Basic Concepts in Information Theory

Fall 2019

Instructor: Shandian Zhe
zhe@cs.utah.edu
School of Computing

THE UNIVERSITY OF UTAH

Coding theory

- Let us start with discrete random variables

Coding theory

- How to represent the information contained in the random variables?

$$h(\mathbf{x}) \geq 0$$

$$h(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y})$$ ***x,y*** are independent

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

$$\Downarrow$$

$$h(\mathbf{x}) = -\log\left(p(\mathbf{x})\right)$$

Entropy

- The average among of information need to transmit

$$H(\mathbf{x}) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log \left(p(\mathbf{x}) \right)$$

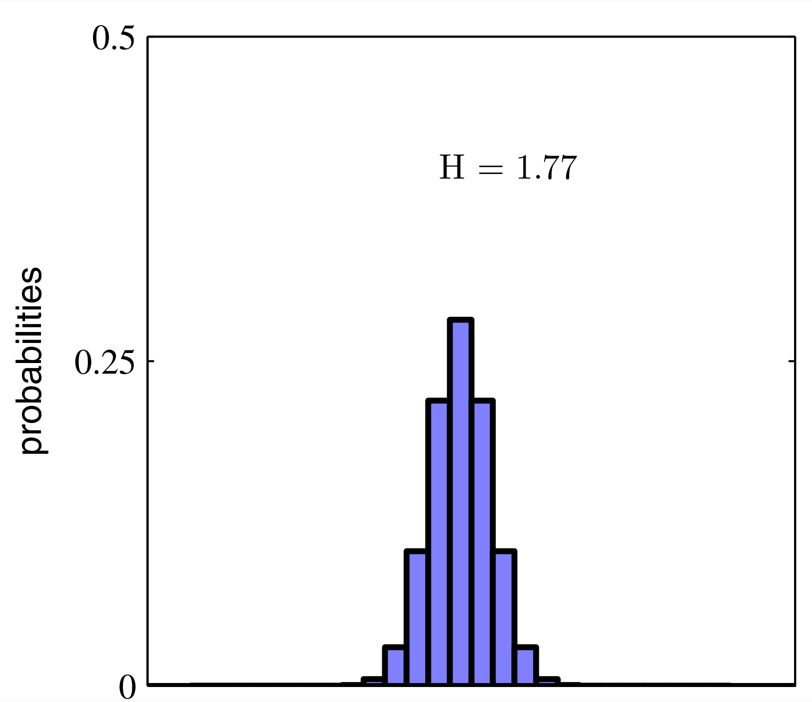
Entropy

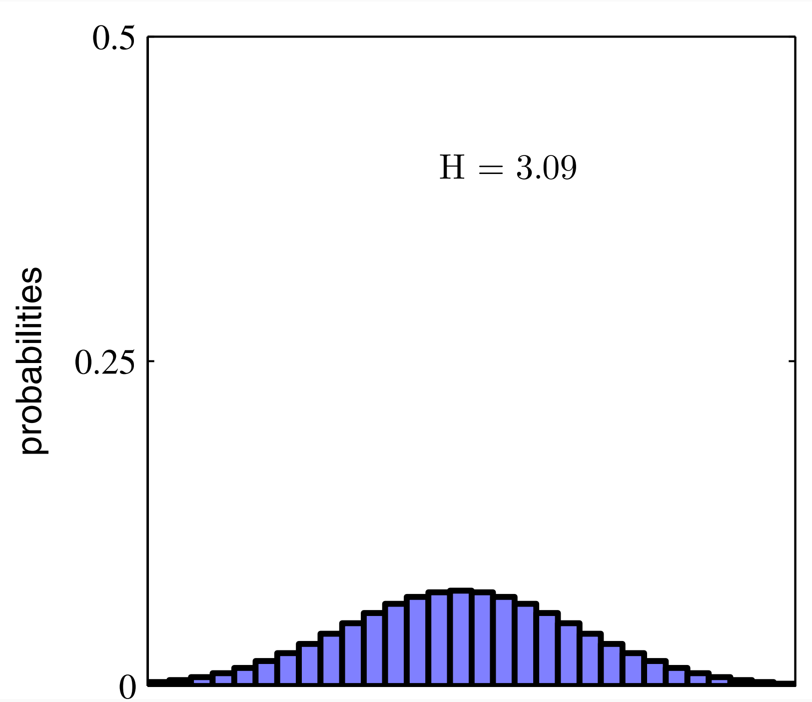
x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

Entropy is also the average code length

Entropy reflects uncertainty





Maximum entropy

- Consider a discrete R.V. with M possible status. We want to find the distribution has the the maximum entropy $\mathrm{H}[p] = -\sum_i p(x_i) \ln p(x_i).$

$$\widetilde{\mathrm{H}} = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right)$$

$$p(x_i) = 1/M$$

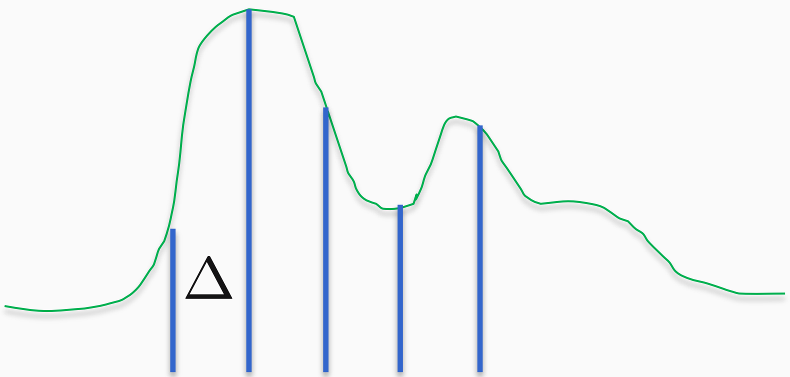
uniform distribution

Differential entropy

- Entropy is naturally defined on discrete random variables.
- But how about continuous variables?

Differential entropy

- Let us divide x into bins of Δ



Mean-value theorem

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\,\mathrm{d}x = p(x_i)\Delta$$

Entropy on discretized probability

$$\mathrm{H}_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta$$

$$\sum_i p(x_i)\Delta = 1$$

Differential entropy

$$H_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta$$

Goes to infinity
Throw out it

$$\lim_{\Delta \to 0} \left\{ \sum_i p(x_i)\Delta \ln p(x_i) \right\} = \int p(x) \ln p(x) \,\mathrm{d}x$$

$$H[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \,\mathrm{d}\mathbf{x}$$

Differential entropy

- The term that is thrown out reflects that to specify a continuous variable very precisely requires many many bits
- Note: differential entropy can be negative!

Differential entropy

- Given a continuous variable *x* with mean μ and variance σ^2, which distribution has the largest entropy?

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1 \\
\int_{-\infty}^{\infty} x p(x)\,\mathrm{d}x &= \mu \\
\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x &= \sigma^2.
\end{aligned}$$

Differential entropy

max

$$-\int_{-\infty}^{\infty} p(x) \ln p(x)\,\mathrm{d}x + \lambda_1 \left(\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x - 1 \right)$$

$$+\lambda_2 \left(\int_{-\infty}^{\infty} x p(x)\,\mathrm{d}x - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x - \sigma^2 \right)$$

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

Gaussian distribution!

Conditional entropy

- Given ***x***, how much information is left for ***y***

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

Prove it by yourself

Kullback-Leibler (KL) divergence

- Also called relative entropy

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}.\end{aligned}$$

If we use q to transmit information for p, how much extra information do we need

Kullback-Leibler (KL) divergence

- KL divergence is widely used to measure the difference between two distributions

$$\mathrm{KL}(p\|q) \geqslant 0$$

=0 iff p = q

Prove it with convexity
And Jensen's inequality

- However, it is not symmetric!

$$\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$$

KL Divergence

- KL divergence plays the key role in approximate inference
- All the deterministic approximate methods aim to minimize the KL divergence between the true and approximate posteriors (or in the reversed direction)
- In general, we have alpha divergence
- We will discuss these in detail later

Mutual information

How many information do the two random variables share?

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] \equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y}))$$
$$= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}$$

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$

Prove it by yourself

What you need to know

- Definition of entropy
- How is differential entropy is derived
- Entropy is an indicator for uncertainty
- KL divergence and properties (especially asymmetric)
- Mutual information