Machine Learning

Lecture 6: Information theory

Hao Tang and Hiroshi Shimodaira

30 September 2022

Ver. 1.0

Topics - you should be able to explain after this week

- How to quantify information / how to measure the amount of information?
- History of information theory *(NE)*
- Information content (aka self-information, Shannon information)
- Entropy
- Conditional entropy
- Mutual information
- Cross entropy
- Kullback-Leibler divergence
- Application of information theory for the training of classifiers

Warming up

- What is meant by "information"?
  - facts provided or learned about something or someone [ODE]
  - what is conveyed of represented by a particular arrangement or sequence of things [ODE]
  - about someone or something consists of facts about them [Cobuild]
  - consisting of the facts and figures that are stored and used by a computer program [Cobuild]

- Which has more information/surprising?

	Event	
USB memory	*2GB*	*32GB*
Weather tomorrow	*rainy*	*snowy*
Next MLG lecture	*Mon, 2nd Oct.*	*Tue, 3rd Oct.*
Roll a dice	*got 1*	*got 6*

How to define the amount of information?

Let $I(x)$ denote the amount of information for event x

Desired properties of $I(x)$:

- Monotonically decreasing function of probability
  - If $p(x) = 1 \;\rightarrow\; I(x) = 0$
  - If $p(x) = 0 \;\rightarrow\; I(x) = \infty$
- Additivity of independent events
  - If $p(x, y) = p(x)p(y) \;\rightarrow\; I(x, y) = I(x) + I(y)$

How to define the amount of information?

Let $I(x)$ denote the amount of information for event x

Desired properties of $I(x)$:

- Monotonically decreasing function of probability
  - If $p(x) = 1 \ \rightarrow \ I(x) = 0$
  - If $p(x) = 0 \ \rightarrow \ I(x) = \infty$
- Additivity of independent events
  - If $p(x, y) = p(x)p(y) \ \rightarrow \ I(x, y) = I(x) + I(y)$

Candidates of $I(x)$:

$\frac{1}{p(x)}$	✗
$\log\left(\frac{1}{p(x)}\right)$	✓

How to define the amount of information?

Let $I(x)$ denote the amount of information for event x

Desired properties of $I(x)$:

- Monotonically decreasing function of probability
  - If $p(x) = 1 \;\rightarrow\; I(x) = 0$
  - If $p(x) = 0 \;\rightarrow\; I(x) = \infty$
- Additivity of independent events
  - If $p(x, y) = p(x)p(y) \;\rightarrow\; I(x, y) = I(x) + I(y)$

Candidates of $I(x)$:

$\frac{1}{p(x)}$	✗
$\log\left(\frac{1}{p(x)}\right)$	✓

Choice of logarithmic base:

$\log_2\left(\frac{1}{p(x)}\right)$ *[bits]*, $\log_e\left(\frac{1}{p(x)}\right)$ *[nats]* *(We use* $\log$ *to denote* $\log_e$ *here)*

How to define similarity between two distributions?

$p_x(x)$ vs $p_y(y)$

- Euclidean distance
- Pearson correlation coefficient
- Any measures based on probability?

History of information theory *(NE)*

1948 Claude E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal

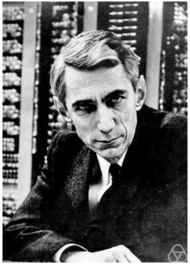
1951 Huffman encoding

1966 Linear Predictive Coding (LPC) by Fumitada Itakura

1972 Discrete Cosine Transform (DCT) by Nasir Ahmed

$\rightarrow$ MPEG video coding, JPEG image compression, MP3 audio compression

1989 Zip file format by Phil Katz



Shannon, Claude - Author: Jacobs, Konrad — Source: Konrad Jacobs, Erlangen — Copyright: MFO. CC BY-SA 2.0 de

Channel coding

- We want to send a message with minimal number of bits.
- We don't know the message ahead of time.

Sending letters

- ASCII codes *(NE)*

Letter	ASCII code		
	Dec	Hex	Bin
'A'	65	41	01000001
'B'	66	42	01000010
'C'	67	43	01000011
⋮	⋮	⋮	⋮
'Z'	90	5A	01011010

- Morse code *(NE)*

Sending letters

- ASCII codes *(NE)*

Letter	ASCII code		
	Dec	Hex	Bin
'A'	65	41	01000001
'B'	66	42	01000010
'C'	67	43	01000011
⋮	⋮	⋮	⋮
'Z'	90	5A	01011010

- Morse code *(NE)*
- Unit of coding
  - Letter
  - Two letters, three letters, ...
  - Word
  - Two words, three words, ...

Sending coin flips

- How many bits do we need to send a coin flip?
- We need 1 bit per message.
- How many bits do we need to send two coin flips?
- We need 2 bits per message

Sending coin flips

- If it's a fair coin, $p(H) = p(T) = 1/2$.
- If there are two fair coins, $p(HH) = p(HT) = p(TH) = p(TT) = 1/4$.
- The number of bits to encode a variable x is

$$\log_2 \frac{1}{p(x)} = -\log_2 p(x). \tag{1}$$

- Low-probability events need more bits, while high-probability events need fewer bits.
- $-\log_2 p(x)$ bits are equivalent to $-\log p(x)$ nats.

Entropy

- The entropy of a distribution p is defined as

$$H(p) = H(x) = \mathbb{E}_{x \sim p(x)}[-\log p(x)]. \tag{2}$$

NB:

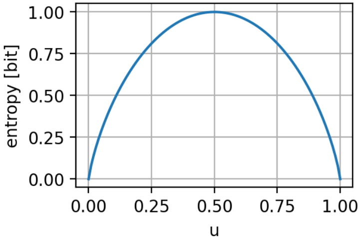
$$\mathbb{E}_{x \sim p(x)}[-\log p(x)] = -\int_{-\infty}^{\infty} p(x) \log p(x) dx \quad \text{or} \quad -\sum_{x \in \Omega} p(x) \log p(x) \tag{3}$$

- Note that $H(x)$ is **not** a function of x.

- The entropy can be interpreted as the expected number of nats needed to a message.

Entropy of a coin

- For a coin with probability u being head, its entropy is $-u \log u - (1-u) \log(1-u)$.
- The entropy peaked at $u = 0.5$.
- In general, the entropy of a distribution is higher when the distribution is closer to uniform.
- Entropy can be seen as a measure of *uncertainty*.



Conditional entropy

- The conditional entropy of x given y is

$$H(x|y) = \mathbb{E}_{x,y\sim p(x,y)}[-\log p(x|y)] \tag{4}$$

- If x and y are independent,

$$\begin{aligned}
H(x|y) &= \mathbb{E}_{x,y\sim p(x,y)}\left[-\log \frac{p(x,y)}{p(y)}\right] && (5)\\
&= \mathbb{E}_{x,y\sim p(x,y)}\left[-\log \frac{p(x)p(y)}{p(y)}\right] && (6)\\
&= \mathbb{E}_{x,y\sim p(x,y)}[-\log p(x)] && (7)\\
&= \mathbb{E}_{x\sim p(x)}[-\log p(x)] && (8)\\
&= H(x) && (9)
\end{aligned}$$

Conditional entropy

- Knowing something reduces the entropy in general.

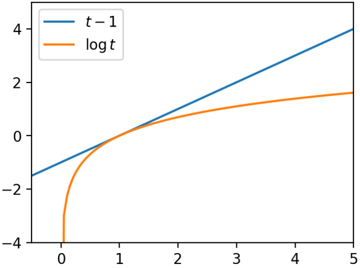
$$H(x|y) \leq H(x) \tag{10}$$

- The proof requires a basic fact

$$\log t \leq t - 1 \qquad \text{for } t > 0. \tag{11}$$

Or,

$$-\log t \geq 1 - t \qquad \text{for } t > 0. \tag{12}$$



Conditional entropy

$$
\begin{aligned}
H(x) - H(x|y) &= \mathbb{E}_{x \sim p(x)}[-\log p(x)] - \mathbb{E}_{x,y \sim p(x,y)}\left[-\log \frac{p(x,y)}{p(y)}\right] && (13) \\
&= \mathbb{E}_{x,y \sim p(x,y)}\left[-log \frac{p(x)p(y)}{p(x,y)}\right] && (14) \\
&\geq \mathbb{E}_{x,y \sim p(x,y)}\left[1 - \frac{p(x)p(y)}{p(x,y)}\right] && (15) \\
&= 1 - \sum_x \sum_y p(x,y) \frac{p(x)p(y)}{p(x,y)} && (16) \\
&= 1 - \sum_x p(x) \sum_y p(y) = 0 && (17)
\end{aligned}
$$

Mutual information

- Since $H(x|y) \leq H(x)$, the **extra** information $H(x) - H(x|y)$ we know about x given y is called the mutual information

$$I(x, y) = H(x) - H(x|y) = H(y) - H(y|x) \tag{18}$$

$$= \mathbb{E}_{x,y \sim p(x,y)} \left[-\log \frac{p(x)p(y)}{p(x,y)} \right] \tag{19}$$

Cross entropy

- Recall that the entropy $\mathbb{E}_{x \sim p(x)}[-\log p(x)]$ can be interpreted as drawing a message x from $p(x)$ and sending it with $-\log p(x)$ nats.
- This assumes that we know p. What happens if we do not?
- We estimate p with some other distribution q.
- The expected number of nats (under p) of encoding messages with distribution q is the cross entropy

$$H(p, q) = \mathbb{E}_{x \sim p(x)}[-\log q(x)]. \tag{20}$$

- NB: the notation $H(p, q)$ is also used to denote joint entropy $H(x, y)$! *(NE)*

Cross entropy

- We need more nats if we encode messages with a distribution q other than the true distribution p.

$$H(p) \leq H(p, q). \tag{21}$$

- The proof uses the inequality $\log t \leq t - 1$ again.

Cross entropy

$$
\begin{aligned}
H(p,q) - H(p) &= \mathbb{E}_{x\sim p(x)}[-\log q(x)] - \mathbb{E}_{x\sim p(x)}[-\log p(x)] && (22)\\
&= \mathbb{E}_{x\sim p(x)}\left[-\log \frac{q(x)}{p(x)}\right] && (23)\\
&\geq \mathbb{E}_{x\sim p(x)}\left[1 - \frac{q(x)}{p(x)}\right] && (24)\\
&= 1 - \sum_x p(x)\frac{q(x)}{p(x)} = 0 && (25)
\end{aligned}
$$

Kullback-Leibler divergence

- The **extra** nats of encoding with the wrong distribution is the Kullback-Leibler divergence

$$
\begin{aligned}
\mathrm{KL}(p \| q) &= H(p, q) - H(p) && (26) \\
&= \mathbb{E}_{x \sim p(x)} \left[-log \frac{q(x)}{p(x)} \right] && (27)
\end{aligned}
$$

- $\mathrm{KL}(p \| q) \geq 0$
- $\mathrm{KL}(p \| p) = 0$
- KL divergence is often used to measure the distance between two distributions.
- However, in general, $\mathrm{KL}(p \| q) \neq KL(q \| p)$.

Mutual information

- Recall that

$$I(x, y) = H(x) - H(x|y) = H(y) - H(y|x) \tag{28}$$
$$= \mathbb{E}_{x,y \sim p(x,y)} \left[-\log \frac{p(x)p(y)}{p(x,y)} \right] \tag{29}$$

- In other words, $I(x, y) = \mathrm{KL}(p \| q)$ where $q(x, y) = p(x)p(y)$.
- Mutual information can be interpreted as the number of extra nats if we assume x and y are independent.

Cross entropy and log loss

- Recall that in multiclass classification,

$$p(y|x) = \frac{\exp(w_y^\top \phi(x))}{\sum_{y' \in \mathcal{Y}} w_{y'}^\top \phi(x)}. \tag{30}$$

- The log loss is

$$-\log p(y^*|x) = -w_{y^*}^\top \phi(x) + \log \left(\sum_{y' \in \mathcal{Y}} w_{y'}^\top \phi(x) \right) \tag{31}$$

where y^* is the label.

Cross entropy and log loss

- Given a data point (x, y^*), we can represent the ground truth as a distribution

$$p(y) = \mathbb{1}_{y=y^*} \tag{32}$$

- The cross entropy between the ground truth and the learned distribution is

$$\begin{aligned}
\mathbb{E}_{y\sim p(y)}[-\log p(y|x)] &= \sum_{y\in\mathcal{Y}} p(y)[-\log p(y|x)] && (33) \\
&= \sum_{y\in\mathcal{Y}} \mathbb{1}_{y=y^*}[-\log p(y|x)] && (34) \\
&= -\log p(y^*|x). && (35)
\end{aligned}$$

- Hence, the log loss is also known as the cross entropy loss.

Textbooks

- M1: Chap.6
- M2: Chap.5

Quizzes

- Derive the entropy of a coin with probability β being head.
- Find β that maximises the entropy of that coin.
- Derive the entropy of a uniform distribution.
- Derive the cross entropy between a discrete distribution against a one-hot distribution.
- Derive the KL-divergence between a discrete distribution against a one-hot distribution.
- Derive the entropy of a Gaussian.
- Derive the cross entropy of two Gaussians.
- Are entropies always positive?