LING530F: Deep Learning for Natural Language Processing (DL-NLP)

Muhammad Abdul-Mageed

muhammad.mageed@ubc.ca

Natural Language Processing Lab

The University of British Columbia

Table of Contents

Many of the current slides are a summary Chapter 3 in Goodfellow et al. (2016). More information can be found therein. Note: The authors credit Pearl (1988) for a lot of the content of the chapter. Other sources used here are credited where approbriate.

Information Theory: Calude Shannon

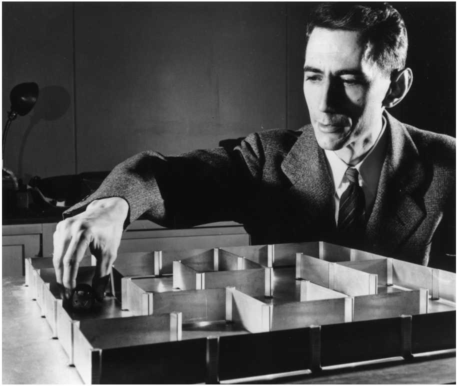

Figure: Claude Shannon. [From Time]. Check about Claude Shannon, e.g. short documentary [here] & lecture by Robert G. Gallager [here].

Information: A Book

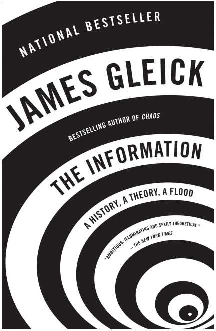

Figure: **Blurb**: A fascinating intellectual journey through the history of communication and information, from the language of Africa's talking drums to the invention of written alphabets; from the electronic transmission of code to the origins of information theory, into the new information age and the current deluge of news, tweets, images, and blogs. . .

What is information theory?

- Focused on **quantifying how much information is present in a signal**
- Originally **invented to study sending messages from discrete alphabets over a noisy channel**
- Communication via **radio transmission** is an example
- Answers **how to design optimal codes**
- Tells **how to calculate the expected length of messages** sampled from specific probability distributions

Information Theory: Basic Intuition

Intuition

- Learning that an **unlikely event** has occurred is **more informative** than learning that a likely event has occurred.
- “The sun rose this morning": **not informative enough** to send as a message
- “There was a solar eclipse this morning": **very informative**

Quantifying Information

Goal: Quantify Info. Such That:

- **Likely events**: have **low information content**, **events guaranteed to happen**: **no information content**
- **Less likely events**: higher information content.
- **Independent events**: have **additive information**. `Finding out that a tossed coin has come up as heads twice conveys twice as much information as finding out that a tossed coin has come up as heads once.`

- **Self-information** deals only with **a single outcome**.
- It is the **surprise** when a random variable is sampled.

1: Self-Information of Event X=x

$$\mathrm{I}(x) = -\log P(x)$$

Example of Self-Information

- When we toss a fair coin, P(x="head"=0.5), $\mathrm{I}(x = 0.5) = -\log_2 P(0.5) = 1$ **bit** of information.
- **Note**: If we use base e, then the unit of measurement is **nats**. (Above gives $\sim$ 0.693 nats).
- **Try it Python**: Base 2: -math.log(0.5,2); Base e: -math.log(0.5).

Shannon Entropy

- Quantify uncertainty in an entire distribution using **Shannon entropy**.
- **SE** of a distribution: **the expected amount of info. in an event drawn from that distribution**. (Denoted $H(P)$):

2: Shannon entropy

$$\textbf{Recall: self\_info.}: \mathrm{I}(x) = -\log P(x)$$

$$\mathrm{H}(x) = \mathbb{E}_{x \sim P}\big[\mathrm{I}(x)\big] = -\mathbb{E}_{x \sim P}\big[\log P(x)\big]$$

- Gives a **lower bound on the number of bits** (or nats) needed on avg to encode symbols drawn from a distribution P.
- **Nearly deterministic distributions**: have low entropy;
- Distributions **closer to uniform**: high entropy

Kullback–Leibler Divergence (KL Divergence) I

KL Divergence

- **Measures how one probability distribution is different from a second probability distribution**.
- **Always greater than or equal to zero**
- A smaller KL divergence value means we can expect more similar behavior of the two distributions.

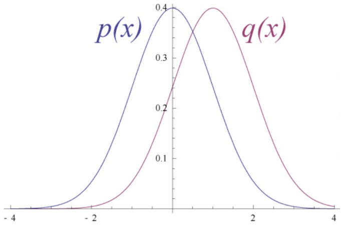

Figure: [From Wikipedia].

KL Divergence II

- With two prob distributions $P(x)$ and $Q(x)$ over the same r.v. x:

3: KL Divergence

$$\mathrm{D}_{KL}(P||Q) =$$

$$\mathbb{E}_{x \sim P}\Big[\log \frac{P(x)}{Q(x)}\Big] = \mathbb{E}_{x \sim P}\Big[\log P(x) - \log Q(x)\Big].$$

- For discrete variables, it is *the extra amount of info. needed* to send a message containing symbols drawn from prob distrib P, when we use a code designed to *minimize* the len of messages drawn from distrib Q.

Properties of KL Divergence

- KL divergence is **non-negative**.
- KL divergence is **not symmetric** (i.e., $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ (and so it is not a measure of distance).
- The **KL divergence is 0 if and only** if P and Q are the same distribution in the case of discrete variables, or equal “almost everywhere" in the case of continuous variables.

- Similar to the KL divergence, but lacking the term on the left:

4: Cross-Entropy

$$\mathrm{H}(\mathrm{Q}, \mathrm{P}) = -\mathbb{E}_{x \sim P} \log Q(x).$$

- **Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence**, because Q does not participate in the omitted term.