语言模型(上)

楔

倘若某日, 有人向你说道: Poincaré第七定理是复变函数中最精妙的定理. 此时你将如何回话? 正常人会请对方叙述Poincaré第七定理, 即便该定理应当不存在; 几乎无人会质疑句中的语法问题, 即便 `Poincaré第七定理` 一词未曾见于大众的语言库.

句中夹杂少数语言库外的词汇几乎不影响我们对语法的判断: 想必这是众所周知的结论. 就任一语言之形式而言, 词汇与语法不可或缺. 词汇协调于语法, 常人不会以自己所知的词汇库限制语言框架, 但一定会为接受源源不断的新名词而留足空间. 由是观之, 语言(尤其是语法)有兼容未知的名词之功能. "为表述未知事物而留足语言空间"是一项深入人心的技能. 在阐述语言模型之前, 先请比对端详下两则例子.

两则例子

例1 请试想如下教科书中的问题: 当一亩石塘中有大量鱼时, 如何快速估算鱼的数量N?

解 一类可行的方法是随机抓捕n条鱼并标记后放回, 当充分混合后随机抓捕M条鱼. 不妨记M条鱼中被做记号者数量为m, 则估算知石塘中约有$N \approx \dfrac{nM}{m}$条鱼. 为同时兼顾精度与效率, 应合理把控M与n的选值.

例2 现适当转化问题: 置有限数量的彩糖于罐中, 现随机取出3颗, 若1颗为红, 另2颗为绿. 今若再取一糖, 糖为红的概率是否为$\dfrac{1}{3}$?

解 非然. 问题之所在显然: 可能存在其他颜色的糖. 从而所求之概率应为某一不超过$\dfrac{1}{3}$的值.

Good–Turing估计之思想

楔中两例有殊. 究其根本, 是以**例1**之实验者已然于全知视角下, 而**例2**之实验者仍置身局中. Good–Turing Estimate之核心如下: 对于未被观测到的事件, 不可认为其发生概率为0. 如Ле́нин所言, 物质是标志客观实在的哲学范畴，这种客观实在是人通过感觉感知的，它不依赖于我们的感觉而存在，为我们的感觉所复写、摄影、反映。

就**例2**讨论, 考虑颜色非红或绿的球, 则红球实际所占的比例应为$\dfrac{1}{3+\delta}$, 绿球实际所占的比例应为$\dfrac{2}{3+\delta}$, 其中δ为某一常数. 由是观之, 概率密度函数渐趋平滑. Good–Turing估计之基本思想如此, 并给出了"最优"估计法.

语言模型

基本概念

记一句话$s = w_1 w_2 \cdots w_m$的先验概率为

$$p(s) = p(w_1)p(w_2 \mid w_1) \cdots p(w_k \mid w_1 w_2 \cdots w_{k-1}) \cdots p(w_m \mid w_1 \cdots w_{m-1}).$$

为简便故, 不妨设$p(w_k \mid w_1 w_2 \cdots w_{k-1})$仅与$\{w_1, \ldots, w_{k-1}\}$有关, 即与前$k-1$基元之具体顺序无关. 由于$w_d$对$w_{d+l}$之影响随$l$之递增而式微, 可再不妨设对$m > l$均有

$$p(w_m \mid w_1 \cdots w_{m-1}) = p(w_m \mid w_{m-l} \cdots w_{m-1}).$$

一般称上述者为Марков假设. 所得的语言模型为l元文法模型(l-gram), 或称作l阶Марков链. 以语句 `我是人类` 为例, 先分解之为如下五部分

$$\texttt{< BOS >}\ 我\ 是\ 人类\ \texttt{< EOS >}.$$

其中< BOS >与< EOS >分别为为起止符, 例如$p(我 \mid \texttt{< BOS >})$为 `我` 字作为起头词之概率. 该句话基于二元文法出现的概率为

$$p(我 \mid \texttt{< BOS >})p(是 \mid 我)p(人类 \mid 是)p(\texttt{< EOS>} \mid 人类).$$

应注意到, Марков假设之局限性:

1. 记$|V|$为词汇量, 则模型之复杂度为$O(|V|^N)$.
2. 自然语言中上下文的相关性跨度悬殊(如语段间联系), 从而过高的n无多裨益.

对汉字而言, 四元语法模型较为合适, 微软拼音即使用之.

数据平滑之必要性

若假定采用2-gram. 今给定训练材料三则:

1. Newton 是 伟大的 物理学家.
2. Hooke 系 物理学家
3. Newton 站在 Hooke 肩膀上

从而由全概率公式知$p(是 \mid \mathrm{Newton}) = 1$. 但注意到$p(是 \mid \mathrm{Hooke}) = 0$, 从而

$$p(\mathrm{Hooke}是伟大的物理学家) = 0.$$

科学史研究工作坚决杜绝历史虚无主义; 上述式子是数据匮乏/稀疏(Sparse Data)引起的零概率怪相.

阅至此处, 读者大概能看出下一步了: 利用某些手段平滑数据, 重点当然为Good–Turing估计.

数据平滑化原则

数据平滑化应遵循如下原则:

1. 减少方差, 即"劫富济贫".
2. 降低困惑度.
3. 保持归一原则, 即事件之所有可能情况概率和为1.

其中, 设语料T由句子$t_1, \ldots, t_m$组成, W_T为T之词数, 则语料之交叉熵为

$$H_p(T) = -\frac{1}{W_T}\sum_{i=1}^{m}\log p(t_i).$$

所谓困惑度即$2^{H_p(T)}$.

> 注: n-gram 对于英语文本的困惑度范围一般为 50 ~ 1000, 对应于交叉熵范围为 6 ~ 10 bits/word.
>
> 注: 信息学中, 若未加说明, $\log$作 $\log_2$ 处理.

加一法

一类平滑化方式为加一法. 该方法原理简单, 对所有情形出现之次数加上某一常数(简便起见取1)即可, 例如将比例$\frac{1}{2} : \frac{1}{3} : \frac{1}{6}$调整至

$$\frac{1+1/2}{3} : \frac{1+1/3}{3} : \frac{1+1/6}{3} = \frac{1}{2} : \frac{4}{9} : \frac{7}{18}$$

相当于一次分式线性变换. 照此,

$$p(是 \mid \mathrm{Newton}) = \frac{1+1}{1+8} = \frac{2}{9}.$$

及

$$
\begin{aligned}
&p(\text{Hooke是伟大的物理学家})\\
=&p(\text{Hooke} \mid \texttt{<BOS>}) \cdots p(\texttt{<EOS>} \mid \text{物理学家})\\
=&\frac{1+1}{3+8} \cdot \frac{0+1}{2+8} \cdot \frac{1+1}{1+8} \cdot \frac{1+1}{1+8} \cdot \frac{1+1}{1+8}\\
\approx&0.0002
\end{aligned}
$$

但语句`肩膀上是物理学家`之显现概率高达0.0025, 不可不言不妥.

折扣法

该方法思想简单, 减少已出现事件之计数, 将多余的"名额"分配给未出现的事件. 现将Good–Turing估计概述如下:

(Good–Turing) 设样本体积为N(此处为所有n-gram), n_r表示出现r次的样本总数, 则

$$
N = \sum_{r=1}^{\infty} n_r \cdot r = \sum_{r=0}^{\infty} (r+1) \cdot n_{r+1}.
$$

从而可记$q_r = (r+1)\dfrac{n_{r+1}}{n_r}$使得

$$
N = \sum_{r=0}^{\infty} n_r \cdot q_r.
$$

Good-Turing 估计适用于大词汇集产生的符合多项式分布的大量的观察数据. 推导如下:

不妨设$X = (X_1, \ldots, X_N)$为总样本, $s = 1, \ldots, K$为所有样本类型. 记$P_s = P(X = s)$. 定义特征函数$\mathbf{1}$, 从而类型s之总数为

$$
C(s) = \sum_{i=1}^{N} \mathbf{1}_{\{X_i = s\}}.
$$

从而

$$
n_r = \sum_{s=1}^{K} \mathbf{1}_{\{C(s) = r\}}.
$$

考虑事件$\{q_r = p_s\}$, 记事件A_s代表所选的样本类型为s, 事件B_r代表所选样本所属的类型共包含r个样本. 因此$\{q_r = p_s\} = A_s \mid B_r$. 从而

$$
\begin{aligned}
E(q_r) &= \sum_{s=1}^{K} p_s P(q_r = p_s) \\
&= \sum_{s=1}^{K} p_s P(A_s \mid B_r) \\
&= \sum_{s=1}^{K} p_s \frac{P(A_s)P(B_r \mid A_s)}{\sum_{k=1}^{K} P(B_r \mid A_k)P(A_k)} \\
&= \sum_{s=1}^{K} p_s \frac{P(A_s)P(C(s)=r)}{\sum_{k=1}^{K} P(C(s)=r)P(A_k)} \\
&= \sum_{s=1}^{K} p_s \frac{P(A_s)\binom{N}{r}p_s^r(1-p_s)^{N-r}}{\sum_{k=1}^{K}\binom{N}{r}p_k^r(1-p_k)^{N-r}P(A_k)} \\
&= \frac{\sum_{s=1}^{K} p_s^{r+1}(1-p_s)^{N-r}}{\sum_{s=1}^{K} p_s^{r+1}(1-p_s)^{N-r}}
\end{aligned}
$$

由于既有样本数为N, 为$E(q_r)$添加角标, 得$E_N(q_r)$. 注意到

$$
\begin{aligned}
E_N(n_r) =& E_N \sum_{s=1}^{K} \mathbf{1}_{\{C(s)=r\}} \\
=& \sum_{s=1}^{K} E_N(\mathbf{1}_{\{C(s)=r\}}) \\
=& \binom{N}{r} \sum_{s=1}^{K} p_s^r (1-p_s)^{N-r}
\end{aligned}
$$

同理有

$$
E_N(n_{r+1}) = \binom{N}{r+1} \sum_{s=1}^{K} p_s^{r+1}(1-p_s)^{N-r-1}.
$$

以及

$$
E_{N+1}(n_{r+1}) = \binom{N+1}{r+1} \sum_{s=1}^{K} p_s^{r+1}(1-p_s)^{N-r}.
$$

从而

$$
\begin{aligned}
E_N(q_r) &= \frac{\binom{N}{r}E_{N+1}(n_{r+1})}{\binom{N+1}{r+1}E_N(n_r)} \\
&= \frac{r+1}{N+1}\frac{E_{N+1}(n_{r+1})}{E_N(n_r)} \\
&\overset{N\gg 1}{=} \frac{r+1}{N}\frac{n_{r+1}}{n_r}
\end{aligned}
$$

是故Good–Turing估计原理得证.