KL divergence or relative entropy

Two pmfs $\mathrm{p}(x)$ and $\mathrm{q}(x)$:

$$D(\mathrm{p} \,\|\, \mathrm{q}) = \sum_{x \in \mathcal{X}} \mathrm{p}(x) \log \frac{\mathrm{p}(x)}{\mathrm{q}(x)} \tag{5}$$

Say $0 \log \frac{0}{\mathrm{q}} = 0$, otherwise $\mathrm{p} \log \frac{\mathrm{p}}{0} = \infty$.

$$D(\mathrm{p} \,\|\, \mathrm{q}) = E_{\mathrm{p}} \left(\log \frac{\mathrm{p}(X)}{\mathrm{q}(X)} \right) \tag{6}$$

$$I(X;\, Y) = D(\mathrm{p}(x, y) \,\|\, \mathrm{p}(x)\, \mathrm{p}(y)) \tag{7}$$

- Measure of how different two probability distributions are
- The average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q.
- $D(\mathrm{p} \,\|\, \mathrm{q}) \geq 0$; $D(\mathrm{p} \,\|\, \mathrm{q}) = 0$ iff $\mathrm{p} = \mathrm{q}$
- Not a metric: not commutative, doesn't satisfy triangle equality

[Slide on $D(p\|q)$ vs $D(q\|p)$]

Cross entropy

- Entropy = uncertainty
- Lower entropy = determining efficient codes = knowing the structure of the language = good measure of model quality
- Entropy = measure of surprise
- How surprised we are when w follows h is pointwise entropy:

$$H(w|h) = -\log_2 \mathrm{p}(w|h)$$

$\mathrm{p}(w|h) = 1$? $\mathrm{p}(w|h) = 0$

- Total surprise:

$$\begin{aligned} H_{\text{total}} &= -\sum_{j=1}^{n} \log_2 \mathrm{m}(w_j|w_1, w_2, \ldots, w_{j-1}) \\ &= -\log_2 \mathrm{m}(w_1, w_2, \ldots, w_n) \end{aligned}$$

Formalizing through cross-entropy

- Our model of language is $\mathrm{q}(x)$. How good a model is it?
- Idea: use $D(\mathrm{p} \parallel \mathrm{q})$, where p is the correct model.
- Problem: we don't know p.
- But we know roughly what it is like from a corpus
- Cross entropy:

$$\begin{aligned} H(X, \mathrm{q}) &= H(X) + D(\mathrm{p} \parallel \mathrm{q}) && (8) \\ &= -\sum_{x} \mathrm{p}(x) \log \mathrm{q}(x) \\ &= E_{\mathrm{p}}(\log \frac{1}{\mathrm{q}(x)}) && (9) \end{aligned}$$

- Cross entropy of a language $L = (X_i) \sim \mathrm{p}(x)$ according to a model m:

$$H(L, \mathrm{m}) = -\lim_{n\to\infty} \frac{1}{n} \sum_{x_{1n}} \mathrm{p}(x_{1n}) \log \mathrm{m}(x_{1n})$$

- If the language is 'nice':

$$H(L, \mathrm{m}) = -\lim_{n\to\infty} \frac{1}{n} \log \mathrm{m}(x_{1n}) \tag{10}$$

  I.e., it's just our average surprise for large n:

$$H(L, \mathrm{m}) \approx -\frac{1}{n} \log \mathrm{m}(x_{1n}) \tag{11}$$

- Since $H(L)$ is fixed if unknown, minimizing cross-entropy is equivalent to minimizing $D(\mathrm{p} \parallel \mathrm{m})$

- Providing: independent test data; assume $L = (X_i)$ is stationary [does't change over time], ergodic [doesn't get stuck]

Entropy of English text

27 letter alphabet

Model	Cross entropy (bits)	
zeroth order	4.76	$(\log 27)$
first order	4.03	
second order	2.8	
Shannon's experiment	1.3 (1.34)	

Perplexity

$$\begin{aligned}\text{perplexity}(x_{1n}, \text{m}) &= 2^{H(x_{1n},\text{m})} \\ &= \text{m}(x_{1n})^{-\frac{1}{n}}\end{aligned}$$