Astroinformatics school - "Rise of the machines"

Curtin University
THE CURTIN INSTITUTE FOR COMPUTATION

4 to 6 February 2019

Presented by Rebecca Lange and Dan Marrable

Decision trees

ASTRONOMY DATA AND COMPUTING SERVICES

- Classification algorithm
  - Predict discrete (category) outputs
  - Will Shiv play cricket?
- Training
  - Recursively split data into subsets based on a single attribute
  - Stop when all subsets are pure (all yes / no)
- Prediction
  - Based on subset where new data is placed

Training data: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Cricket
1	Sunny	High	Weak	No
2	Sunny	High	Strong	No
3	Overcast	High	Weak	Yes
4	Rain	High	Weak	Yes
5	Rain	Normal	Weak	Yes
6	Rain	Normal	Strong	No
7	Overcast	Normal	Strong	Yes
8	Sunny	High	Weak	No
9	Sunny	Normal	Weak	Yes
10	Rain	Normal	Weak	Yes
11	Sunny	Normal	Strong	Yes
12	Overcast	High	Strong	Yes
13	Overcast	Normal	Weak	Yes
14	Rain	High	Strong	No

New data:

15	Rain	High	Weak	?

9 yes / 5 no

Outlook

Sunny

Day	Outlook	Humidity	Wind
1	Sunny	High	Weak
2	Sunny	High	Strong
8	Sunny	High	Weak
9	Sunny	Normal	Weak
11	Sunny	Normal	Strong

2 yes / 3 no
split further

Overcast

Day	Outlook	Humidity	Wind
3	Overcast	High	Weak
7	Overcast	Normal	Strong
12	Overcast	High	Strong
13	Overcast	Normal	Weak

4 yes / 0 no
pure subset

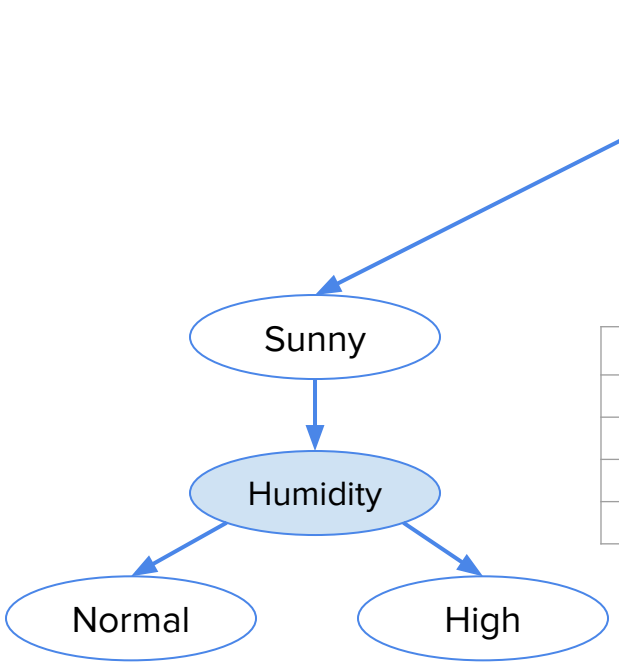
Rain

Day	Outlook	Humidity	Wind
4	Rain	High	Weak
5	Rain	Normal	Weak
6	Rain	Normal	Strong
10	Rain	Normal	Weak
14	Rain	High	Strong

3 yes / 2 no
split further

Source: Victor Lavrenko, Decision Trees, University of Edinburgh 2014.

9 yes / 5 no



Day	Outlook	Humidity	Wind
3	Overcast	High	Weak
7	Overcast	Normal	Strong
12	Overcast	High	Strong
13	Overcast	Normal	Weak

4 yes / 0 no
pure subset

Day	Outlook	Humidity	Wind
4	Rain	High	Weak
5	Rain	Normal	Weak
6	Rain	Normal	Strong
10	Rain	Normal	Weak
14	Rain	High	Strong

3 yes / 2 no
split further

Day	Humidity	Wind
9	Normal	Weak
11	Normal	Strong

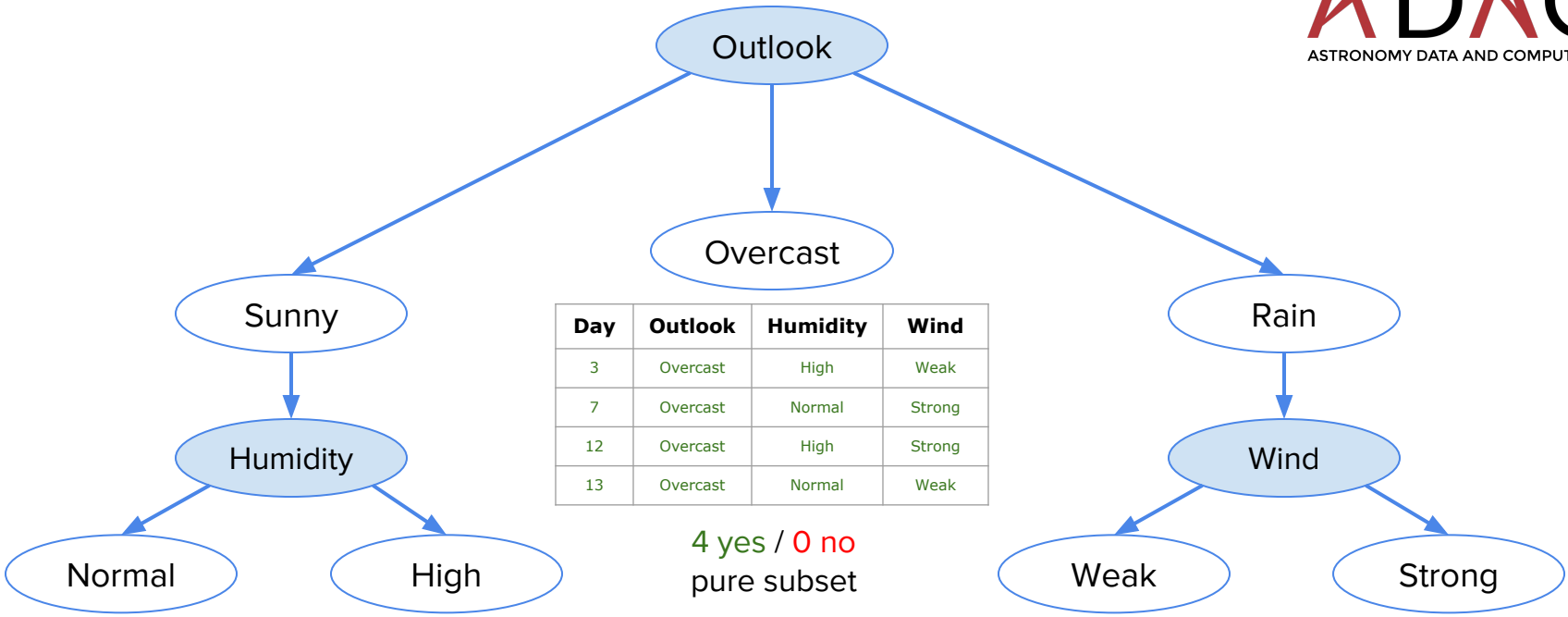
2 yes / 0 no
pure subset

Day	Humidity	Wind
1	High	Weak
2	High	Strong
8	High	Weak

0 yes / 3 no
pure subset

Source: Victor Lavrenko, Decision Trees, University of Edinburgh 2014.

9 yes / 5 no



Day	Outlook	Humidity	Wind
3	Overcast	High	Weak
7	Overcast	Normal	Strong
12	Overcast	High	Strong
13	Overcast	Normal	Weak

4 yes / 0 no
pure subset

Day	Humidity	Wind
9	Normal	Weak
11	Normal	Strong

2 yes / 0 no
pure subset

Day	Humidity	Wind
1	High	Weak
2	High	Strong
8	High	Weak

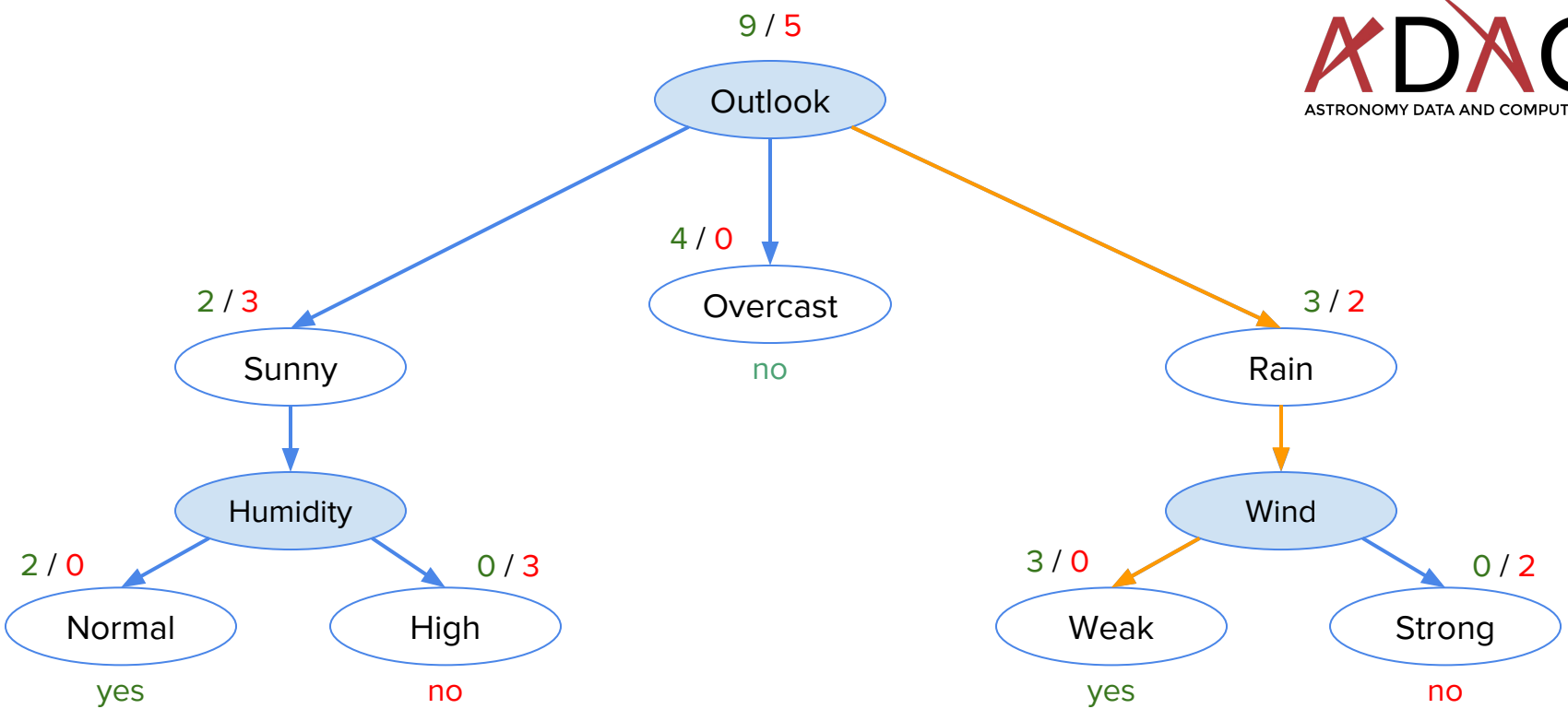
0 yes / 3 no
pure subset

Day	Outlook	Humidity
4	Rain	High
5	Rain	Normal
10	Rain	Normal

3 yes / 0 no
pure subset

Day	Outlook	Humidity
6	Rain	Normal
14	Rain	High

0 yes / 2 no
pure subset



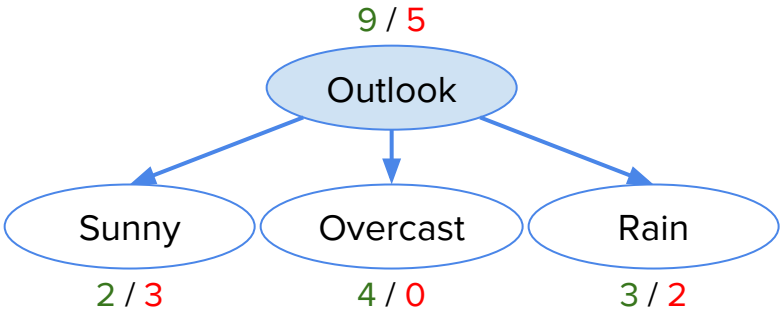
New data:

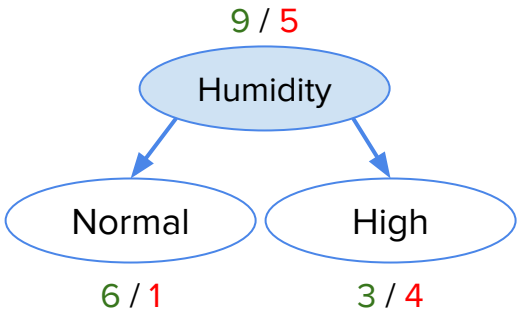
Day	Outlook	Humidity	Wind	Cricket
15	Rain	High	Weak	?

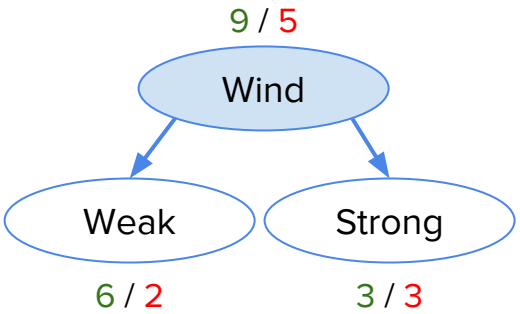
→ yes

Decision trees

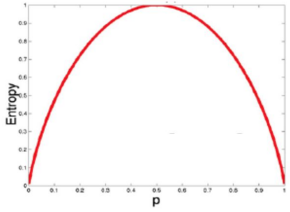
- Which attribute provides the best split?







- Split attributes based on a purity measure (e.g. entropy, gini)
- Information gain = average weighted entropy of each subset
- Example: http://www.saedsayad.com/decision_tree.htm



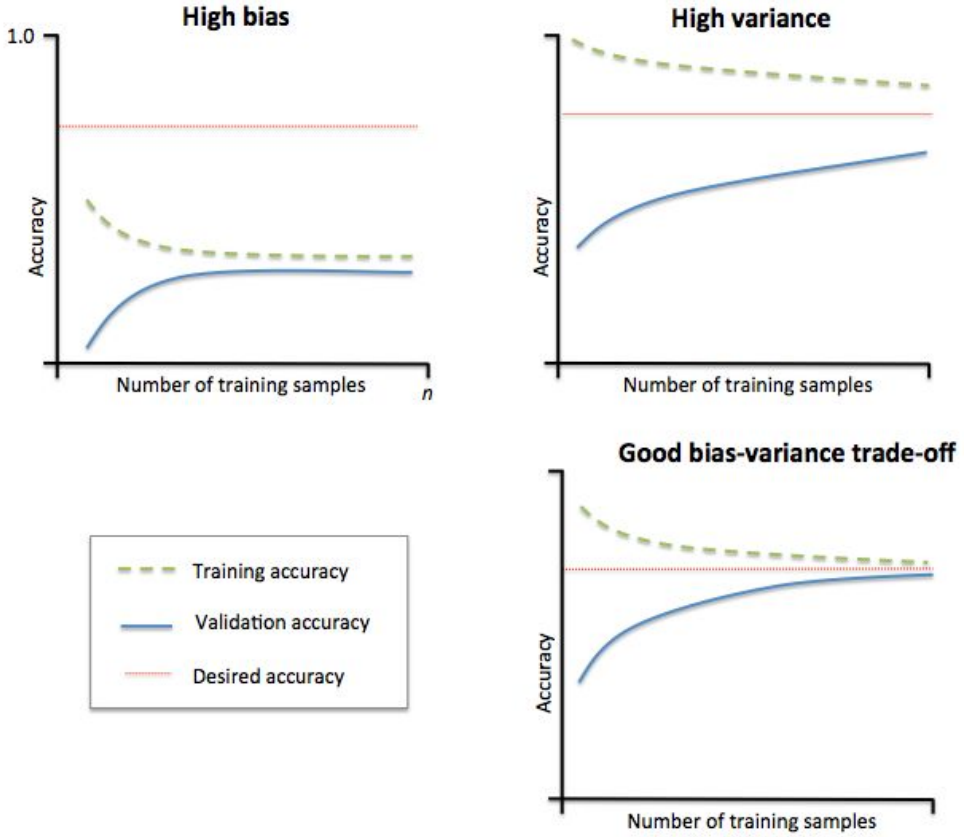
Learning curves

Underfit (high bias)

- Try more features
- Decrease regularisation

Overfit (high variance)

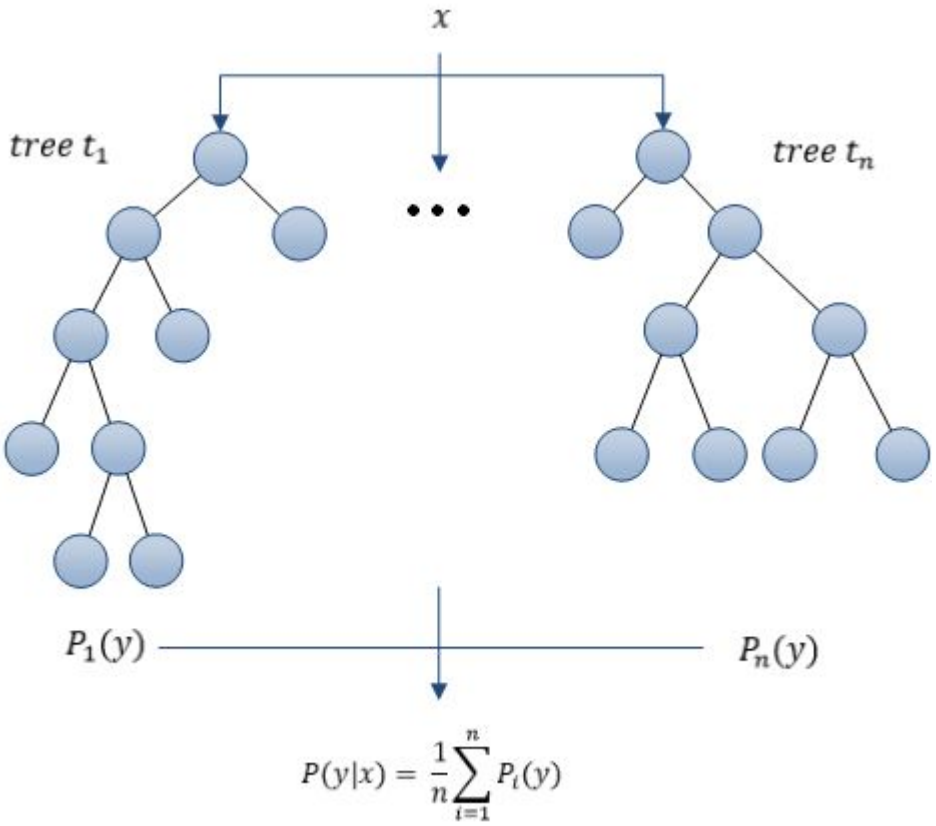
- Get more data
- Use less features
- Increase regularisation



Source: Andrew Ng, Coursera - Machine Learning. 2012

Random forests

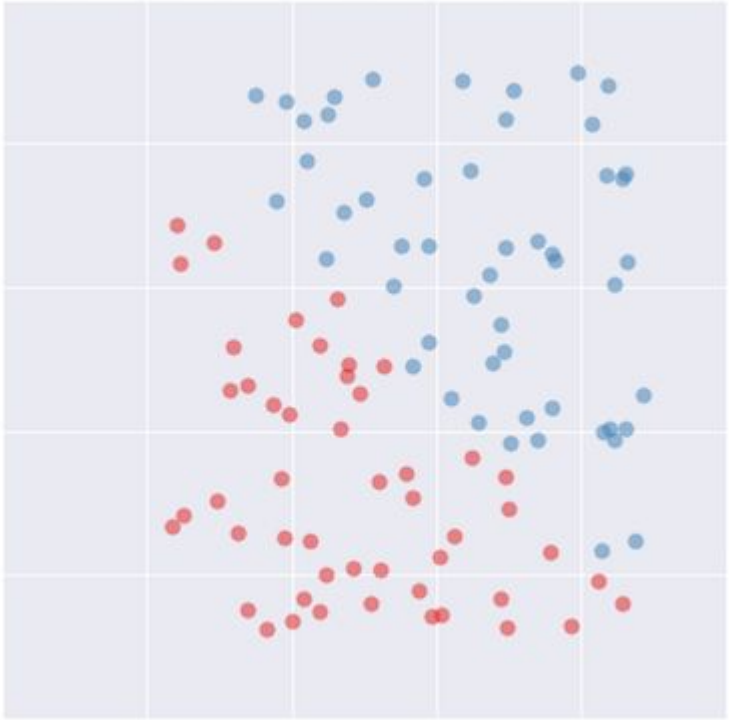
- Train multiple decision trees (forest
  - training examples
  - features
- Final prediction is based on the ave
- Avoids overfitting problems with sir



Disadvantages

ASTRONOMY DATA AND COMPUTING SERVICES

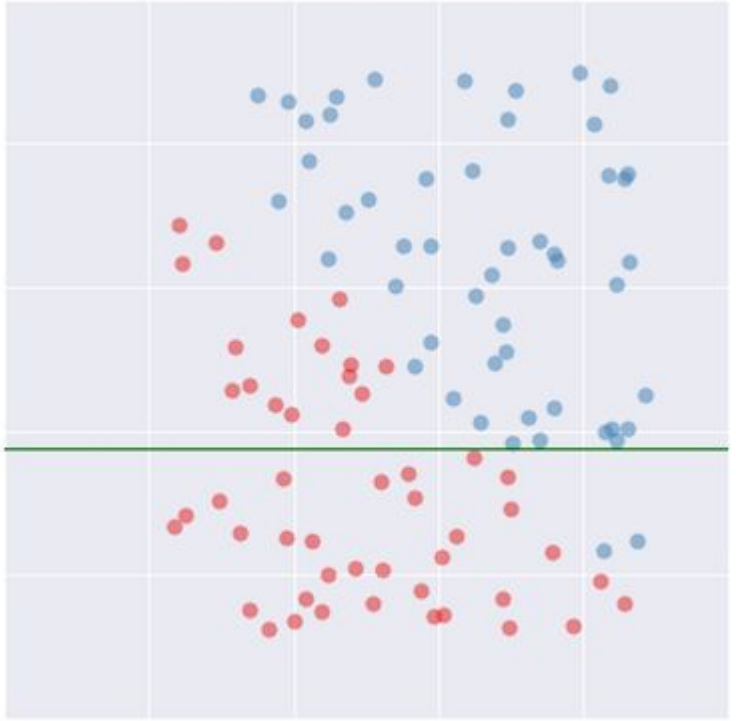
- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model



Disadvantages

ASTRONOMY DATA AND COMPUTING SERVICES

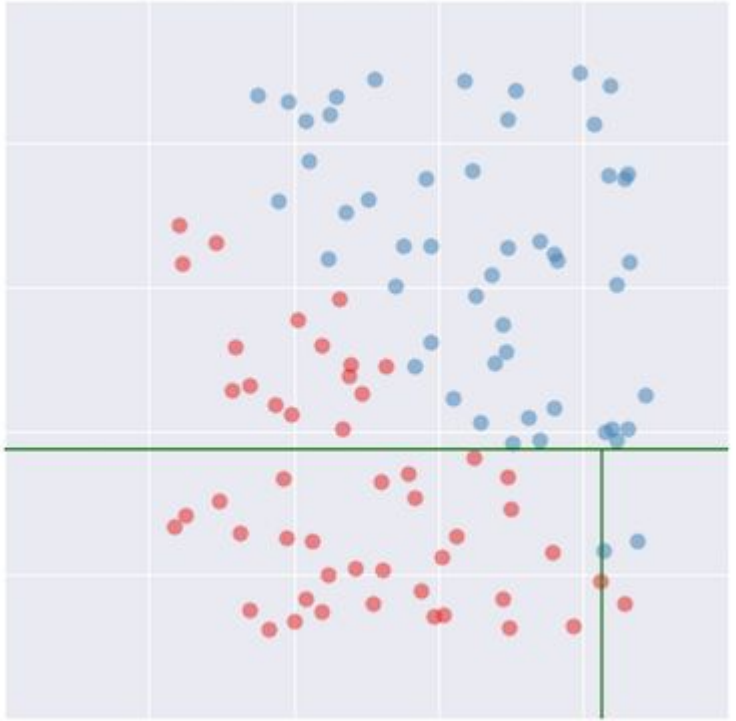
- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model



Disadvantages

ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

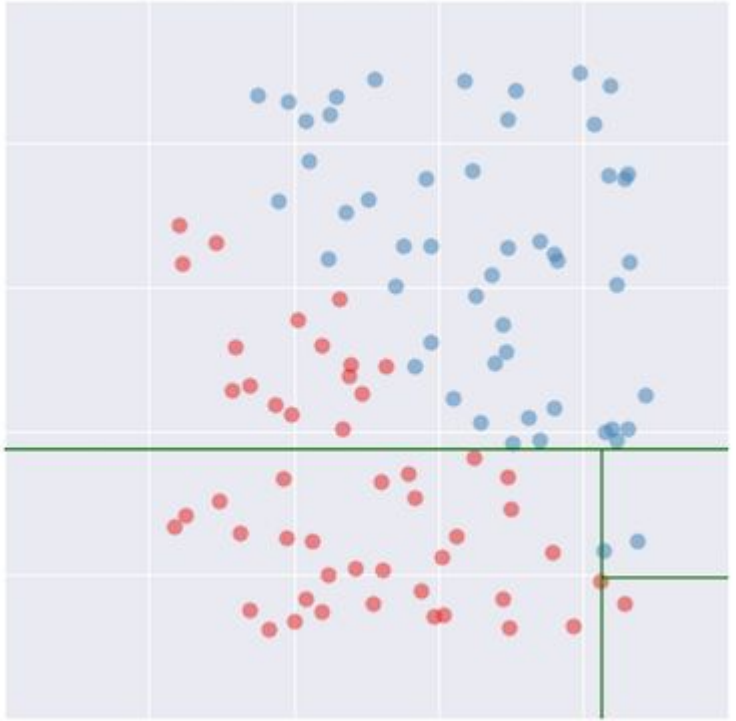
- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model



Disadvantages

ASTRONOMY DATA AND COMPUTING SERVICES

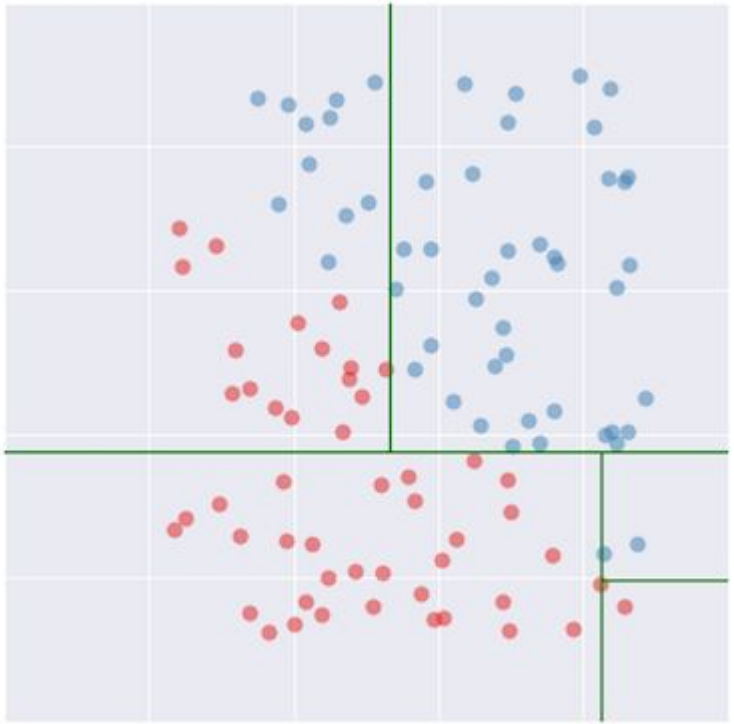
- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model



Disadvantages

ASTRONOMY DATA AND COMPUTING SERVICES

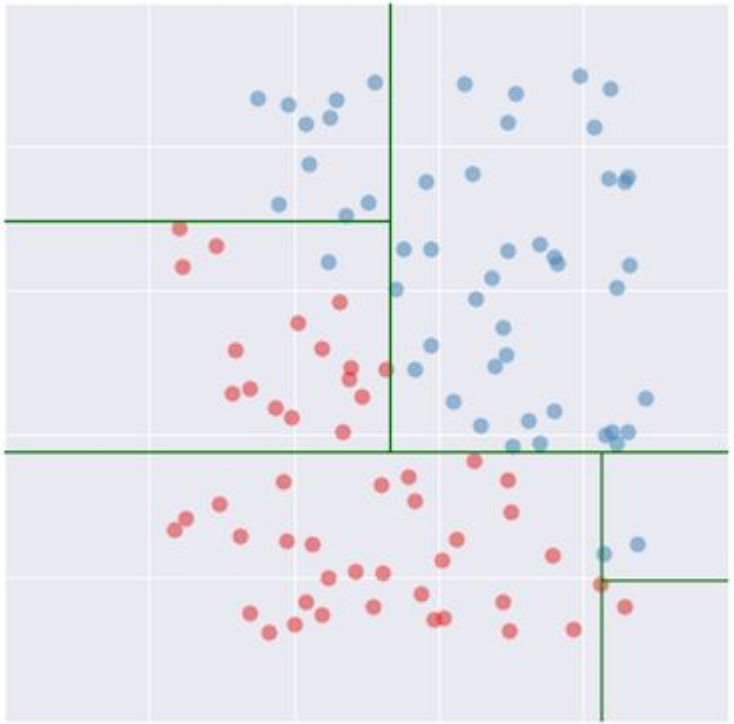
- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model



Disadvantages

ASTRONOMY DATA AND COMPUTING SERVICES

- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model



Disadvantages

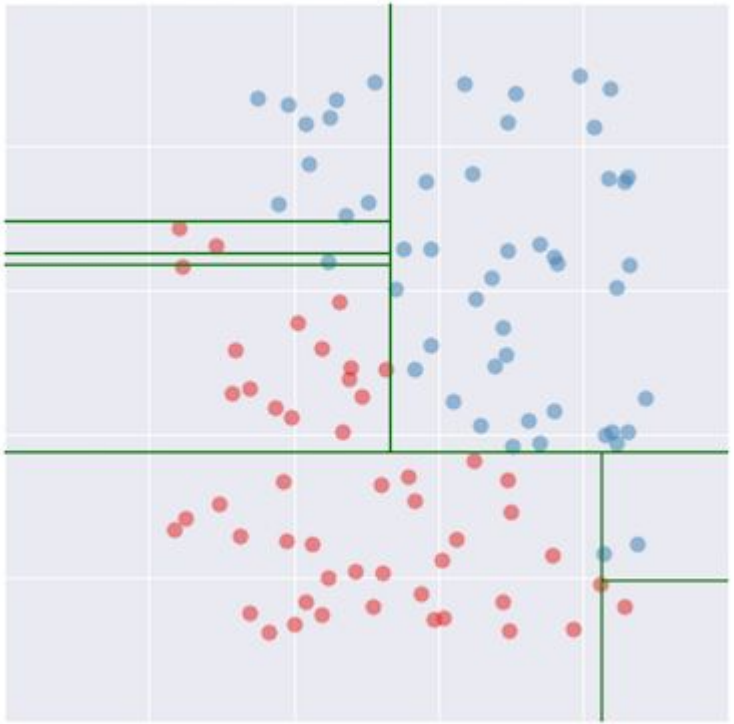
ASTRONOMY DATA AND COMPUTING SERVICES

- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model

Disadvantages

ASTRONOMY DATA AND COMPUTING SERVICES

- Decision trees split one attribute at a time
- Greedy algorithm
- Can result in an overly complex model



Disadvantages

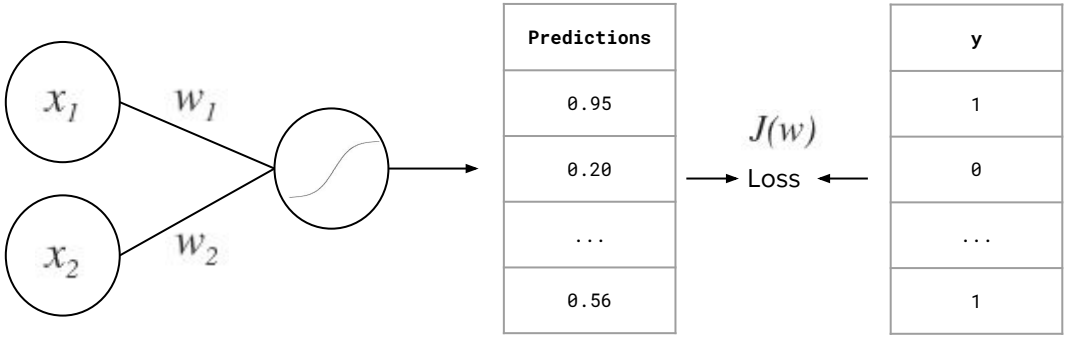
Better: a model that can predict using less parameters, e.g. a function instead of a decision tree.

Logistic Regression

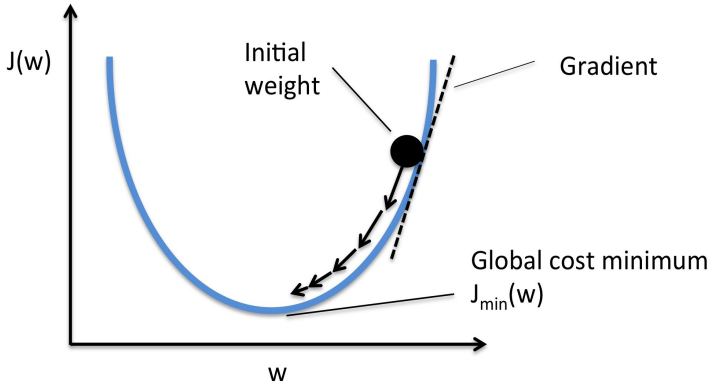
- Linear model

$$z = w_1 x_1 + w_2 x_2 + b$$

3 parameters



Optimisation



Logistic Regression

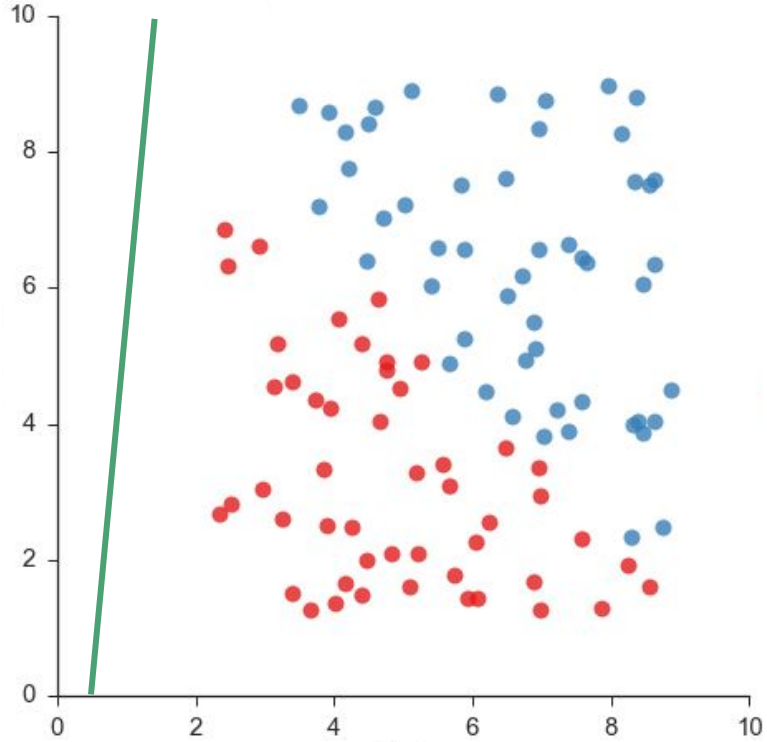
ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

Iteration: 1

Model: $0.09x_1 - 0.01x_2 - 0.024$

Loss: 0.641

Accuracy: 0.51



Logistic Regression

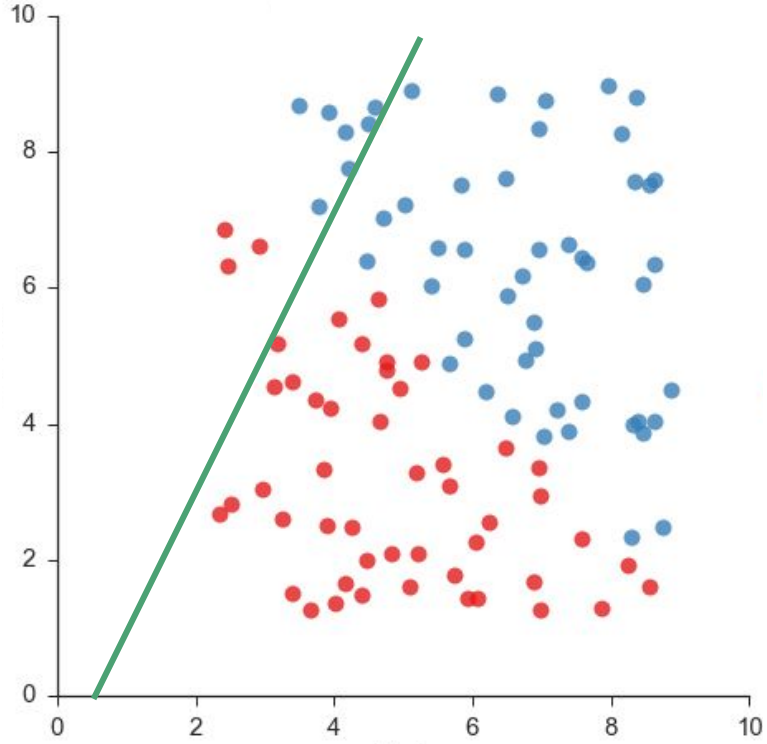
ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

Iteration: 2

Model: $0.17x_1 - 0.08x_2 - 0.07$

Loss: 0.623

Accuracy: 0.66



Logistic Regression

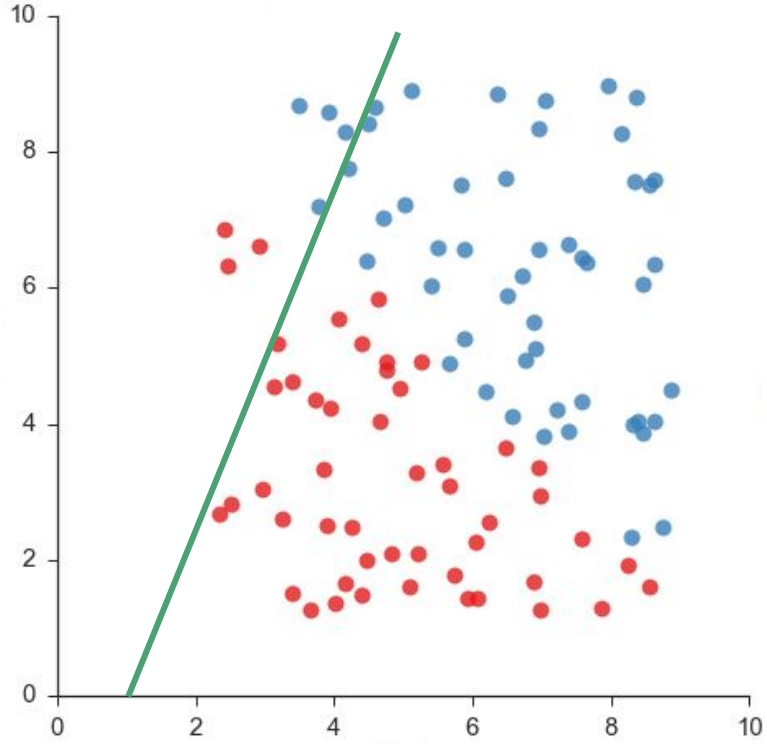
ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

Iteration: 5

Model: $0.42x_1 - 0.16x_2 - 0.46$

Loss: 0.583

Accuracy: 0.68



Logistic Regression

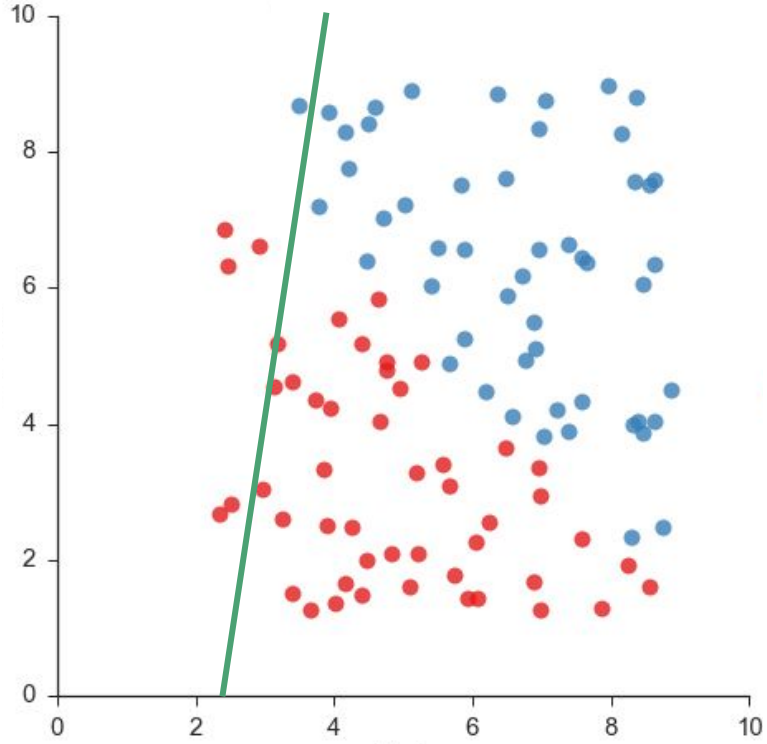
ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

Iteration: 7

Model: $0.64x_1 - 0.08x_2 - 1.65$

Loss: 0.527

Accuracy: 0.77



Logistic Regression

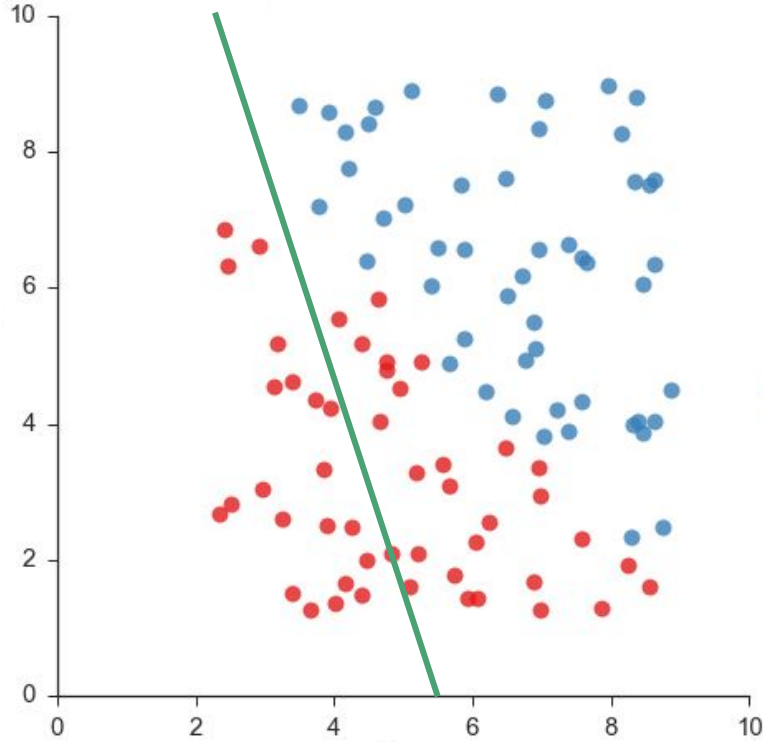
ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

Iteration: 9

Model: $1.15x_1 - 0.37x_2 - 6.39$

Loss: 0.351

Accuracy: 0.82



Logistic Regression

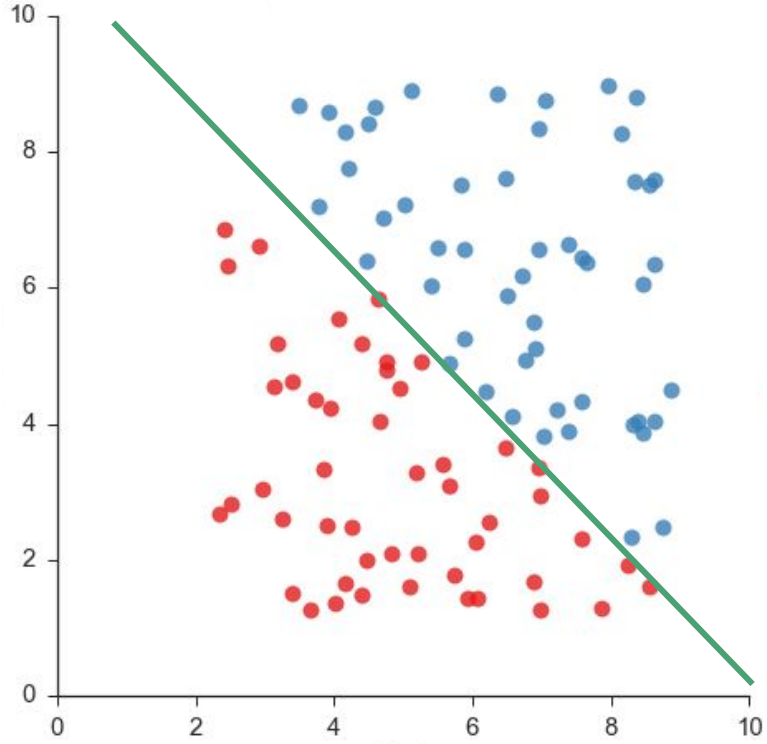
ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

Iteration: 18

Model: $2.13x_1 - 2.05x_2 - 21.88$

Loss: 0.091

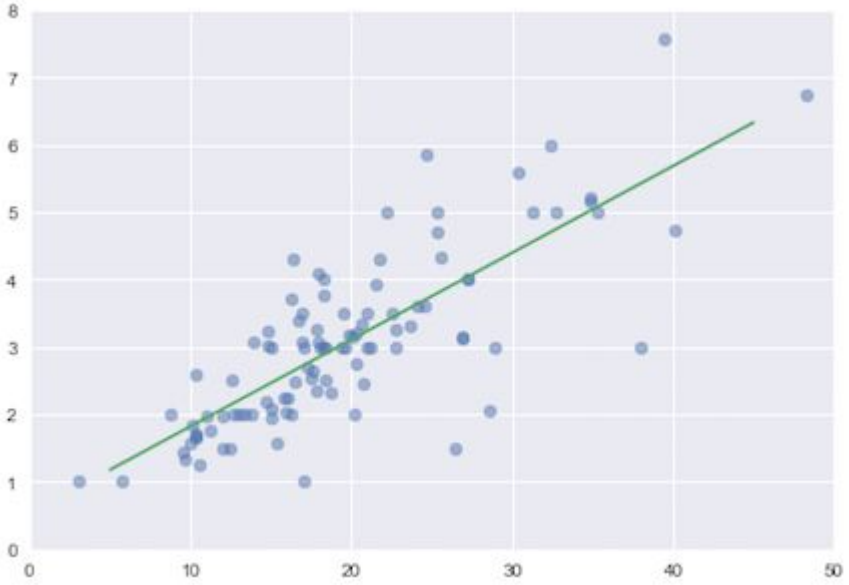
Accuracy: 0.99



Linear regression

ADACS
ASTRONOMY DATA AND COMPUTING SERVICES

- Predict a continuous value
- Find the line of best fit
- Training process similar to logistic regression
  - different loss function
  - mean squared error



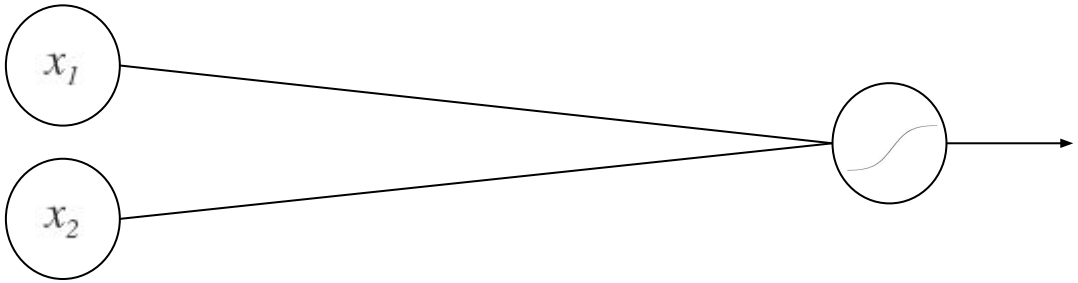
- Linear model

$$y = 0.12x + 0.53$$

Artificial Neural Networks

ASTRONOMY DATA AND COMPUTING SERVICES

Logistic regression

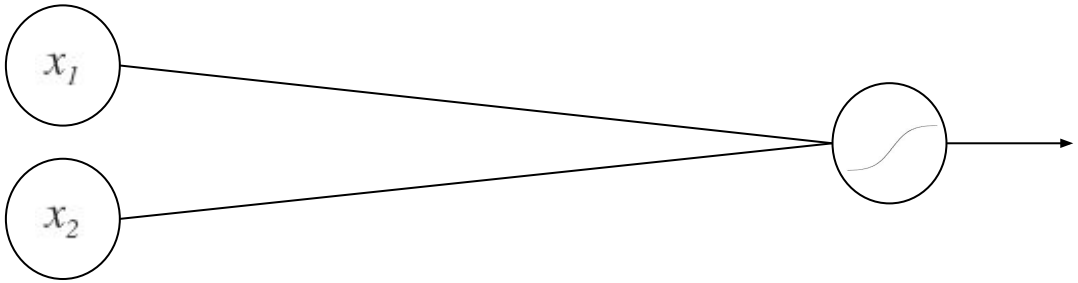


Artificial Neural Networks

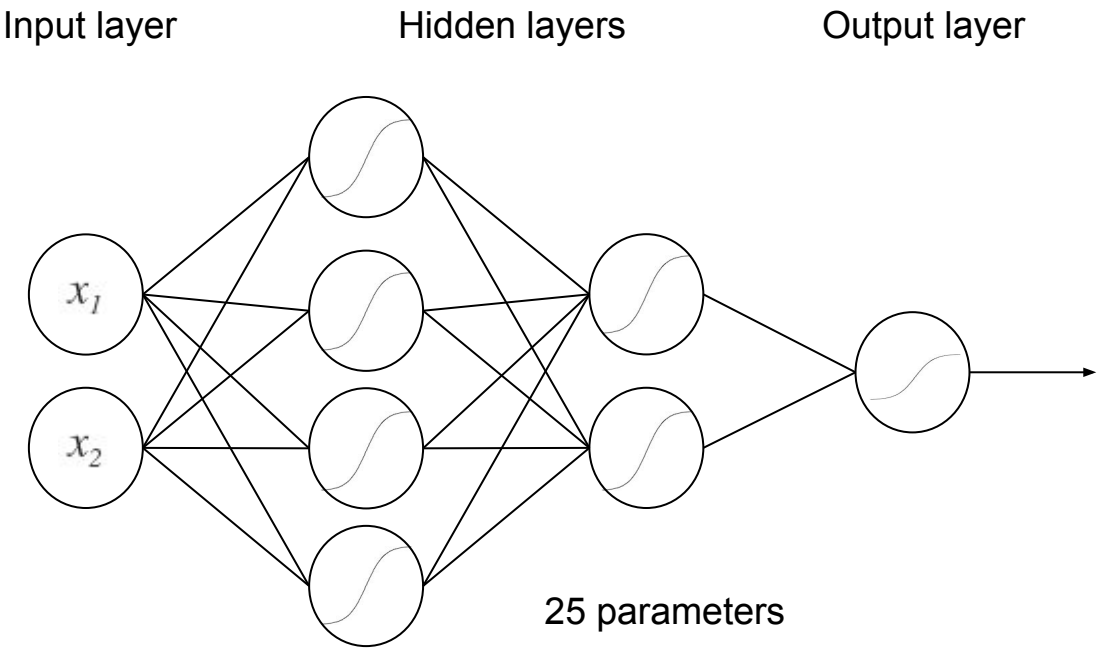
ASTRONOMY DATA AND COMPUTING SERVICES

Input layer

Output layer



Artificial Neural Networks

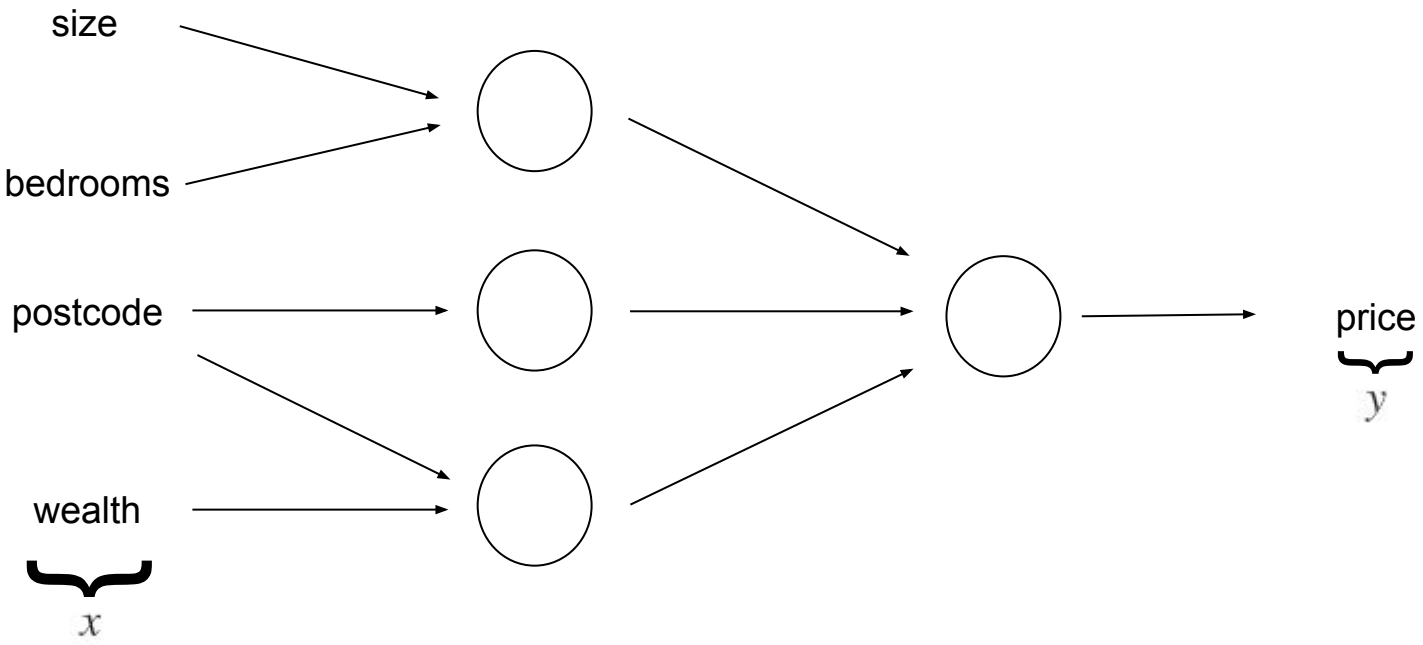


Demo: http://playground.tensorflow.org/

Artificial Neural Networks

ASTRONOMY DATA AND COMPUTING SERVICES

Housing price prediction

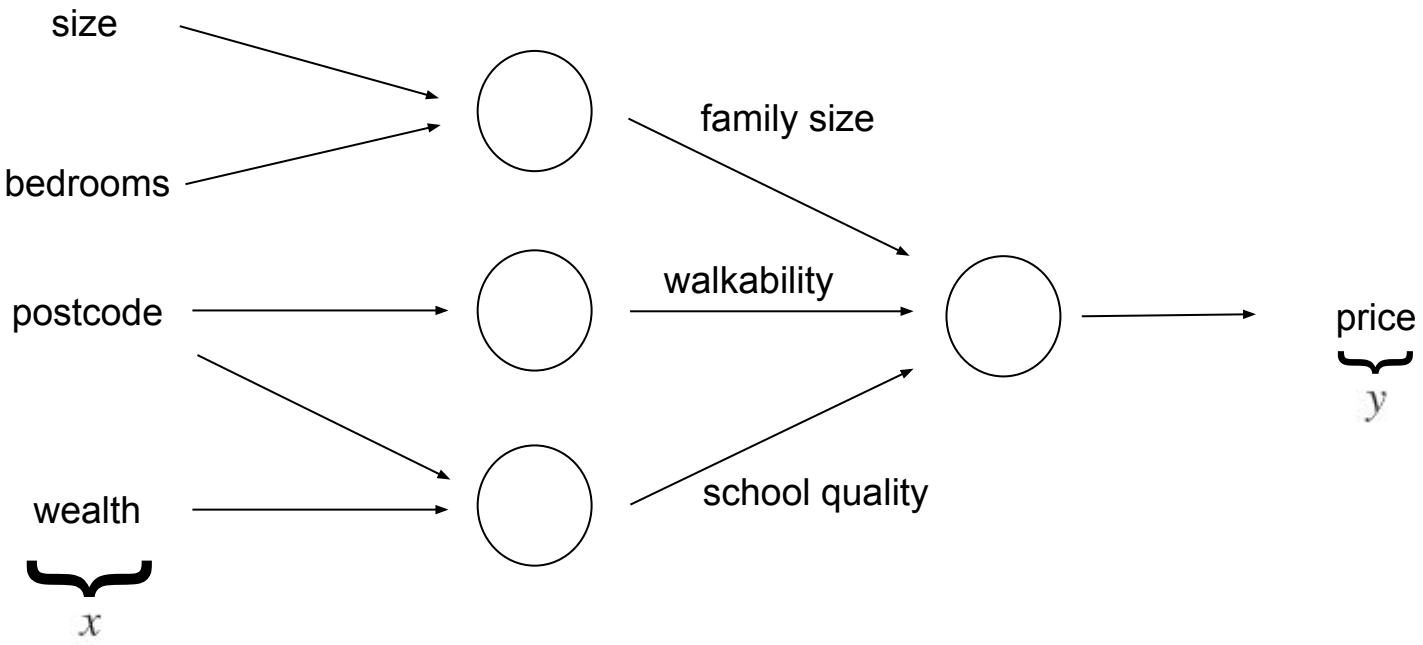


Source: Andrew Ng, Coursera - Neural Networks and Deep Learning. 2017

Artificial Neural Networks

ASTRONOMY DATA AND COMPUTING SERVICES

Housing price prediction



Source: Andrew Ng, Coursera - Neural Networks and Deep Learning. 2017