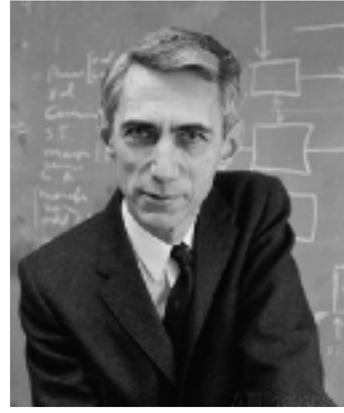
Similarity measures between distributions



Claude Shannon

- Shannon postulated that any measure of the informativeness of an event, x, should satisfy three conditions:
  1. An event with probability 1 yields no information
  2. The probability of an event and the information it yields vary inversely with each other
  3. The total information coming from independent events is purely additive

  - Which he used to define *self-information*: $I(x) = -\log P(x)$

- Shannon entropy: $H(P) = \mathbb{E}_{x\sim P}[I(x)] = -\mathbb{E}_{x\sim P}[\log P(x)] = -\sum_{x\sim P} P(x)\log P(x)$

- Kullback-Leibler (KL) divergence: $D_{KL}(P\,||\,Q) = \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right]$

- Cross entropy: $H(P,Q) = H(P) + D_{KL}(P\,||\,Q) = -\mathbb{E}_{x\sim P}[\log Q(x)] = -\sum_{x\sim P} P(x)\log Q(x)$

Machine learning problem formulation

- The machine learning approach expresses NLP as an optimization problem:

$$\hat{\mathbf{Y}} = \underset{\mathbf{y} \in f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})}{\operatorname{argmax}} \ \Psi(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})$$

where $\mathbf{x} \in X$ is the input

$\mathbf{y} \in Y$ is the output

$\Psi(\cdot) \rightarrow \mathbb{R}$ is a function expressing the learning objective

$f(\cdot)$ is the function, or model, that maps $\mathbf{x}$ to $\mathbf{y}$

$\boldsymbol{\theta}$ parameterizes $f(\cdot)$