Introduction to Information Theory and its applications in Machine Learning

Amaury Gouverneur

WASP - Mathematical Foundations of AI Cluster

September 21, 2021

Outline

Shannon information content

Shannon entropy

KL-divergence

Mutual information

Application in Machine Learning

The 1948 paper

The 1948 paper

- *A Mathematical Theory of Communication (1948)*

The 1948 paper

- *A Mathematical Theory of Communication (1948)*

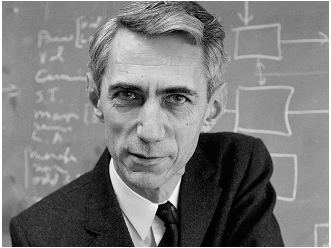


Figure: Claude Shannon

Shannon information content

"Can we define a quantity which will measure, in some sense, how much information is “produced” by such a process?"

Shannon information content

"Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process?"

Consider a discrete random variable X.
The Shannon information content for the outcome $X = x$ is defined as:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

Shannon information content

> *"Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process?"*

Consider a discrete random variable X.
The Shannon information content for the outcome $X = x$ is defined as:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

Desiderata in measuring information:

Shannon information content

> *"Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process?"*

Consider a discrete random variable X.
The Shannon information content for the outcome $X = x$ is defined as:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

Desiderata in measuring information:

1. Deterministic outcomes contain no information

Shannon information content

"Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process?"

Consider a discrete random variable X.
The Shannon information content for the outcome $X = x$ is defined as:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

Desiderata in measuring information:

1. Deterministic outcomes contain no information
2. Information content increases with decreasing probability

Shannon information content

> *"Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process?"*

Consider a discrete random variable X.
The Shannon information content for the outcome $X = x$ is defined as:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

Desiderata in measuring information:

1. Deterministic outcomes contain no information
2. Information content increases with decreasing probability
3. Information content is additive for independent R.V.s.

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

1. Deterministic outcomes contain no information.

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

1. Deterministic outcomes contain no information.

$$P(X = x) = 1 \implies h(X = x) = 0$$

Shannon information content

Verification of the properties:

$$h(X=x) = \log_2 \frac{1}{P(X=x)} = -\log_2 P(X=x)$$

1. Deterministic outcomes contain no information.

$$P(X=x) = 1 \implies h(X=x) = 0$$

$$h(X=x) = -\log_2(\frac{1}{1}) = 0$$

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

2. Information content increases with decreasing probability.

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

2. Information content increases with decreasing probability.

$$P(X = x) < P(X = x') \implies h(X = x) > h(X = x')$$

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

2. Information content increases with decreasing probability.

$$P(X = x) < P(X = x') \implies h(X = x) > h(X = x')$$

$$\frac{d}{dp} \log_2 \frac{1}{p} = -\frac{1}{p \ln 2} < 0 \quad \text{for} \quad p > 0$$

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

3. Information content is additive for independent R.V.s.

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

3. Information content is additive for independent R.V.s.

$$P(X = x, Y = y) = P(X = x)P(Y = y) \implies h(X = x, Y = y) = h(X = x) + h(Y = y)$$

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

3. Information content is additive for independent R.V.s.

$$P(X = x, Y = y) = P(X = x)P(Y = y) \implies h(X = x, Y = y) = h(X = x) + h(Y = y)$$

$$h(X = x, Y = y) = \log_2 \frac{1}{P(X = x)P(Y = y)}$$

Shannon information content

Verification of the properties:

$$h(X=x) = \log_2 \frac{1}{P(X=x)} = -\log_2 P(X=x)$$

3. Information content is additive for independent R.V.s.

$$P(X=x, Y=y) = P(X=x)P(Y=y) \implies h(X=x, Y=y) = h(X=x) + h(Y=y)$$

$$\begin{aligned} h(X=x, Y=y) &= \log_2 \frac{1}{P(X=x)P(Y=y)} \\ &= \log_2 \frac{1}{P(Y=y)} + \log_2 \frac{1}{P(Y=y)} \end{aligned}$$

Shannon information content

Verification of the properties:

$$h(X = x) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x)$$

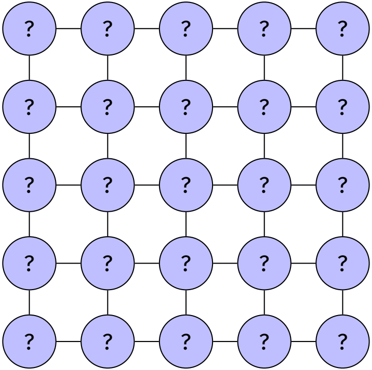
3. Information content is additive for independent R.V.s.

$$P(X = x, Y = y) = P(X = x)P(Y = y) \implies h(X = x, Y = y) = h(X = x) + h(Y = y)$$

$$\begin{aligned} h(X = x, Y = y) &= \log_2 \frac{1}{P(X = x)P(Y = y)} \\ &= \log_2 \frac{1}{P(Y = y)} + \log_2 \frac{1}{P(Y = y)} \\ &= h(X = x) + h(Y = y) \end{aligned}$$

Submarine example

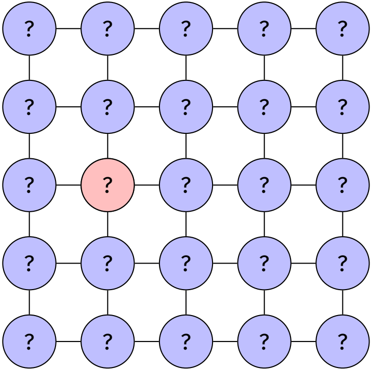
total: 0 bits



The goal is to find the submarine

Submarine example

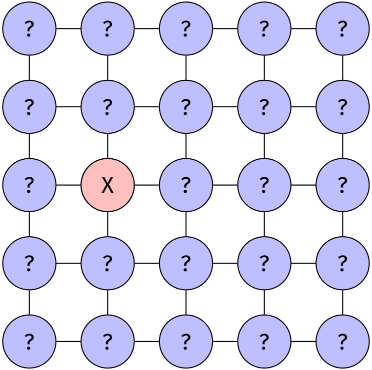
total: 0 bits



Let’s chose to uncover the red "?"

Submarine example

total: 0.0588 bits

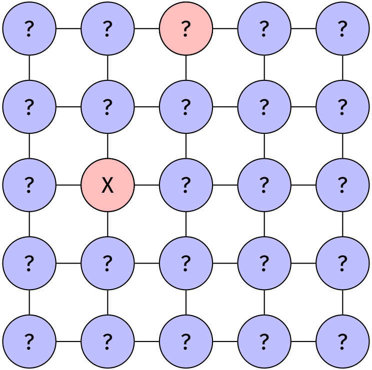


We missed. The probability to miss was 24/25. The information we gained is

$$
\begin{aligned}
h(\text{miss} \quad w/25) &= \log_2(25/24) \\
&= 0.0588
\end{aligned}
$$

Submarine example

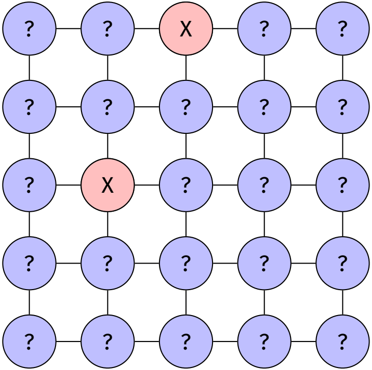
total: 0.0588 bits



Let’s chose to uncover the red "?"

Submarine example

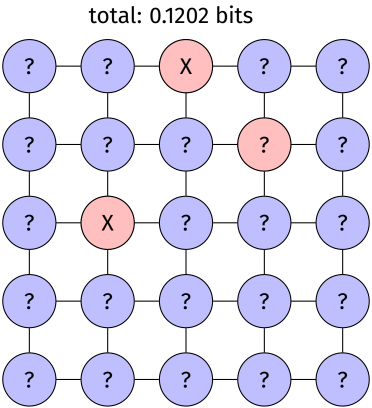
total: 0.1202 bits



We missed. The probability to miss was $23/24$. The information we gained is

$$
\begin{aligned}
h(\text{miss} \quad w/24) &= \log_2(24/23) \\
&= 0.0614
\end{aligned}
$$

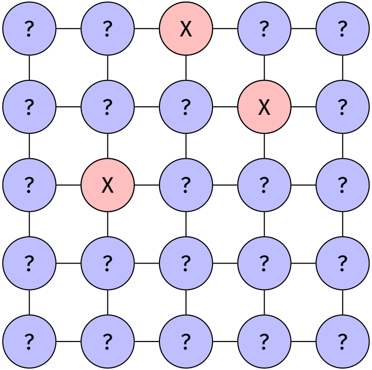
Submarine example



Let’s chose to uncover the red "?"

Submarine example

total: 0.1844 bits

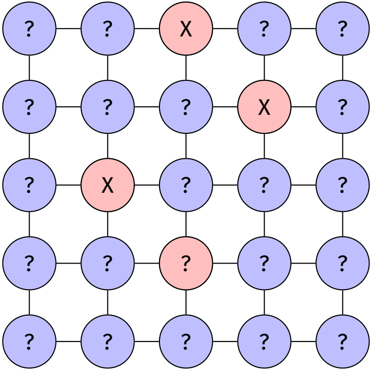


We missed. The probability to miss was $22/23$. The information we gained is

$$
\begin{aligned}
h(\text{miss} \quad w/23) &= \log_2(23/22) \\
&= 0.0641
\end{aligned}
$$

Submarine example

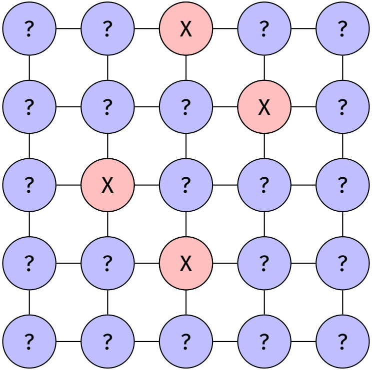
total: 0.1844 bits



Let's chose to uncover the red "?"

Submarine example

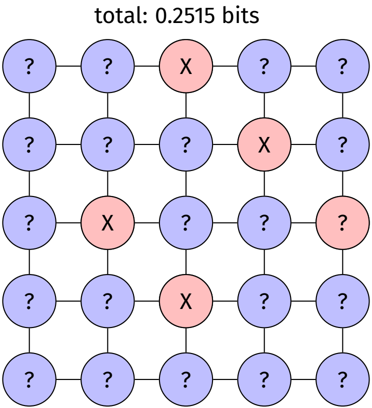
total: 0.2515 bits



We missed. The probability to miss was $21/22$. The information we gained is

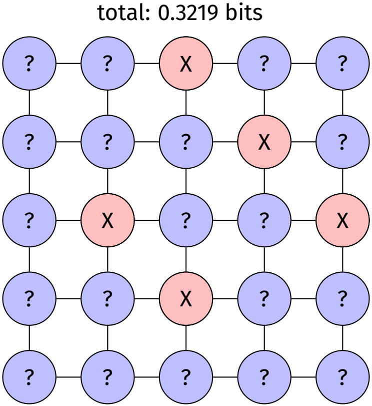
$$
\begin{aligned}
h(\text{miss} \quad w/22) &= \log_2(22/21) \\
&= 0.0671
\end{aligned}
$$

Submarine example



Let's chose to uncover the red "?"

Submarine example

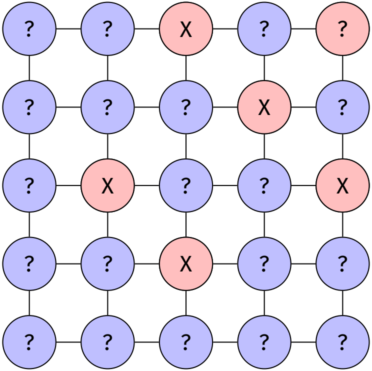


We missed. The probability to miss was 20/21. The information we gained is

$$h(\text{miss} \quad w/21) = \log_2(21/20)$$
$$= 0.0703$$

Submarine example

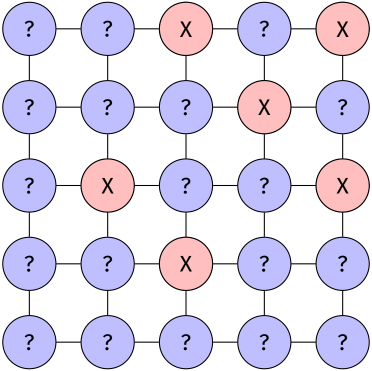
total: 0.3219 bits



Let’s chose to uncover the red "?"

Submarine example

total: 0.3959 bits

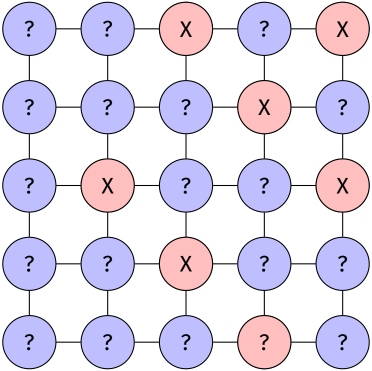


We missed. The probability to miss was 19/20. The information we gained is

$$
\begin{aligned}
h(\text{miss} \quad w/20) &= \log_2(20/19) \\
&= 0.0740
\end{aligned}
$$

Submarine example

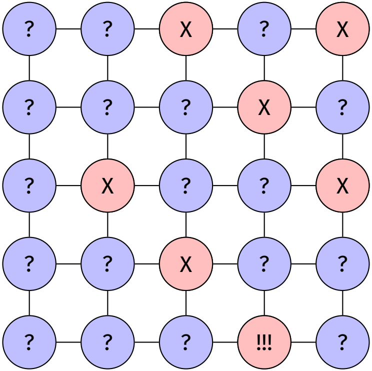
total: 0.3959 bits



Let’s chose to uncover the red "?"

Submarine example

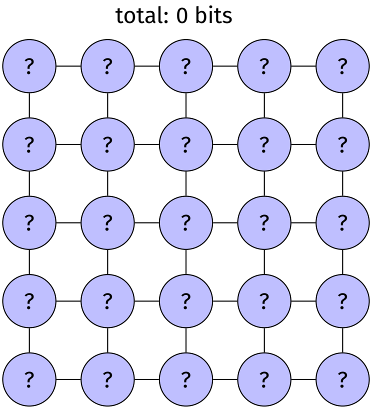
total: 4.6439 bits



We found it! The probability to hit was $1/19$. The information we gained is

$$
\begin{aligned}
h(\text{hit} \quad w/22) &= \log_2(19/1) \\
&= 4.248
\end{aligned}
$$

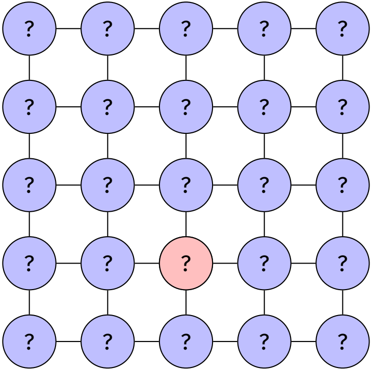
Another try



Let's try again

Another try

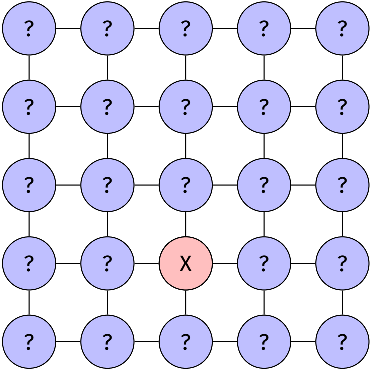
total: 0 bits



Let’s chose to uncover the red "?"

Another try

total: 0.0588 bits

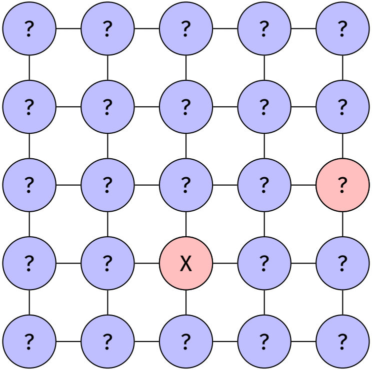


We missed. The probability to miss was $24/25$. The information we gained is

$$h(\text{miss} \quad w/25) = \log_2(25/24)$$
$$= 0.0588$$

Another try

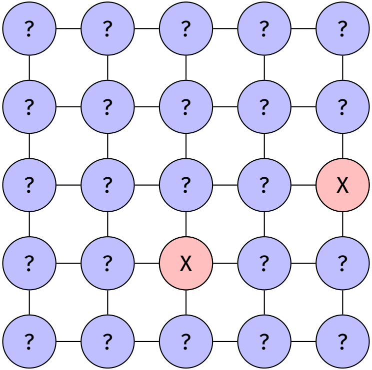
total: 0.0588 bits



Let’s chose to uncover the red "?"

Another try

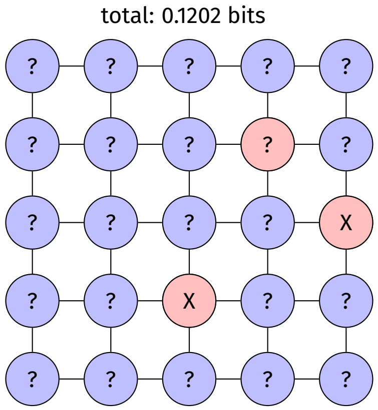
total: 0.1202 bits



We missed. The probability to miss was $23/24$. The information we gained is

$$\begin{aligned} h(\text{miss} \quad w/24) &= \log_2(24/23) \\ &= 0.0614 \end{aligned}$$

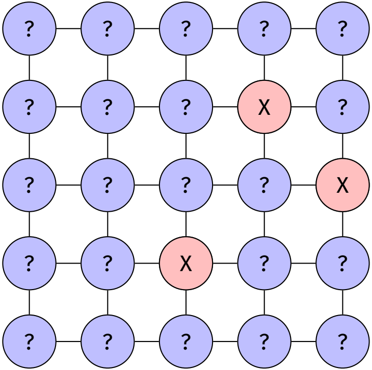
Another try



Let’s chose to uncover the red "?"

Another try

total: 0.1844 bits

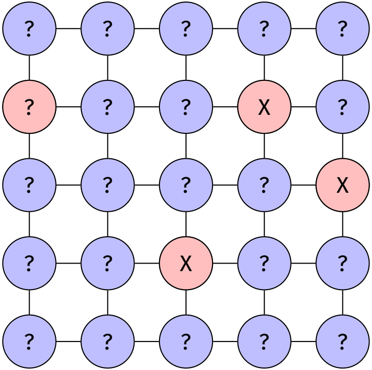


We missed. The probability to miss was $22/23$. The information we gained is

$$
\begin{aligned}
h(\text{miss} \quad w/23) &= \log_2(23/22) \\
&= 0.0641
\end{aligned}
$$

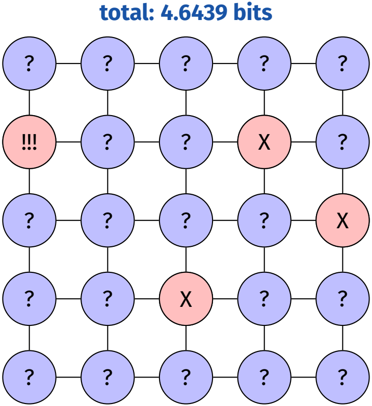
Another try

total: 0.1844 bits



Let's chose to uncover the red "?"

Another try

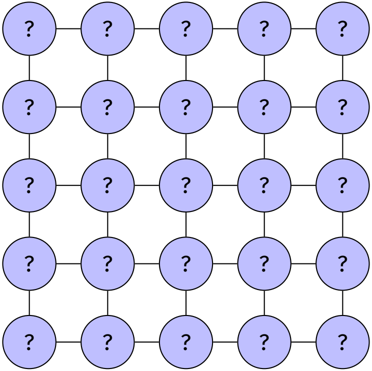


We found it! The probability to hit was $1/22$. The information we gained is

$$h(\text{hit} \quad w/22) = \log_2(22/1)$$
$$= 4.4594$$

One more try

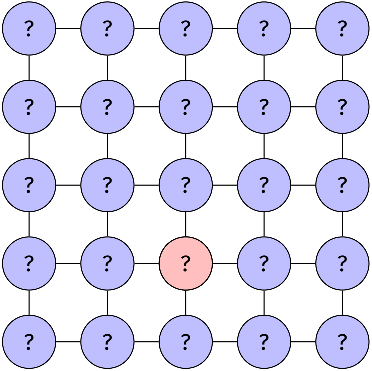
total: 0 bits



Let's try again

One more try

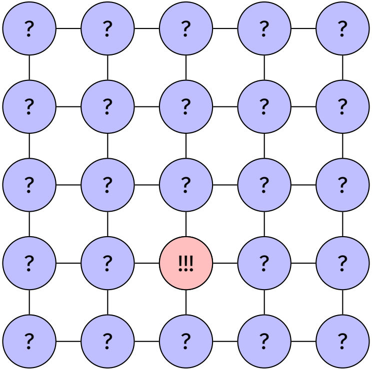
total: 0 bits



Let's chose to uncover the red "?"

One more try

total: 4.6439 bits



We found it! The probability to hit was $1/25$. The information we gained is

$$h(\text{hit} \quad w/25) = \log_2(25/1)$$
$$= 4.6439$$

Shannon entropy

Average information content:

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) h(x = X)$$

Shannon entropy

Average information content:

$$
\begin{aligned}
H(X) &= \sum_{x \in \mathcal{X}} P(X = x) h(x = X) \\
&= \sum_{x \in \mathcal{X}} P(X = x) \log_2\left(\frac{1}{P(X = x)}\right)
\end{aligned}
$$

Shannon entropy

Average information content:

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} P(X=x) h(x=X) \\ &= \sum_{x \in \mathcal{X}} P(X=x) \log_2\left(\frac{1}{P(X=x)}\right) \\ &= -\sum_{x \in \mathcal{X}} P(X=x) \log_2(P(X=x)) \end{aligned}$$

Shannon entropy

Example for a weighted coin

Let

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

Shannon entropy

Example for a weighted coin

Let

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Then

$$H(X) = -p\log(p) - (1-p)\log(1-p)$$

Shannon entropy

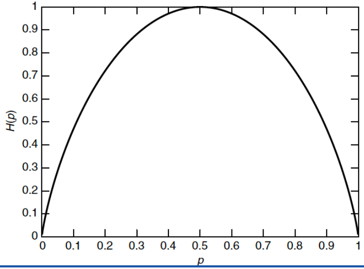
Example for a weighted coin

Let

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

Then

$$H(X) = -p\log(p) - (1-p)\log(1-p)$$



Joint entropy

Multivariate generalization of the Shannon entropy.

$$H(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(X=x, Y=y) \log_2\left(\frac{1}{P(X=x)P(Y=y)}\right)$$

Conditional entropy

Entropy of the conditional distribution

$$H(X|Y=y) = \sum_{x \in \mathcal{X}} P(X=x|Y=y) \log_2\left(\frac{1}{P(X=x|Y=y)}\right)$$

Conditional entropy

Entropy of the conditional distribution

$$H(X|Y=y) = \sum_{x \in \mathcal{X}} P(X=x|Y=y) \log_2\left(\frac{1}{P(X=x|Y=y)}\right)$$

Conditional entropy

Conditional entropy

Entropy of the conditional distribution

$$H(X|Y=y) = \sum_{x\in\mathcal{X}} P(X=x|Y=y)\log_2\left(\frac{1}{P(X=x|Y=y)}\right)$$

Conditional entropy

$$H(X|Y) = \sum_{y\in\mathcal{Y}} P(Y=y)H(X|Y=y)$$

Conditional entropy

Entropy of the conditional distribution

$$H(X|Y=y) = \sum_{x \in \mathcal{X}} P(X=x|Y=y)\log_2(\frac{1}{P(X=x|Y=y)})$$

Conditional entropy

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} P(Y=y)H(X|Y=y) \\ &= \sum_{y \in \mathcal{Y}} P(Y=y) sum_{x \in \mathcal{X}} P(X=x|Y=y)\log_2(\frac{1}{P(X=x|Y=y)}) \end{aligned}$$

Conditional entropy

Entropy of the conditional distribution

$$H(X|Y=y) = \sum_{x \in \mathcal{X}} P(X=x|Y=y) \log_2\left(\frac{1}{P(X=x|Y=y)}\right)$$

Conditional entropy

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} P(Y=y) H(X|Y=y) \\ &= \sum_{y \in \mathcal{Y}} P(Y=y) sum_{x \in \mathcal{X}} P(X=x|Y=y) \log_2\left(\frac{1}{P(X=x|Y=y)}\right) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(X=x, Y=y) \log_2\left(\frac{1}{P(X=x|Y=y)}\right) \end{aligned}$$

Conditional entropy

Properties of the conditional entropy:

Conditional entropy

Properties of the conditional entropy:

1. $H(X,Y) = H(X|Y) + H(Y)$

Conditional entropy

Properties of the conditional entropy:

1. $H(X,Y) = H(X|Y) + H(Y)$
2. $H(X|Y) = 0$ if X is deterministic knowing Y

Conditional entropy

Properties of the conditional entropy:

1. $H(X,Y) = H(X|Y) + H(Y)$
2. $H(X|Y) = 0$ if X is deterministic knowing Y
3. $H(X|Y) = H(X)$ if X and Y are independent

Kullback-Leibler divergence

A useful "measure" of difference between two distributions. Let P and Q be two distributions,

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{P(x)}{Q(x)}$$

Kullback-Leibler divergence

A useful "measure" of difference between two distributions. Let P and Q be two distributions,

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{P(x)}{Q(x)}$$

Properties of the KL-divergence:

Kullback-Leibler divergence

A useful "measure" of difference between two distributions. Let P and Q be two distributions,

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{P(x)}{Q(x)}$$

Properties of the KL-divergence:

1. $D_{KL}(P||Q) \geq 0$

Kullback-Leibler divergence

A useful "measure" of difference between two distributions. Let P and Q be two distributions,

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{P(x)}{Q(x)}$$

Properties of the KL-divergence:

1. $D_{KL}(P||Q) \geq 0$
2. $D_{KL}(P||Q) = 0$ only if $P = Q$

Kullback-Leibler divergence

A useful "measure" of difference between two distributions. Let P and Q be two distributions,

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{P(x)}{Q(x)}$$

Properties of the KL-divergence:

1. $D_{KL}(P||Q) \geq 0$
2. $D_{KL}(P||Q) = 0$ only if $P = Q$
3. It's not a metric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ in general

KL-divergence

Illustration

Mutual information

$$I(X;Y) = D_{KL}(P(x,y)||P(x)P(y))$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(x,y)||P(x)P(y)) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(X=x, Y=y) \log_2\left(\frac{P(X=x, Y=y)}{P(X=x)P(Y=y)}\right)
\end{aligned}
$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(x,y)||P(x)P(y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(X=x,Y=y)\log_2\left(\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}\right) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(x,y)||P(x)P(y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(X=x,Y=y)\log_2\left(\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}\right) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(x,y)||P(x)P(y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(X=x,Y=y)\log_2\left(\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}\right) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(x,y)||P(x)P(y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(X=x,Y=y)\log_2\left(\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}\right) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

Data-processing inequality.

Mutual information

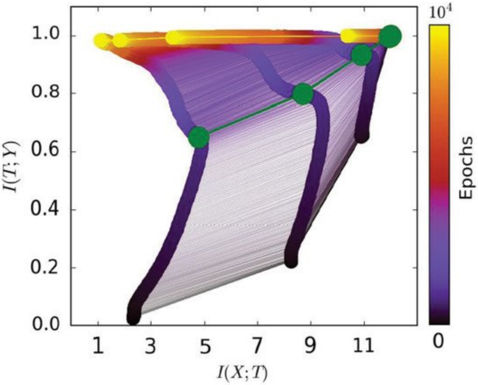
$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(x,y)||P(x)P(y)) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(X=x,Y=y)\log_2\left(\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}\right) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

Data-processing inequality. If $X \rightarrow Y \rightarrow Z$ forms a Markov chain,

$$I(X;Y) \geq I(X;Z)$$

Application in Machine Learning

The information plane and the information bottleneck. ***Shwartz-Ziv, Ravid, and Naftali Tishby. "Opening the black box of deep neural networks via information." (2017)***



DNN as an encoder-decoder

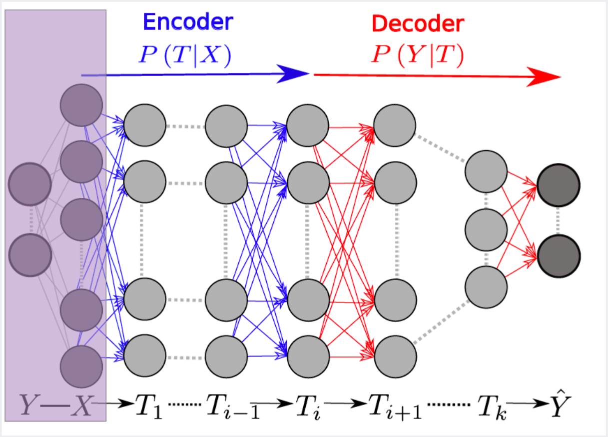


Figure: The DNN layers form a Markov chain of successive internal representations of the input layer X.

The Information Plane

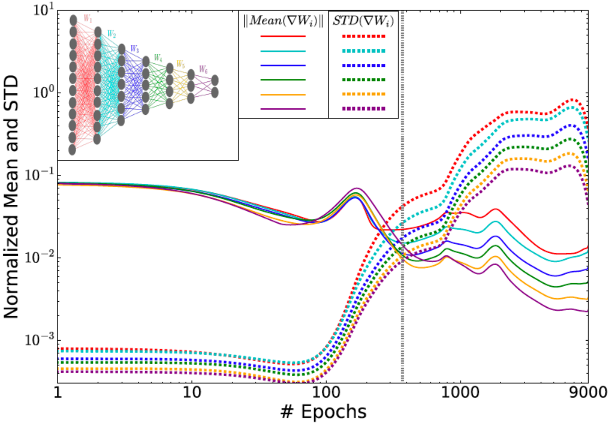
First optimization phase

Learning part 1

Second optimization phase

Learning part 2

The drift and diffusion phases of SGD optimization



The end

Thank you for listening. Any **questions**?