Introduction to Data Mining

Your Name

August 3, 2024

Introduction

What is Data Science

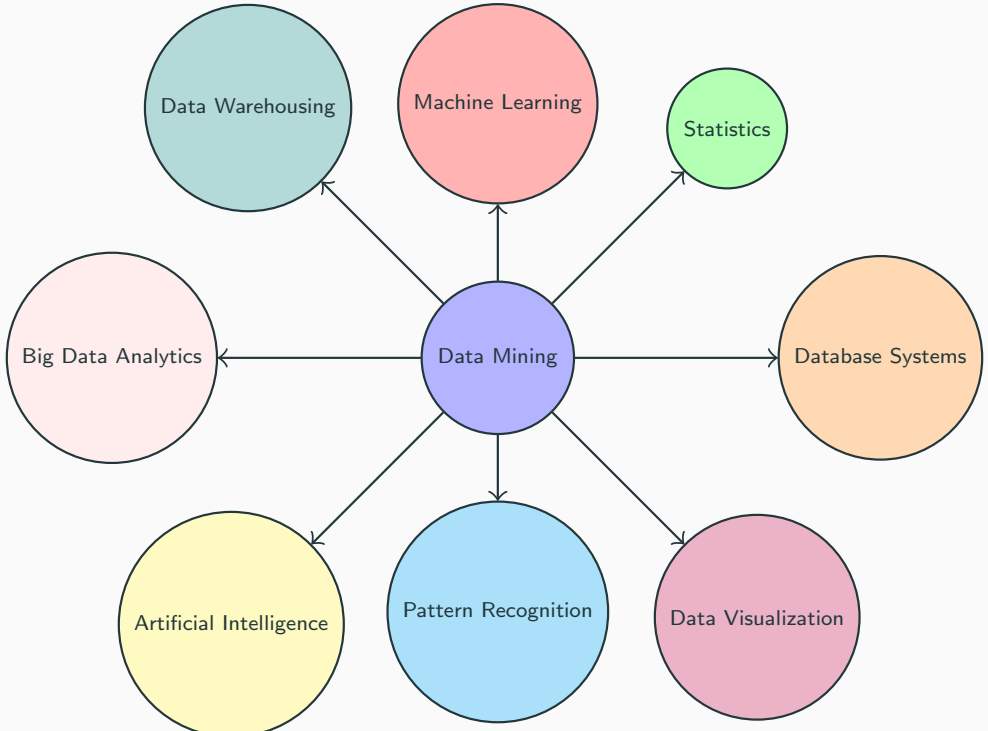
Definition

Data science is an interdisciplinary field that uses various techniques and tools to **analyze** and **interpret** complex data. It integrates principles from mathematics, statistics, computer science, and domain-specific knowledge to understand and solve real-world problems. Data science involves data cleaning, preparation, advanced modeling, and extracting insights from data to aid **decision-making** and **strategic planning**.

What is Data Science

Definition

Data science is an interdisciplinary field that uses various techniques and tools to **analyze** and **interpret** complex data. It integrates principles from mathematics, statistics, computer science, and domain-specific knowledge to understand and solve real-world problems. Data science involves data cleaning, preparation, advanced modeling, and extracting insights from data to aid **decision-making** and **strategic planning**.



What makes a DATA Scientist

- Uses their data and analytical ability to **find** and **interpret** rich data sources.

What makes a DATA Scientist

- Uses their data and analytical ability to **find** and **interpret** rich data sources.
- Manage **large amounts** of data.

What makes a DATA Scientist

- Uses their data and analytical ability to **find** and **interpret** rich data sources.
- Manage **large amounts** of data.
- Create visualizations to aid in understanding data.

What makes a DATA Scientist

- Uses their data and analytical ability to **find** and **interpret** rich data sources.
- Manage **large amounts** of data.
- Create visualizations to aid in understanding data.
- Build Mathematical models using the data.

What makes a DATA Scientist

- Uses their data and analytical ability to **find** and **interpret** rich data sources.
- Manage **large amounts** of data.
- Create visualizations to aid in understanding data.
- Build Mathematical models using the data.
- Present and communicate the data insights.

What is Data

What is Data

Child Interpretation

Outlook	Temperature	Windy	Play
sunny	hot	no	no
sunny	hot	yes	no
sunny	mild	no	yes
cloudy	hot	no	yes
rainy	mild	no	yes
rainy	cold	yes	no

Child Interpretation

Outlook	Temperature	Windy	Play
sunny	hot	no	no
sunny	hot	yes	no
sunny	mild	no	yes
cloudy	hot	no	yes
rainy	mild	no	yes
rainy	cold	yes	no

- It's sunny, mild, and windy... should I play?

Features

- Method 1 :

Outlook	Temperature	Windy	Play
1	1	0	0

Features

- Method 1 :

Outlook	Temperature	Windy	Play
1	1	0	0

$$F = (1, 1, 0, 1)$$

Features

- Method 1 :

Outlook	Temperature	Windy	Play
1	1	0	0

$$F = (1, 1, 0, 1)$$

- Method 2:

Sunny	Cloudy	Rainy	Hot	Mild	Cold	Windy	Play
1	0	0	1	0	0	0	0

$$F = (1, 0, 0, 1, 0, 0, 0, 0)$$

Measurements

	deg	feel	precip.	wsw	uv	thunder
	22	25	13	13	9	0
units	°	°	%	km/h	index	%

Table 1: Example of data as measurement

Others

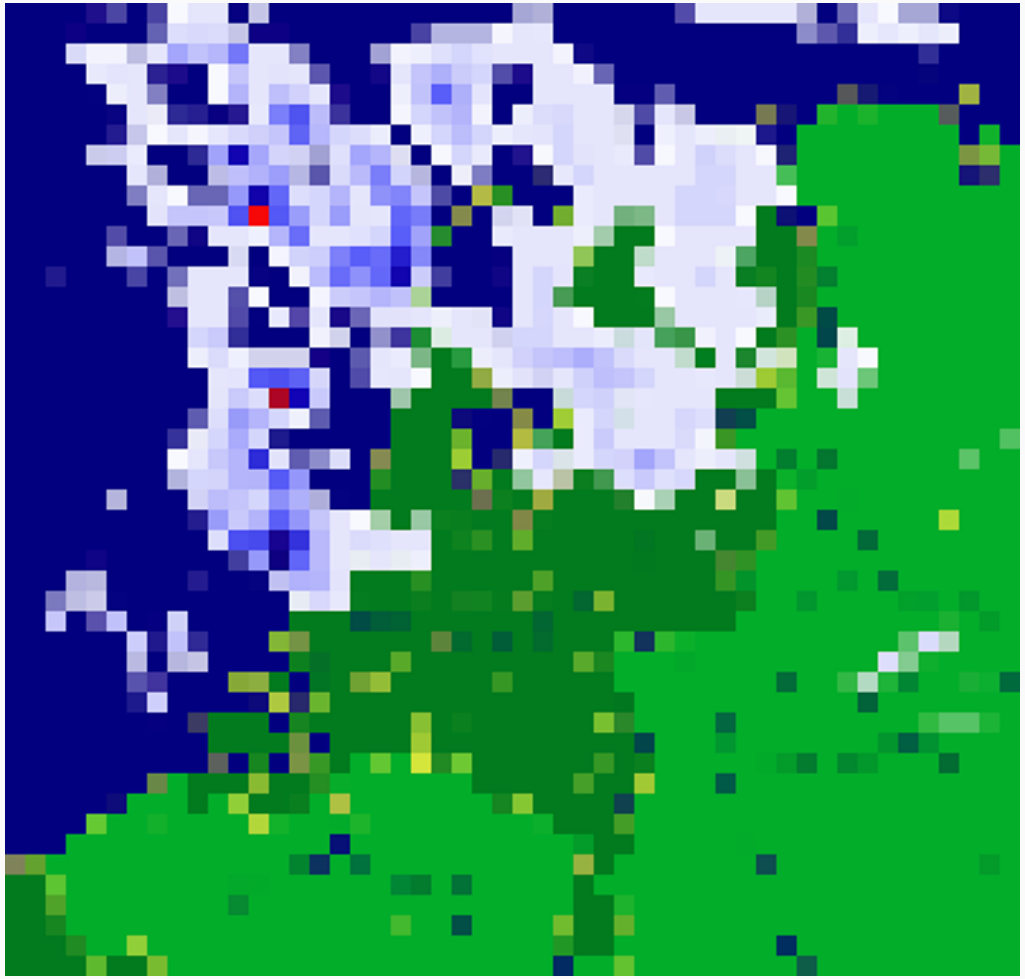


Figure 1: Weather Measurements

Interpretting Data

Back to our basic Example

Outlook	Temperature	Windy	Play
sunny	hot	no	no
sunny	hot	yes	no
sunny	mild	no	yes
cloudy	hot	no	yes
rainy	mild	no	yes
rainy	cold	yes	no

- Can we think of a set of rules to get outside and **play**?

Rules for Prediction

Objective

We want to predict our target play given the features we have available.

Rules for Prediction

Objective

We want to predict our target play given the features we have available.

- If it's **Windy** $\longrightarrow$ No play.

Rules for Prediction

Objective

We want to predict our target play given the features we have available.

- If it's **Windy** $\longrightarrow$ No play.

- If it is **hot** and **no wind** $\longrightarrow$ No play.

Rules for Prediction

Objective

We want to predict our target play given the features we have available.

- If it's **Windy** $\longrightarrow$ No play.
- If it is **hot** and **no wind** $\longrightarrow$ No play.
- If it's **not windy** and **not hot** $\longrightarrow$ Play

Formally

- We have our data $\mathbf{X}$:
  - (with features: outlook, temp and windy).
- Our data consists of smaller instances, 'some instance' is written as: $\mathbf{x}$.
- If we want to specifically point at a particular instance (say our first row), we write: $\mathbf{x}_1$.
- We can see our model as a function f, that when given any instance $\mathbf{x}$, gives us a prediction $\hat{y}$.

$$\hat{y} = f(x)$$

- The application of the model to some instance in our data can be written as $f(\mathbf{x}) = \hat{y}$.
- Our hope is that $\hat{y}$ is the same as our target: y.

Recapitulation

- Features $\mathbf{X}$:
  - (outlook, temp., windy)
- Target:
  - (play)
- Some instance: $\mathbf{x}$
- Some target: y
- First Row x_1:
  - (sunny, hot, no)
- First target:
  - (no)
- Model: if it's not windy and not hot $\rightarrow$ play ($f(\mathbf{x})$)
- Predictions by f: $\hat{y}_i$
- Prediction for $\mathbf{x}_1$: $\hat{y}_1$ (no)

Predictive Model

Model

What makes a model?

```
def play_predictor(data):
    if data['windy'] == 'no' and data['temp'] != 'hot':
        return 'play'
    else:
        return 'no play'
```

Evaluating the model

- How do we evaluate our model.

Outlook	Temperature	Windy	Play
sunny	hot	no	no
sunny	hot	yes	no
sunny	mild	no	yes
cloudy	hot	no	yes
rainy	mild	no	yes
rainy	cold	yes	no

Evaluating the model

- How do we evaluate our model.

Outlook	Temperature	Windy	Play
sunny	hot	no	no
sunny	hot	yes	no
sunny	mild	no	yes
cloudy	hot	no	yes
rainy	mild	no	yes
rainy	cold	yes	no

- We got $\frac{5}{6} = 0.83\%$

Evaluating the model

- How do we evaluate our model.

Outlook	Temperature	Windy	Play
sunny	hot	no	no
sunny	hot	yes	no
sunny	mild	no	yes
cloudy	hot	no	yes
rainy	mild	no	yes
rainy	cold	yes	no

- We got $\frac{5}{6} = 0.83\%$
- Did we cover all conditions?

Testing

- Let's consider a new data

Outlook	Temperature	Windy	Play
cloudy	hot	yes	?
rainy	mild	no	?

Testing

- Let's consider a new data

Outlook	Temperature	Windy	Play
cloudy	hot	yes	?
rainy	mild	no	?

- Actual values are
  1. Yes
  2. No.

Accuracy

Our accuracy for this test is 0%.

- Should update our model?

Realistic Use case

Predicting Housing Prices

- Would you be able to determine the price of a house?
  - **Expert Knowledge**.

Predicting Housing Prices

- Would you be able to determine the price of a house?
  - **Expert Knowledge**.

- Many observations required to gain experience.

Predicting Housing Prices

- Would you be able to determine the price of a house?
  - **Expert Knowledge**.

- Many observations required to gain experience.

- Can you come up with a few features to predict the price of a house?