Classical Machine Learning: Classification and Regression (II)

- Learn the concept, theory, toy example, and scikit-learn usage of a few interesting base classifiers.
- Learn the concept, theory, toy example, and scikit-learn usage of ensemble classifiers (rationale, parallel ensembles: bagging, random forest, and extra trees).

Base Classifier: Decision Tree

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

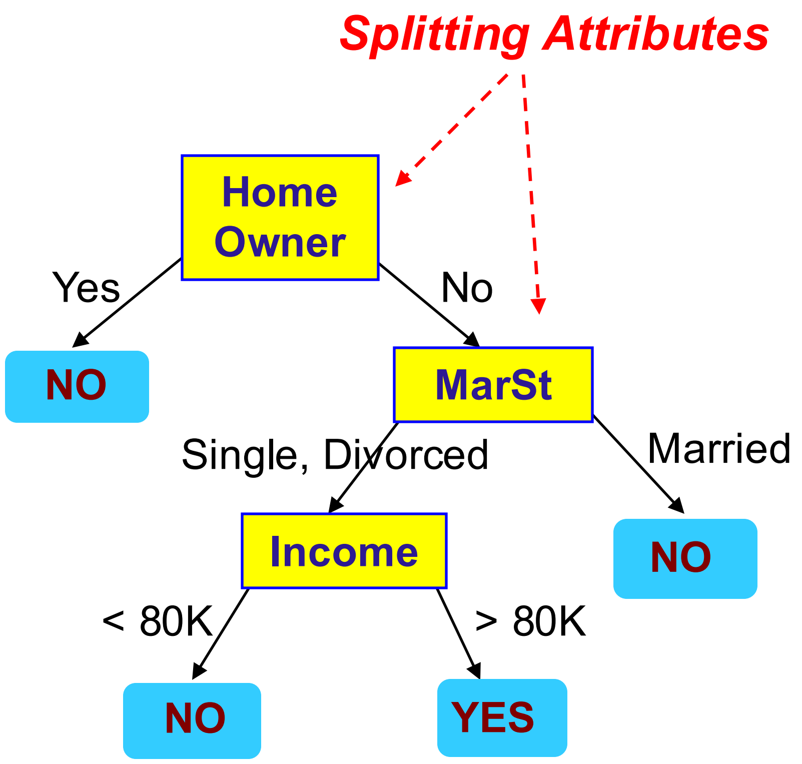
Example of a Decision Tree

Review

categorical categorical continuous class

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

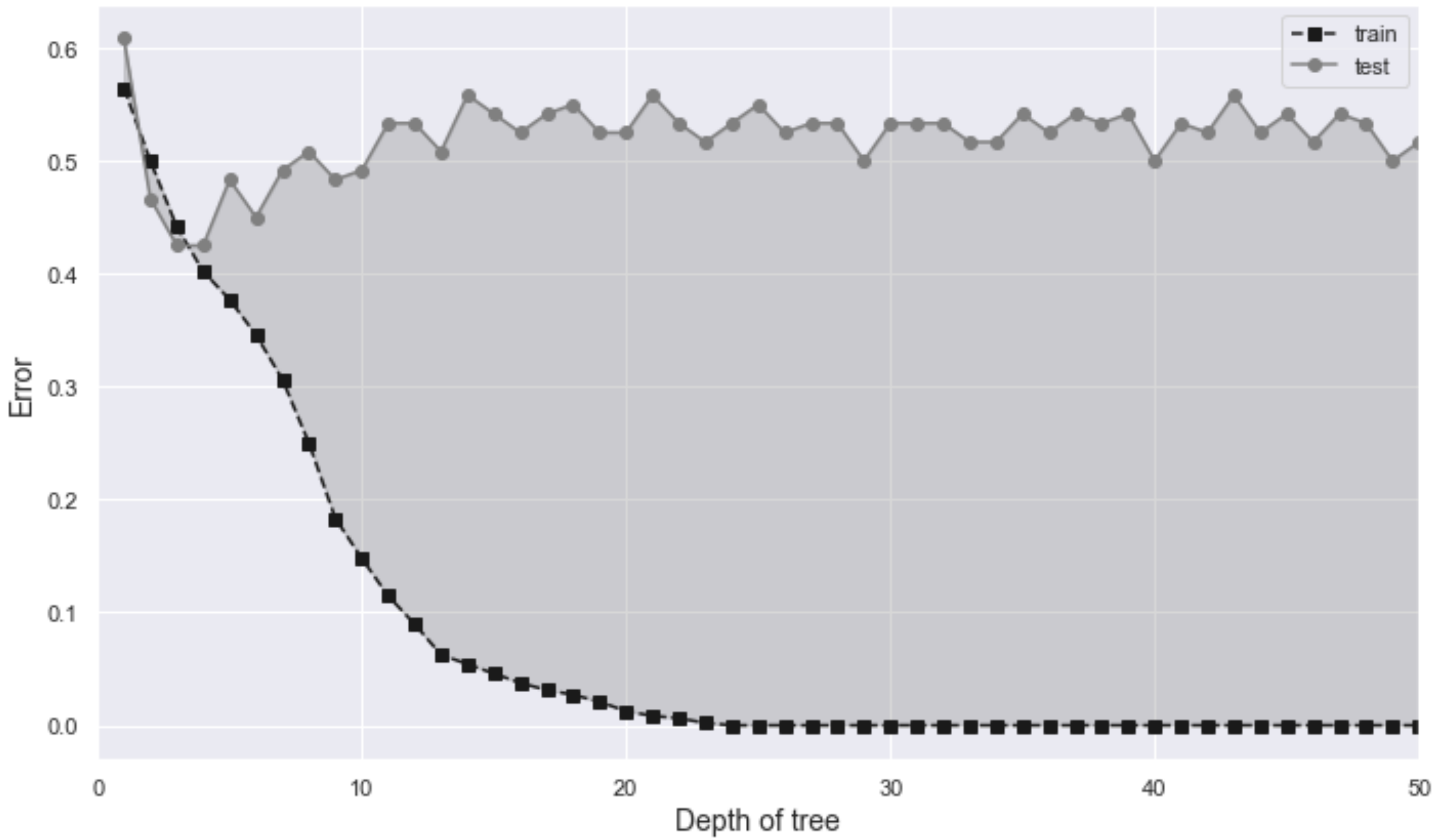
Training Data



Model: Decision Tree

Decision Tree: theoretical minimum and example

- The phrase "theoretical minimum" is taken from a very successful book series written by Leonard Susskind, a great physicist at Stanford University.
- "Theoretical minimum" means just the minimum theories and equations you need to know in order to proceed to the next level.
- See Decision_Tree.pdf



Fun time: what have you observed as the depth of the tree increases (多選)? (1) training accuracy increases (2) training accuracy decreases (3) test accuracy increases (4) test accuracy decreases.

Summary

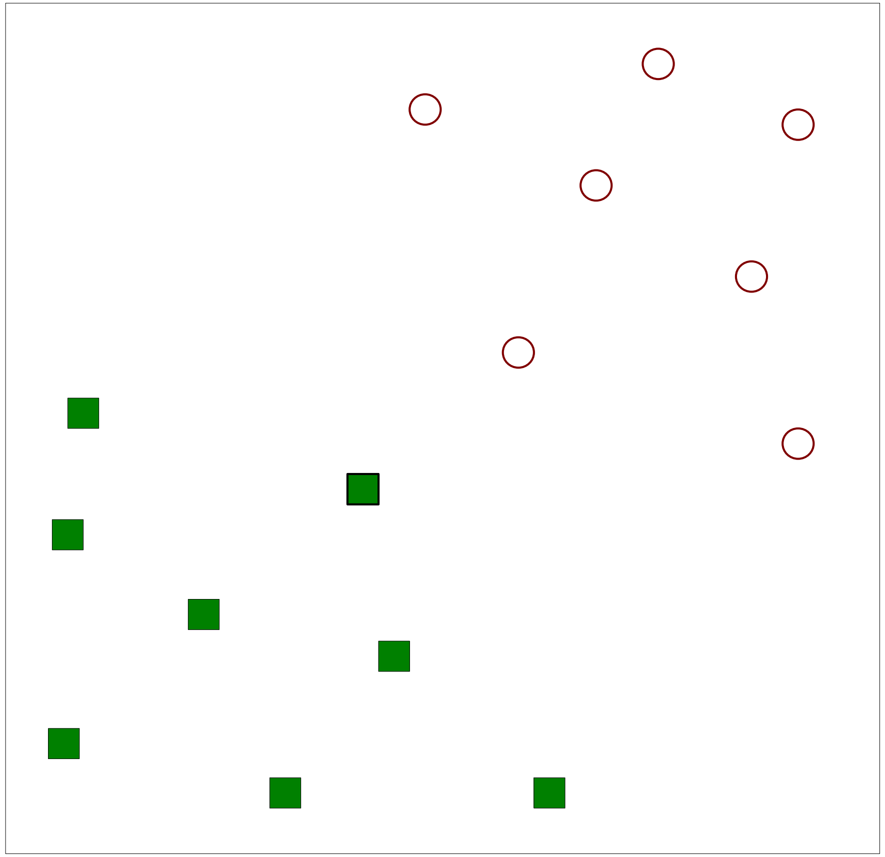
Classification Algorithm Walkthrough: Decision Tree

- Decision tree is simple and useful for interpretation.
- Decision tree uses a greedy algorithm with a best-split attribute to recursively split the tree.
- The “Gini” criteria, or the “Entropy” criteria is the most commonly used index to determine the best split.
- Shallow decision trees are weak learners and are not competitive in terms of prediction accuracy
- Deep decision trees tend to overfit data.
- An ensemble of randomized decision trees such as random forests is a powerful algorithm for classification. This will be covered in the sequel.

Base Classifier: Support Vector Machine (SVM)

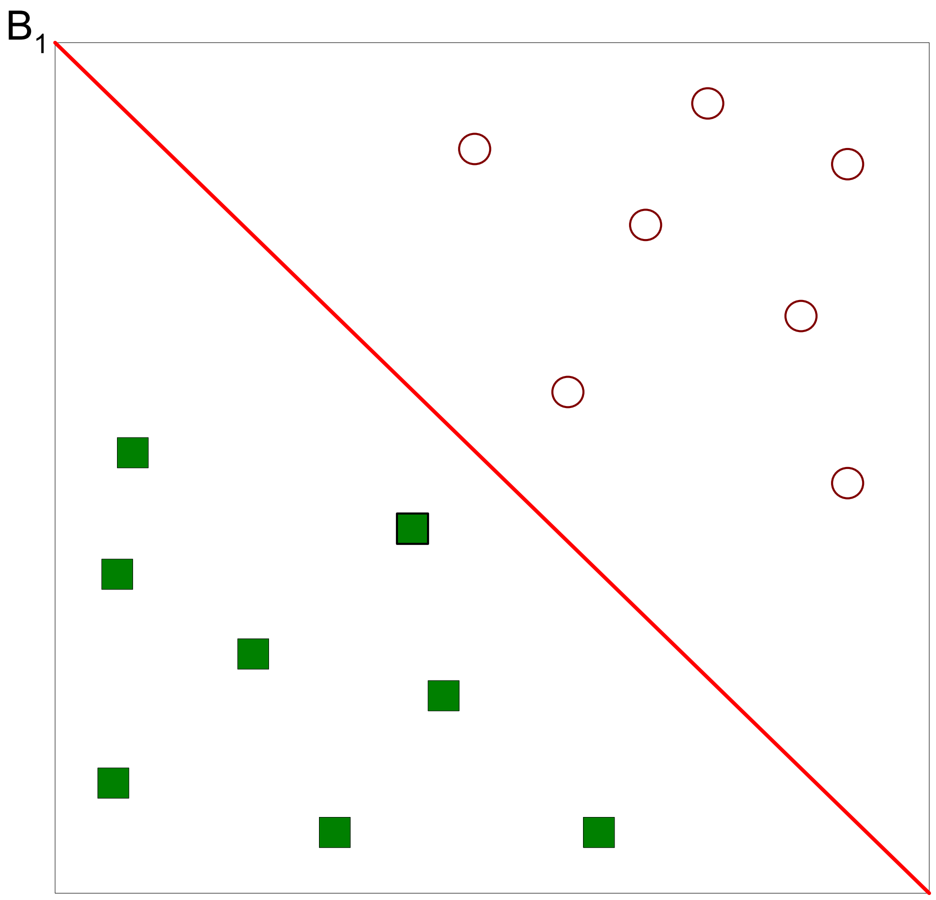
C-S David Chen, Department of Civil Engineering, National Taiwan University

Support Vector Machines



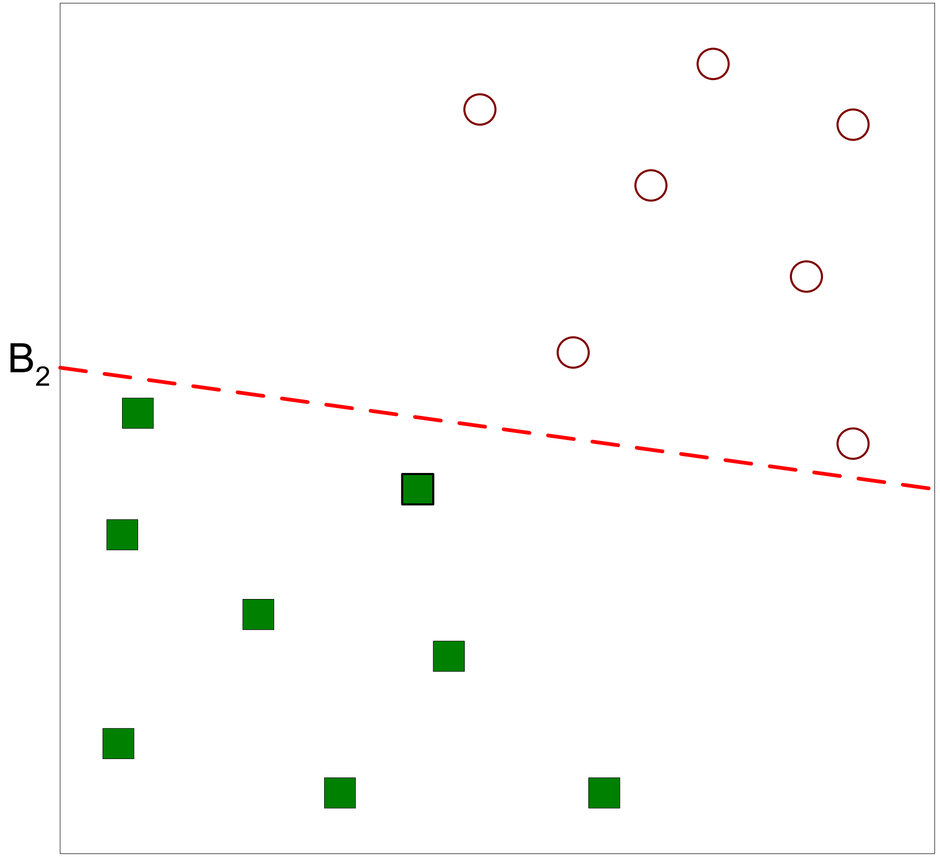
- Find a linear hyperplane (decision boundary) that will separate the data

Support Vector Machines



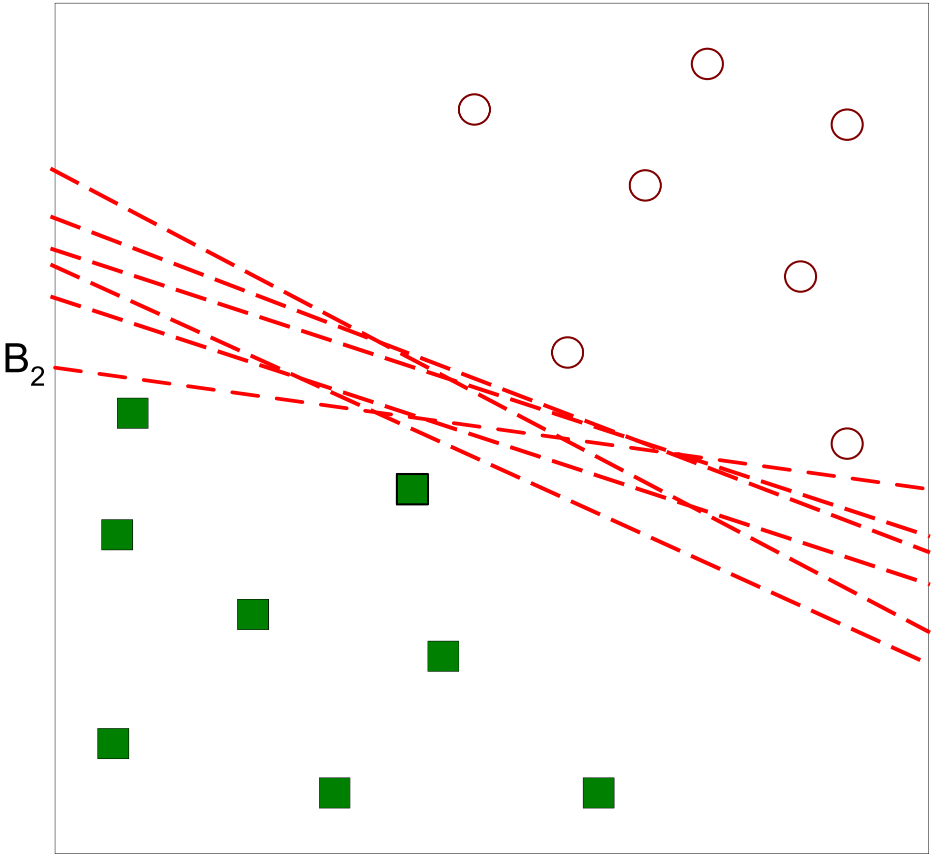
- One Possible Solution

Support Vector Machines



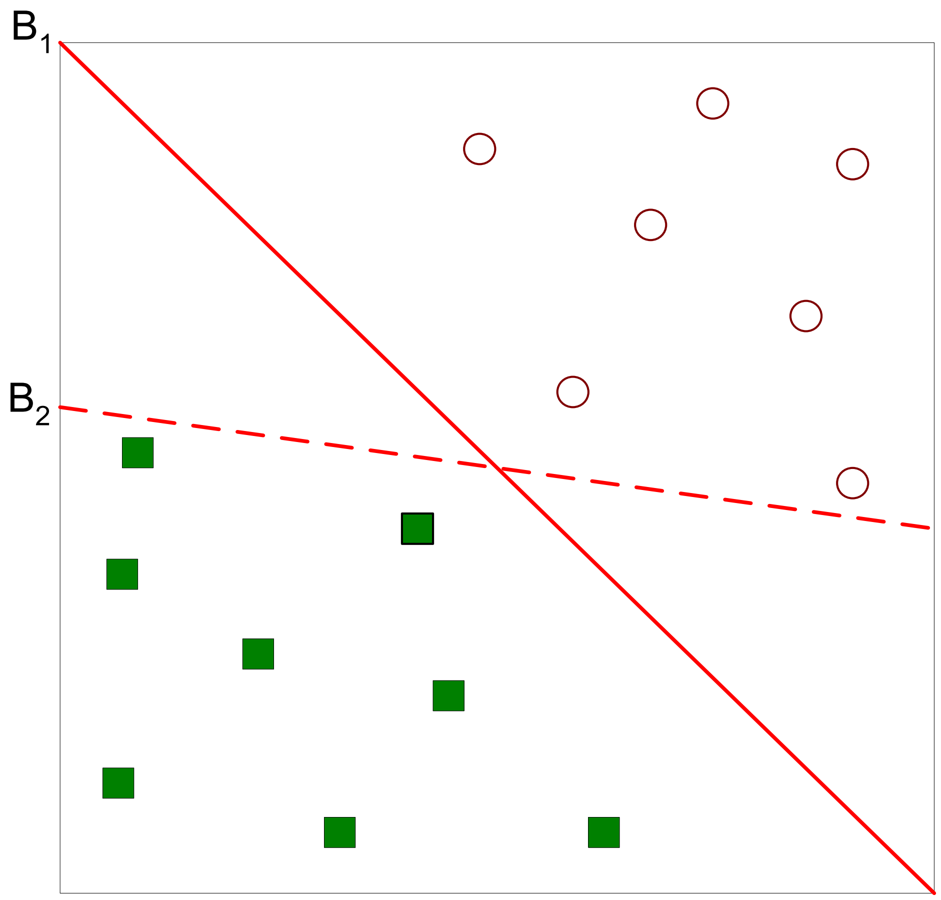
- Another possible solution

Support Vector Machines



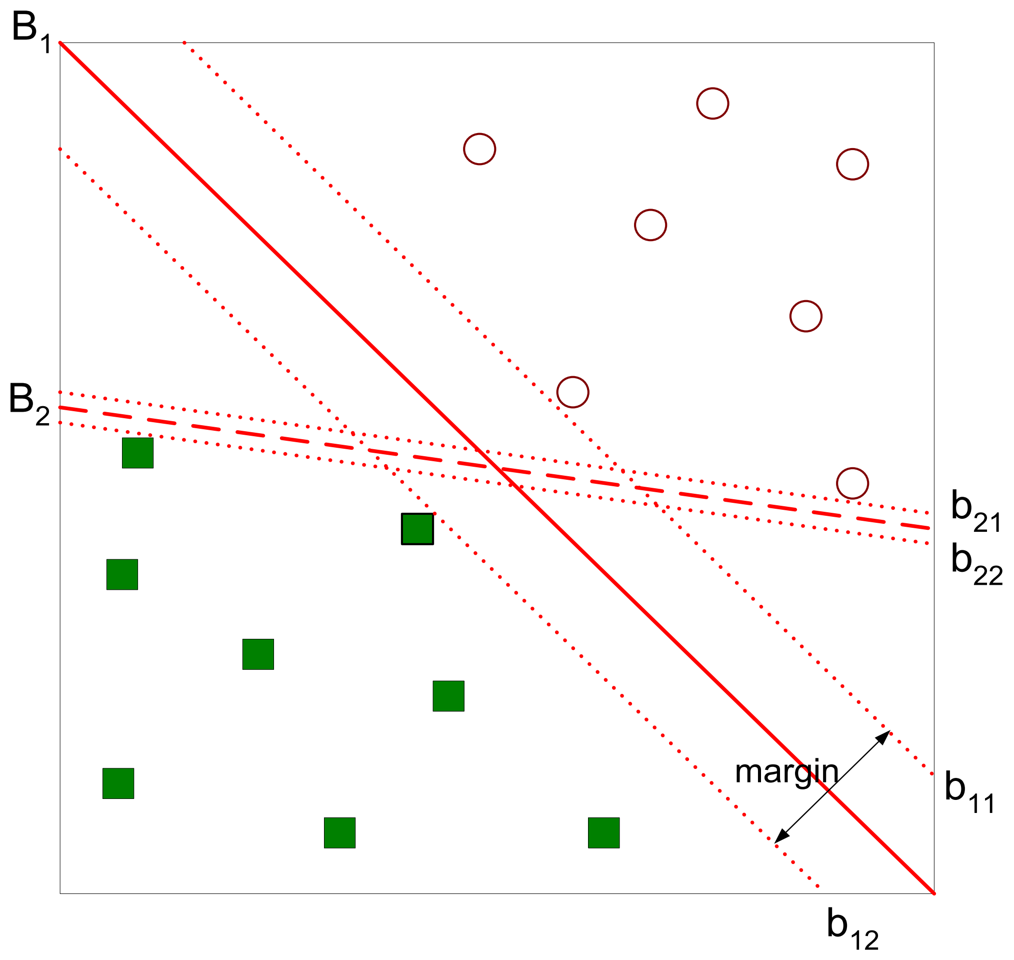
- Other possible solutions

Support Vector Machines



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

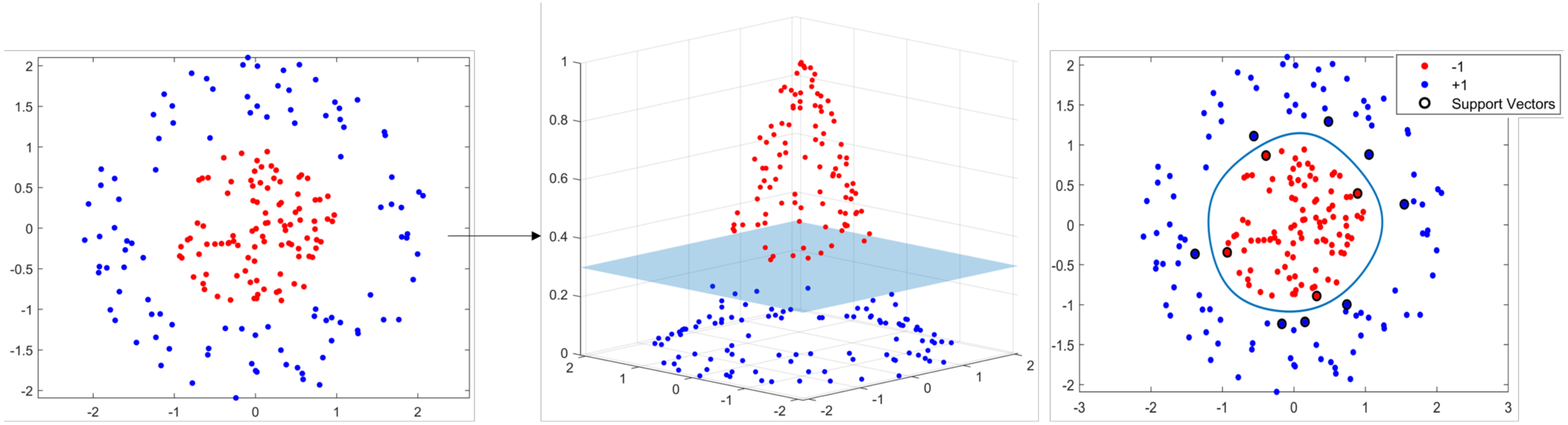
Support Vector Machines



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

Support Vector Machine (SVM): Summary

- It is one of the **classical supervised machine learning algorithm** that excels in pattern recognition and data classifications.
- It is a **mathematical entity** that selects the maximum-margined N-dimensional separating hyperplane → maximize its ability to characterize unseen samples
- **Hyperplane selection**: utilize various kernel functions to transfer low-dimensioned, non-linear, and possibly non-separable training data to higher-dimensional feature spaces → linearly separable



(a) 2D non-linear training data

(b) 3D mapped data points using the Gaussian kernel and separating hyperplane

(c) non-linear SVM with the Gaussian kernel training result

Classification Algorithm Walkthrough: Other Base Classifiers

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

Classification algorithm shortlist

Base_classifiers.ipynb

- Linear Machine Learning Algorithms
  - Logistic Regression
  - Linear Discriminant Analysis
- Nonlinear Machine Learning Algorithms
  - k-Nearest Neighbors
  - Naïve Bayes
  - Classification and Regression Trees (CART or just decision trees)
  - Support Vector Machine

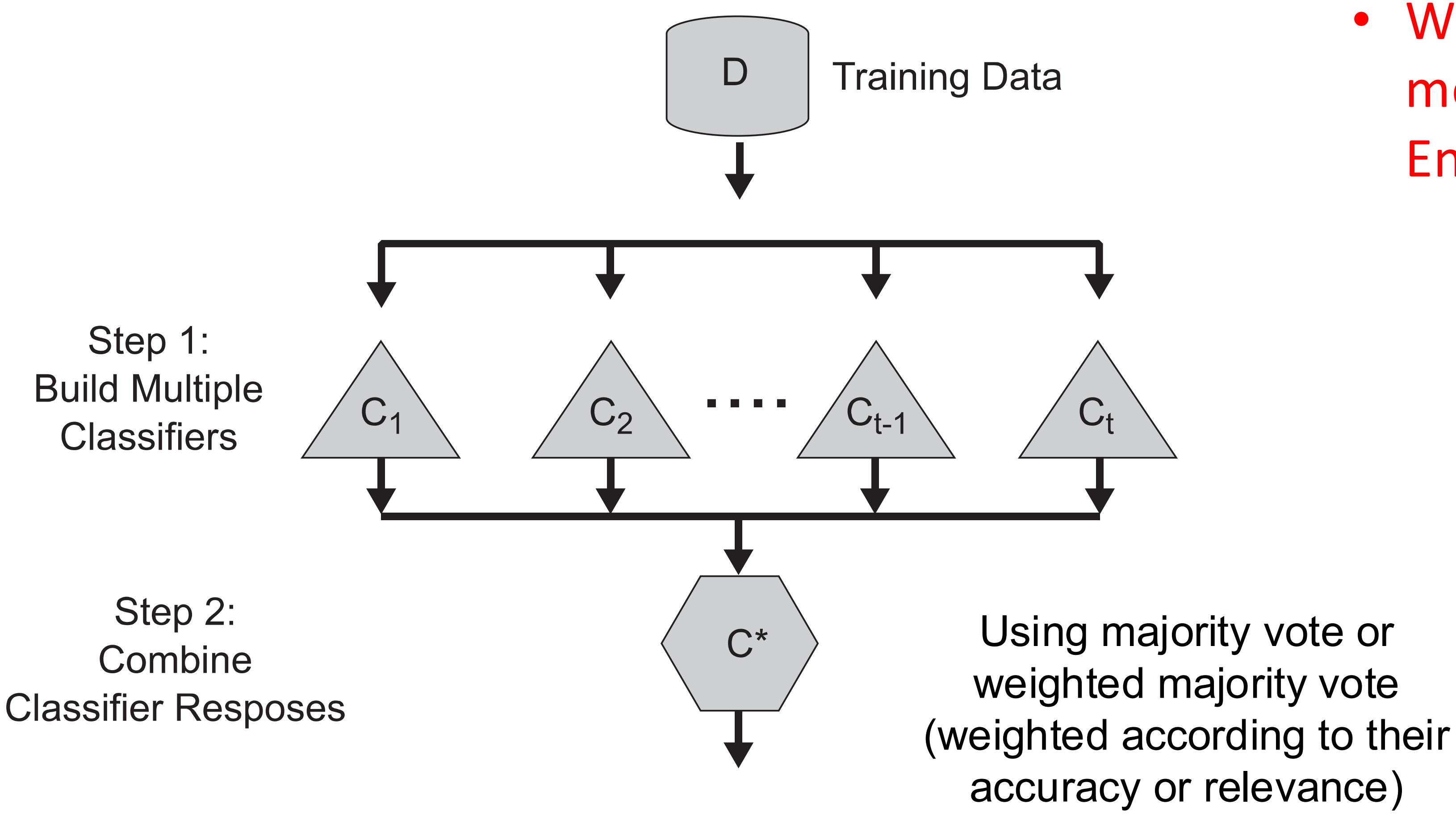
Classification Algorithm Walkthrough: Ensemble Classifiers

Ensemble Methods

- Construct a set of base classifiers learned from the training data

- Predict class label of test records by combining the predictions made by multiple classifiers (e.g., by taking majority vote)

General Approach of Ensemble Learning



- Why do ensemble methods work? See Ensemble_Rationale.pdf

Fun Time: Which statement is true?

- The ensemble classifier outperforms the base classifier of any error rate
- The ensemble classifier outperforms the base classifier when $e > 0.5$
- The ensemble classifier outperforms the base classifier when $e < 0.5$.

Base Classifiers for Ensemble Learning

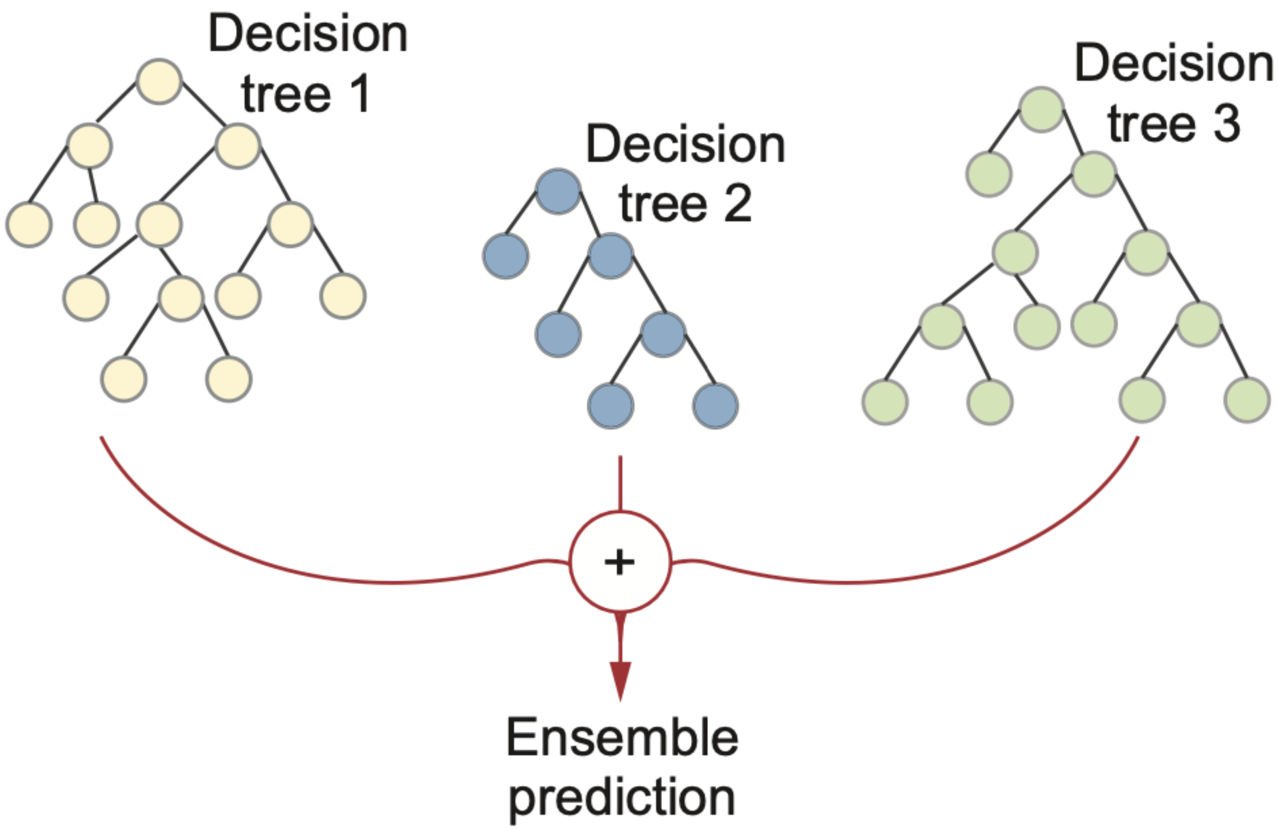
Ensemble Methods work best with **unstable base classifiers**

- Classifiers that are sensitive to minor perturbations in training set, due to *high model complexity*
- Ensemble methods try to reduce the variance of complex models (with low bias) by *aggregating* responses of multiple base classifiers
- Examples: decision trees, ANNs, …

Classification Algorithm Walkthrough: Parallel Ensemble Classifiers – Bagging, Random Forest and Extra Trees

Parallel Ensembles

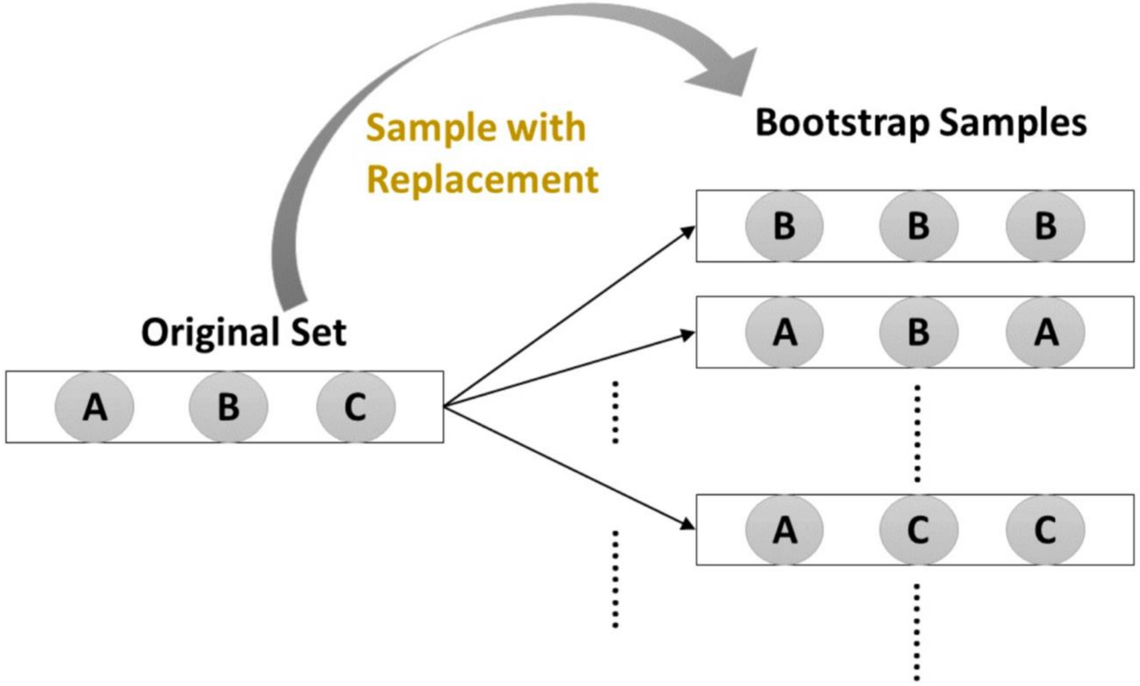
- Trained using the same base machine-learning algorithm.
- Ensemble diversity is created from a single algorithm with random data or feature sampling to train each base model.
- **Ensembles in this family**: bagging, random forest, extra trees etc.



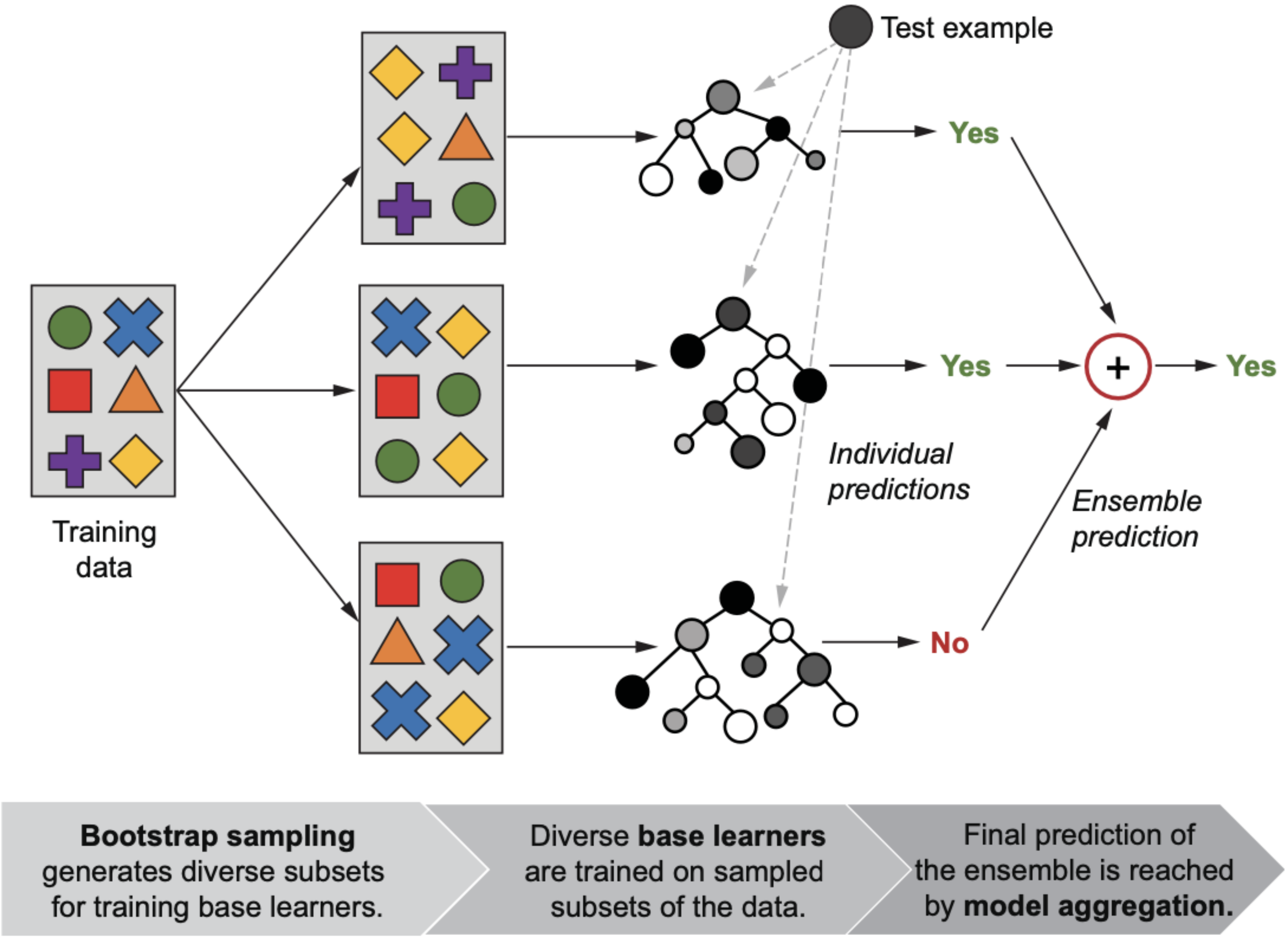
Bagging (Bootstrap AGGregatING)

- Bootstrap sampling: sampling with replacement

inference about a population from sample data (sample → population) can be modelled by resampling the sample data and performing inference about a sample from resampled data (resampled → sample).



Bagging Illustration



G. Kunapuli (2023) Ensemble Methods for Machine Learning, Manning.

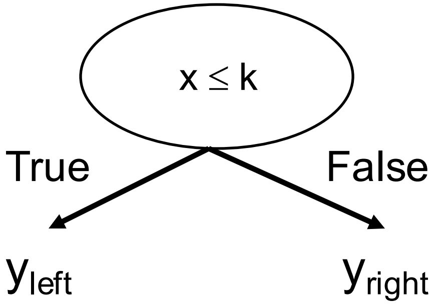
Bagging Example

- Consider 1-dimensional data set:

Original Data:

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	1	1	-1	-1	-1	-1	1	1	1

- Classifier is a decision stump (decision tree of size 1)
  - Decision rule: $x \le k$ versus $x > k$
  - Split point k is chosen based on entropy



Fun Time: what is the best accuracy a stump can reach for this simple 1D example? (1) 50% (2) 60% (3) 70% (4) 80%

Bagging Example

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

x <= 0.35 → y = 1
x > 0.35 → y = -1

Bagging Example

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

x <= 0.35 ➔ y = 1
x > 0.35 ➔ y = -1

Bagging Round 2:

x	0.1	0.2	0.3	0.4	0.5	0.5	0.9	1	1	1
y	1	1	1	-1	-1	-1	1	1	1	1

x <= 0.7 ➔ y = 1
x > 0.7 ➔ y = 1

Bagging Round 3:

x	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.35 ➔ y = 1
x > 0.35 ➔ y = -1

Bagging Round 4:

x	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.3 ➔ y = 1
x > 0.3 ➔ y = -1

Bagging Round 5:

x	0.1	0.1	0.2	0.5	0.6	0.6	0.6	1	1	1
y	1	1	1	-1	-1	-1	-1	1	1	1

x <= 0.35 ➔ y = 1
x > 0.35 ➔ y = -1

Bagging Example

Bagging Round 6:

x	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 7:

x	0.1	0.4	0.4	0.6	0.7	0.8	0.9	0.9	0.9	1
y	1	-1	-1	-1	-1	1	1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 8:

x	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.8	0.9	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 9:

x	0.1	0.3	0.4	0.4	0.6	0.7	0.7	0.8	1	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ➔ y = -1
x > 0.75 ➔ y = 1

Bagging Round 10:

x	0.1	0.1	0.1	0.1	0.3	0.3	0.8	0.8	0.9	0.9
y	1	1	1	1	1	1	1	1	1	1

x <= 0.05 ➔ y = 1
x > 0.05 ➔ y = 1

Bagging Example

- Summary of Trained Decision Stumps:

Round	Split Point	Left Class	Right Class
1	0.35	1	-1
2	0.7	1	1
3	0.35	1	-1
4	0.3	1	-1
5	0.35	1	-1
6	0.75	-1	1
7	0.75	-1	1
8	0.75	-1	1
9	0.75	-1	1
10	0.05	1	1

Bagging Example

- Use majority vote (sign of sum of predictions) to determine class of ensemble classifier

Round	x=0.1	x=0.2	x=0.3	x=0.4	x=0.5	x=0.6	x=0.7	x=0.8	x=0.9	x=1.0
1	1	1	1	-1	-1	-1	-1	-1	-1	-1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	-1	-1	-1	-1	-1	-1	-1
4	1	1	1	-1	-1	-1	-1	-1	-1	-1
5	1	1	1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	1	1	1
8	-1	-1	-1	-1	-1	-1	-1	1	1	1
9	-1	-1	-1	-1	-1	-1	-1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
Sum	2	2	2	-6	-6	-6	-6	2	2	2
Sign	1	1	1	-1	-1	-1	-1	1	1	1

Predicted Class

Original Data:

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	1	1	1	-1	-1	-1	-1	1	1	1

Bagging: theoretical minimum and python example

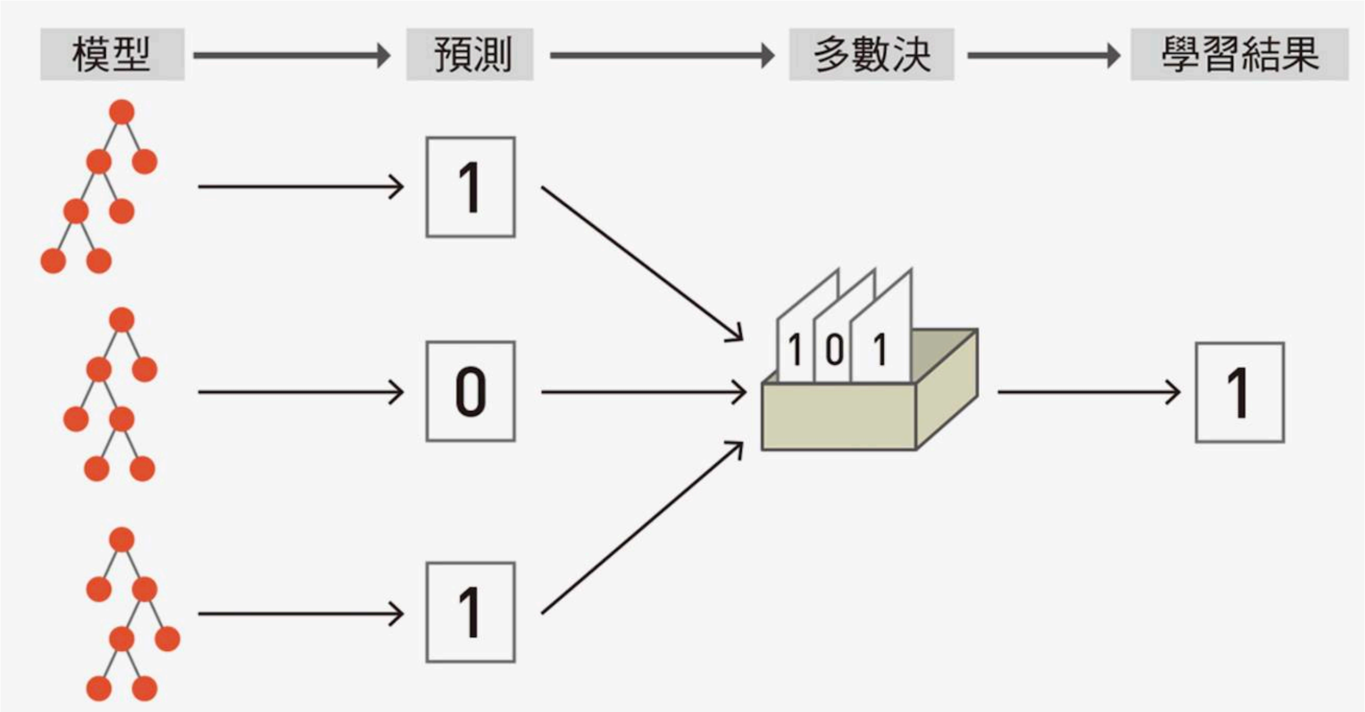
- The phrase “theoretical minimum” is taken from a successful book series by Leonard Susskind, a great physicist at Stanford University.

- “Theoretical minimum” means <u>just the minimum theories and equations you need to know to proceed to the next level.</u>

- See Ensemble_Bagging.pdf

Random Forest Algorithm

- Construct an ensemble of decision trees by manipulating training set as well as features
  - Use bootstrap sample to train every decision tree (similar to Bagging)
  - Use the following tree induction algorithm:
    - At every internal node of the decision tree, randomly sample p attributes ($p < d$) for selecting split criterion



Random Forest: theoretical minimum and python example

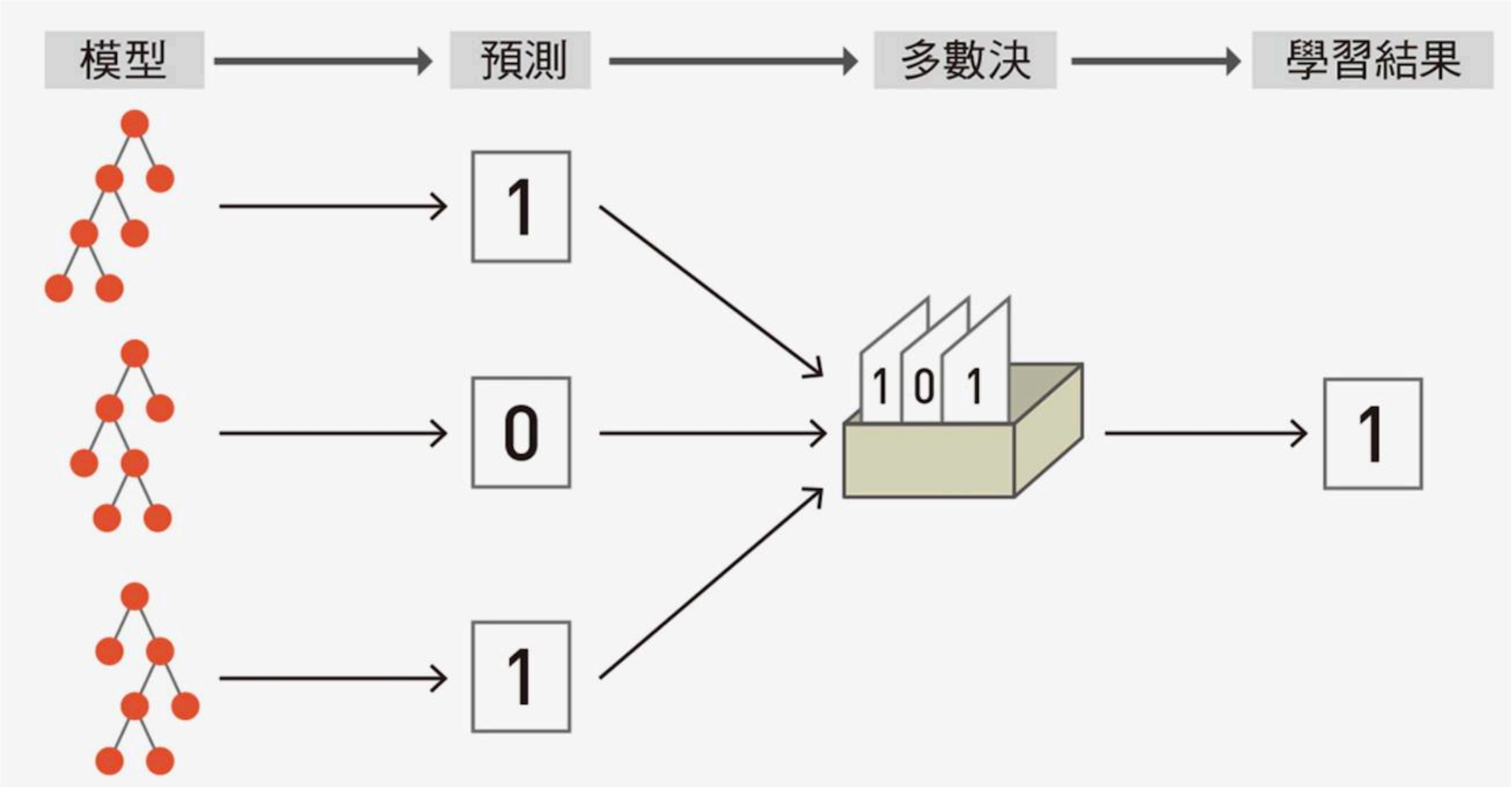
- The phrase “theoretical minimum” is taken from a successful book series by Leonard Susskind, a great physicist at Stanford University.

- “Theoretical minimum” means <u>just the minimum theories and equations you need to know to proceed to the next level.</u>

- See Ensemble_RF_ET.pdf

Feature Importance: Extra Bonus of Random Forest

- Random forest measures a feature's importance by looking at how much the tree nodes that use that feature to reduce impurity on average (across all trees in the forest).
- The feature that can reduce more impurity, the more important.



Feature Importance: Extra Bonus of Random Forest

Automation in Construction 118 (2020) 103274

ELSEVIER

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Machine learning-based seismic capability evaluation for school buildings

Nai-Wen Chi[a], Jyun-Ping Wang[b], Jia-Hsing Liao[c], Wei-Choung Cheng[d], Chuin-Shan Chen[b,*]

Fun Time: what is the most important feature of seismic capability for old school buildings in Taiwan?

1. Total floor area of the building
2. Spectral acceleration demand
3. Tensile strength of steel
4. Amount of walls in Y direction
5. The built year

Summary: Ensemble Rationale, Bagging, Random Forest and Extra Trees

- For the ensemble classifiers to outperform the base classifiers, two conditions must be met:
  - The base classifier should do better than random guessing. (This is easy in general)
  - The base classifiers should be independent of each other. (This is hard!)
- Three well-known **parallel** ensemble methods are Bagging, Random Forest, and Extra Trees.
- **Bagging** creates different subsets of data (this is called bootstrapping), trains one model per subset, and aggregates all predictions to get the final prediction.

Summary: Ensemble Rationale, Bagging, Random Forest and Extra Trees

- **Random Forest** is similar to Bagging. Random Forest differs from Bagging by further randomly choosing candidate features to decide a node's split criteria.
- One benefit of using Random Forest is that it provides a natural mechanism for scoring features based on their importance.
- **Extra Trees** is similar to Random Forest, which randomly chooses candidate features. Extra Trees differ from Random Forest by further randomly deciding the split threshold.