引言

1.1 语言模型

定义：一种对词元序列的概率分布

自回归语言模型：基于之前生成的词元每次预测一个词元

$$x_i \sim p(x_i|x_{i-1})^{\frac{1}{T}}, i = 0, 1, \ldots, L, T\text{为温度}$$

注意：对条件分布应用温度参数T，并使用迭代采样（除非T=1）不等同于对长度为L的退火分布一次性采样

拓展&回忆：在知识蒸馏中也有使用温度参数

1.2 大模型相关历史回顾

熵：$H(p) = \sum_x p(x) log \frac{1}{p(x)}$

交叉熵：$H(p, q) = \sum_x p(x) log \frac{1}{q(x)}$

$N-gram$模型：$p(x_i|x_{1:i-1}) = p(x_i|x_{i-(n-1),i-1})$

交叉熵的上界是熵

即证$sup\mathbf{H}(p, q) = \mathbf{H}(p)$

$proof$：令$\mathbf{Q(x)} = \int_{-\infty}^{x} q(x)dx$

$\mathbf{H}(p, q) = \int_R p(x) ln \frac{1}{q(x)} = \int_R p(x) ln \frac{1}{Q^{'}(x)} dx$

根据变分法，得出$\delta(p(x) ln \frac{1}{Q^{'}(x)}) = 0$，又$Q(-\infty) = 0, Q(\infty) = 1$

$q(x) = Q^{'}(x) = p(x)$此时$H(p, q) = H(p)$又$\delta^2((p(x) ln \frac{1}{Q^{'}(x)})$半负定，此为最大值。证毕。

神经语言模型：RNNs,Transformers

大模型的能力

概述

在一些问题表现好，另一些任务表现不好

语言模型适应性

1.训练+微调

2.上下文学习（zero-shot,one-shot,few-shot）

困惑度

定义

$$P(X)=P(x_1,x_2,\dots,x_n)^{-\frac{1}{N}}=\exp\left(\frac{1}{L}\sum_i\log\frac{1}{p(x_i|x_{i-1})}\right)$$

直观理解：可以理解为每个$token$的平均分支因子$(branching\ factor)$
这里的分支因子可以理解为每个位置模型认为有多少种可能词出现（个人认为不太严谨）

两类错误

1.召回错误：把正确识别为错误，$P(X)\to\infty$
2.精确度错误：未能识别出错误，$P(X)$不会趋近无穷
把垃圾分布r混入p,得到q
$q(x_i|x_{i-1})=(1-\epsilon)p(x_i|x_{i-1})+\epsilon r(x_i|x_{i-1})$
由于$r(x_i|x_{i-1})>0\to q>(1-\epsilon)p\to p<\frac{q}{1-\epsilon}\approx(1+\epsilon)q$
$perlexity_q(x_{1:L})<(1+\epsilon)perplexity_p(x_{1:L})$
这里因为困惑度是$\frac{1}{p(x_i|x_{i-1})}$,$q(x_i|x_{i-1})$几何平均数
注：不等式在ϵ很小时近似取等