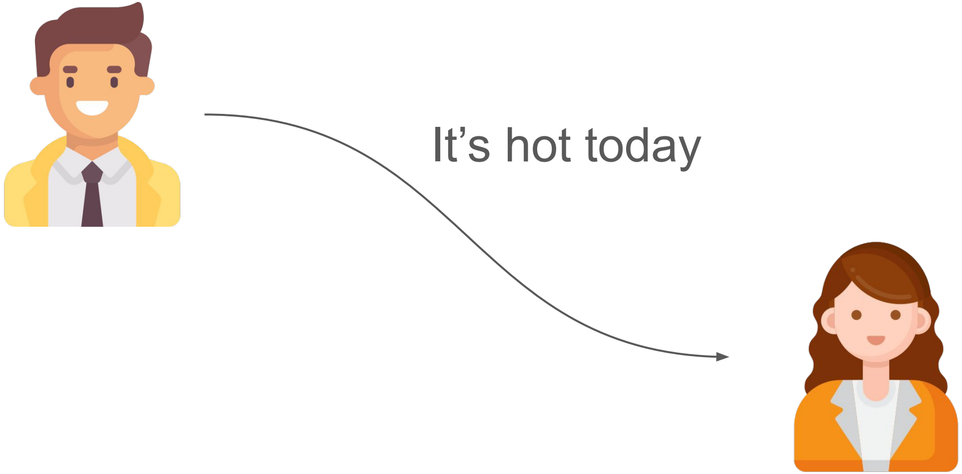
KL Divergence

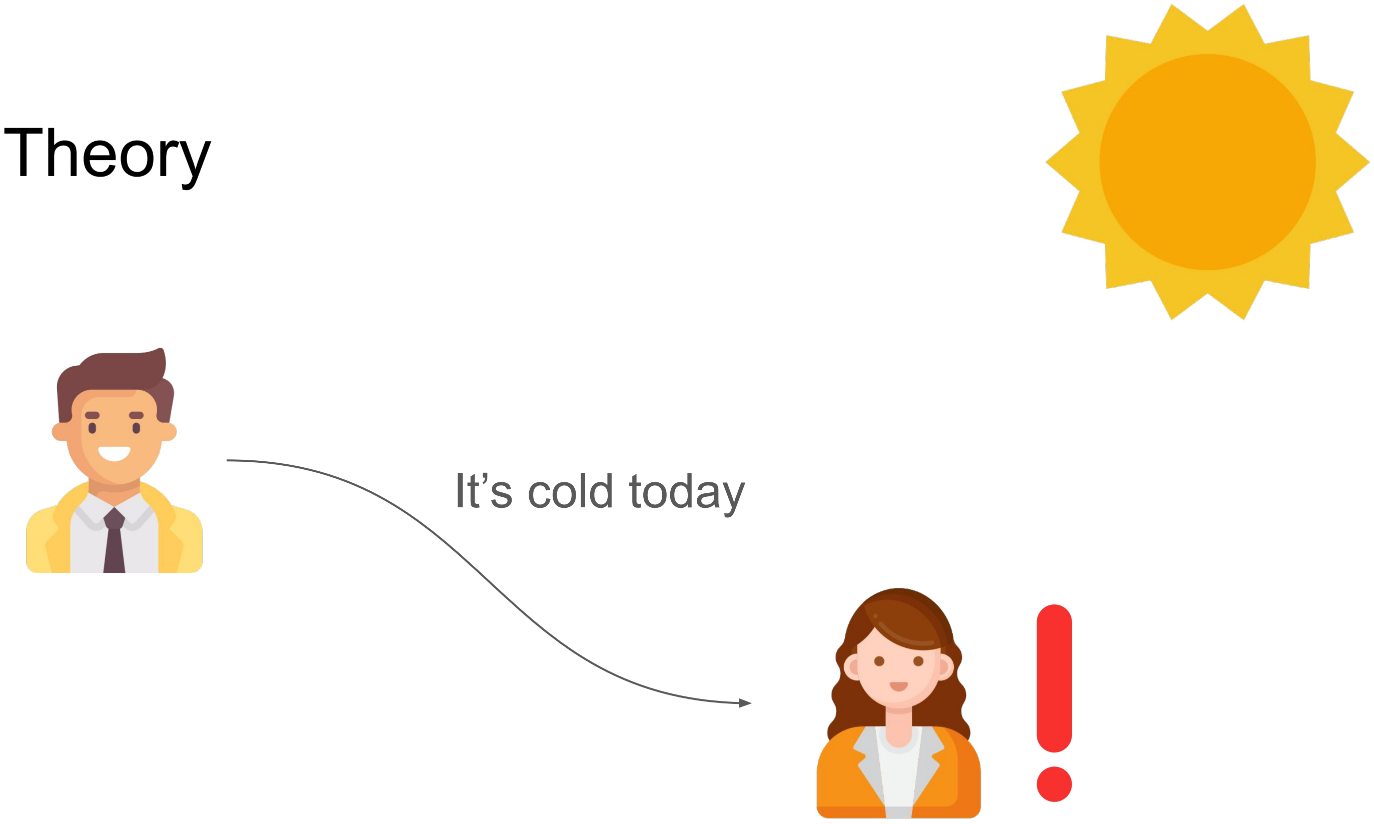
Deep Learning

Information Theory



Not surprising = High chance = No new information

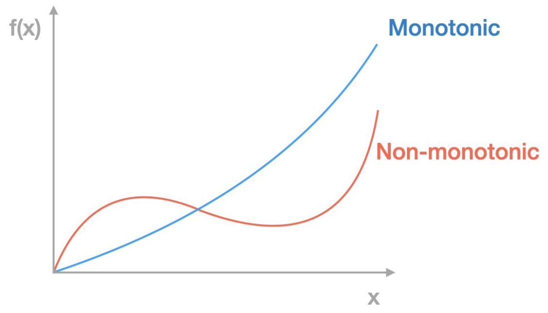
Information Theory



Surprising = Low chance = Information

Function h

- We are seeking a function **h** that quantifies the amount of information contained in a random variable **x**.
- For example, we have a random variable x
  - Hot and Cold
  - *p(hot) = 0.999999999, p(cold) = 0.000000001*
- Should be like *h(cold) > h(hot)*
- Monotonic



Function h

- Random variables X, Y
  - X is Hot or Cold weather
  - Y is Dr. Kim’s Class or No class
  - X and Y are independent
- $h(x, y) = h(x) + h(y)$
- $p(x, y) = p(x) * p(y)$
- The function h that satisfies all the conditions above is log function!
- $\boldsymbol{h(x) = -\log_2 p(x)}$

Yes, it's about Entropy!!!

Example

$$h(x) = -log_2 p(x)$$

```
h(hot)  = -log p(hot)  = -log (0.99999999) = 0.000000014

h(cold) = -log p(cold) = -log (0.00000001) = 26.5754247591
```

So, amount of information on average?

```
p(hot)*h(hot)+p(cold)*h(cold) = 0.99999999*0.000000014 + 0.00000001*26.5754247591

= 2.79754247451e-07
```

$$H[x] = -\sum_x p(x) log_2 p(x) = E_p[-log_2 p(x)]$$

What if?

$h(hot) = -\log p(hot) = -\log (0.53) = 0.916$

$h(cold) = -\log p(cold) = -\log (0.47) = 1.089$

So, amount of information on average?

$p(hot)*h(hot)+p(cold)*h(cold) = 0.53*0.916 + 0.47*1.089$

$= 0.99731$

8 sided dice - case 1

$\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}$

Entropy?

$H[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = \mathbf{3}$ (bits)

What does this mean? When we express entropy in bits, it's essentially a measure of the average length, in binary digits (bits), needed to encode the information about the uncertainty or randomness of an event.

$$H[x] = -\sum_x p(x) log_2 p(x) = E_p[-log_2 p(x)]$$

8 sided dice - case 1

$\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, 1/16, 1/64, 1/64, 1/64, 1/64$

$H[x] = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - 1/16 \log 1/16 - 4/64 \log 1/64 = 2$ (bits)

What if we code the information like 0, 10, 110, 1110, 111100, 111101, 111110, 111111?

Average code length

$= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + 1/16 \times 4 + 4 \times 1/64 \times 6 = 2$

Entropy is a lower bound of average coding length.

4 sided dice

Actual distribution

$p = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

Incorrect distribution (Fool's idea)

$q = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$

The fool coded the variables like **0, 10, 110, 111**.

(We know the ideal coding is 00, 01, 10, 11.

4 sided dice

Average coding length is

$\frac{1}{4} * 1 + \frac{1}{4} * 2 + \frac{1}{4} * 3 + \frac{1}{4} * 3 = 2.25$

Its entropy is

$-\frac{1}{4} * \log(0.5) - \frac{1}{4} * \log(0.25) - \frac{1}{4} * \log(0.125) - \frac{1}{4} \log(0.125) = 2.25$

If the actual distribution is used

$-\frac{1}{4} * \log(0.25) - \frac{1}{4} * \log(0.25) - \frac{1}{4} * \log(0.25) - \frac{1}{4} \log(0.25) = 2$

Therefore, because of the difference between q and p, additional cost to transfer the information, 2.25 - 2 = 0.25

Cost caused by incorrect modeling

$$(-\sum_x p(x)log_2 q(x)) - (-\sum_x p(x)log_2 p(x)) = (-\sum_x p(x)log_2 \frac{q(x)}{p(x)})$$

Continuous variable

$$KL(p||q)$$

$$= -\int p(x) \ln q(x) dx - (-\int p(x) \ln p(x) dx)$$

$$= -\int p(x) \ln \{\frac{q(x)}{p(x)}\} dx$$

KL divergence

KL divergence quantifies how much information is lost when a distribution q is used to approximate another distribution p.

A higher value indicates a greater discrepancy between the two distributions.