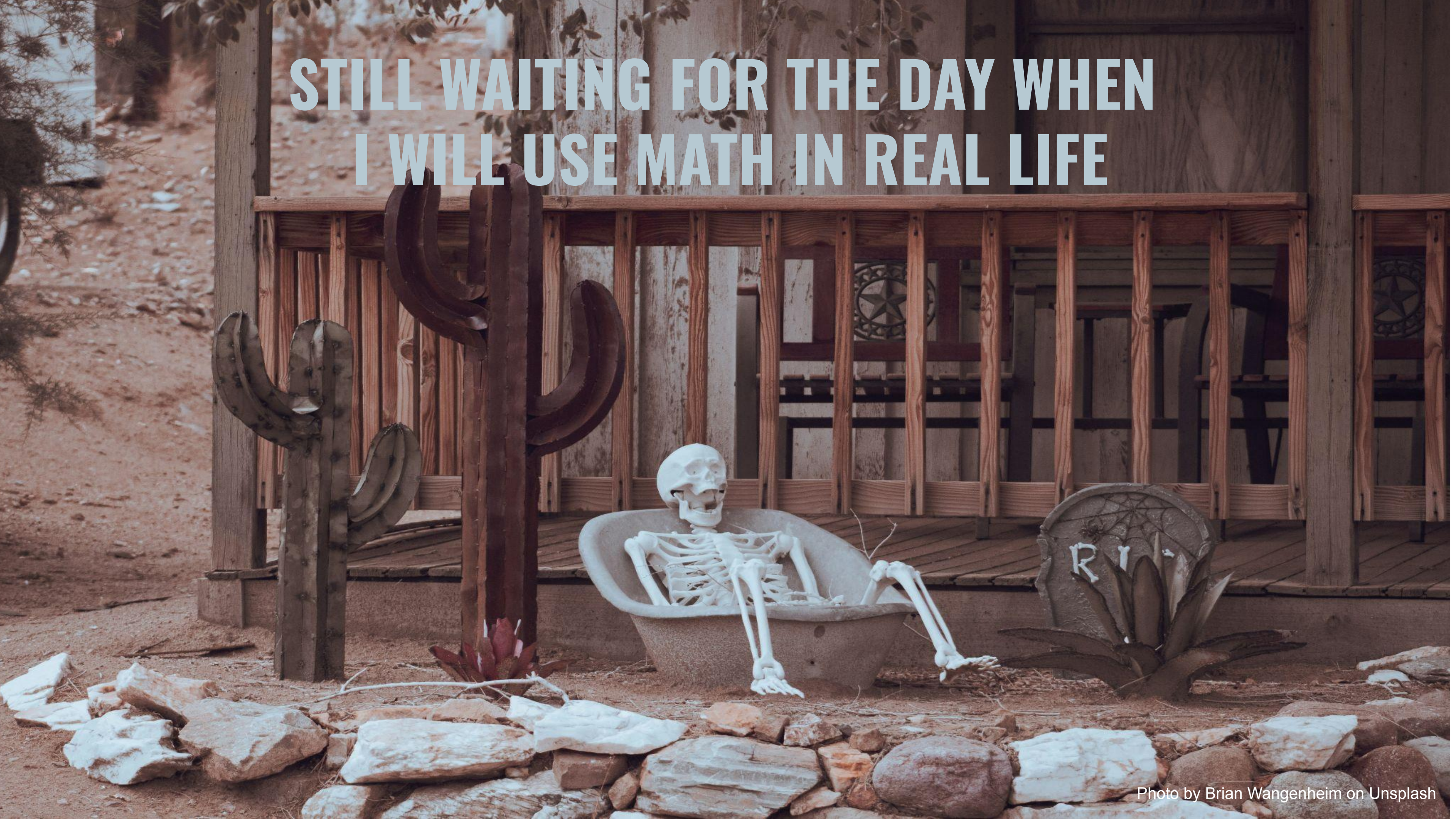
Mechanistic Interpretability 101

Episode 2

Probability, Statistics & Information Theory

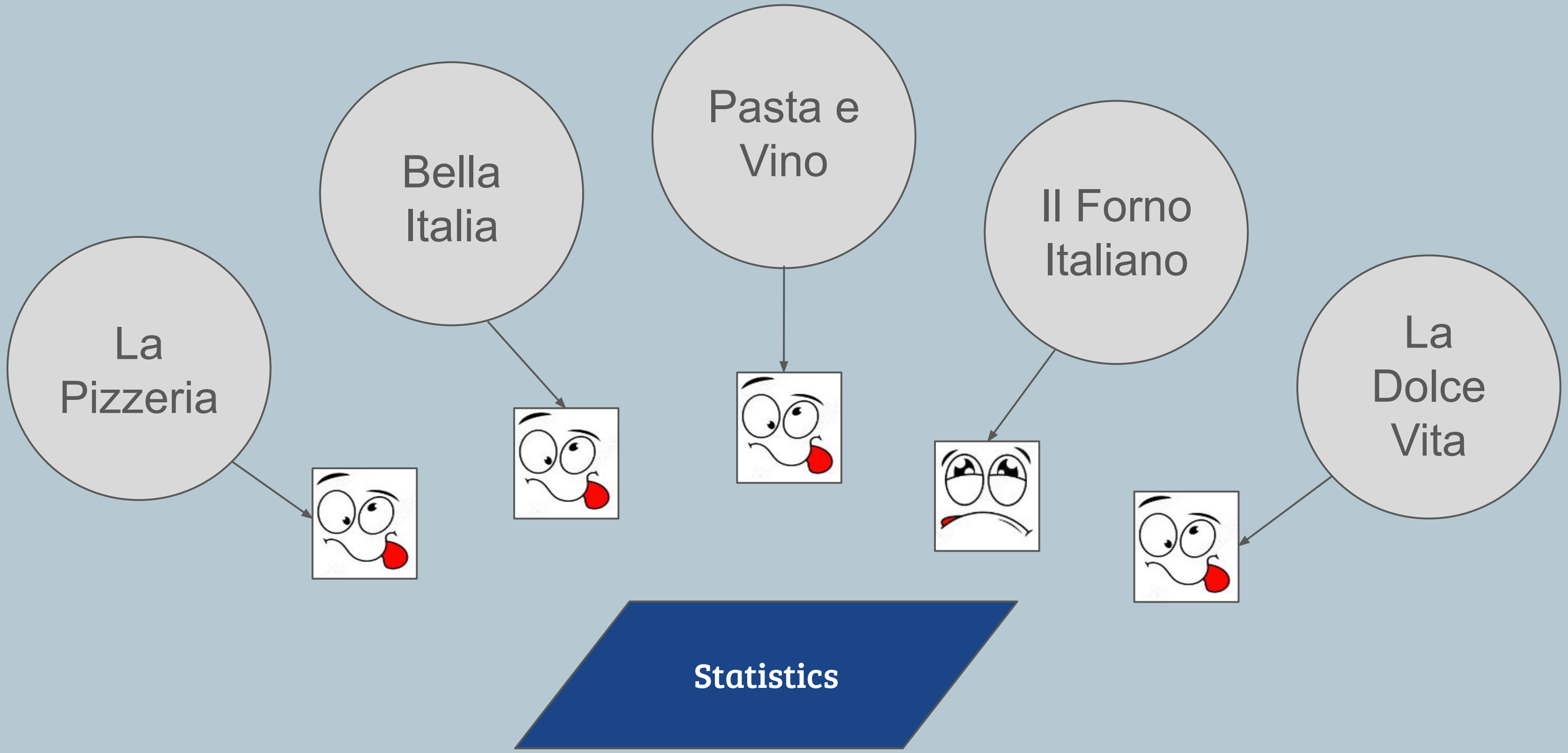

STILL WAITING FOR THE DAY WHEN
I WILL USE MATH IN REAL LIFE
Photo by Brian Wangenheim on Unsplash


VIA 9GAG.COM
JOEY DOESN'T SHARE FOOD!!

How do you select a Restaurant?

PROBABILITY

Friend's Recommendations



Option A

Option B

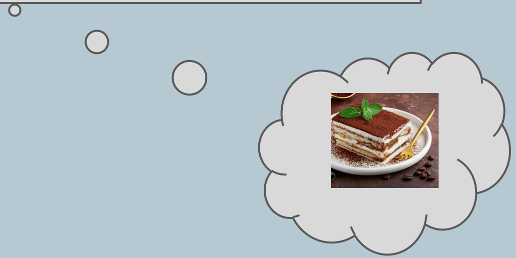
Which
Is
the
one?

What is friend's opinion?

What are the reviews saying?

Which place is closer?

Do they have Tiramisu?



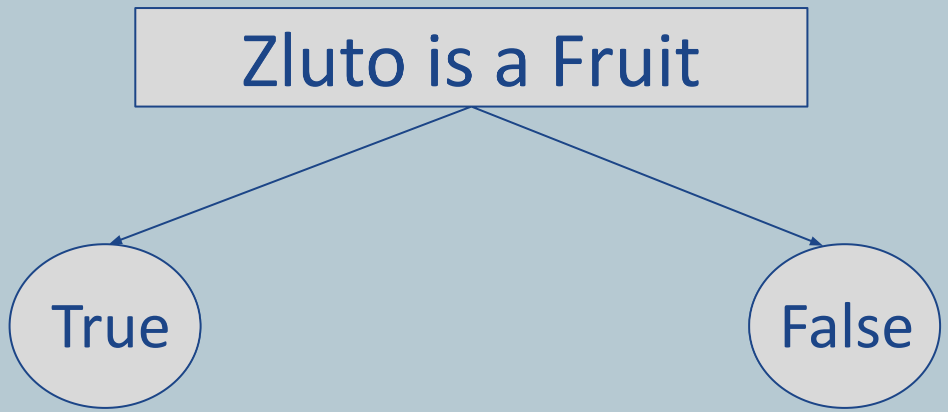
Information Theory

DECISIONS,

DECISIONS,

DECISIONS...

Probability



$$P(\text{True}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} = \frac{1}{2}$$

$$P(\text{False}) = \frac{1}{2}$$

Rules of Probability

The Rule of Non-Negativity:

$0 \leq P(\text{True}) \leq 1$ $0 \leq P(\text{False}) \leq 1$

The Rule of Total Probability:

$P(\text{True}) + P(\text{False}) = 1$

The Complementary Rule:

$P(\text{False}) = 1 - P(\text{True})$

The Addition Rule for Disjoint Events:

$P(\text{True or False}) = P(\text{True}) + P(\text{False})$

The Multiplication Rule for Independent Events:

$P(\text{Zluto is True and Lumo is True}) = P(\text{Zluto is True}) \times P(\text{Lumo is True})$

The Conditional Probability Rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The Law of Total Probability:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n)$$

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Mutually exclusive events:

$$P(A \cap B) = P(A) \times P(B)$$

Not (necessarily) mutually exclusive events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probability Distributions

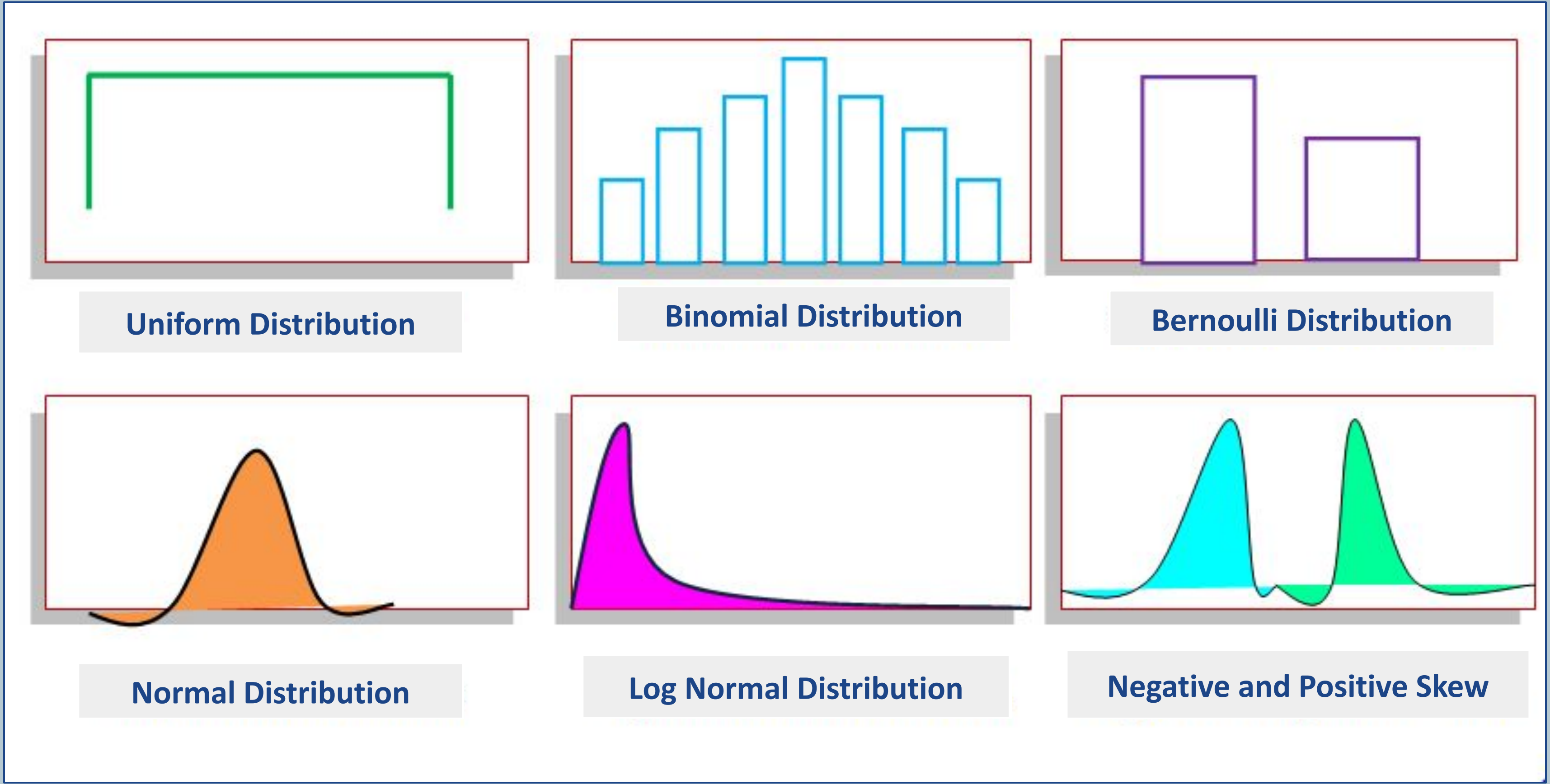


Image Source: Types of distributions in statistics blog by Gladys Kiruba

Bayesian Inference

Bayes' Theorem:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

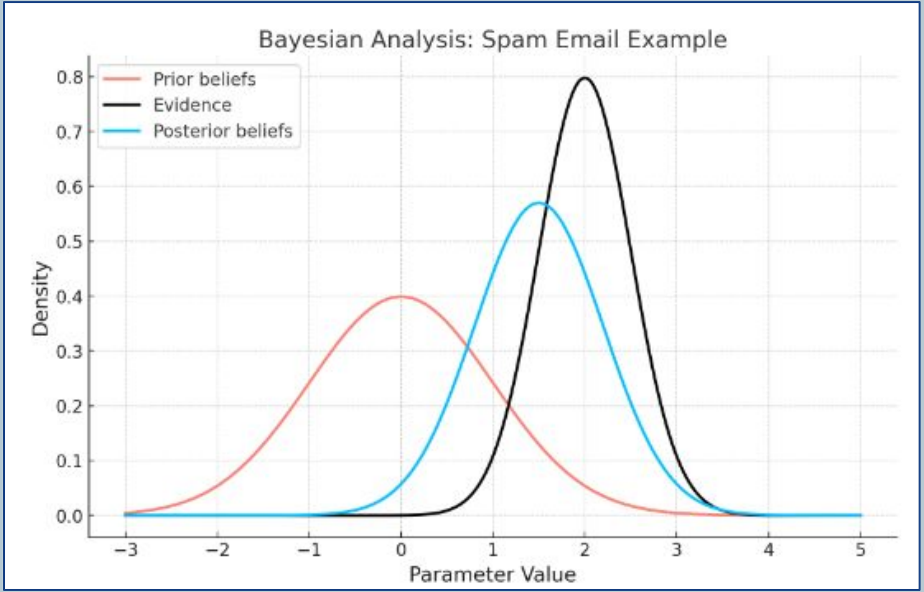
Hypothesis 1 (H1): Email is spam
Hypothesis 2 (H2): Email is not spam

P(H1): Prior Probability of email being a spam is 20%
P(H2): Prior Probability of email not being a spam is 80%

Likelihood P(D|H1): Probability that the word "lottery" appears in a spam email is 70% P(D|H1)= 0.7
Likelihood P(D|H2): Probability that the word "lottery" appears in a non-spam email is 5% P(D|H2)= 0.05

P(D) = P(D|H1).P(H1) + P(D|H2).P(H2) = (0.7x0.2)+(0.05x0.8) = 0.18

P(H1|D) = P(D|H1).P(H1) / P(D) = (0.7x0.2) / 0.18 = 0.77



Markov Chain

Markov Property:

$$P(X_{n+1} = x | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x | X_n = x_n)$$

Transition Matrix:

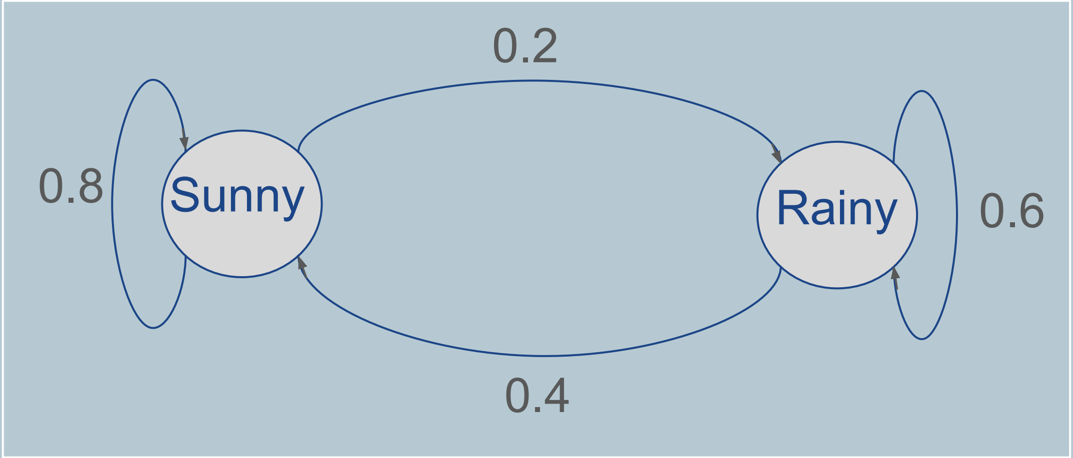
$$P = \begin{pmatrix} P(S_1 \to S_1) & P(S_1 \to S_2) & \ldots & P(S_1 \to S_k) \\ P(S_2 \to S_1) & P(S_2 \to S_2) & \ldots & P(S_2 \to S_k) \\ \vdots & \vdots & \ddots & \vdots \\ P(S_k \to S_1) & P(S_k \to S_2) & \ldots & P(S_k \to S_k) \end{pmatrix}$$

Initial State vector $v_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Transition Matrix $P = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$

Weather for tomorrow $v_1 = v_0 \times P = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix}$

Weather for Day after tomorrow $v_2 = v_1 \times P = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix} \times \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.72 \\ 0.28 \end{pmatrix}$



Statistics

Expected Value (Mean):

$$E(X) = \sum_{x_i} x_i P(X = x_i),\ \ When\ X\ is\ a\ discrete\ random\ variable$$

$$E(g(X)) = \sum_{x_i} g(x_i) P(X = x_i),\ \ (g\ is\ an\ arbitrary\ function)$$

$$E(X) = \int_a^b x f_X(x) dx \quad (a \leq X \leq b)\ ,When\ X\ is\ a\ continuous\ random\ variable$$

Variance:

$$Var(X) = E[(X - \mu)^2],\ \ X\ is\ a\ random\ variable$$

Standard Deviation:

$$\sigma(X) = \sqrt{Var(X)}$$

Expectation, Variance & Standard Deviation

Expected Value (Mean):

$$E[\hat{Y}] = \frac{1}{5} \times (300,000 + 320,000 + 310,000 + 330,000 + 315,000) = 315,000$$

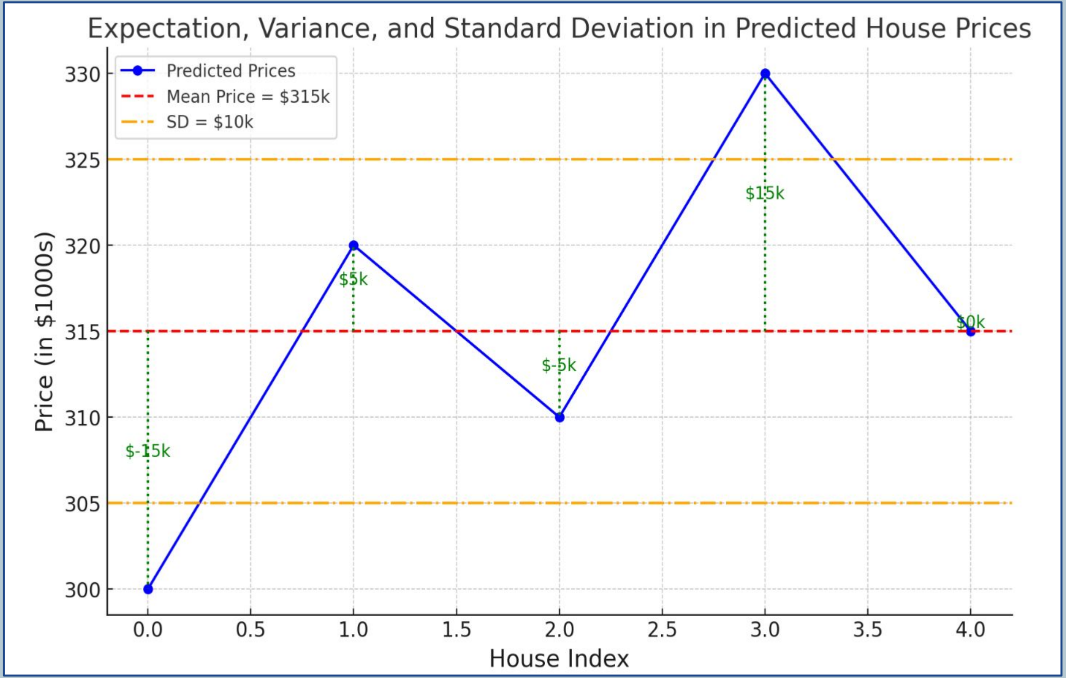
Variance:

$$\mathrm{Var}(\hat{Y}) = \frac{1}{5} \times (225,000,000 + 25,000,000 + 25,000,000 + 225,000,000 + 0)$$

$$\mathrm{Variance} = \frac{1}{5} \times 500,000,000 = 100,000,000$$

Standard Deviation:

$$\mathrm{SD}(\hat{Y}) = \sqrt{\mathrm{Var}(\hat{Y})} = \sqrt{100,000,000} = 10,000$$



Information Theory

Chance of picking a Yellow ball from Basket A,
X = 0.1

Given $P(X) = 0.1$, the entropy $H(X)$ is:

$$H(X) = -0.1 \cdot \log_2(0.1)$$

$$H(X) = 0.332 \text{ bits}$$

Chance of picking a Blue ball from Basket A ,
Y = 0.9

Given $P(Y) = 0.9$, the entropy $H(Y)$ is:

$$H(Y) = -0.9 \cdot \log_2(0.9)$$

$$H(Y) = 0.137 \text{ bits}$$

Total Entropy of the system:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

$$H(\text{System}) = H(X) + H(Y)$$

$$H(\text{System}) = 0.332 + 0.137 = 0.469 \text{ bits}$$

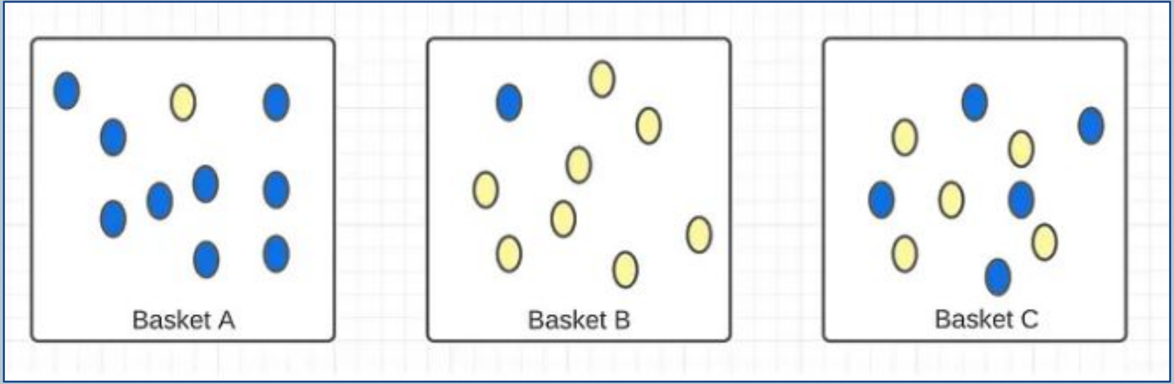
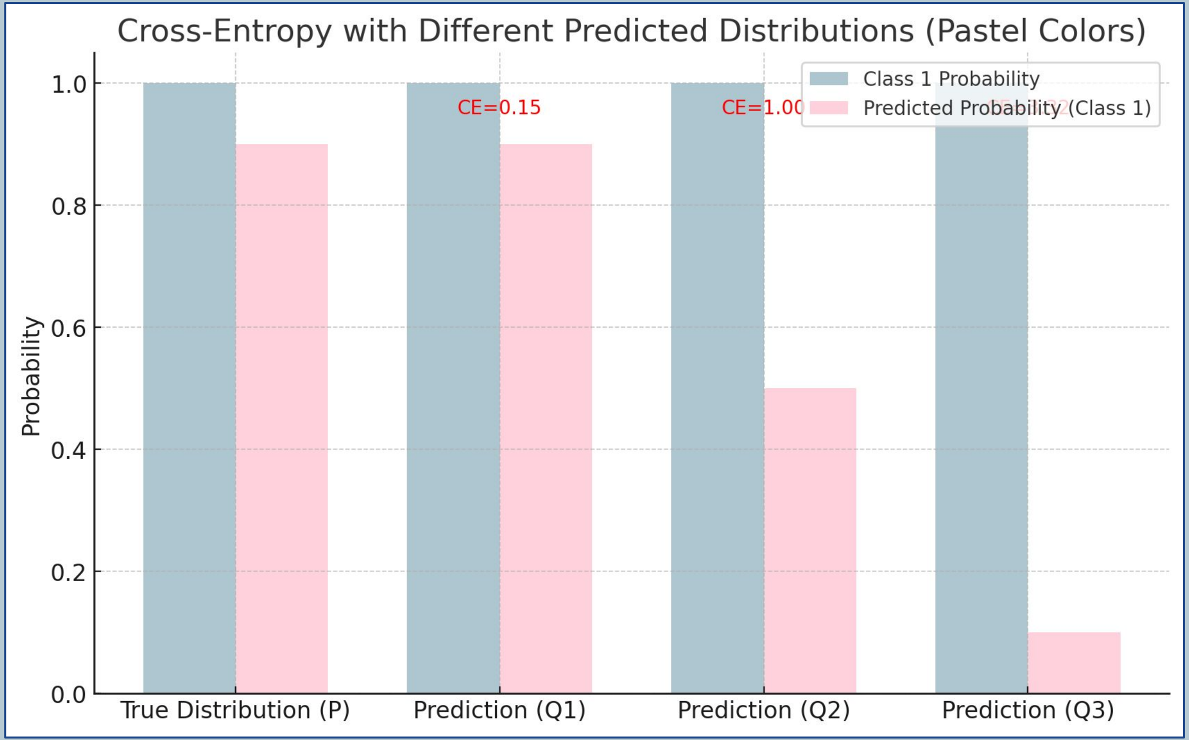


Image Source: Entropy; A method for Data Science & Machine Learning by Goku Adekunle

Information Theory

Cross Entropy:

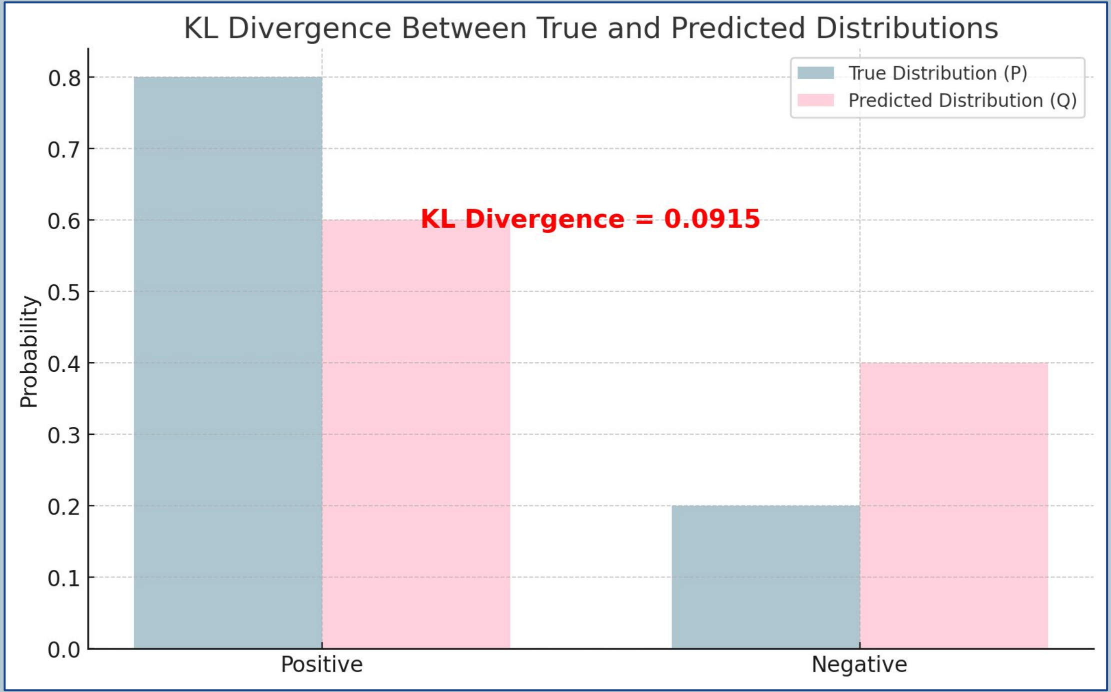
$$H(P,Q) = -\sum_i P(x_i)\log(Q(x_i))$$



Information Theory

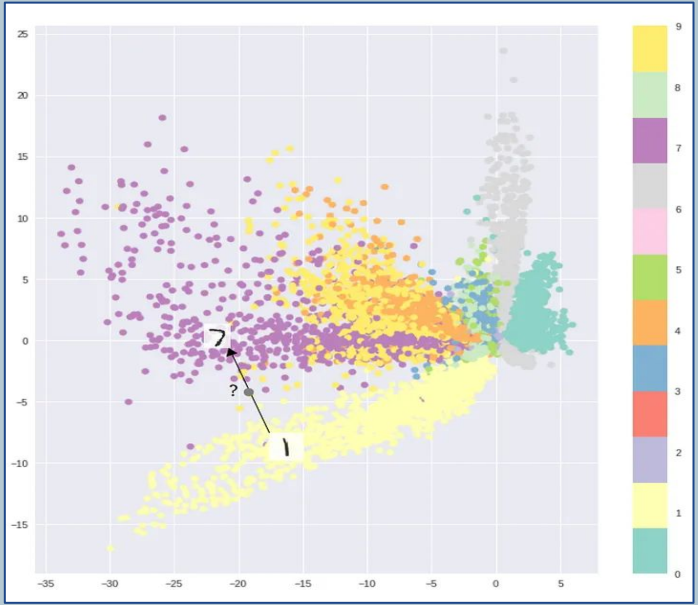
Kullback-Leibler Divergence:

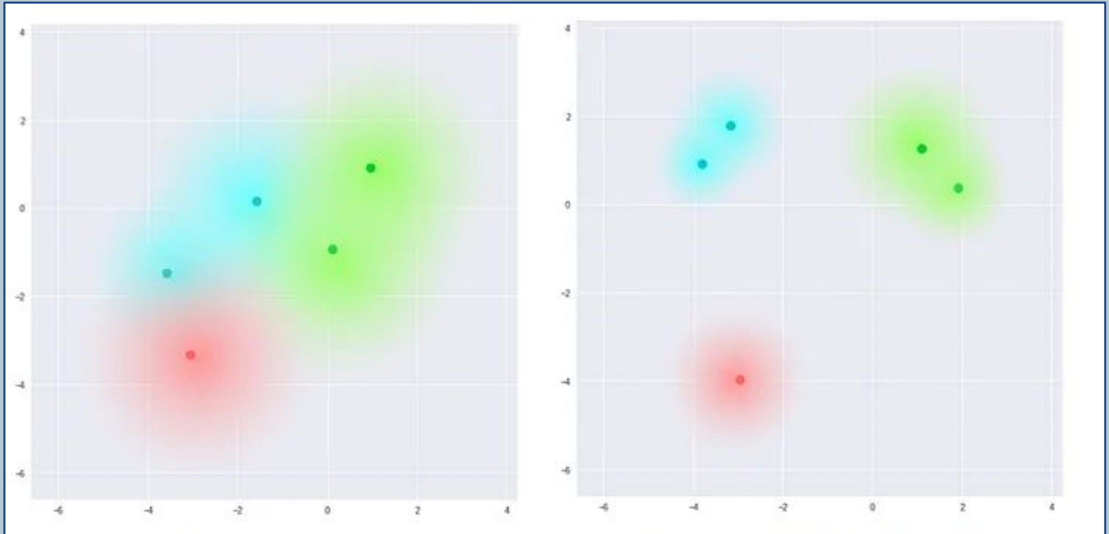
$$D_{KL}(P\|Q) = \sum_i P(x_i) \log\left(\frac{P(x_i)}{Q(x_i)}\right)$$



Information Theory

KL Divergence





What we require

What we may inadvertently end up with

Topics covered in this session

- Probability
- Rules of Probability
- Probability distributions and their applications
- Bayesian Inference
- Markov Chain
- Statistics: Mean, Variance and Standard Deviation
- Entropy
- Cross Entropy
- KL Divergence