Agenda
- Summary
- Overview
- Applications
- extra topics

---

1. DSML libraries
   - Data – Numpy, pandas
   - Plotting – matplotlib & seaborn
2. Prob and stats
3. Co-ordinate geometry ← linear algebra
4. Calculus & optimization
   - GD

Plotting

- 1V
  - Numerical – hist, KDE, boxplot
  - Categorical – countplot, pies, bar charts
- 2V
  - Num – Num – scatter, hating, line plot, hist, bike
  - Cat – cat – heat, stacked, count plot
  - Num – cat – bar – dodger, violin, dodger, box, dodger

Health
→ diabetic (Y) → two → scatter
→ X → weight, height
Num – Num – Cat

Prob / stats

at population level $P(\text{diabetic}) = 0.2$

$w = 120$ kg.

$P(\text{diabetic} \mid w = 120] = 0.6$

$h = 7.5$ ft

$P(\text{diabetic} \mid w = 120 \text{ & } h = 7.5\text{ft}) = 0.15$

10,000 data points

$x_q \Rightarrow w_q, h_q$

$P(\text{diabetic} \mid w = w_q, h = h_q)$

(present in data)
→ Prob by count (Data analyst)
→ not present in data.
data science

Linear algebra

$y = w^T x + w_0$

$w^T x_q + w_0$

x_q

DS vs SDE

Email → Spam or Not spam

SDE
↳ rules → congratulations, click here, having discount → spam

New email → spam / not spam.
are there words present

DS

Emails → labelled spam or not
↓ X, Y
$f(x)$
model
hypothesis function

SDE

Input → SDE → Decision
patterns / rules

Data, Labels → DS (ML algo) → model (input) → Decision.

Tom Mitchell

Task (T) ← Experience (E)
↓
Performance (P)

Task

- Classification
  ↳ Y (target) will be part of some class
  eg – email spam or not.
  digit written or not
  cancer terminal or not
  image house, dog or cat
- Regression.
  ↳ Y (target) will be continuous
  eg → stock prices, house prices
- Clustering
  ↳ group similar samples.
  eg customer segmentation based on past purchase behaviour.

- Recommendation
  eg Netflix
- Forecasting.
  eg – weather tomorrow.

---

learning method

Supervised learning (label in training data)

→ spam or not (classification) → 2 classes (binary)
↳ 2+ classes (multi-)
→ price (regression)

Classification

$D \{(x_i, y_i)\}$ $x_i \in R^d$, $y_i \in \{0, 1\}$ (target)

Regression

$D \{(x_i, y_i)\}$ $x_i \in R^d$, $y_i \in R$ (target)

- Unsupervised
  ↳ clustering

- Reinforcement learning