DATA ANALYTICS

Statistics
AI, Machine Learning } Data Analysis
Data Base – data warehouse OLAP

01 Introduction to Statis Machine Learning

Data → Analysis → Model → Predictions

Learn from sample (NOT BIAS)

	เรียนรู้แบบมีผู้สอน Supervised Learning	จับกลุ่มตามลักษณะ Unsupervised Learning
data=>I Discrete	Categorization	Clustering
Continous data => R	regression สัมการสุดลงเชิงเส้น	dimensionality reduction การลดขนาดมิติข้อมูล

Model => Visualization

function => $f(\text{apple}) = \{\text{หวานน้ำ}_{x_1}, \text{แดง}_{x_2}\}$; x data ที่ส่งมา, y output

Lost function => ค่าสูญเสีย / ค่าผิดพลาด; ยิ่งน้อยยิ่งข้อมูลแม่น(ดี)

ex. Linear Regression plot

Machine Learning จะเรียนรู้ model ที่มาจากข้อมูลที่เราส่งมาจากการ predict ที่มีค่า Loss function น้อยๆ จะเรียนรู้ได้แม่น

02 Data Concept & Big Data

+Collect Data Object & Attributes

+ Attribute "column"

rows -> data Objects "tuple", columns -> attributes

+ Attribute Values : หน่วยที่บอกลักษณะ

& Attribute same: ความสูง เป็นฟุต/เมตรก็ได้ มีได้หลายหน่วย

different: ID ไม่มีค่า max-mi

Age มีค่า max-min

Discrete vs Continous Attribute

- Discrete เป็นจำนวนเต็ม I เป็นค่าไม่ต่อเนื่อง

  special case ▸ Binary Attribute

- Continous เป็นจำนวนจริง R realnumber

Type of Data Measurement

Qualitative (ปริมาณ)

+ Nominal "name of things" เอาไปคำนวณไม่ได้ นับได้ว่ามีเท่าไหร่ =, ≠ ex. eye_colour = {blue, black, brown}

Binary : 2 state "0 and 1" Yes/No

- Symmetric : both เท่าเทียมสำคัญเท่ากัน ex gender, hands
- Asymmetric : ความสำคัญของ 2 สิ่งไม่เท่ากัน ex. HIV positive - HIV negative

+ Ordinal พามา ranking, order ได้ =, ≠, >, < ex. Size = {S, M, L} ; S < M < L

Quantitative (คำนวณได้)

+ Interval ช่วง range "Not true zero" =, ≠, >, <, +, − ex. Year, Temperature °C °F °K

+ Ration ค่าที่เอาไปคำนวณได้จริง =, ≠, >, <, +, −, *, / ex. length, counts, weight, Temp K, Quantities

Type of Data

Record → Table, Relation, Data Matrix

Data Matrix ▸▸ มองเป็นมิติ ขนาด m x n

m row → object (tuple)

n column → attribute

Document Data ▸▸ เก็บแบบ keyword

Transaction Data ▸▸ บันทึกข้อมูลทั้งชุดรายการที่สนใจ

ex. บันทึกการ shopping ของลูกค้าว่าคนหนึ่งซื้ออะไรไปบ้าง

Record Data ▸▸ เก็บเป็นรายการโดยกำหนด attribute

- Graph & Network → world wide web, หน้าเว็บเพจ
- Ordered → sequence, time-series, ลำดับสำคัญ!

ex. ร้าน bakery เก็บข้อมูล ordered data ของช่วงเช้า, เย็น ตามTimeline (Item/Even)

Big Data 4 V

- Volumn High Volumn data เพิ่มขึ้นแบบ expo (Big Scale)
- Velocity Speed เข้าถึงข้อมูลได้รวดเร็ว ไป-กลับรวดเร็ว

  +Parallelization ทำงานแบบคู่ขนาน ตรงกับ Sequential (ตามลำดับ)
- Variety complexity ex. Relation data, Text data (web), Big Public Data many feature per item, irregular structure
- Veracity trustworthy, reliability, completeness

03 Data Analytics & Introduction to Statistics

Data + IT + Statistical Analysis + Quantitative Method + math + computer based model

⇓

managers gain improved insight make better

Scope of Business Analytics

▶ Descriptive "past & present" what has happen?

พรรณนา => เอาข้อมูลอดีต + ปัจจุบันมาวิเคราะห์ สร้างเป็น visualization

▶ Predictive "future" what could happen? จะเกิดอะไรขึ้นต่อ

พยากรณ์ => วิเคราะห์ข้อมูลในอดีต เพื่อคาดการณ์ล่วงหน้า

regression, machine learning, neural net works

▶ Presriptive "outcome" what should we do? เราควรทำอะไร

ให้เกิดประโยชน์สูงสุด => วิเคราะห์หาวิธีการทำให้เกิดประโยชน์สูงสุด

ทำลอง Algorithms on possible

Introduction to Statistics : Sample สุ่มมาแบบ NOT BIAS (SAVE TIME & MONEY)

2 Type → Descriptive : บอกแนวโน้มดี-ไม่ดี

- Mode ฐานนิยม
- Median ค่ากลาง มัธยฐาน
- Mean ค่าเฉลี่ย $\frac{\Sigma x}{n}$
- SD ส่วนเบี่ยงเบนมาตรฐาน $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

  SD ยิ่งมาก ค่าการกระจายมาก

  SD ยิ่งน้อย ค่าการกระจายน้อย

→ Inferential : พยากรณ์, คาดการณ์ คำนวณจาก sample

05 Exploration Data Analysis & Data Preprocessing

GET DATA → EDA → Preprocessing → MODEL

"EDA" การวิเคราะห์การสำรวจข้อมูล

WELL → Accuracy ความแม่นยำ
Completeness ความสมบูรณ์
Cosistensy ความสอดคล้อง
Timeless ช่วงเวลา
Believability ความน่าเชื่อถือ
Value added คุณค่า ประโยชน์
Interpretability ข้อมูลแปลผลได้
Accessibility การเข้าถึง

Problem Dirdy data → Noisy & Outlier : Salary="-10", Voice flie
→ Missing Value : incomplete "", n/a
→ Duplicate Data : 1คน หลายemail
→ Inconsistent Data : data ต่าง database
Age, Birth
Tel → String, Int

Preparation → cleaning : fill, identify, remove outlier
→ integration: เอามาเก็บในที่เดียวกัน
→ reduction & feature selection regression
→ discretization : แปลงทศนิยมเป็นจำนวนเต็ม (continuous → discrete)
→ transformation : Z-score $z = \frac{x-\mu}{\sigma}$
ขัดช่องข้อมูล

Type of Data Format

- CSV : Comma Separated Value
- Excel
- JSON : JavaScript Object Notation ~ Dictionary {key : Value}
- SQL : Structured Query Language
- And More! flie image, voice

PYTHON

- numpy --> import numpy as np
- Scipy --> import scipy as sp
- matplotlib --> import matplolib.pyplot as plt
- seaborn --> import seaborn as sns
- pandas --> import pandas as pd

```
dataframe = pd.read_csv('filename.csv')
dataframe.columns
```

out: Index([...,...,...], dtype='object')

```
dataframe.shape
```

out: (2018, 18)
row column

06 Data Visualization

เห็นภาพ + understanding

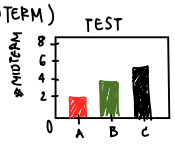
Table	vs.	Graph
✓specific values ค่าเฉพาะ		✓shape, bar, line
✓มีความแม่นยำ precise		✓relationship multiple value ความสัมพันธ์ของหลายๆข้อมูล
✓compare related values ต้องการเปรียบเทียบค่า		✓show trends
✓different units of measure เปรียบเทียบข้อมูลที่มีหลายหน่วย		✓large date sets

Bar Graph การกำหนด scale !

แกน x ข้อมูลที่ซ้ำกันเยอะๆ
แกน y ข้อมูลที่แตกต่างกันมากๆ

seaborn : sns.displot(df['alcohol'])
pandas : df['alcohol'].plot.hist()

```
movies = ["A", "B", "C"]
num = [2, 4, 6]
xs = range(len(movies))
plt.bar(xs, num, color = ('r','g','b'))
plt.xticket(xs, movies)
plt.ylabel("# MIDTERM")
plt.title("TEST")
plt.show()
```



Pie Chart 100% .

```
restyarant_salad = df.groupby(['restyarant'])['salad'].count()
plt.pie(restyarant_salad, labels = restyarant_salad.index)
df.groupby(['rank'])['salary'].count().plot(kind = 'pie')
```

bar, line, ...

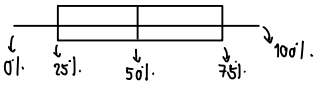
Scatter Plot

```
sns.regplot(x = 'sodium', y = 'protein', data = df, color = 'pink')
plt.scatter(df['service'], df['salary'], color = 'pink')
```

Bubble Plot ต้องใช้ marker

Box Plot

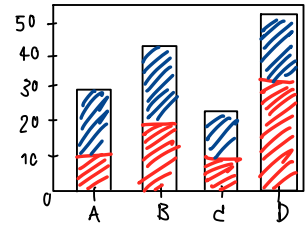
Quartiles → 0% min - 25% - 50% median - 75% - 100% max
Range + Median + Percentags ตัด outlier ออก




```
sns.boxplot(x = 'protein', y = 'restyarant', data = df)
```

Stacked Charts

```
x = ['A','B','C','D']
y1 = [10,20,10,30]
y2 = [20,25,15,25]
plt.bar(x, y1, color = 'r')
plt.bar(x, y2, bottom = y1, color = 'b')
plt.show()
```



Line Charts

```
df.groupby(['rank'])['salary'].count().plot(kind = 'line')
```

Note ✓.matplotlip inline ช่วยขึ้น scal