Reading 4

☆ Article: Bayes Rule

- Baye's Rule is arguably the most important rule in data science.
  - ↳ It is the mathematical rule that describes how to update a belief, given some evidence
  - ↳ It describes the act of learning.

Equation:

$$\underbrace{P(A|B)}_{\text{posterior}} = \underbrace{P(A)}_{\text{prior}} \times \frac{\overbrace{P(B|A)}^{\text{likelihood}}}{\underbrace{P(B)}_{\text{marginal}}}$$

- posterior probability: updated probability after the evidence is considered
- prior probability: the probability before the evidence is considered
- liklihood: probability of the evidence, given the belief is true
- marginal probability: probability of the evidence under any circumstance

Conditional Probability

- $P(A|B)$ → "probability of A given B"
  - ↳ $P(A|B) \neq P(B|A)$

Bayes' Rule in detail:

- helpful to think in terms of 2 events:
  - ↳ a hypothesis (which can be true or false)
  - ↳ evidence: (which can be present or absent)

$$P(Hypothesis \mid Evidence) = P(Hypothesis) \times \frac{P(Evidence \mid Hypothesis)}{P(Evidence)}$$

- It can be applied to any type of events, with any number of discrete or continuous outcomes.
- Think of $\frac{P(Evidence \mid Hypothesis)}{P(Evidence)}$ as the "strength" of the evidence.

Worked Out Example

- Your neighbor is watching their favorite soccer team. You hear them cheering, and want to estimate the probability their team has scored.
  - ↳ Step 1: Write down the posterior probability of a goal, given cheering.
  - ↳ Step 2: estimate the prior probability of a goal as 2%.
  - ↳ Step 3: estimate the likelihood probability of cheering, given there's a goal as 90%.
  - ↳ Step 4: estimate the marginal probability of cheering, this could be because:
    - ↳ a goal has been scored (2% of the time, times 90% probability)
    - ↳ or any other reason, such as the other team missing a penalty or having a player sent off (98% of the time, times perhaps 1% probability)
- Now piece everything together:

$$P(Goal \mid Cheer) = P(Goal) \times \frac{P(Cheer \mid Goal)}{P(Cheer \mid Goal) + P(Cheer \mid No\ goal)}$$

$$= 0.02 \times \frac{0.9}{(0.02 \times 0.9) + (0.98 \times 0.01)}$$

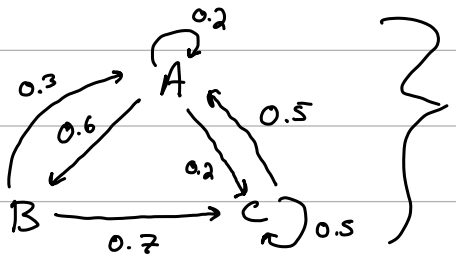
$$= 64.7\%$$

- Use cases for Bayes's Rule:
  - ↳ Understanding probability problems (including those in medical research)
  - ↳ Statistical modelling and inference
  - ↳ Machine learning algorithms (such as Naive Bayes, Expectation Maximisation)
  - ↳ Quantative modelling and forecasting

☆ Video 1 Markov Chain Clearly Explained

Ex A restaurant serves only 3 foods, but on any given day they only serve 1 item, and it depends on what they served yesterday.

Foods: {A, B, C}



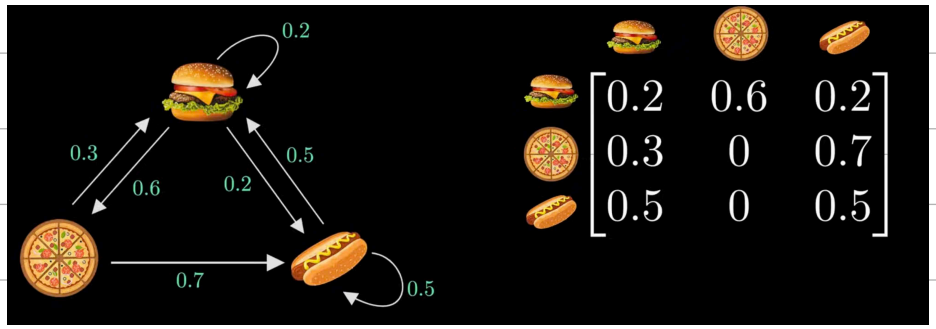
This is a Markov Chain (each number is a probability)

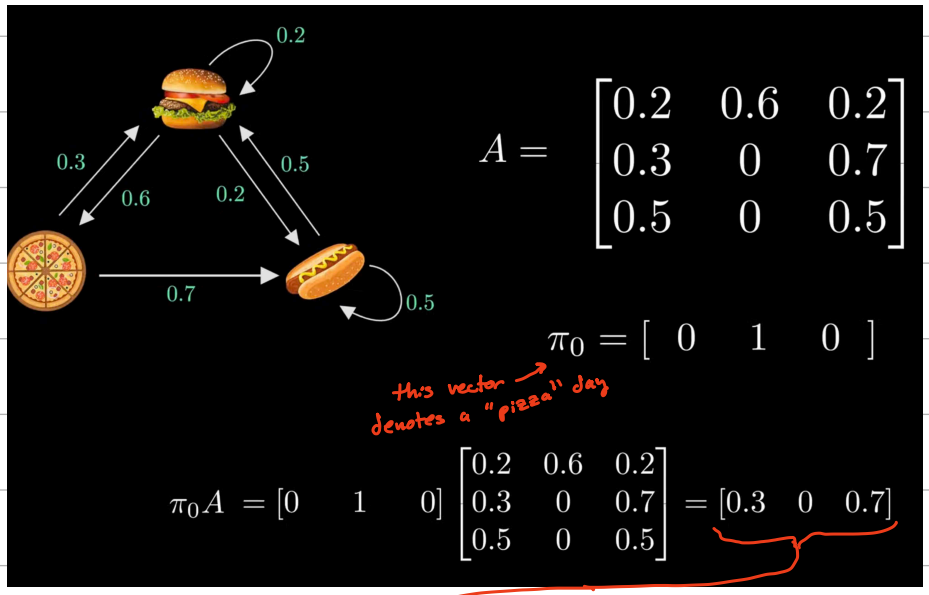
Properties of Markov Chain:

1. The future state only depends on the current state (Markov Property)
   - ↳ $P(X_{n+1} = x \mid X_n = x_n)$
2. The sum of the weights of the outgoing edges on any state is equal to 1.
   - ↳ This is because they are probabilities.

- Stationary Distribution or Equilibrium State is the probability distribution of the states after a walk of ∞ steps.

- You can use an adjancey matrix to show the relationship in the graph.
  - ↳ This is also a transition matrix:





↳ future probabilities corresponding to current pizza state → Now this vector is your new π_0 vector to multiply with A. → Keep repeating

- If at any point in this process you find $\pi_k A = \pi_k$ then you have found the stationary distribution.
  - ↳ Notice this is analgous to eigenvectors! $\pi_k A = \pi_k A$, $A\vec{v} = \lambda\vec{v}$

- This eigenvector's components must add up to one since it denotes a probability distribution.

- Any other eigenvectors with eigenvalue $\lambda = 1$, are also stationary states.
  - ↳ usually all eigenvectors with eigenvalue $\lambda = 1$ are very similar and justify each other.
  - ↳ The e-vec $[x_1\ x_2\ x_3]$ tells you the probability of hamburger on any given day is x_1, etc.

☆ Video: Hidden Markov Model

- There are some hidden states that we don't know about, but those hidden states directly affect some observed states that we do know about.

- Example: Suppose your professor is either:
  - happy or sad (hidden states)
  - wears red, blue, or green (observed states)

- This is how they are related!

or

Transitions

(t) ← current day

previous day (t-1)	H	S
H	0.7	0.3
S	0.5	0.5

Emissions

	R	G	B
H	0.8	0.1	0.1
S	0.2	0.3	0.5

You Observe!

Professor wore Green → Blue → Red

- You want to infer the sequence of moods on these 3 days:

↳ So you need to maximise the probability function:

$$\underset{m_1, m_2, m_3}{MAX} \; P\left(\begin{matrix} C_1 = G, & C_2 = B, & C_3 = R \\ M_1 = m_1, & M_2 = m_2, & M_3 = m_3 \end{matrix}\right)$$

color on 1st day

mood on day 1

- Using probability that function is maximised when the following product is maximised:

$$P[C_3 \mid C_2, C_1, m_3, m_2, m_1] \times$$
$$P[C_2 \mid C_1, m_3, m_2, m_1) \times$$
$$P[C_1 \mid m_3, m_2, m_1] \times$$
$$P[m_3 \mid m_2, m_1] \times$$
$$P[m_2 \mid m_1] \times$$
$$P[m_1]$$

but via Markov Assumption (mood on any given day only is affected by mood on previous day, and color on given day only affected by mood that day) we can simplify to the following product:

$$P(C_3 \mid m_3) \times$$
$$P(C_2 \mid m_2) \times$$
$$P(C_1 \mid m_1) \times$$
$$P(m_3 \mid m_2) \times$$
$$P(m_2 \mid m_1) \times$$
$$P(m_1 \mid S) \times$$

"start"

- We have all these probabilities in our graph and matrices at the start and can compute this product is largest when moods are $\{S, S, H\}$

So these are our most probable moods.

- Real world applications:

NLP:

↳ Words are observed

↳ part-of-speech are hidden states