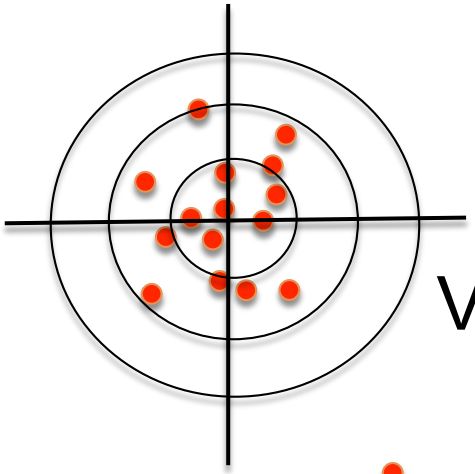


GoogleNet .....
AlexNet .....
ResNet
ZFNet .....
TreeLSTM
....

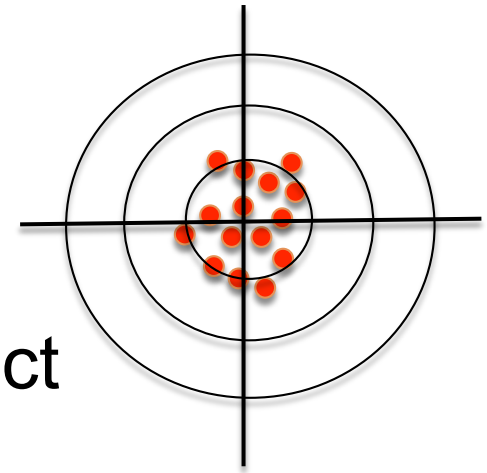
Bias & Variance



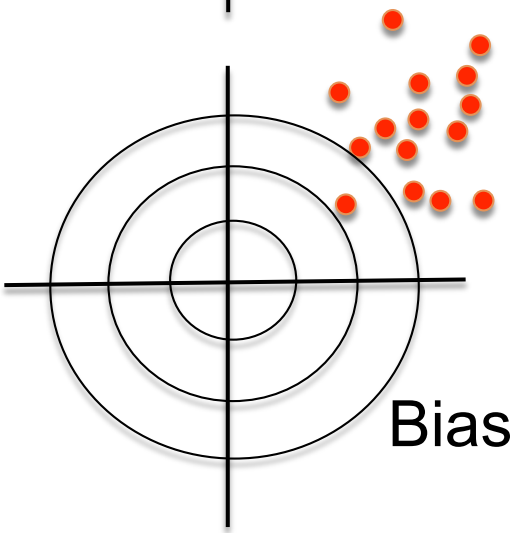
Bias



Variance



Perfect



Bias+Variance

Bias-Variance Trade-Off

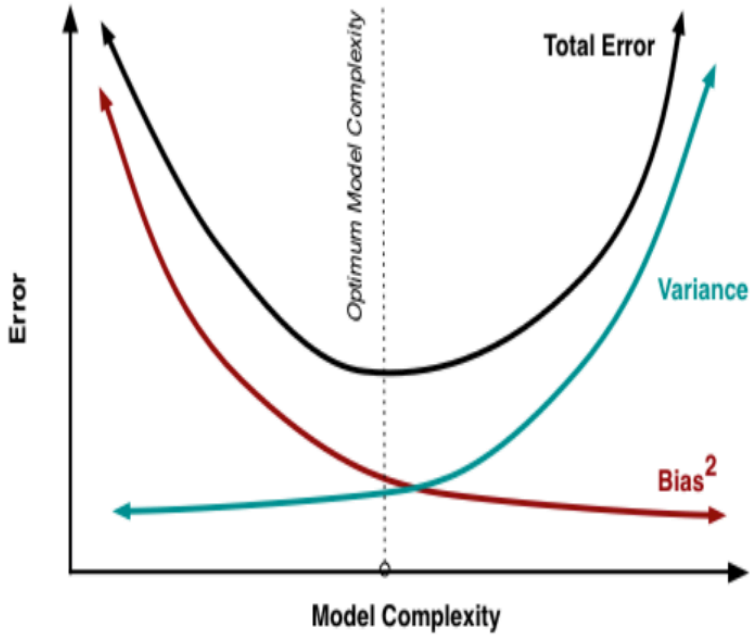
As model complexity increases we see the characteristic U-shape of total error

We reduce bias to an optimum where increasing variance starts to dominate

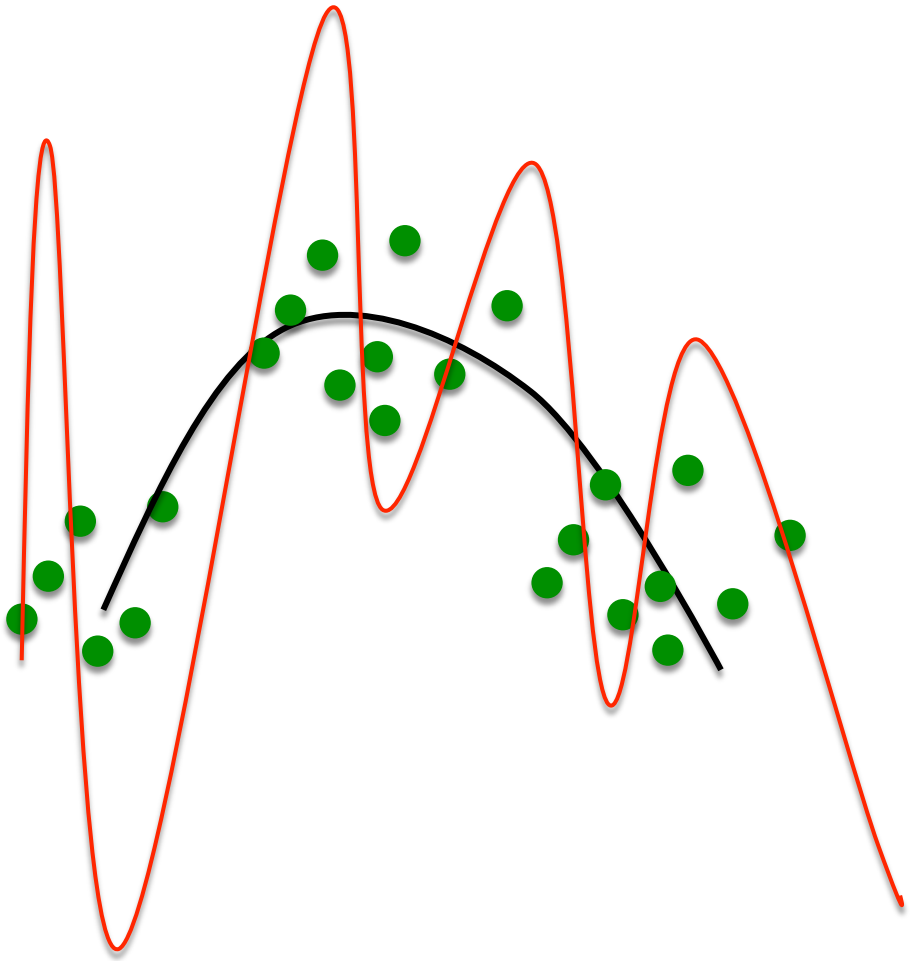
Finding the right balance is a key machine learning skill

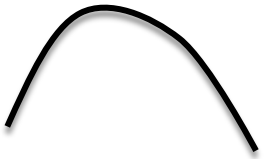
High bias, low variance = **underfitting**

Low bias, high variance = **overfitting**



Underfitting, overfitting

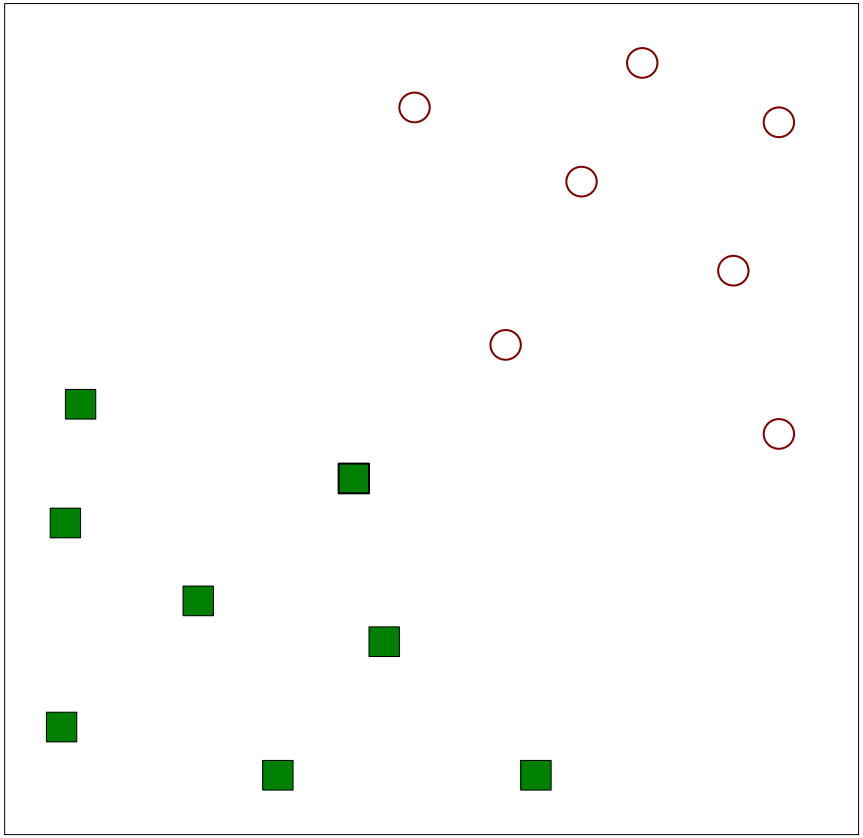




High bias, low variance
= **underfitting**

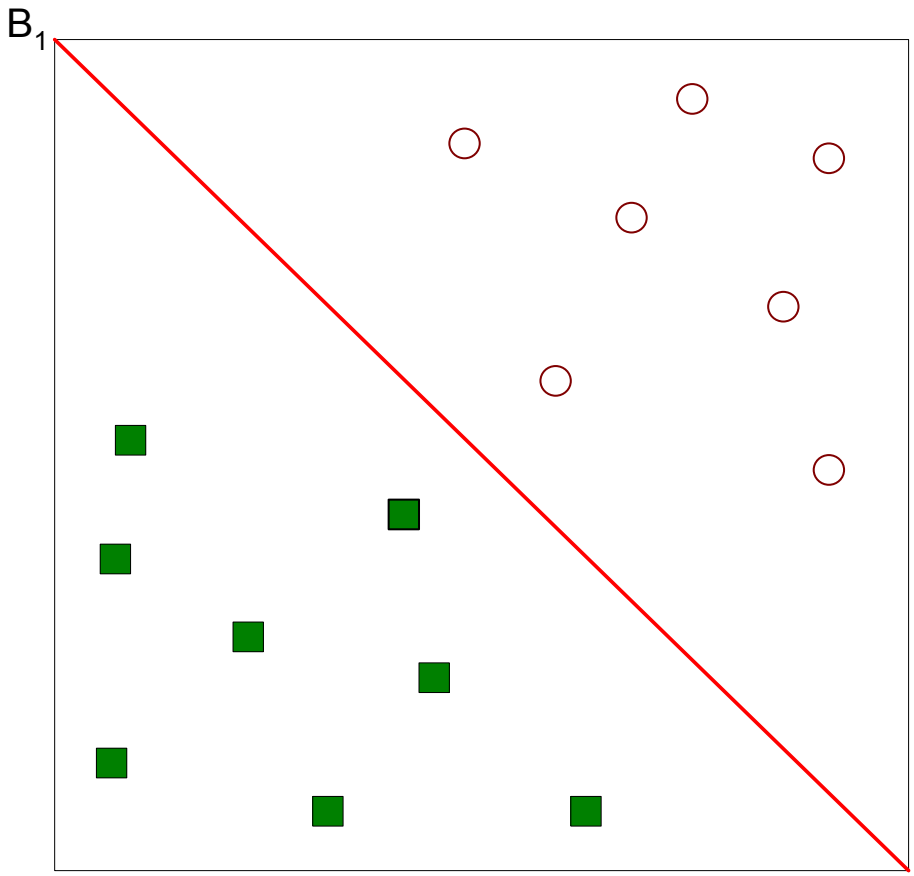
Low bias, high variance
= **overfitting**

Support Vector Machines



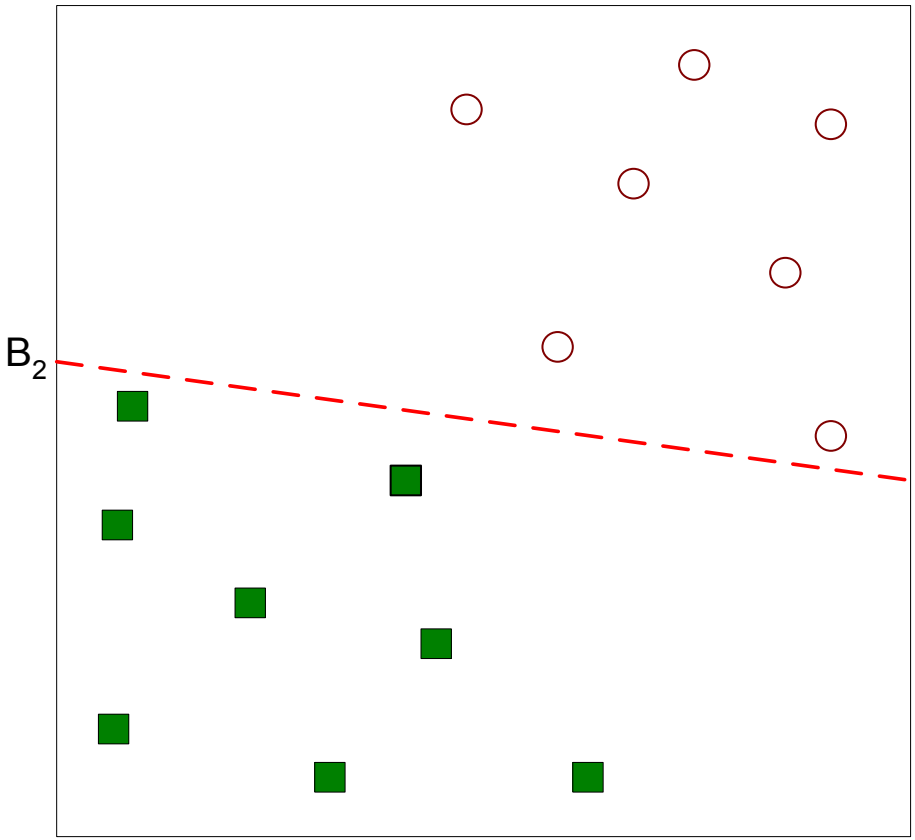
- Find a linear hyperplane (decision boundary) that will separate the data

Support Vector Machines



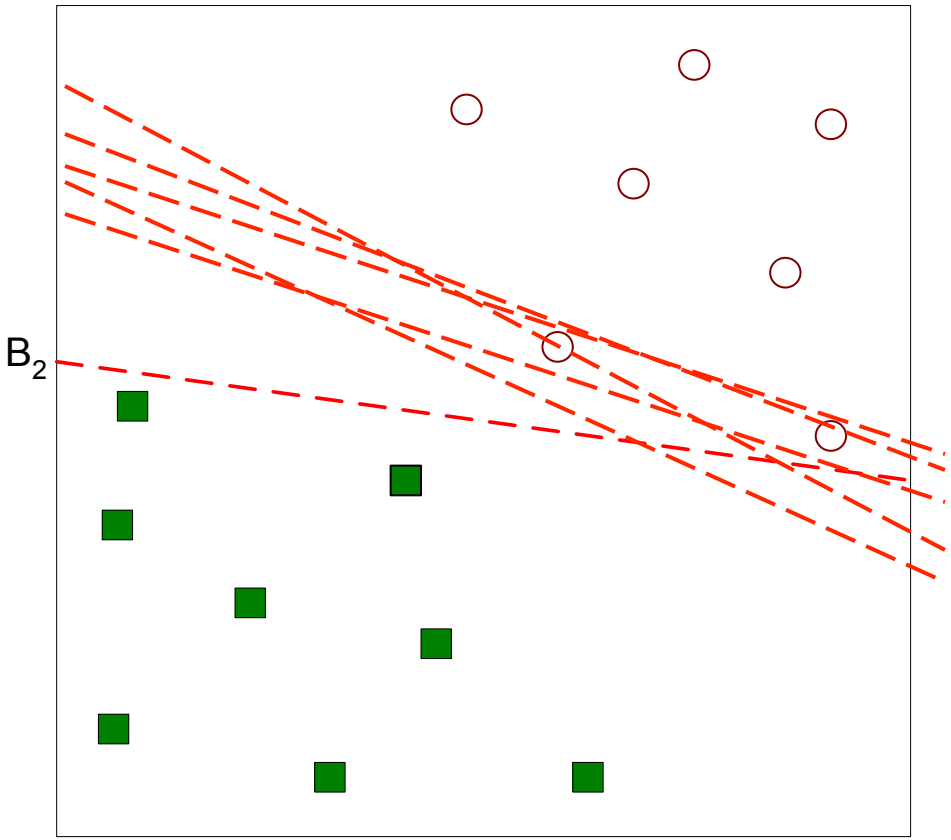
- One Possible Solution

Support Vector Machines



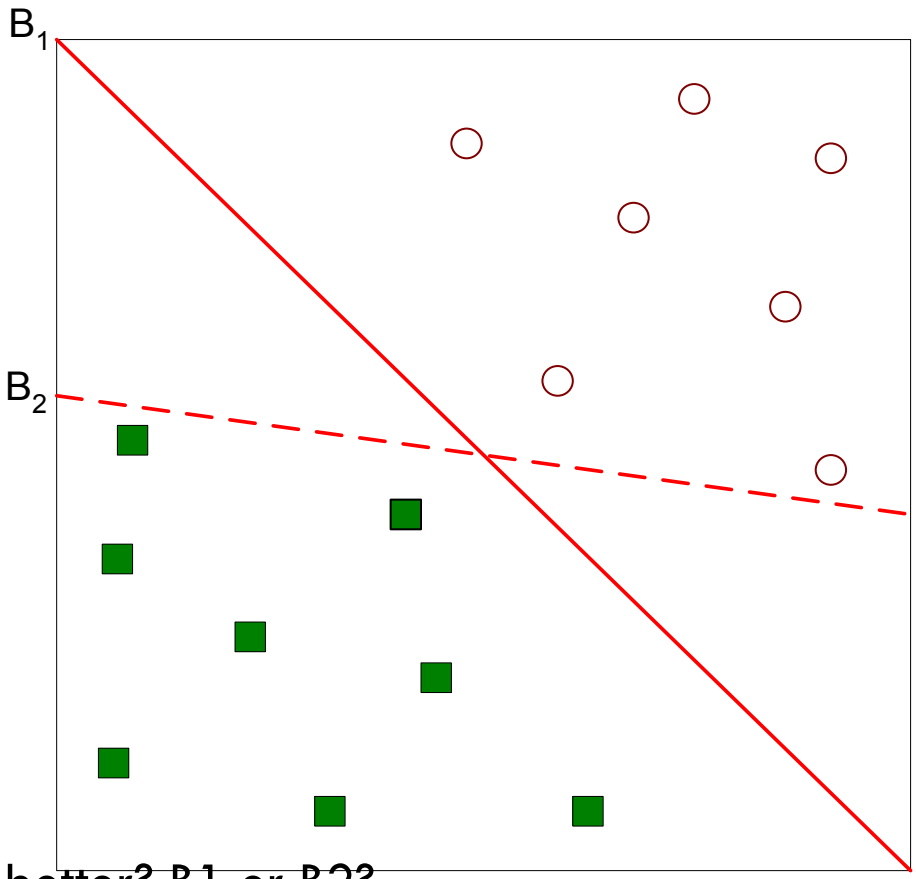
- Another possible solution

Support Vector Machines



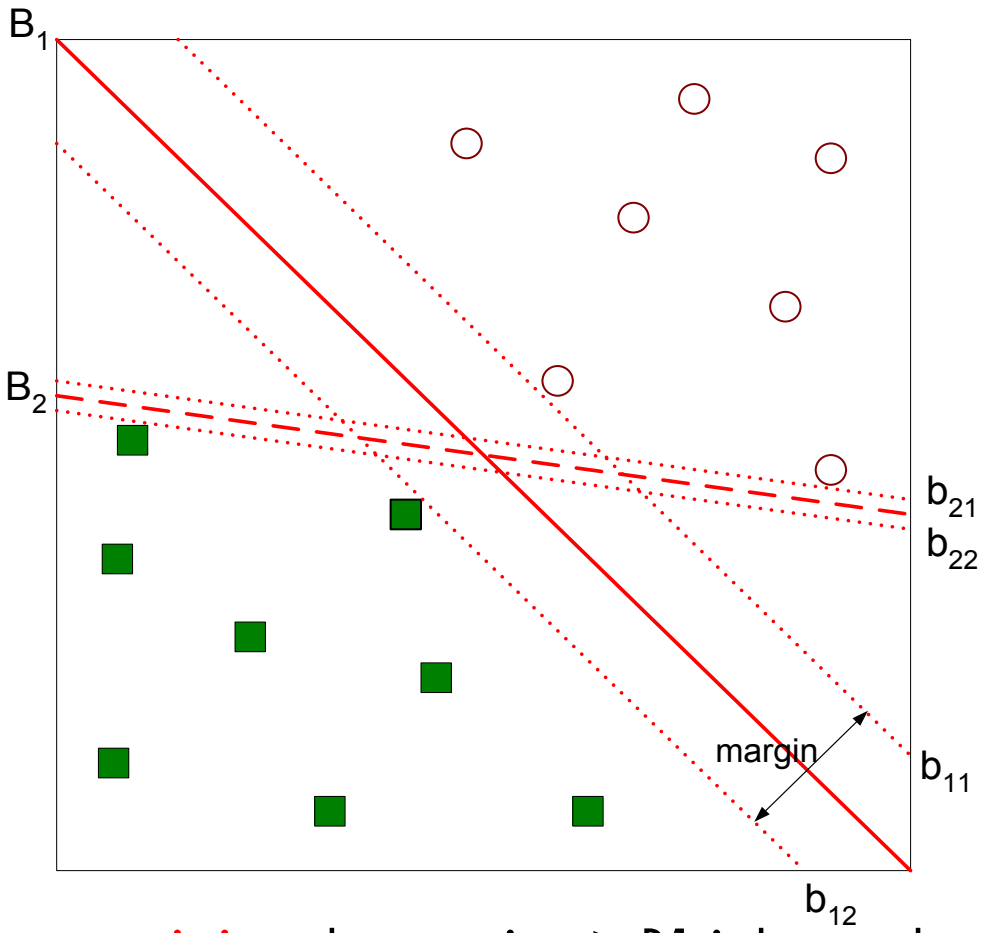
- Other possible solutions

Support Vector Machines



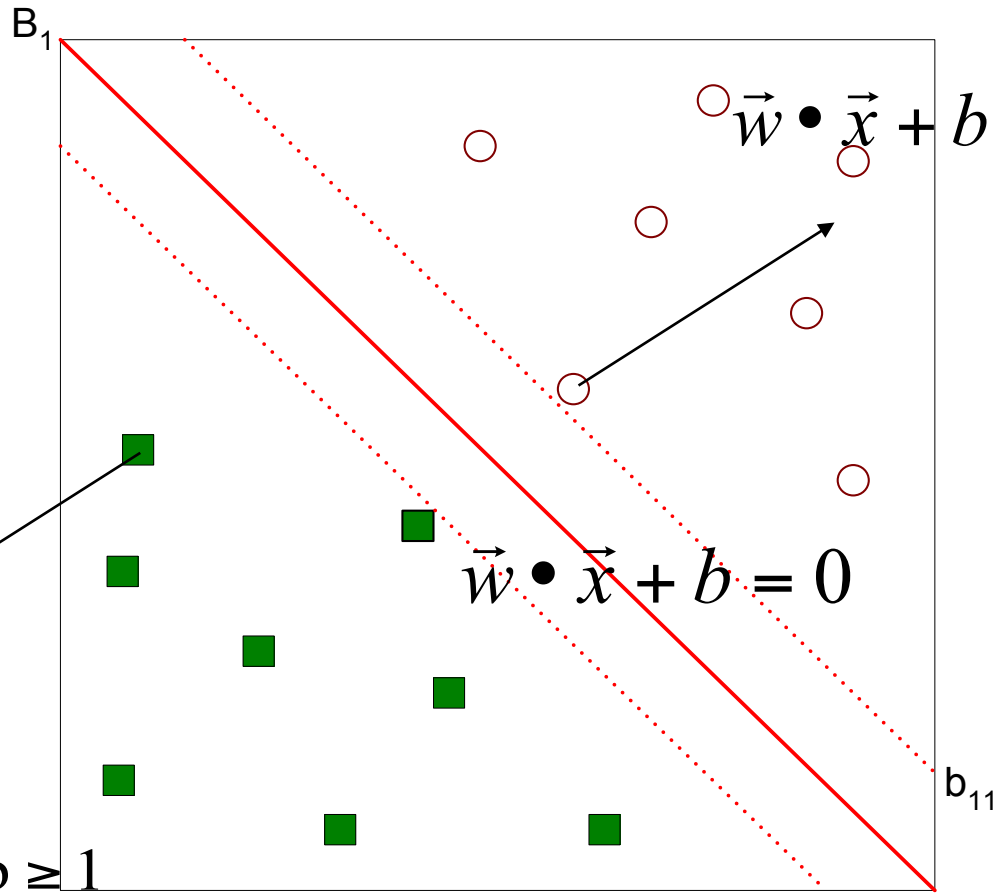
- Which one is better? B1 or B2?
- How do you define better?

Support Vector Machines



- Find hyperplane maximizes the margin => B1 is better than B2

Support Vector Machines



$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

$$\text{Margin} = \frac{2}{\| \vec{w} \|^2}$$

Support Vector Machines

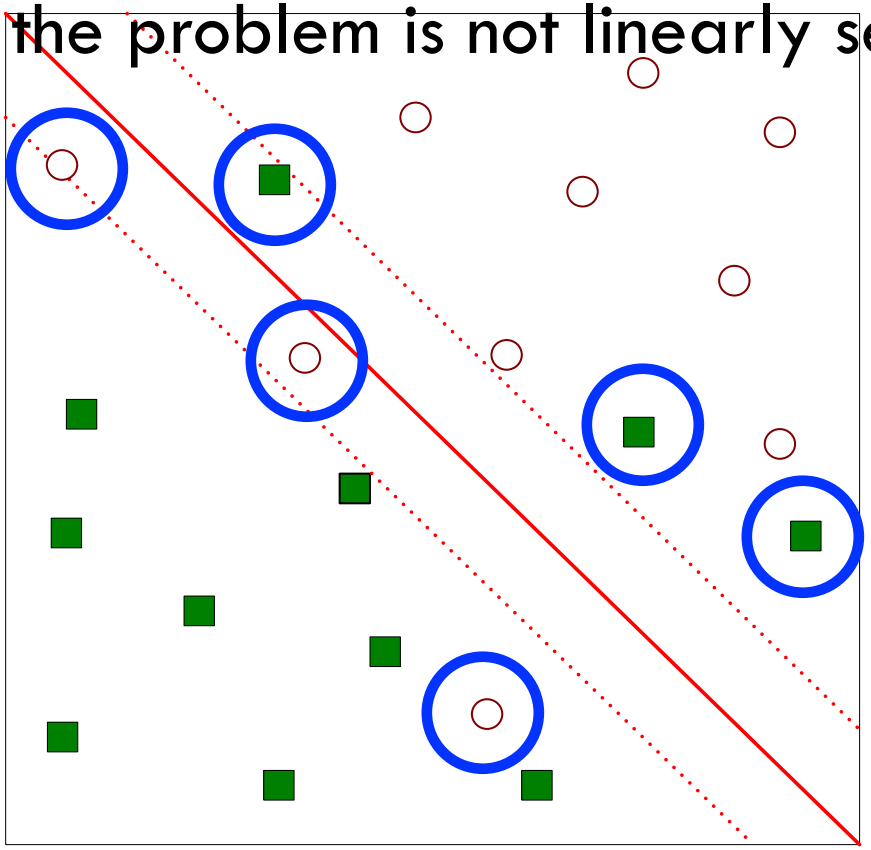
- We want to maximize: $\text{Margin} = \frac{2}{\| \vec{w} \|^2}$
  - Which is equivalent to minimizing: $L(w) = \frac{\| \vec{w} \|^2}{2}$
  - But subjected to the following constraints:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$$

    - This is a constrained optimization problem
      - Numerical approaches to solve it (e.g., quadratic programming)

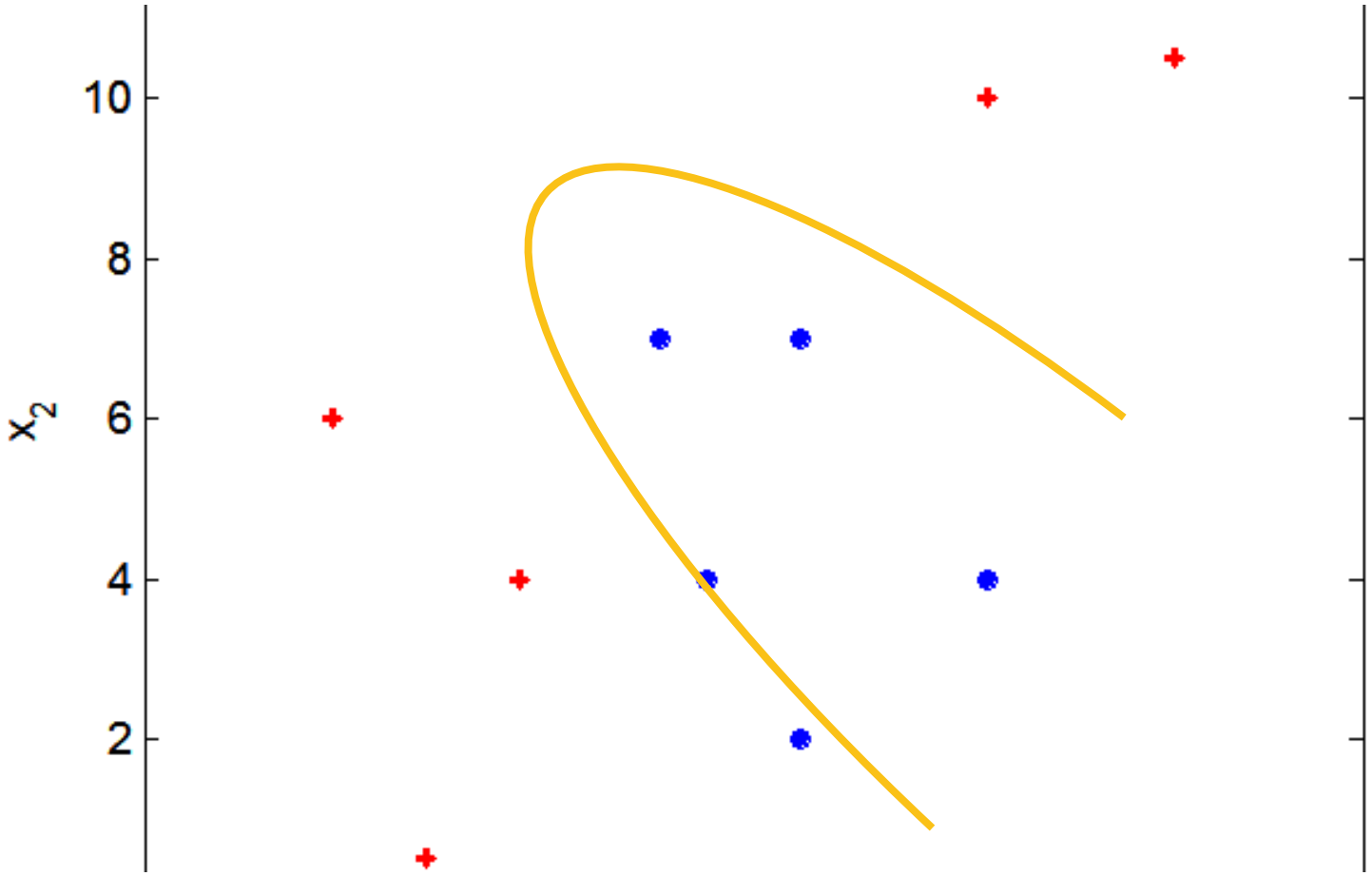
Support Vector Machines

- What if the problem is not linearly separable?



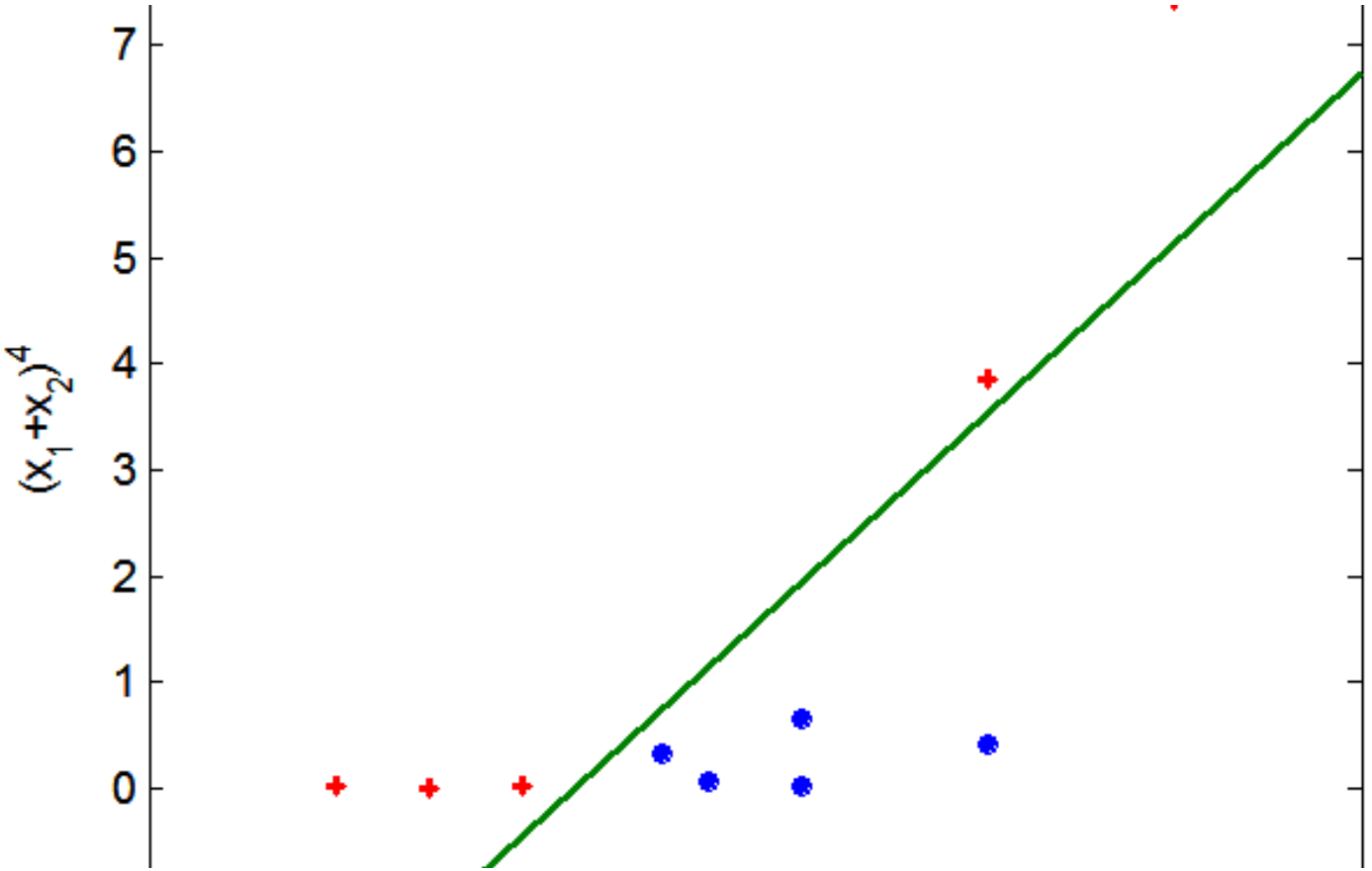
Nonlinear Support Vector Machines

- What if decision boundary is not linear?



Nonlinear Support Vector Machines

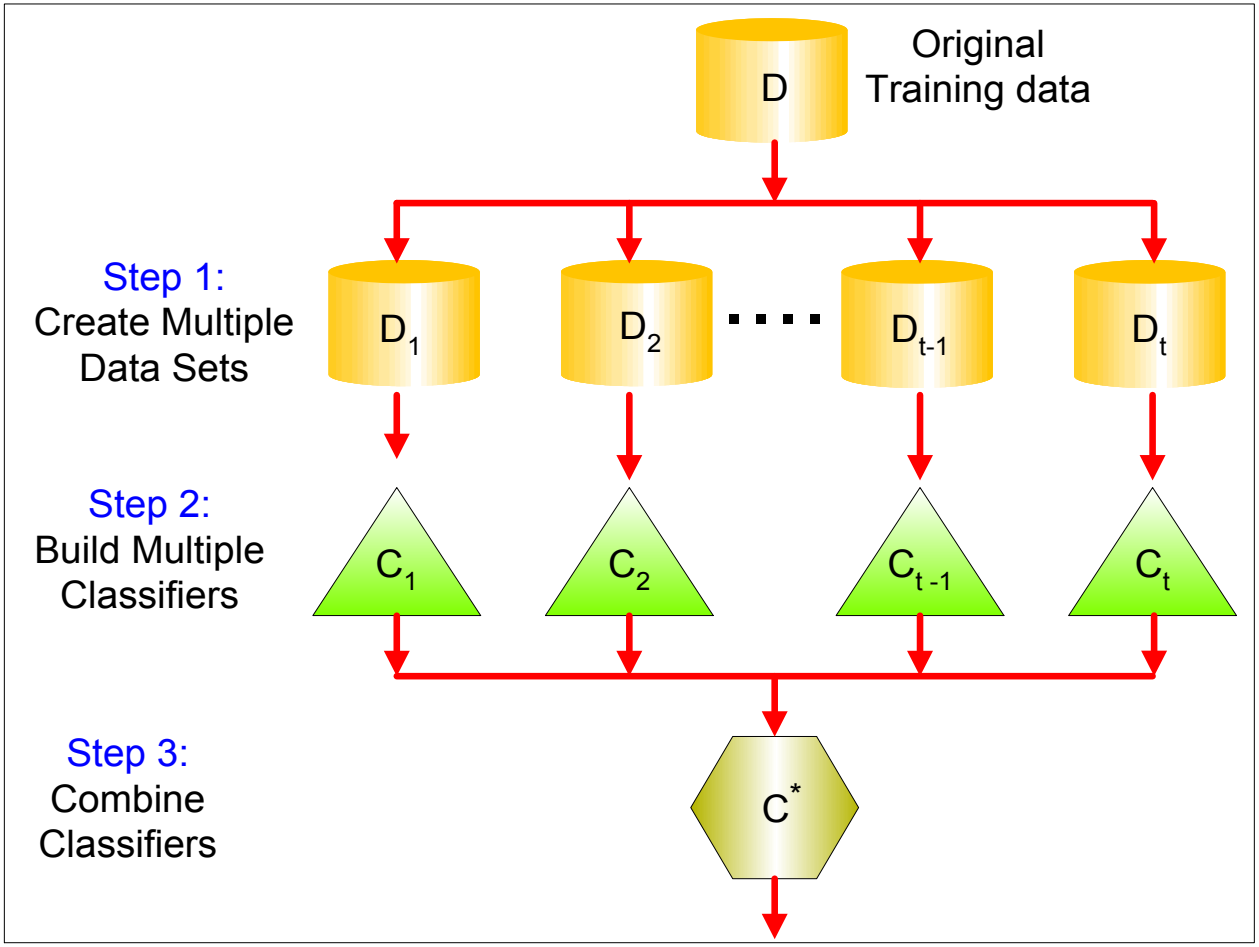
- Transform data into higher dimensional space



Ensemble Methods

- Construct a set of classifiers from the training data
- Predict class label of previously unseen records by aggregating predictions made by multiple classifiers

General Idea



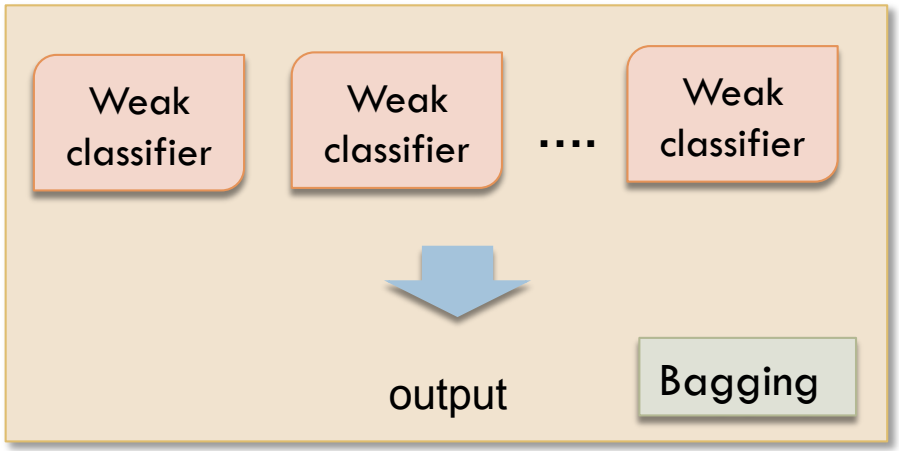
Why does it work?

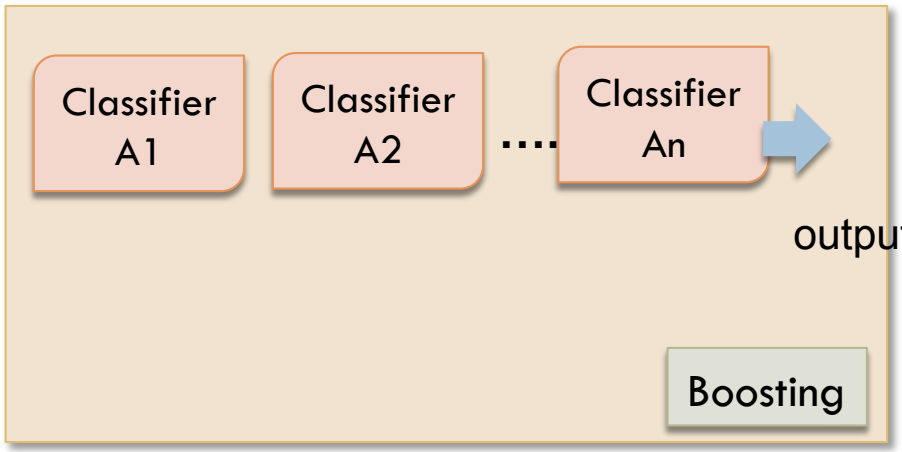
- Suppose there are 25 base classifiers
  - Each classifier has error rate, $\varepsilon = 0.35$
  - Assume classifiers are independent
  - Probability that the ensemble classifier makes a wrong prediction:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

Examples of Ensemble Methods

- How to generate an ensemble of classifiers?
  - Bagging
    - Breiman, 1996
    - Bootstrap Aggregating = Bagging
  - Boosting





Bagging

- Training
  - Create k bootstrap samples $S[1], S[2], \ldots, S[k]$
  - Build a distinct classifier on each $S[i]$ to produce k classifiers, using the same learning algorithm.
- Testing
  - Classify each new instance by voting of the k classifiers (equal weights)

Bagging (cont.)

- When does it help?
  - When learner is unstable
    - Small change to training set causes large change in the output classifier
    - True for decision trees, neural networks; not true for *k*-nearest neighbor, naïve Bayesian, class association rules
  - Experimentally, bagging can help substantially for **unstable learners**, may somewhat degrade results for stable learners

Bagging Predictors, Leo Breiman, 1996

Boosting

- A family of methods:
  - **AdaBoost (Adaptive Boosting)** (Freund & Schapire, 1996)
    - sensitive to noisy data and outliers
- Training
  - Produce a sequence of classifiers (the same base learner)
  - Each classifier is dependent on the previous one, and focuses on the previous one's errors
  - Examples that are incorrectly predicted in previous classifiers are given **higher weights**
- Testing
  - For a test case, the results of the series of classifiers are combined to determine the final class of the test case.

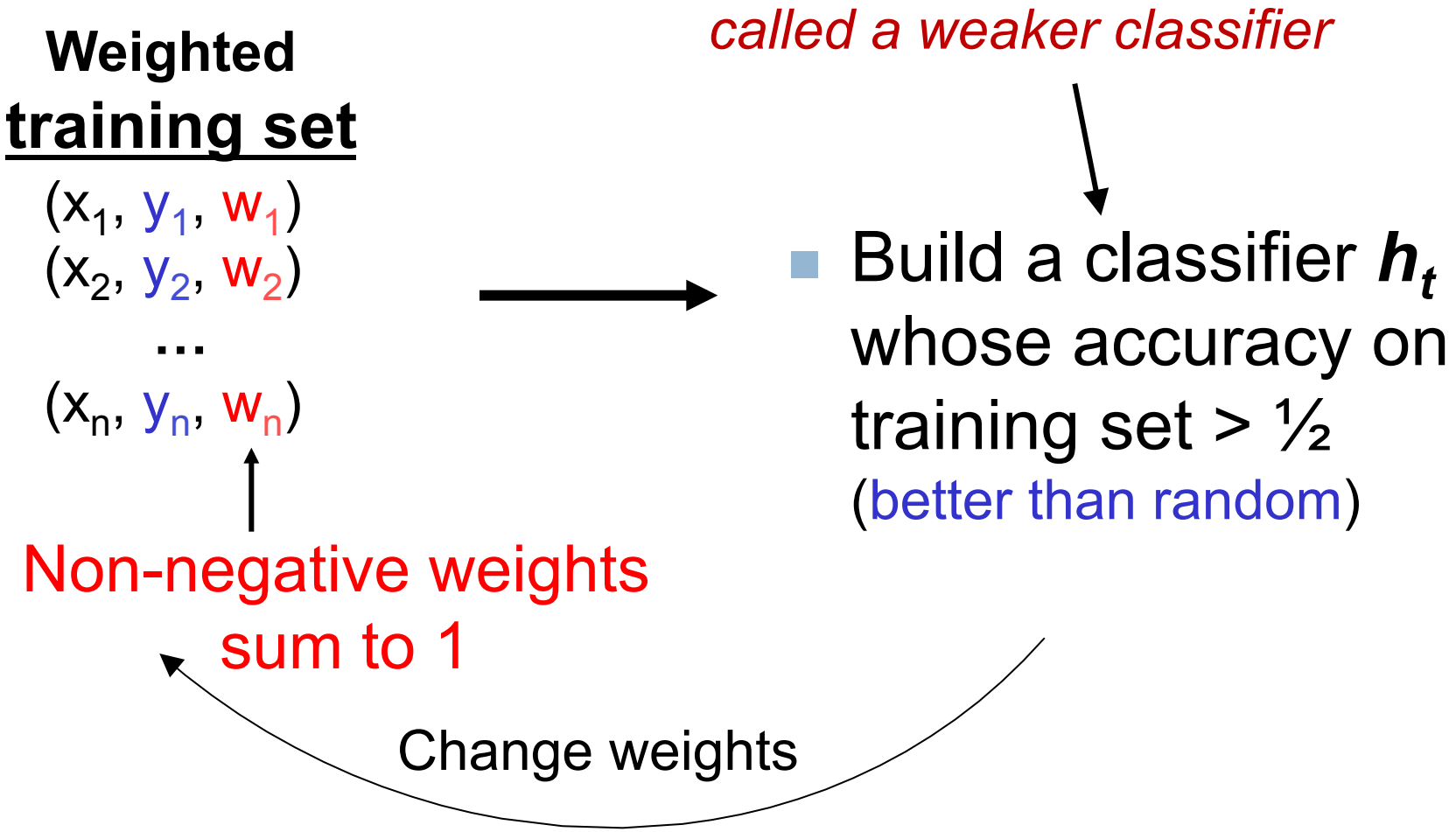
Boosting

- Records that are wrongly classified will have their weights increased
- Records that are classified correctly will have their weights decreased

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

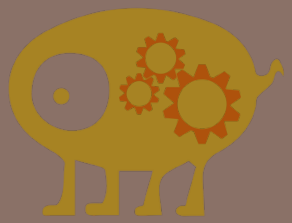
- Example 4 is hard to classify
- Its weight is increased, therefore it is more likely to be chosen again in subsequent rounds

AdaBoost


Weighted
training set
(x_1, y_1, w_1)
(x_2, y_2, w_2)
...
(x_n, y_n, w_n)
called a weaker classifier
Build a classifier h_t whose accuracy on training set > ½ (better than random)
Non-negative weights sum to 1
Change weights

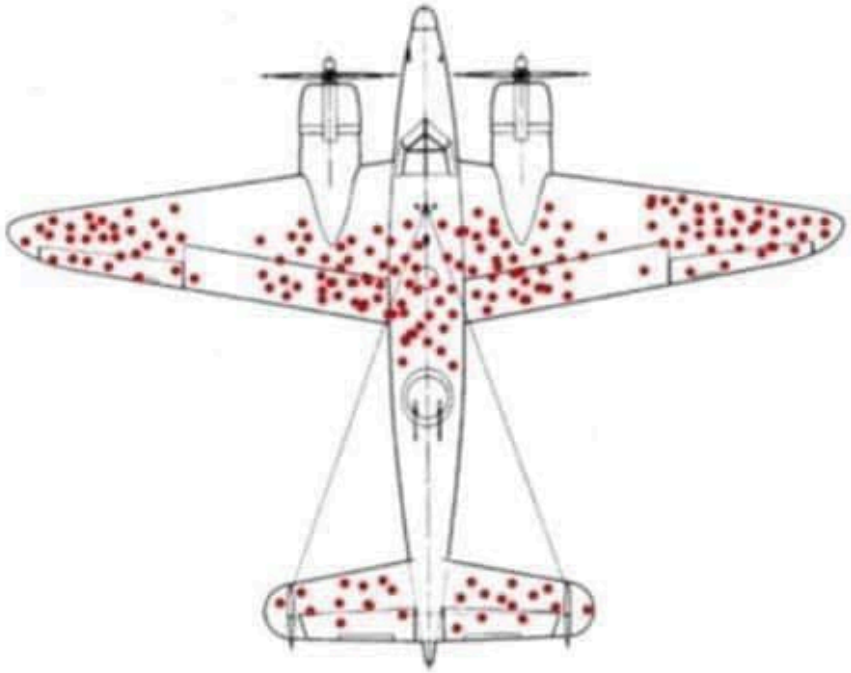
Does AdaBoost always work?

- The actual performance of boosting depends on the data and the base learner.
  - It requires the base learner to be unstable as bagging.
- Boosting seems to be susceptible to noise.
  - When the number of outliners is very large, the emphasis placed on the hard examples can hurt the performance.



SUPERVISED, SEMI-SUPERVISED, UNSUPERVISED

Problem induced labeling



Positive / Negative Data

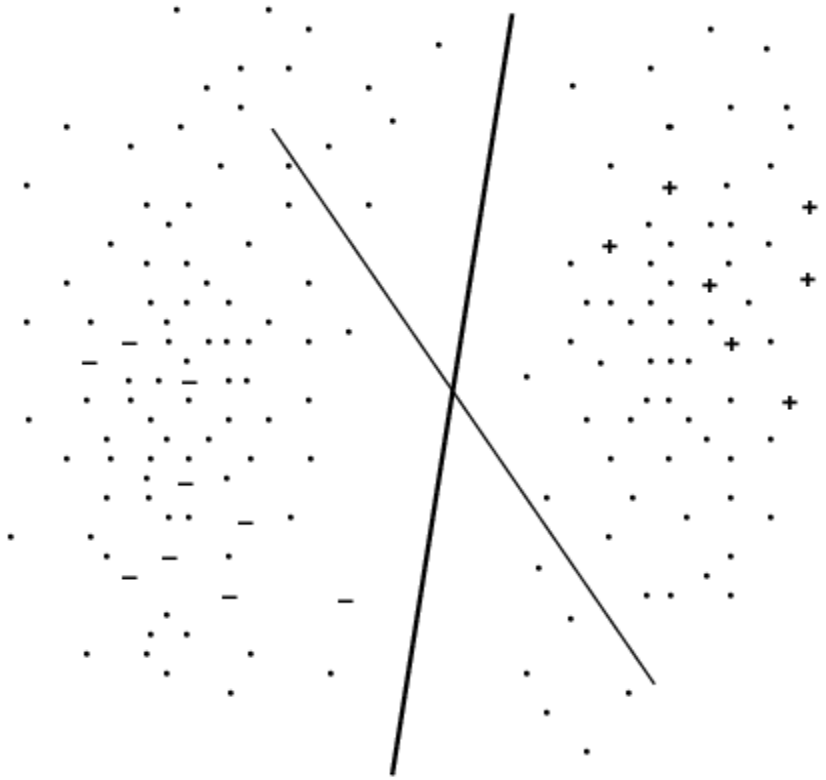
Supervised Learning with Unlabeled Data

- *Supervised v.s. Unsupervised*
- Assigning labels to training set is
  - expensive
  - time consuming
- Abundance of unlabeled data
  - suggests possible use to improve learning
  - Learning with a small set of labeled examples and a large set of unlabeled examples (LU learning)
  - Learning with positive and unlabeled examples (no labeled negative examples) (PU learning).

Why Unlabeled Data helpful?

- Consider positive and negative examples
  - as two separate distribution
  - with very large number of samples available parameters of distribution can be estimated well
  - needs only few labeled points to decide which Gaussian is associated with positive and negative class
- In text domains
  - categories can be guessed using term co-occurrences

Why Unlabeled Data?



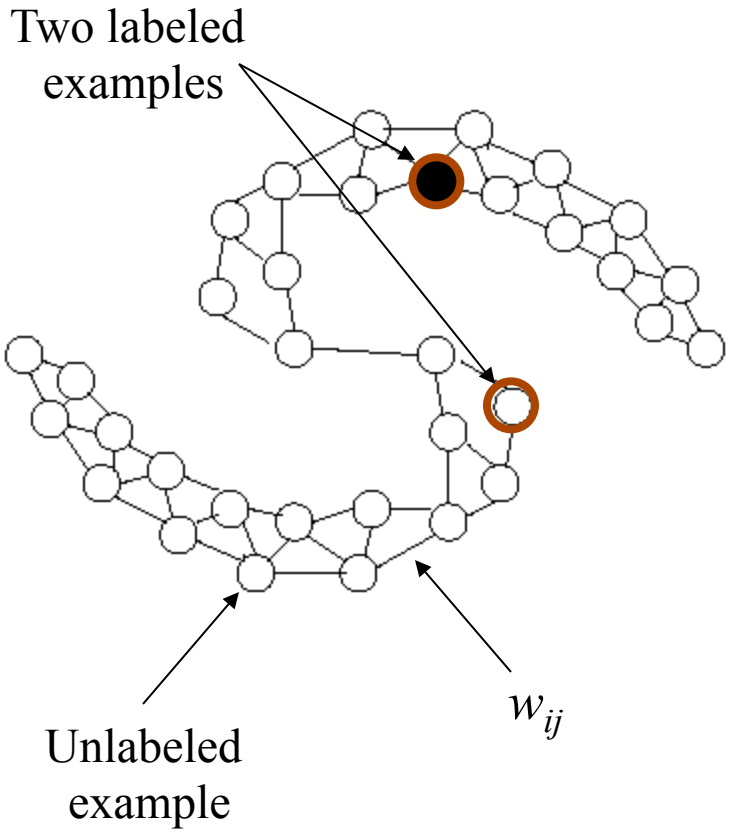
Semi-supervised learning

usage	supervised learning	semi-supervised learning	unsupervised learning
$\{(x, y)\}$ labeled data	yes	yes	no
$\{x\}$ unlabeled data	no	yes	yes

- Label propagation
- Transductive learning
- Co-training
- Active learning
- Transfer learning?

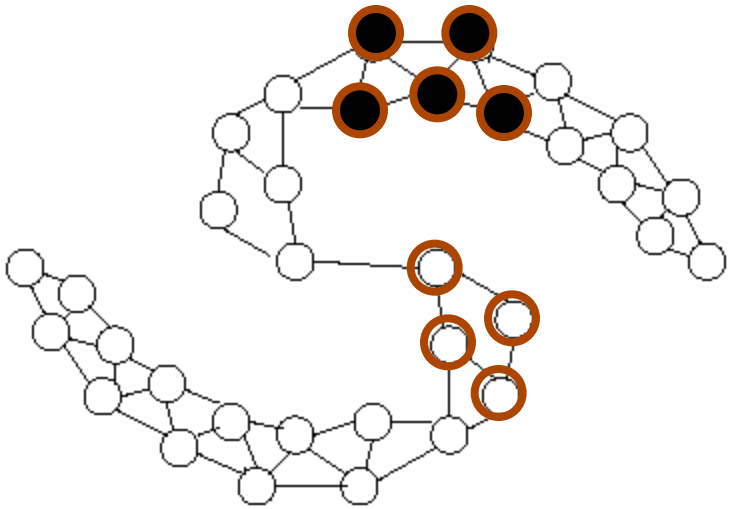
Label Propagation

- A toy problem
  - Each node in the graph is an example
    - Two examples are labeled
    - Most examples are unlabeled
  - Compute the similarity between examples S_{ij}
  - Connect examples to their most similar examples
- How to predicate labels for unlabeled nodes using this graph?



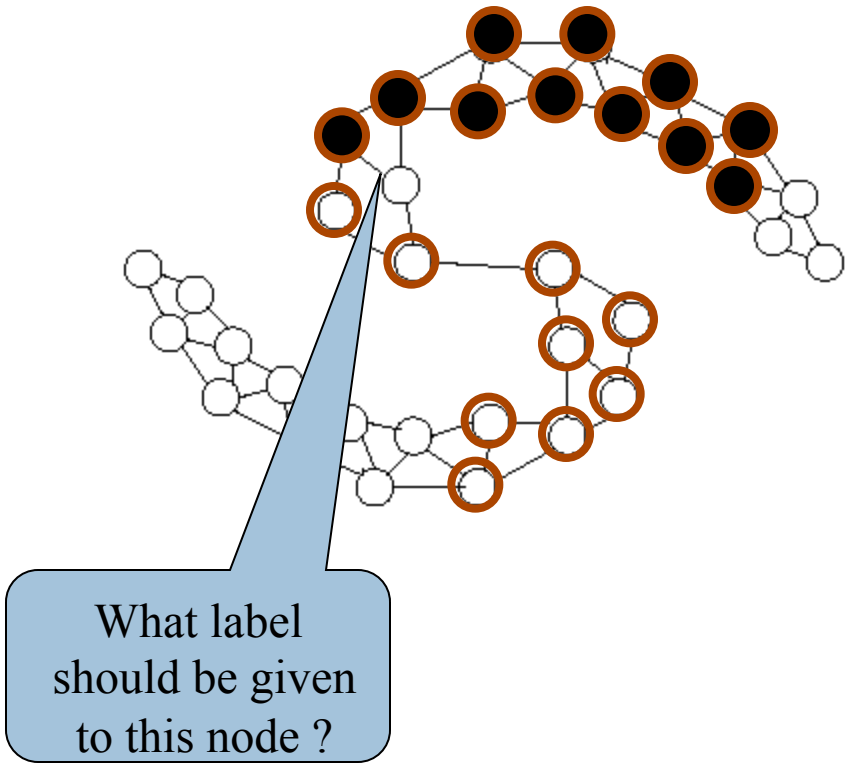
Label Propagation

- Forward propagation



Label Propagation

- Forward propagation
- Forward propagation
- Forward propagation
  - How to resolve conflicting cases



Learning from Positive & Unlabeled data (PU learning)

- Positive examples: One has a set of examples of a class *P*, and
- Unlabeled set: also has a set *U* of unlabeled (or mixed) examples with instances from *P* and also not from *P* (*negative examples*).
- Build a classifier: Build a classifier to classify the examples in *U* and/or future (test) data.
- **Key feature of the problem**: no labeled negative training data.
- We call this problem, PU-learning.

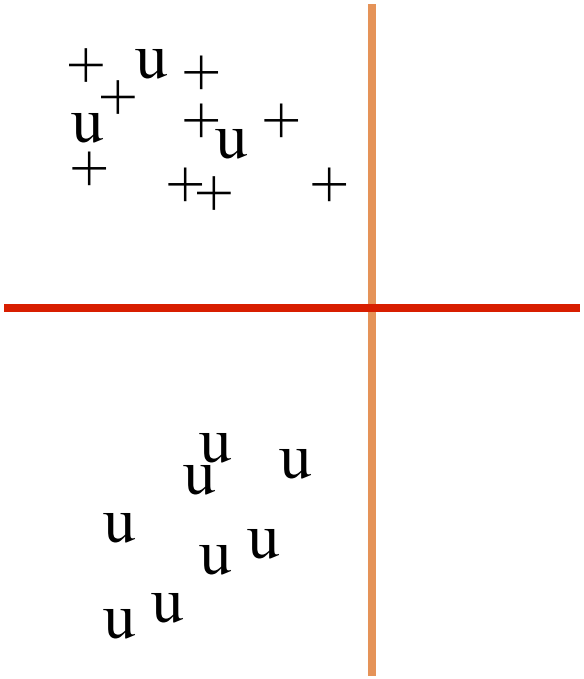
Applications of the problem

- With the growing volume of online texts available through the Web and digital libraries, one often wants to find those documents that are related to one's work or one's interest.
- For example, given a ICML proceedings,
  - find all machine learning papers from AAAI, IJCAI, KDD
  - No labeling of negative examples from each of these collections.
- Similarly, given one's bookmarks (positive documents), identify those documents that are of interest to him/her from Web sources.

Direct Marketing

- Company has database with details of its customer – positive examples, but no information on those who are not their customers, i.e., no negative examples.
- Want to find people who are similar to their customers for marketing
- Buy a database consisting of details of people, some of whom may be potential customers – unlabeled examples.

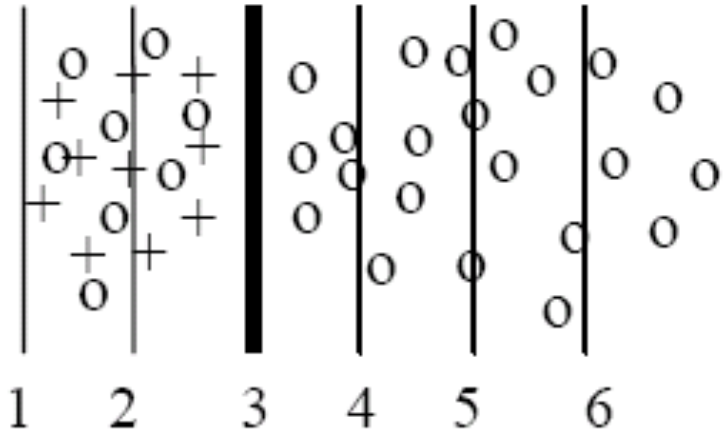
Are Unlabeled Examples Helpful?



"Not learnable" with only positive examples. However, addition of unlabeled examples makes it learnable.

An illustration

- Assume a linear classifier. Line 3 is the best solution.



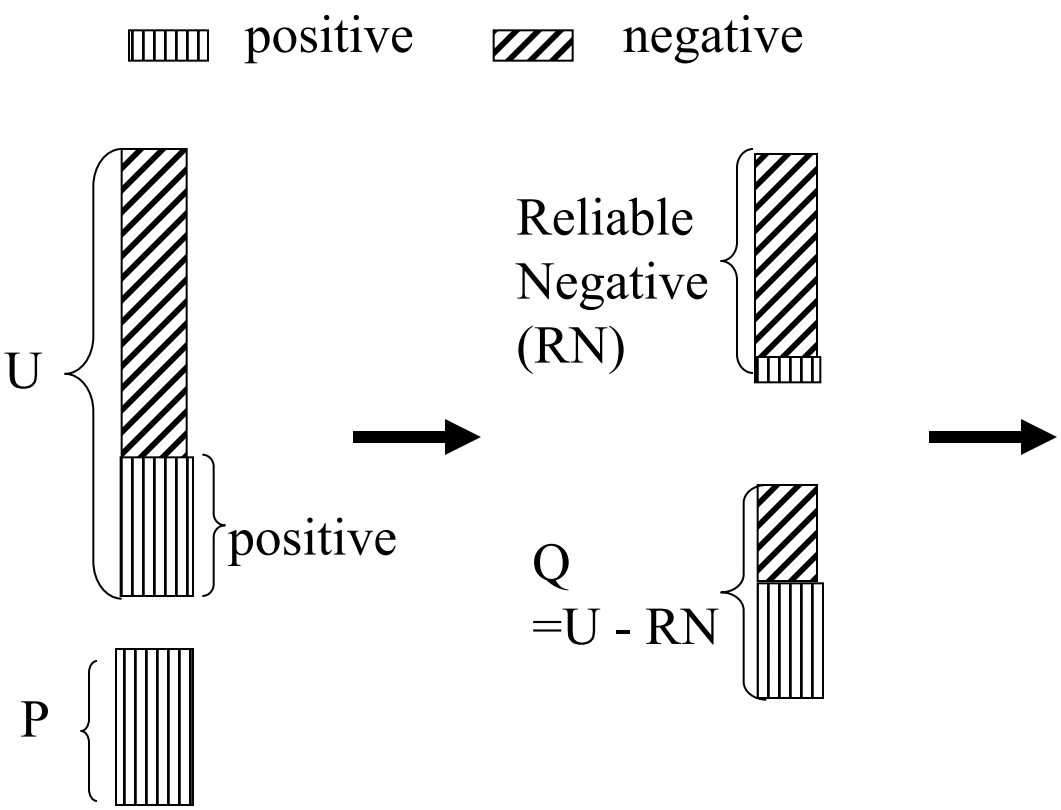
An illustration of the constrained optimization problem

Existing 2-step strategy

- Step 1: Identifying a set of reliable negative documents from the unlabeled set.
  - S-EM [Liu et al, 2002] uses a Spy technique,
  - PEBL [Yu et al, 2002] uses a 1-DNF technique
  - Roc-SVM [Li & Liu, 2003] uses the Rocchio algorithm.
  - …
- Step 2: Building a sequence of classifiers by iteratively applying a classification algorithm and then selecting a good classifier.
  - S-EM uses the Expectation Maximization (EM) algorithm, with an error based classifier selection mechanism
  - PEBL uses SVM, and gives the classifier at convergence, i.e., no classifier selection.
  - Roc-SVM uses SVM with a heuristic method for selecting the final classifier.

Step 1

Step 2



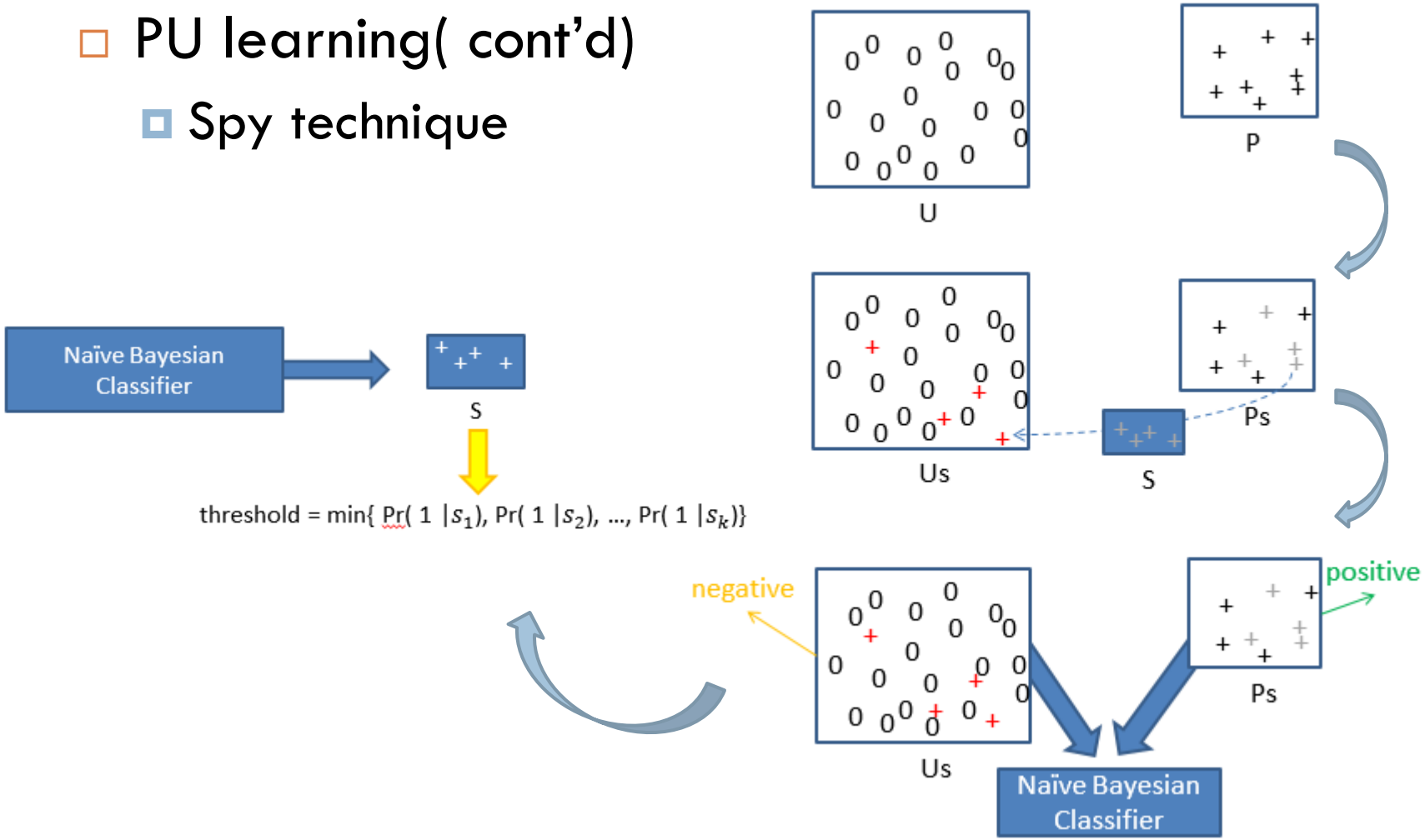
Using P, RN and Q to build the final classifier iteratively

or

Using only P and RN to build a classifier

Spy

- PU learning(cont'd)
  - Spy technique



Step1: The Spy technique

- Sample a certain % of positive examples and put them into unlabeled set to act as "spies".
- Run a classification algorithm *assuming all unlabeled examples are negative*,
  - we will know the behavior of those actual positive examples in the unlabeled set through the "spies".
- We can then extract reliable negative examples from the unlabeled set more accurately.

Step 1: Other methods

- 1-DNF method:
  - Find the set of words W that occur in the positive documents more frequently than in the unlabeled set.
  - Extract those documents from unlabeled set that do not contain any word in W. These documents form the reliable negative documents.

- Rocchio method from information retrieval.
- Naïve Bayesian method.

Do they follow the theory?

- Yes, heuristic methods because
  - Step 1 tries to find some initial reliable negative examples from the unlabeled set.
  - Step 2 tried to identify more and more negative examples iteratively.
- The two steps together form an iterative strategy of increasing the number of unlabeled examples that are classified as negative while maintaining the positive examples correctly classified.

Co-training Algorithm

[Blum and Mitchell, 1998]

Given: labeled data L,

unlabeled data U

Loop:

Train h1 (e.g., hyperlink classifier) using L

Train h2 (e.g., page classifier) using L

Allow h1 to label *p* positive, *n* negative examples from U

Allow h2 to label *p* positive, *n* negative examples from U

Add these most confident self-labeled examples to L

Co-training Algorithm

- Labeled data are used to infer two Naïve Bayes classifiers, one for each view
- Each classifier will
  - examine unlabeled data
  - pick the most confidently predicted positive and negative examples
  - add these to the labeled examples
- Classifiers are now retrained on the augmented set of labeled examples
- *As well as the text on the pages being relevant to the classification of the page, the hyperlinks also contain useful information*

Co-training: Experimental Results

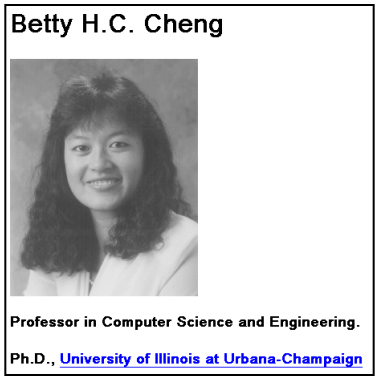
- begin with 12 labeled web pages (academic course)
- provide 1,000 additional unlabeled web pages
- average error: learning from labeled data 11.1%
- average error: co-training 5.0%

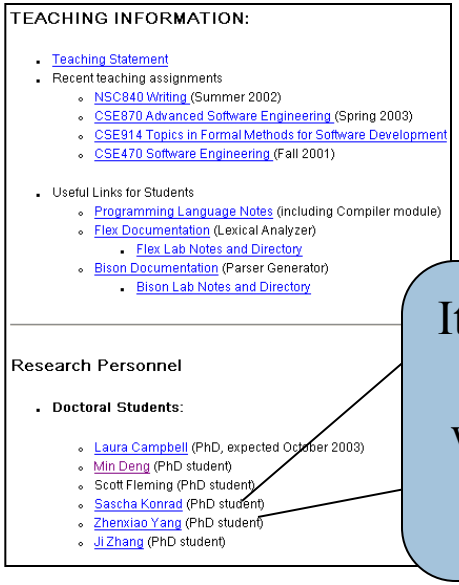
	Page-base classifier	Link-based classifier	Combined classifier
Supervised training	12.9	12.4	11.1
Co-training	6.2	11.6	5.0

An example: Co-training for Semi-supervised Learning

- Consider the task of classifying web pages into two categories: category for students and category for professors
- Two aspects of web pages should be considered
  - Content of web pages
    - "I am currently the second year Ph.D. student …"
  - Hyperlinks
    - "My advisor is …"
    - "Students: …"

Co-training for Semi-Supervised Learning





It is easier to classify this web page using hyperlinks

It is easy to classify the type of this web page based on its content

Co-training

- Two representation for each web page

Content representation:

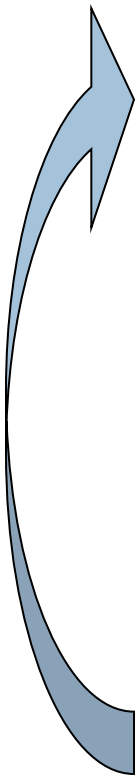
(doctoral, student, computer, university…)

Hyperlink representation:

Inlinks: Prof. Cheng

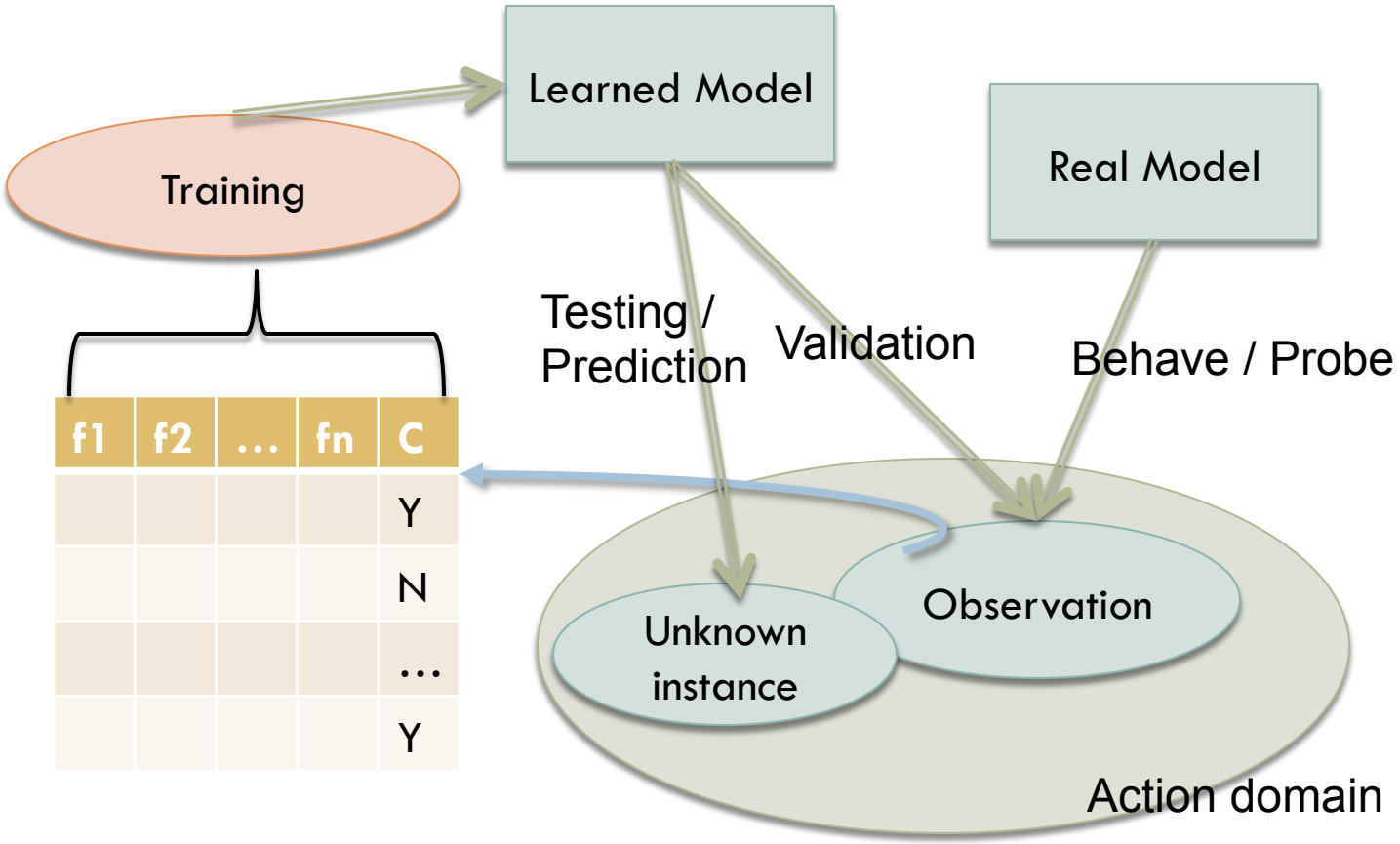
Oulinks: Prof. Cheng

Co-training: Classification Scheme



1. Train a content-based classifier using labeled web pages
2. Apply the content-based classifier to classify unlabeled web pages
3. Label the web pages that have been confidently classified
4. Train a hyperlink based classifier using the web pages that are initially labeled and labeled by the classifier
5. Apply the hyperlink-based classifier to classify the unlabeled web pages
6. Label the web pages that have been confidently classified

Learning flow



Top-10 Machine Learning Algorithms for Data Scientist