Mathematical Foundations: Information Theory

Prof. Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2021)
http://cs182.sist.shanghaitech.edu.cn

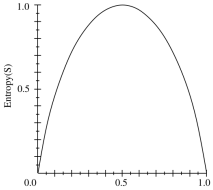
Information Content

- Information content (a.k.a. self-information, surprisal, Shannon information) of an event x_i from the random variable X with probability $p(X = x_i)$ follows from some properties:
  - $I(X = x_i)$ is monotonically decreasing in $p(X = x_i)$
  - $I(X = x_i) \geq 0$
  - when $p(X = x_i) = 1$, $I(X = x_i) = 0$; and when $p(X = x_i) = 0$, $I(X = x_i) = +\infty$
  - $I(X = x_1, X = x_2) = I(X = x_1) + I(X = x_2)$ for independent events x_1 and x_2
- Information content is mathematically chosen as

$$I(X = x_i) = -\log_b p(X = x_i)$$

Entropy: Intuitive Notion

- Measures the impurity, uncertainty, irregularity, surprise
- Entropy: the expected (i.e., average) amount of information conveyed by identifying the outcome of a random trial
- Suppose we have two discrete classes
  - S: a sample of training examples
  - $p_{\oplus}$: proportion of positive examples in S
  - $p_{\ominus}$: proportion of negative examples in S
- Optimal purity (impurity/uncertainty$=0$): either
  - $p_{\oplus} = 1$, $p_{\ominus} = 0$
  - $p_{\oplus} = 0$, $p_{\ominus} = 1$
- Least pure (maximum impurity/uncertainty):
  - $p_{\oplus} = 0.5$, $p_{\ominus} = 0.5$



Entropy: Formal Definition

- X: discrete random variable with alphabet $\mathcal{X} = \{x_1, \ldots, x_n\}$ and probability mass function $p(x) = Pr(X = x)$, $x \in \mathcal{X}$
- entropy (or Shannon entropy)

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x)$$

  - in our previous example: $H(X) = -p_{\oplus} \log_b p_{\oplus} - p_{\ominus} \log_b p_{\ominus}$
  - convention: $0 \cdot \log_b 0 = 0$

Example

S is a collection of 14 examples, 9 positive and 5 negative

$$\text{Entropy}([9+, 5-]) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

  - all members of S belong to the same class $\Rightarrow$ entropy$= 0$
  - equal number of $+$ve and -ve examples $\Rightarrow$ entropy$= 1$
  - otherwise, entropy is between 0 and 1
- for continuous X with probability density function $p(x)$, differential entropy:

$$h(X) = -\int_{x \in \mathcal{X}} p(x) \log_b p(x) dx$$

Entropy: Formal Definition...

Question

What units is entropy measured in?

- depends on the base b of the log
  - $b = e$: nats, b = 2: bits (adopted here)
- entropy can be changed from one base to another
  - $H_b(X) = (\log_b a) H_a(X)$

Note

two different probability distributions p and q can lead to the same entropy $H(p) = H(q)$

- e.g., $p = \{.5, .25, .25\}, q = \{.48, .32, .2\}$
- $H(p) = H(q) = 1.5$

In general, when X can take n values

- $H(X) \leq \log n$, with $H(X) = \log n$ if $p(x) = 1/n$
- $H(X) \geq 0$, with $H(X) = 0$ if there is a x_k with $p(x_k) = 1$

Example: Coding

	a	b	c	d
$P(X)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

- Code:

	a	b	c	d
code	00	01	10	11

- Expected length to encode one symbol from X: 2 bits
- Consider another code:

	a	b	c	d
code	0	10	110	111

- Expected length:

$$\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1.75\text{bits}$$

Relationship with Coding

- Information theory: optimal length code assigns $-\log_2 p$ bits to message having probability p
- Expected number of bits to encode $\oplus$ or $\ominus$ of a random member of S:

$$p_{\oplus}(-\log_2 p_{\oplus}) + p_{\ominus}(-\log_2 p_{\ominus}) = \text{Entropy}(S)$$

- $\text{Entropy}(S)$ = expected number of bits needed to encode class ($\oplus$ or $\ominus$) of a randomly drawn member of S under the optimal, shortest-length code
- In the worst case ($p_{\oplus} = 0.5$), requires one bit to encode each example
- If there is less uncertainty (e.g., $p_{\oplus} = 0.8$), we can use less than 1 bit each

Joint Entropy

- entropy: single random variable
- joint entropy of a pair of discrete random variables X, Y with a joint distribution $p(x,y)$:
$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y)$$
- if X and Y are two independent sample spaces, then $H(X,Y) = H(X) + H(Y)$

Conditional Entropy $H(Y \mid X)$

- uncertainty we have about Y, given that we know X

$$\begin{aligned} H(Y \mid X) &= \sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log p(y \mid x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y \mid x) \end{aligned}$$

  - $H(Y \mid X) \neq H(X \mid Y)$ in general
- chain rule: $H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$
  - interpretation: the uncertainty (entropy) about both X and Y is equal to the uncertainty (entropy) we have about X, plus whatever we have about Y , given that we know X
- when X and Y are independent, $H(Y \mid X) = H(Y)$
- $H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1, \ldots, X_{n-1})$
  - when $X_1, \ldots, X_n$ are i.i.d., $H(X_1, \ldots, X_n) = nH(X_1)$

Kullback-Leibler Divergence (or Relative Entropy)

Motivation:

- suppose there is a r.v. X with true distribution p
- recall we can represent the r.v. with a code that has average length $H(X)/H(p)$
- however, we do not know p; instead we assume the distribution of the r.v. is q
- then the code would need more bits to represent the r.v., and the difference in the number of bits is denoted as $KL(p \parallel q)$

KL-divergence from $p(x)$ to $q(x)$:

$$KL(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- convention: $0 \cdot \log \frac{0}{q} = 0$ and $p \cdot \log \frac{p}{0} = \infty$

Cross entropy: the average coding length under the wrong distribution assumption q

$$CE(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

- $KL(p \parallel q) = CE(p, q) - H(X)$

KL-Divergence...

Example

$\mathcal{X} = \{0, 1\}, p(0) = 1 - r, p(1) = r, q(0) = 1 - s, q(1) = s$

$$KL(p \parallel q) = (1 - r) \log \tfrac{1-r}{1-s} + r \log \tfrac{r}{s}$$

Information inequality

- $KL(p \parallel q) \geq 0$, with equality if and only if $p(x) = q(x) \quad \forall x$
- in the example above: if $r = s$, then $KL(p \parallel q) = KL(q \parallel p) = 0$

Proof:

$$KL(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)}$$

$$\geq - \log(\sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)}) = - \log(\sum_{x \in \mathcal{X}} q(x)) = - \log(1) = 0$$

The equality is attained when $p(x) = cq(x)$. Since $\sum_{x \in \mathcal{X}} p(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, we have $c = 1$.

KL-Divergence...

Often used as a "distance" measure between distributions, but

- not symmetric
  - $KL(p \parallel q) \neq KL(q \parallel p)$ in general
  - in the example above, $KL(q \parallel p) = (1-s)\log\frac{1-s}{1-r} + r\log\frac{s}{r}$
  - may use the symmetric KL-divergence instead

$$KL(p \parallel q) + KL(q \parallel p)$$

- does not satisfy the triangle inequality
  - $KL(p \parallel q) \leq KL(p \parallel r) + KL(r \parallel q)$ does not hold in general
- not a distance between distributions
  - may use the Jensen-Shannon divergence (JSD) instead
  - $\text{JSD}(p, q) = \frac{1}{2}KL(p \parallel r) + \frac{1}{2}KL(r \parallel q)$, where $r = (p+q)/2$
  - $\sqrt{\text{JSD}}$ is a metric!

Mutual Information

Question

How much information does one random variable (Y) tell about another one (X)?

- given: two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$
- mutual information $I(X; Y)$:
  - KL-divergence between the joint distribution $p(x, y)$ and the product distribution $p(x)p(y)$

$$I(X;Y) = KL(p(x,y) \parallel p(x)p(y)) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual Information...

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log\left(\frac{1}{p(x)} \cdot \frac{p(x,y)}{p(y)}\right) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x \mid y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x \mid y) \\
&= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x \mid y)\right) \\
&= H(X) - H(X \mid Y)
\end{aligned}
$$

- interpretation: mutual information is the reduction in the uncertainty of X due to the knowledge of Y
- when X, Y are independent, $p(x,y) = p(x)p(y)$, so $I(X;Y) = 0$
  - if they are independent, Y can tell us nothing about X

Mutual Information...

- chain rule: $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
  - by symmetry $I(X;Y) = I(Y;X)$
  - X says as much about Y as Y says about X
- as $H(X,Y) = H(X) + H(Y \mid X)$, so

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

  - information that X tells about Y = uncertainty in X + uncertainty about Y – uncertainty in both X and Y
- $I(X;X) = H(X) + H(X) - H(X \mid X) = H(X)$
  - mutual information of a r.v. with itself is the entropy of the r.v.
- $I(X;Y) = KL(p(x,y) \parallel p(x)p(y))$, so by the information inequality, $I(X;Y) \geq 0$ with equality if and only if X and Y are independent
  - in other words, $H(X,Y) \leq H(X) + H(Y)$

Entropy and Mutual Information