Eric Zacharia

1a. The expected value of the <u>Normal distribution</u> is its mean, μ, and its variance is its stdev squared, σ^2.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Entropy, $H = E\left(\lg_2 \frac{1}{p(x)}\right) = -\sum_x p(x) \lg_2 p(x)$ for a discrete r.v.

Plug

Entropy, $H(x) = -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx$ for a continuous r.v.

$$= -\int_{-\infty}^{\infty} f(x)\left(-\tfrac{1}{2}\ln \sigma\sqrt{2\pi} - \frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

$$= \int_{-\infty}^{\infty} f(x)\, \tfrac{1}{2}\ln \sigma\sqrt{2\pi} + f(x)\frac{(x-\mu)^2}{2\sigma^2}\, dx$$

$$= \tfrac{1}{2}\ln \sigma\sqrt{2\pi} \underbrace{\int_{-\infty}^{\infty} f(x)\, dx}_{1} + \frac{1}{2\cancel{\sigma^2}} \underbrace{\int_{-\infty}^{\infty} f(x)(x-\mu)^2\, dx}_{\cancel{\sigma^2}}$$

$$= \tfrac{1}{2}\ln \sigma\sqrt{2\pi} + \tfrac{1}{2} = 1.25 \ln \sigma + \tfrac{1}{2} = \underline{\ln \sigma + C}$$

1b. Cross entropy loss for the test sequence of words:

$$-\frac{1}{n}\sum_{i=K}^{n}\lg_2 \hat{P}(w_i|c_i)$$

Perplexity of a model (w.r.t. corpus)

$$PP(W) = P(w_1 w_2 \ldots w_n)^{-\frac{1}{n}} = \sqrt[n]{\frac{1}{P(w_1 w_2 \ldots w_n)}}$$

by the chain rule,

$$\sqrt[n]{\prod_{i=1}^{n}\frac{1}{P(w_i|w_1 w_2 \ldots w_{i-1})}}$$

w.r.t. the test sequence, this becomes

$$\sqrt[n]{\prod_{i=K}^{n}\frac{1}{P(w_i|c_i)}}, \text{ where } c_i = w_{i-K+1:i-1}$$

((K-1)-gram context)

Cross entropy loss is derived from the log perplexity.

$$\log PP(W) = \log\left(\sqrt[n]{\prod_{i=K}^{n}\frac{1}{P(w_i|c_i)}}\right)$$

$$= -\frac{1}{n}\sum_{i=K}^{n}\lg_2 P(w_i|c_i)$$

1c. $p(cw)$ = "The probability of the K-gram corresponding to $\underbrace{\text{context of size } K-1}_{c}$ followed by a $\underbrace{\text{word}}_{w}$."

$p(w|c)$ = "The conditional probability of seeing a $\underbrace{\text{word}}_{w}$ after seeing a $\underbrace{\text{context of size } K-1}_{c}$."

The entropy of English based on a K-gram language model:

$$H_K = -\sum_{c,w} p(cw) \lg_2 p(w|c)$$

Random process in which each word-context pair (w,c) is generated independently with prob. $p(cw)$:

$$G_K = -\sum_{c,w} p(cw) \lg_2 p(cw)$$

To say "In general, $H_K < G_K$" implies

$\Rightarrow$ "$\cancel{-}\cancel{\sum_{c,w} p(cw) \lg_2}\, p(w|c) \underset{>}{<} \cancel{-}\cancel{\sum_{c,w} p(cw) \lg_2}\, p(cw)$"

flip

$\Rightarrow$ implies "$p(w|c) > p(cw)$"

implies that seeing a word after its context is more probable than a K-gram that corresponds to a context followed by a word, and that makes sense.

1c.

$$H_K = -\sum_{c,w} p(cw) \lg_2 p(w|c) = -\sum_{c,w} p(cw) \lg_2 \frac{p(cw)}{p(c)} = -p(c) \lg_2 \frac{p(c)}{p(c)} = 0$$

$$G_K - G_{K-1} = -\sum_{c,w} p(cw) \lg_2 p(w_{i-K+1:i-1} w_i) - \left(-\sum_{c,w} p(cw) \lg_2 p(w_{i-K+2:i-1} w_i)\right)$$

$$= -p(c) \lg_2 p(w_{i-K+1:i-1}) + p(c) \lg_2 p(w_{i-K+2:i-1}) = 0$$

$$\therefore \underline{H_K = G_K - G_{K-1}}$$