语言模型

- 给定文本序列 $x_1, x_2, \cdots$ 语言模型的目标是估计联合概率 $p(x_1, \cdots, x_T)$
- 语言模型的应用包括：
  - 做预训练模型（BERT，GPT等）
  - 生成文本（给定前面几个词，不断使用 $x_t \sim p(x_t \mid x_1, \cdots, x_{t-1})$ 来生成后续文本
  - 判断多个序列哪个更常见
- 使用计数来建模
  - 假设序列长度为2，我们预测 $p(x, x') = p(x) \cdot p(x' \mid x) = \frac{n(x)}{n} \cdot \frac{n(x, x')}{n(x)}$ 其中n是总词数，n(x)和n(x,x')是单个单词和连续词对的出现次数
  - 容易拓展到长度为3的情况：$p(x, x', x'') = p(x)\, p(x' \mid x)\, p(x'' \mid x, x') = \frac{n(x)}{n} \cdot \frac{n(x, x')}{n(x)} \cdot \frac{n(x, x', x'')}{n(x, x')}$
- N元语法（n-gram）
  - 序列很长时，因为文本量不够大很可能 $n(x_1, x_2, \cdots, x_T) \leq 1$ 使用马尔可夫假设可缓解这个问题
  - 一元语法：$p(x_1, x_2, x_3, x_4) = p(x_1)\, p(x_2)\, p(x_3)\, p(x_4)$
  - 二元语法：$= p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2) \cdot p(x_4 \mid x_3)$
  - 三元语法：$= p(x_1) \cdot p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2)\, p(x_4 \mid x_2, x_3)$