Statistical Test for Final Project
Harvard Case Solution & Analysis

TASK: Construct a model to predict whether a person makes over $50.000 per year.

TheCaseSolutions.com

Pre-Processing

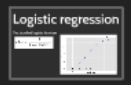
Feature	Type
age	INT
workclass	STRING
fnlwgt	INT
education	STRING
education-num	INT
marital-status	STRING
occupation	STRING
relationship	STRING
race	STRING
sex	STRING
capital-gain	INT
capital-loss	INT
hours-per-week	INT
native-country	STRING
label	STRING

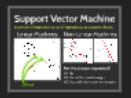
Input must be converted to normalized real numbers

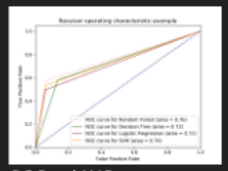
Training & Testing

Training size	7.500
Test size	2.500
Total	10.000

Classifiers







ROC and AUC

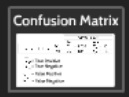
Testing

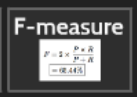
Model	Accuracy
Decison tree	79.28%
Logistic Regression	83.08%
SVM	84%
Random Forest	85.05%

Accuracies of the classifiers

Further Analysis

of the Random Forest Classifier





Analysis of new data

Positives: 189
Negatives: 1811

Summary

- Task: Determine if a person will earn more than $50.000/year
- Training data set: 7500 entries
- Testing data set: 2500 entries
- Methods tested: (Decision Tree, Logistic Regression, SVM, Random Forest)
- Method chosen: **Random Forest**
- **Result: 189 out of 2000 will earn more than** $50.000/year

THANKS FOR LISTENING

Pre-Pr

Statistical Test for Final Project
Harvard Case Solution & Analysis

Feature	Type
age	INT
workclass	STRING
fnlwgt	INT
education	STRING
education-num	INT
marital-status	STRING
occupation	STRING
relationship	STRING
race	STRING
sex	STRING
capital-gain	INT
capital-loss	INT
hours-per-week	INT
native-country	STRING
label	STRING

TASK: Construct a model to predict whether a person makes over $50.000 per year.

TheCaseSolutions.com

Pre-Processing

Feature	Type
age	INT
workclass	STRING
fnlwgt	INT
education	STRING
education-num	INT
marital-status	STRING
occupation	STRING
relationship	STRING
race	STRING
sex	STRING
capital-gain	INT
capital-loss	INT
hours-per-week	INT
native-country	STRING
label	STRING

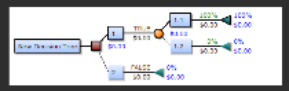
Input must be converted to normalized real numbers

Training & Testing

Training size	7.500
Test size	2.500
Total	10.000

Classifiers

Decision tree classifier

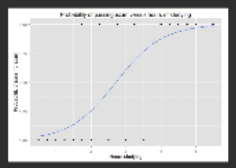


- Flowchart-like structure
- Nodes represent "tests" on an attribute
- Branches represent the outcome of the test
- Leafs represent a class label
- The paths from root to leaf represent classification rules.

Logistic regression

The standard logistic function

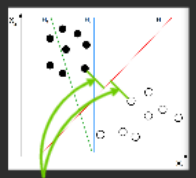
$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



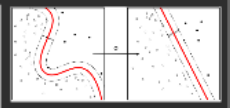
Support Vector Machine

Constructs a hyperplane or set of hyperplanes to separate classes

Linear Machines



Non-Linear Machines

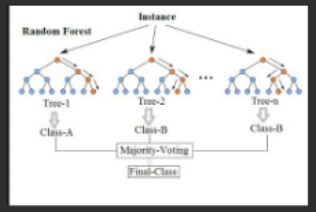


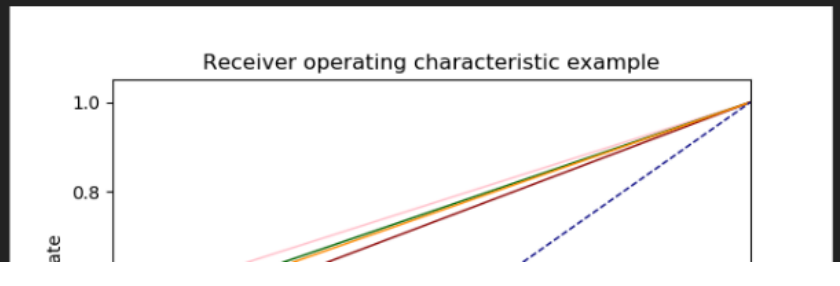
Are the classes separated?

H1: No
H2: Yes, with a small margin
H3: Yes, with the maximum margin.

Random Forest

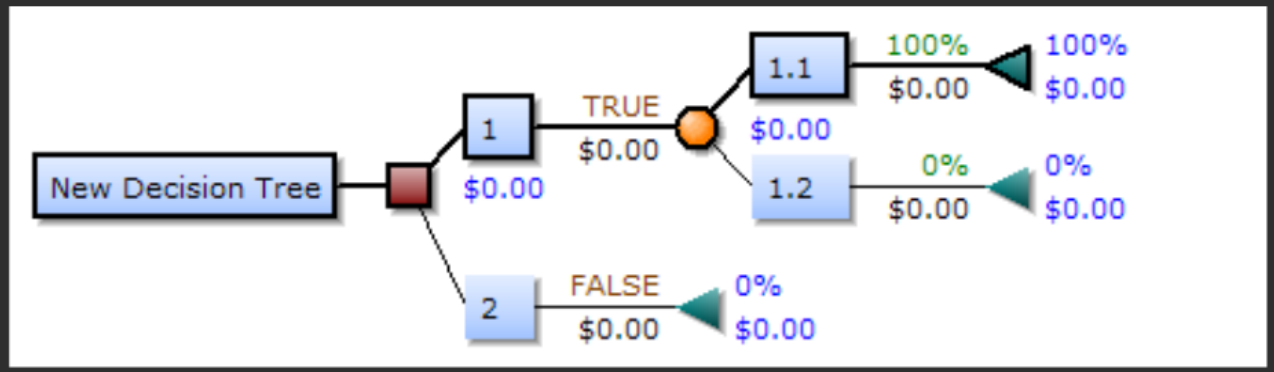
Constructs many decision trees to classify using majority voting





Testing

Decision tree classifier

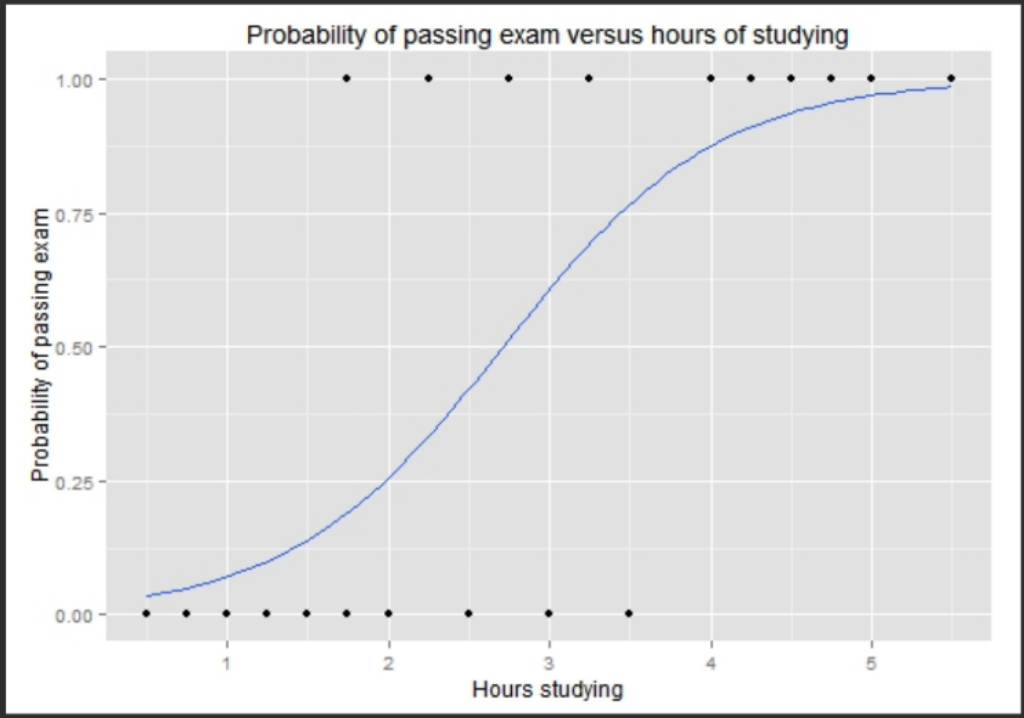


- Flowchart-like structure
- Nodes represent "tests" on an attribute
- Branches represent the outcome of the test
- Leafs represent a class label
- The paths from root to leaf represent classification rules.

Logistic regression

The standard logistic function

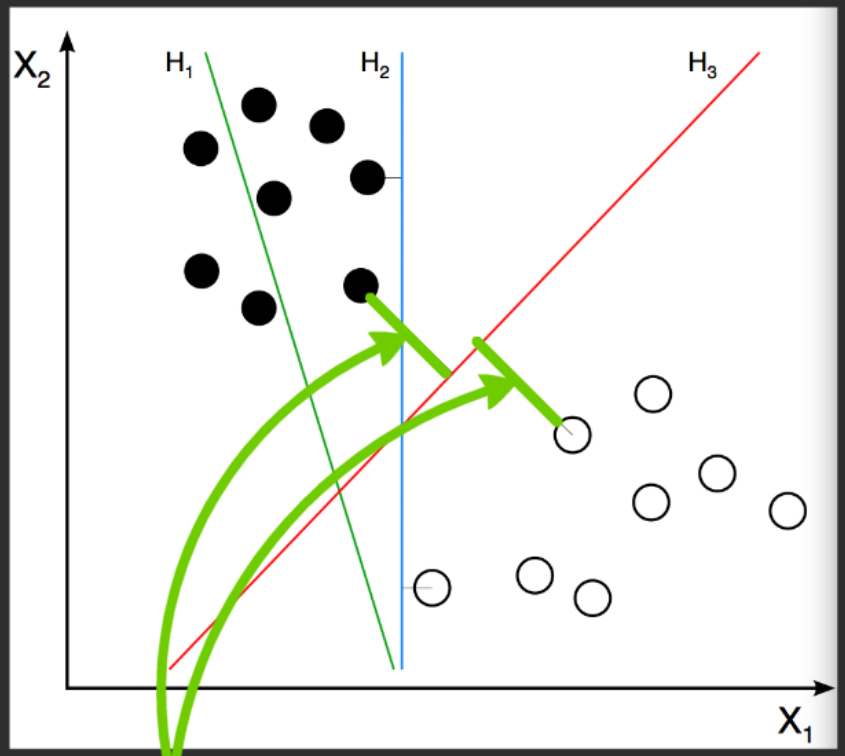
$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



Support Vector Machine

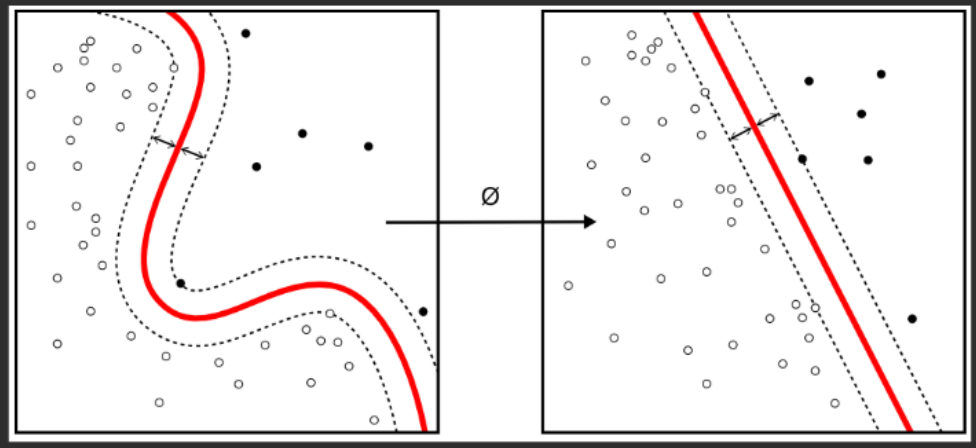
Constructs a hyperplane or set of hyperplanes to separate classes

Linear Machines



Offset

Non-Linear Machines



Are the classes separated?

H1: No

H2: Yes, with a small margin

H3: Yes, with the maximum margin.

Random Forest

Constructs many decision trees to classify using majority voting