CIE 5133 機器學習與深度學習導論
線上課程

開始之前 (10.27.2021)

- 請將你的麥克風靜音
- 請找個安全、舒適的空間
- 聽講時有任何問題請到 slido #073374 留言
- HW2 will be due today and TAs are ready to help!
- Midterm competition starts today!! 30%
- Happy Learning!!!

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

Classical Machine Learning: Classification and Regression (II)

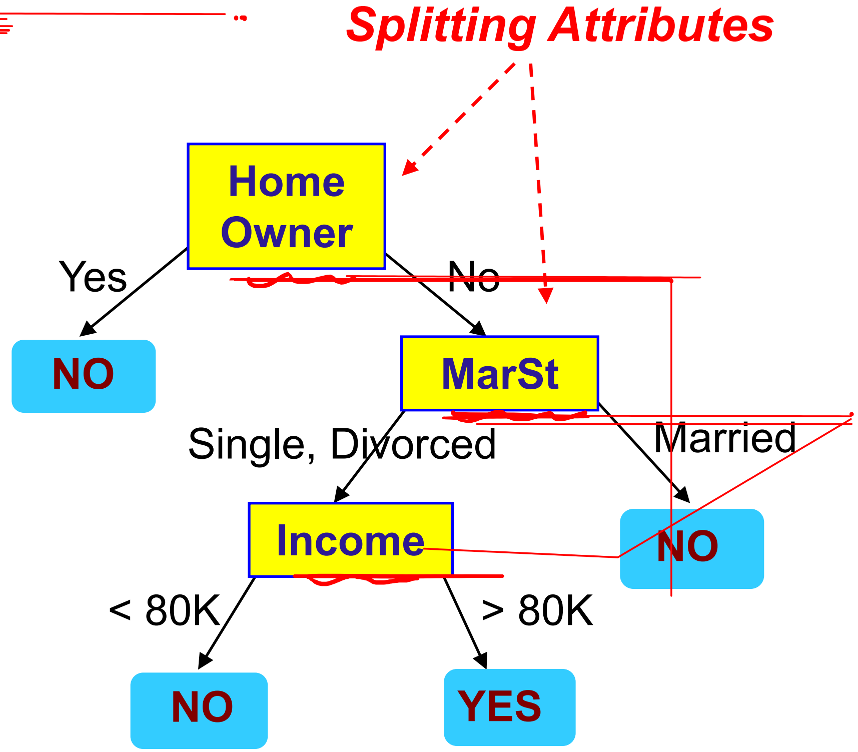
Learning Objectives

- Learn the basic concepts of a few interesting base classifiers.
- Learn the basic concepts of ensemble classifiers.

Summary and Recap
10.13.2021

Classification algorithm walkthrough: decision tree

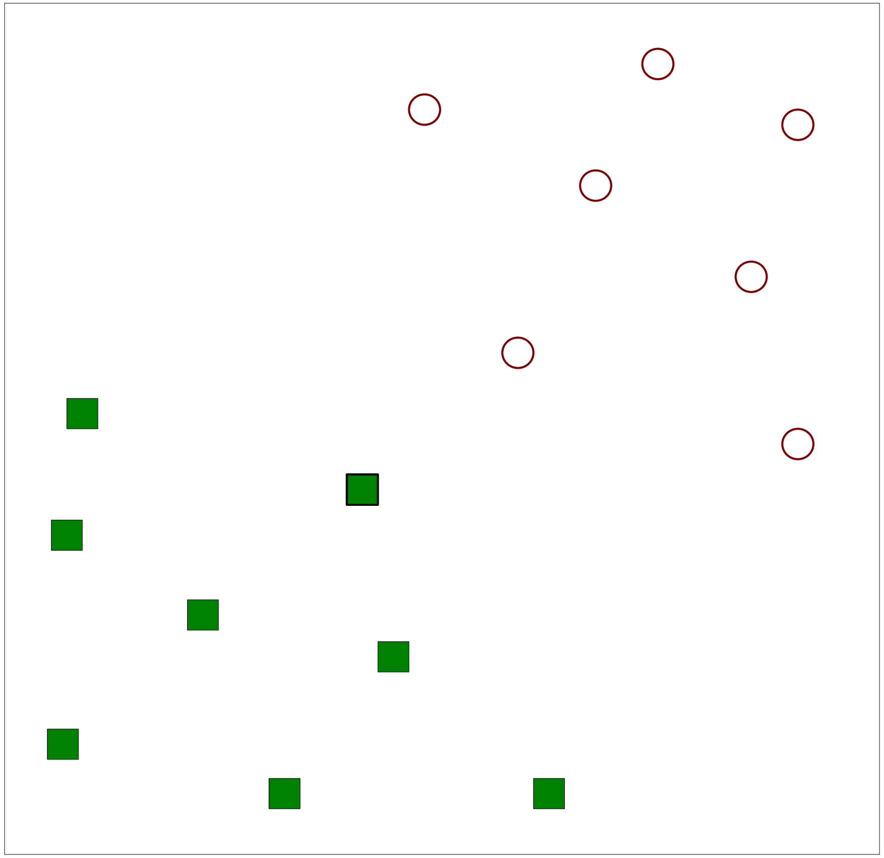
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



- The "Gini" criteria, or the "Entropy" criteria is the most commonly used index to determine the best split.
- Deep decision tree tends to overfit data (be alert to depth!).

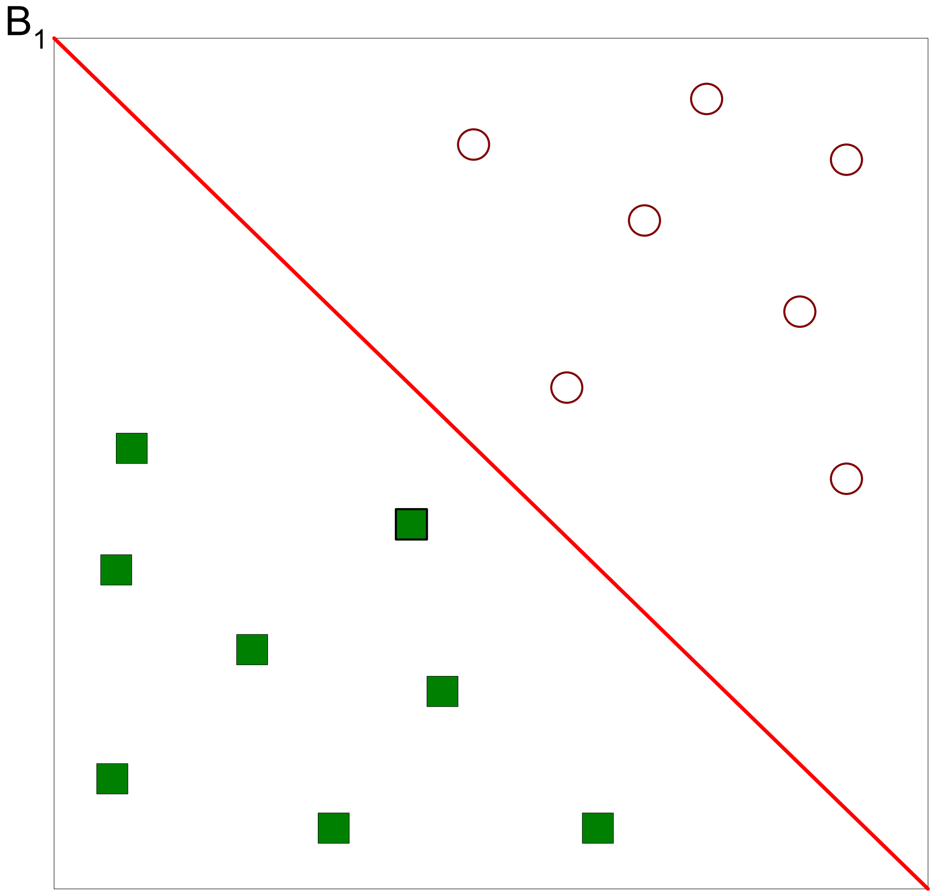
Classification Algorithm Walkthrough: Support Vector Machine (SVM)

Support Vector Machines



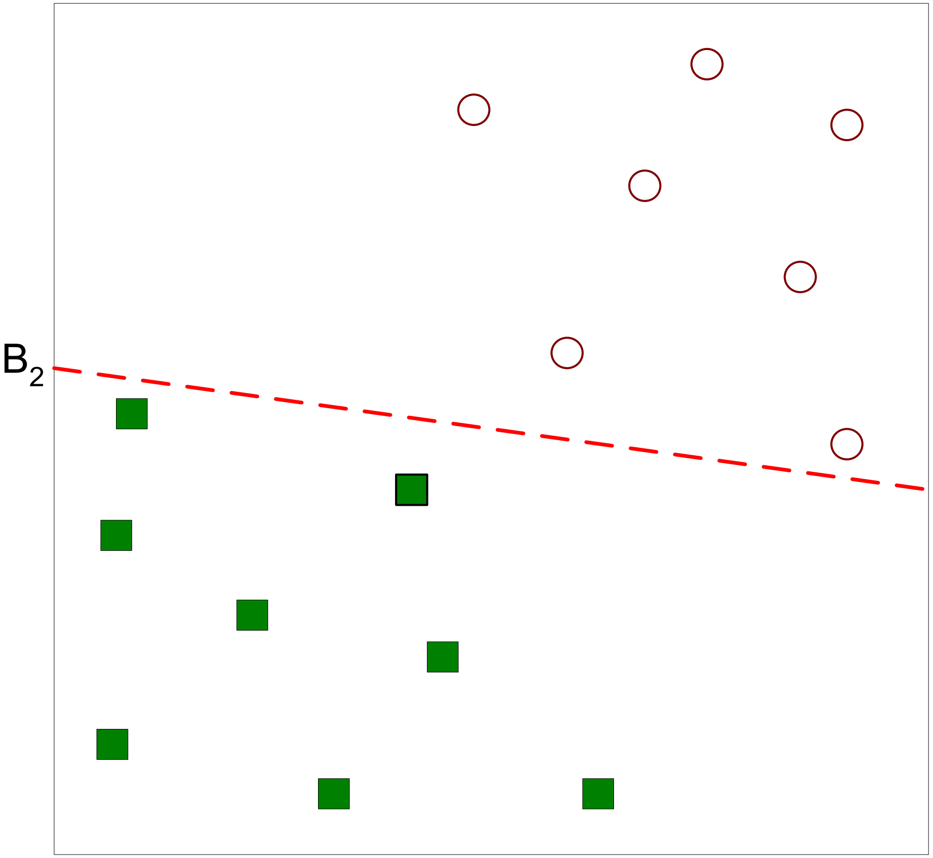
- Find a linear hyperplane (decision boundary) that will separate the data

Support Vector Machines



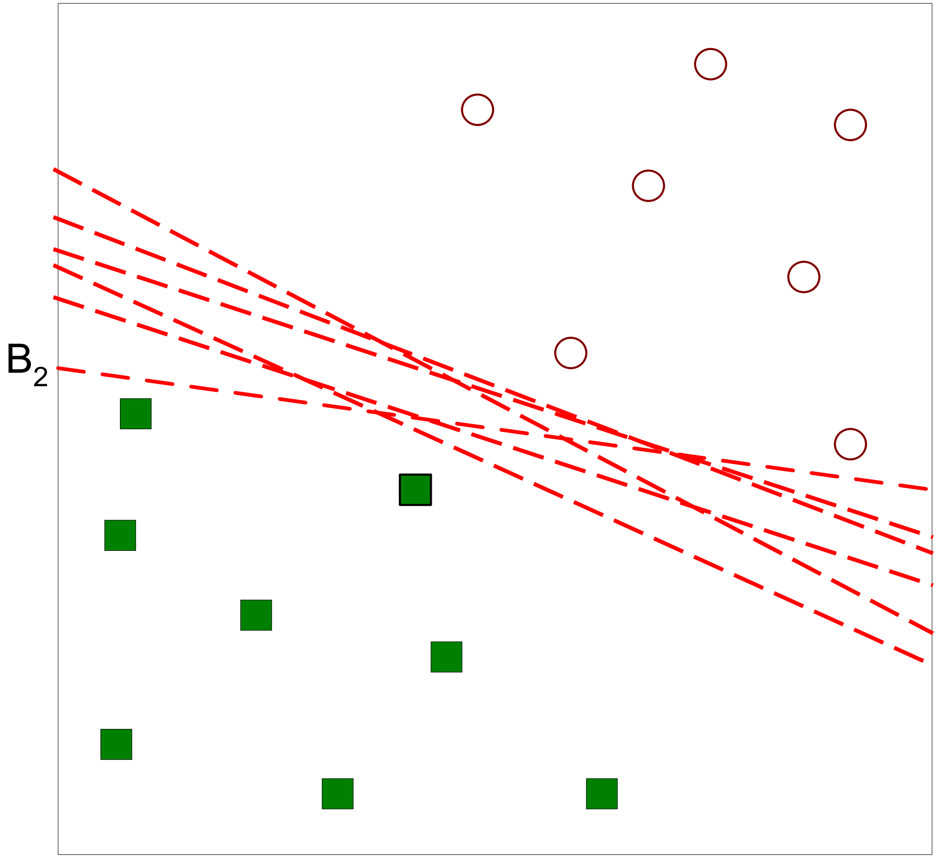
- One Possible Solution

Support Vector Machines



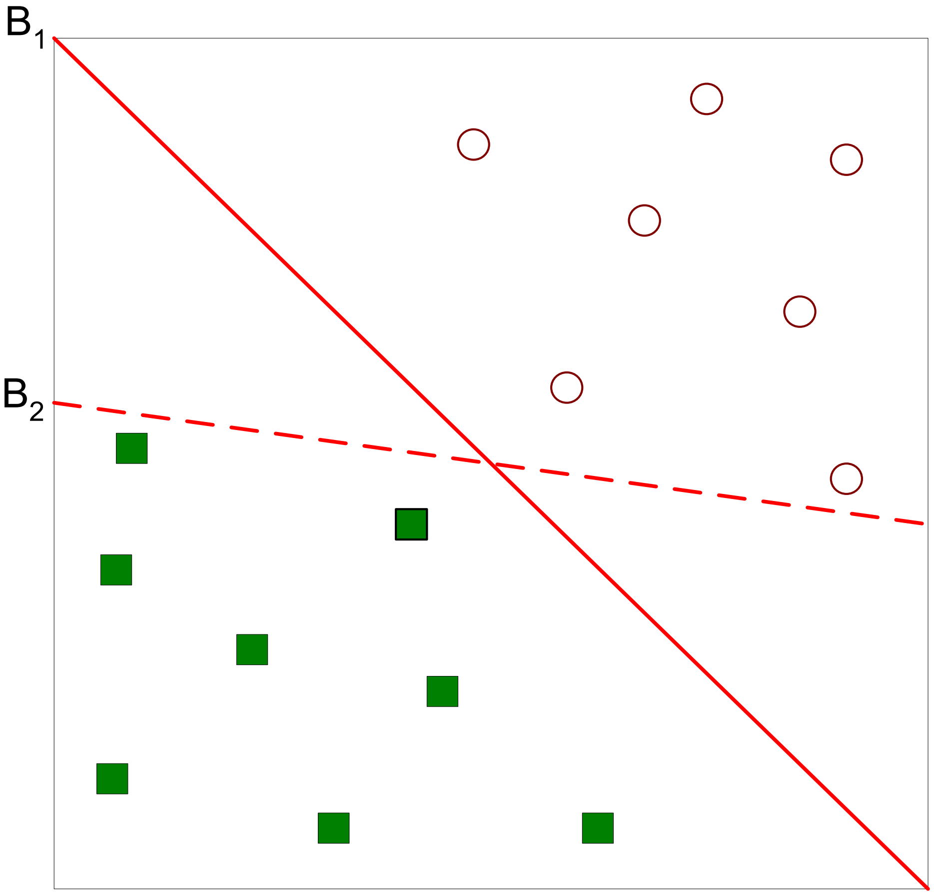
- Another possible solution

Support Vector Machines



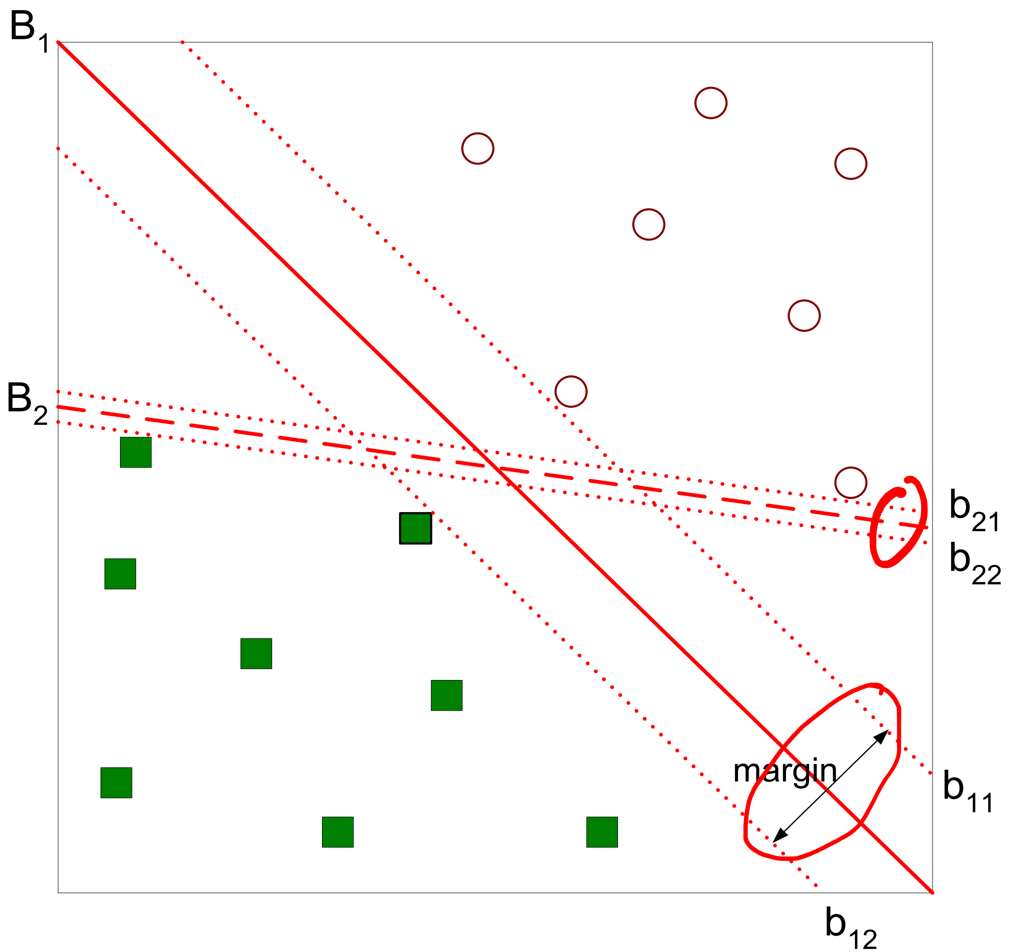
- Other possible solutions

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: 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 Support_Vector_Machine.pdf

Summary

Classification Algorithm Walkthrough: Support Vector Machine (SVM)

- **SVM classification is based on relatively few support vectors and is robust to noise.**
- **SVM can handle irrelevant and redundant data better than many other techniques.**
- **The integration with kernel methods makes SVM very versatile, able to adapt to many types of data.**

Classification Algorithm Walkthrough: Other Base Classifiers

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

6

C&RT

CART

decision tree

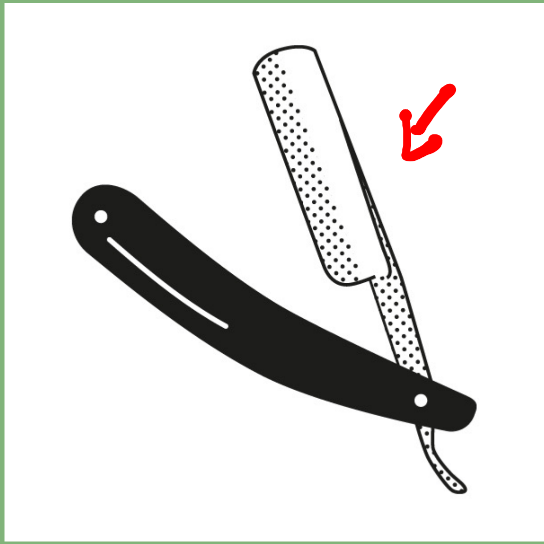
Classification algorithm shortlist

Fun Time: After cross validation, if a simple, linear machine learning model gives the same performance as a complex, nonlinear model, which model should we use?

(1) Simple model (2) Complex model

073374

Learning Principle: Occam's Razor



Occam's Razor: The simplest model that fits the data is also the most plausible (合理).

- Occam's razor in machine learning means simpler model has a better chance of being right. However, if a complex explanation of the data performs better, we will take it.
- The argument that simpler has a better chance of being right goes as follows. With complex hypotheses, there would be enough of them to fit the data set regardless of what the labels are, even if these are completely random. Therefore, fitting the data does not mean much.
- If, instead, we have a simple model with few hypotheses and we still found one that perfectly fits. This is surprising, and therefore it means some thing.

Classification Algorithm Walkthrough: Ensemble Classifiers

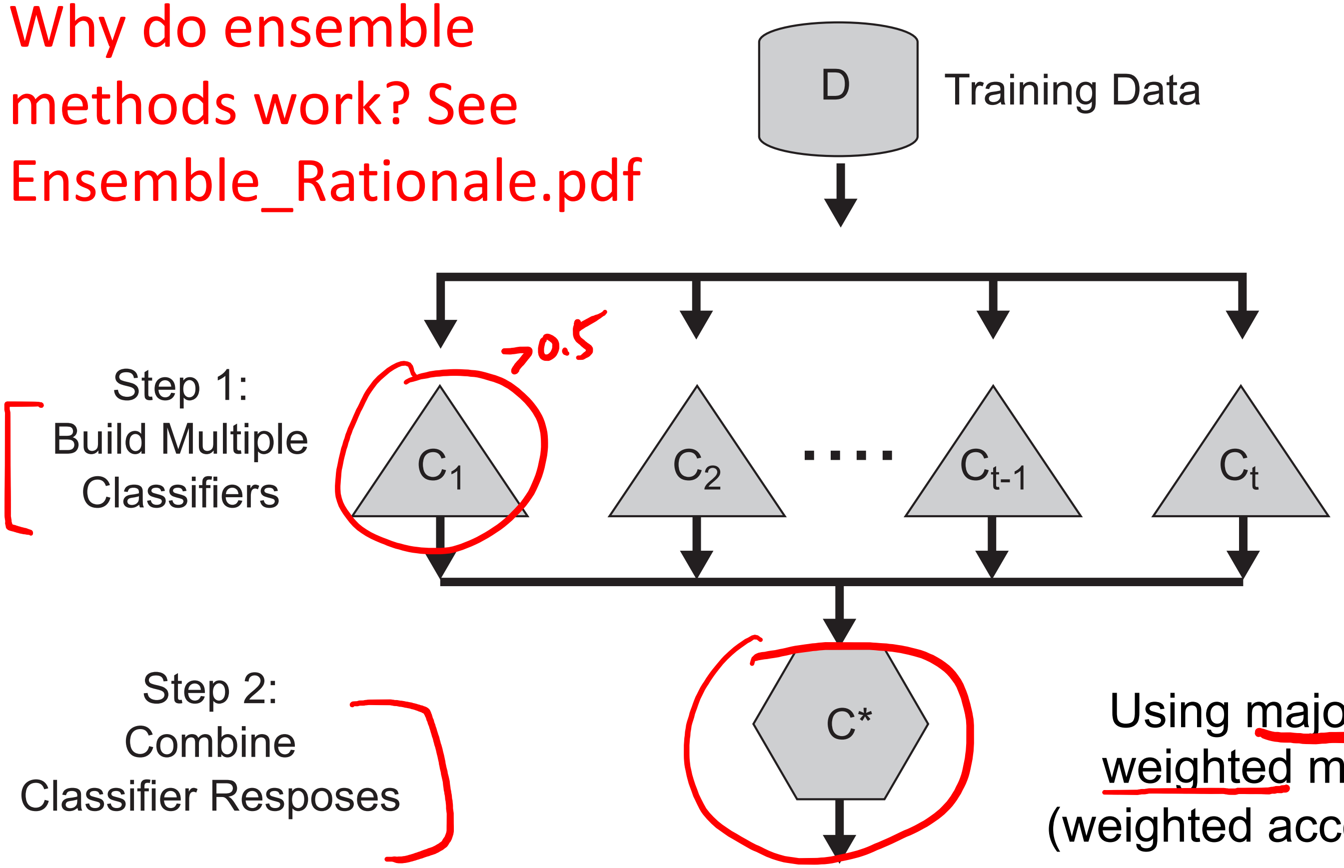
Ensemble Methods

- 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



Using majority vote or weighted majority vote (weighted according to their accuracy or relevance)

Base Classifiers for Ensemble Learning

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

  high ~~bias~~ variance low bias

  – 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: Unpruned decision trees, ANNs, …

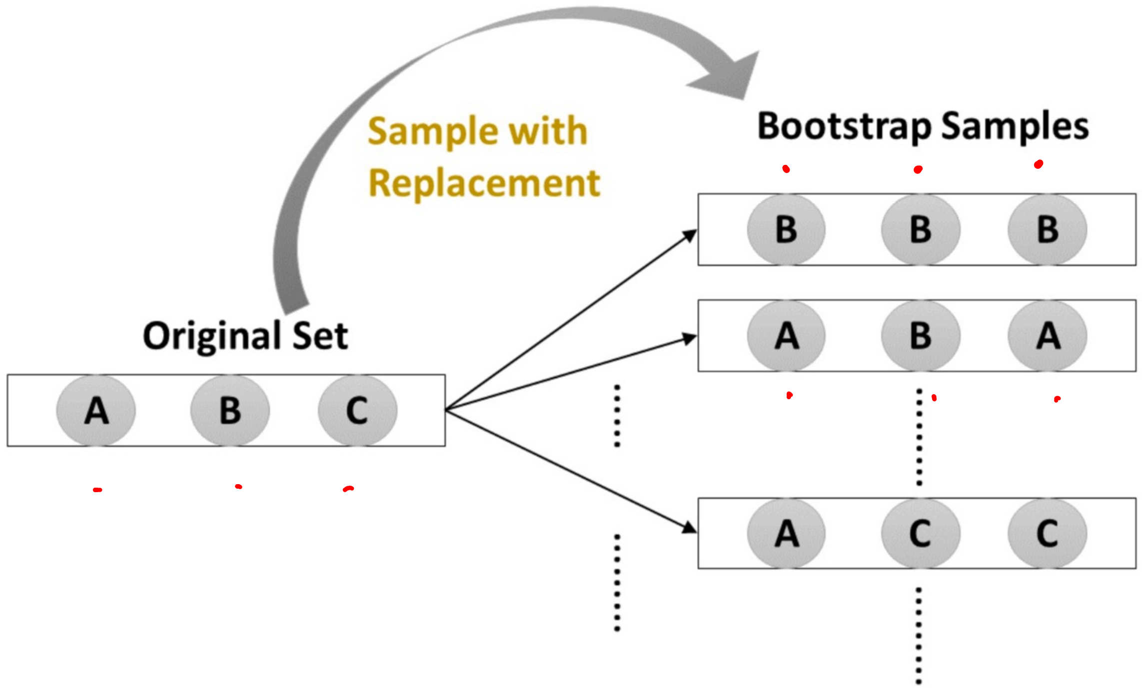
Sturctural Data

Classification Algorithm Walkthrough: Ensemble Classifiers - Bagging

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).



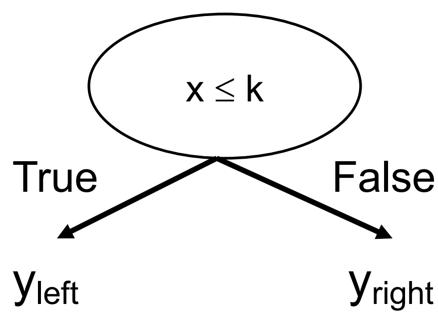
Bagging Example

- 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: $x \leq k$ versus $x > k$
  - Split point k is chosen based on entropy



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 ≤ k versus x > k
  - Split point k is chosen based on entropy

Fun Time: what is the best accuracy that a decision tree (a decision 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

-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

x=0.35

x=0.7

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

Wow!

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

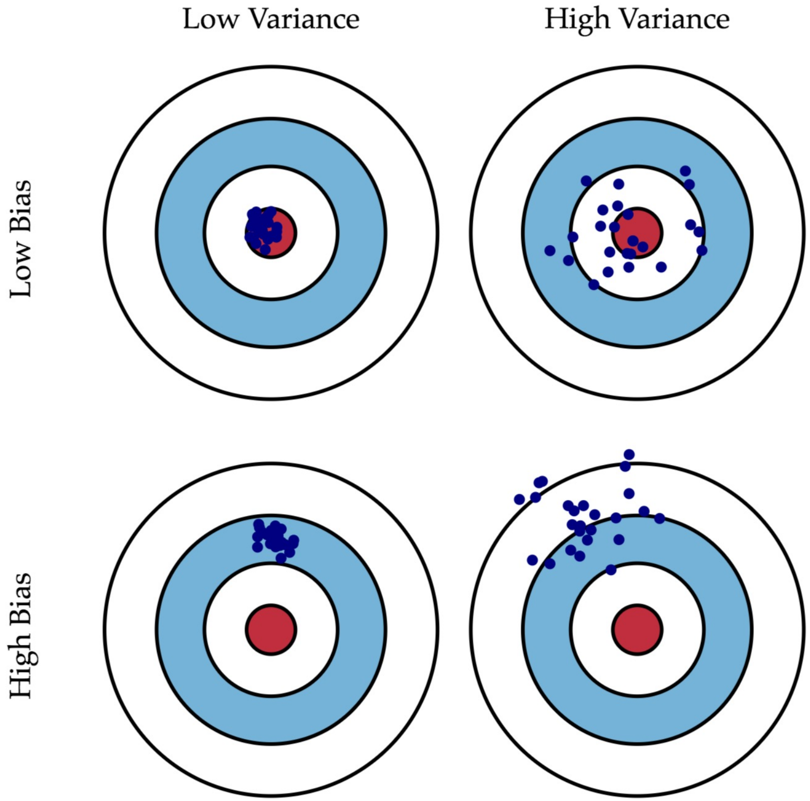
70% → 100%

sli.do #073374

Fun Time

Bagging is mainly meant to reduce the error from:

1. **bias**
2. **variance**



Bagging: 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 Ensemble_Bagging.pdf

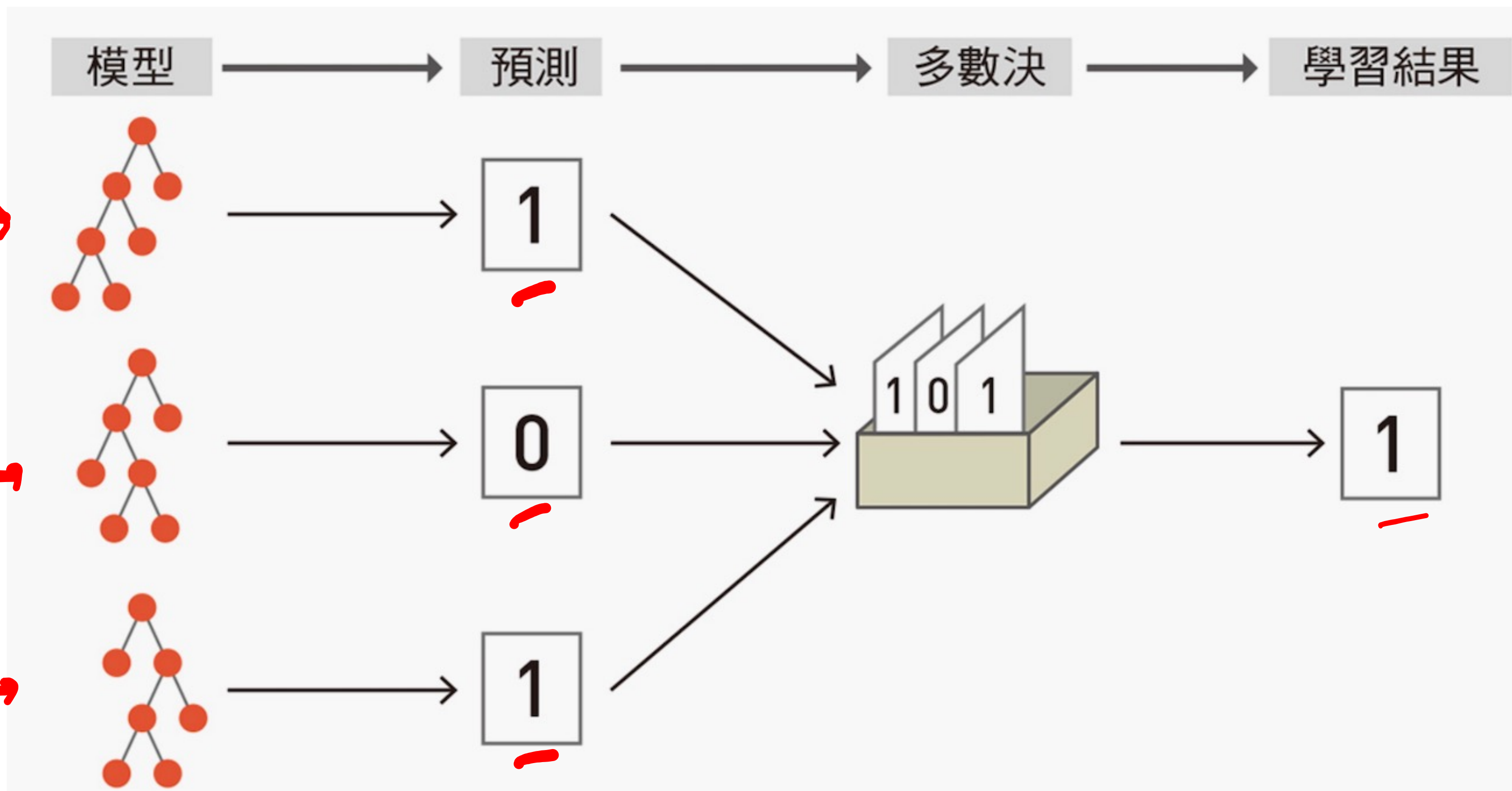
Classification Algorithm Walkthrough: Ensemble Classifiers – Random Forest

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 decision tree, randomly sample p attributes ($p < d$) for selecting split criterion
    - Repeat this procedure until all leaves are pure (unpruned tree)

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 p attributes ($p < d$) to split tree

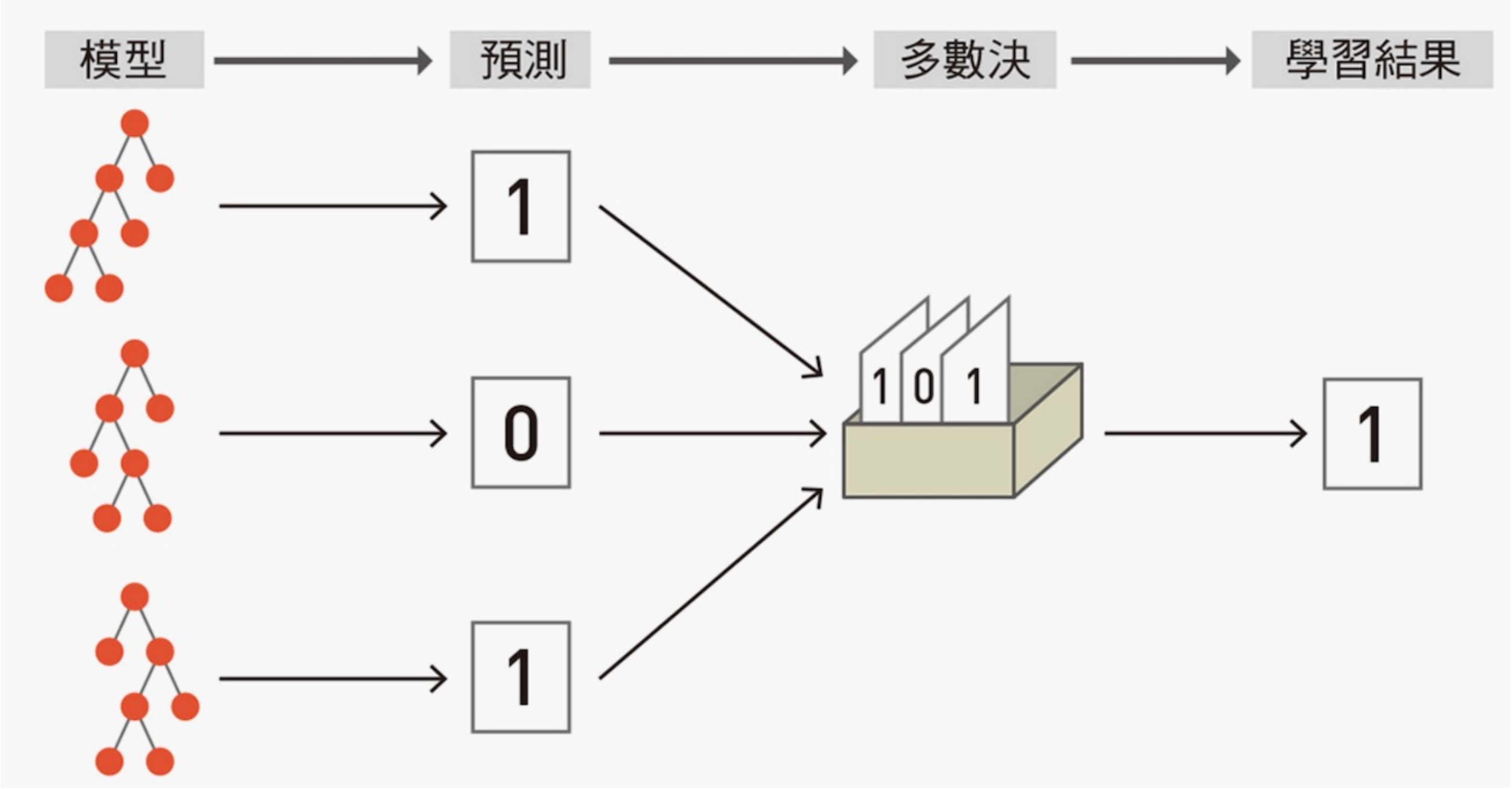


Random Forest: theoretical minimum and python example

- 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 Ensemble_Random_Forest.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

Contents lists available at ScienceDirect

Automation in Construction

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

ELSEVIER

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 for 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: Characteristics of Random Forest

- Base classifiers are unpruned trees and hence are *unstable classifiers*
- Base classifiers are *decorrelated* (due to randomization in training set as well as features)
- Random forests reduce variance of unstable classifiers without negatively impacting the bias
- Selection of hyper-parameter p
  - Small value ensures lack of correlation
  - High value promotes strong base classifiers
  - Common default choices: $\sqrt{d}, \ \log_2(d+1)$