Intro to LLMs (solution)

Activity 2

Q1)(a) When T approaches 1, $p(x_i|x_{1:i-1})^{1/T} \rightarrow p(x_i|x_{1:i-1})$. Implying that there is no effect to amplify or deamplify the distribution: It will just remain the same.

Q1)(b) The model will become too strict in prediction since the probability would become too deterministic. The model won't explore much randomly and will be straight-forward in generation.

Q1)(c) Infinity would just flatten the distribution entirely; All outcomes would become equally probable. This would be the case even after normalization.

Q2)(a) $$PP(x_{1:L}) = \exp\left(\frac{1}{L}\sum_{i=1}^{L} \log\left(\frac{1}{p(x_i|x_{1:i-1})}\right)\right)$$

"the cat sat"

$$PP(x_{1:3}) = \exp\left\{\frac{1}{3}\left[\sum_{i=1}^{3} \underbrace{\log\left(\frac{1}{0.4}\right) + \log\left(\frac{1}{0.6}\right) + \log\left(\frac{1}{0.8}\right)}_{\ln}\right]\right\}$$

$$PP(x_{1:3}) = \exp(0.55) \approx 1.7334$$

low perp → more confident prediction.

(b) $$PP(x_{1:3}) = \exp\left[\frac{1}{3}\left\{\ln(1/0.3) + \ln(1/0.4) + \ln(1/0.6)\right\}\right]$$

$$= \exp(0.87)$$

$$PP(x_{1:3}) \approx 2.404$$

(c) $\exp\left(\frac{1}{3} \times \left(\ln\left(\frac{1}{3}\right) + \ln\left(\frac{1}{3}\right) + \ln\left(\frac{1}{3}\right)\right)\right)$

$= \exp(\ln(3))$

$\boxed{PP(x_{1:3}) = 3}$

high perp → model is confused to predict next word.

(d) Perplexity measures branching factor, essentially the number of choices a model has to choose the next at each step. low perp means that the model is certain of its prediction for the next word and has less choices (more confident model) while it is vice versa for high perp.