0. 写在前面

1 本课程总体结构

章节		教学内容
第一章 引言 (刘均, 2)		概念与研究背景；主要任务；挑战与研究方向；相关资源
第二章：自然语言的统计特性 (刘均, 1)		Zipf定律、Heaps定律、Benford 定律。
第三章：语言模型	词袋模型 (刘均, 3)	语言模型；词袋模型（BoW）；TF-IDF。 NLU任务：情感分析、文本聚类。
	概率语言模型 (李辰, 6)	概率语言模型；n-gram 模型；最大似然估计；平滑技术。 NLU任务：分词、语义关系抽取。
	主题模型 (刘均, 6)	生成模型；主题模型的图表示；LSA、PLSA、LDA；NMF等。 NLU任务：话题检测、推荐。
	神经网络语言模型 (李辰, 6)	分布式表示；C&W、CBOW、Skip-Gram、Glove等。 NLU任务：对话、实体消歧。
第四章：机器翻译	概述 (李辰, 1)	面临的挑战；发展历程；方法类别及特点；MT评估。
	统计机器翻译 (李辰, 3)	统计MT；Noisy Channel模型；IBM模型。
	神经网络机器翻译 与大语言模型 (刘均, 4)	RNN与LSTM简介；Encoder-Decoder框架；Attention模型；大语言模型。

- 这门课由于由两位老师授课，个人感觉结构比较混乱
- 由于时间紧任务重经费无，所以笔记还是按PPT内容和以上结构展开，即使有很多不合理的地方

2 考试有关事项

1. 概率模型

1.1. 马可夫模型

1.1.1. 概述

1 Markov模型概述

1. 定义：描述系统状态的统计模型，用过去一段时间的状态描述当前状态(而忽略更早状态)
2. 类型：

类型	含义	示例
显马可夫模型(VMM)	状态是可以直接观察到	天气预报(晴/雨/雪)
隐马可夫模型(HMM)	状态只能通过直接观测到的值推断	语音识别(声波→发音)

2 Markov模型的数学描述

1. 状态描述：系统有$S_1, S_2, ..., S_N$共N个状态/任意t时的状态表示为q_t；例如$q_t=S_j$
2. 马可夫假设：
   - k阶马可夫：
     $$P\left(q_t=S_j|q_{t-1}=S_{i_1}, q_{t-2}=S_{i_2}, ...\right)=P\left(q_t=S_j|q_{t-1}=S_{i_1}, ..., q_{t-k}=S_{i_k}\right)$$
     $$\Big\Downarrow k=1(\text{一阶马可夫})$$
     $$P\left(q_t=S_j|q_{t-1}=S_{i_1}, q_{t-2}=S_{i_2}, ...\right)=P\left(q_t=S_j|q_{t-1}=S_{i_1}\right)$$
   - 时间无关：让上述过程独立于时间，即$\forall t$都有$P\left(q_t=S_j|q_{t-1}=S_{i_1}, ..., q_{t-k}=S_{i_k}\right)$
3. 状态转移概率：
   - 含义：一阶马可夫中，从i状态转移到j状态的概率，即$P(q_t=S_j|q_{t\text{-}1}=S_i)=a_{ij}$
   - 性质：$\sum_{j=1}^{N} a_{ij}=1$

1.1.2. 隐马可夫过程及Viterbi算法

0 问题描述

1. 一个隐马可夫的示例：From Wikipedia

Item	描述
条件	只考虑三种活动(散步/购票/购物)，两种天气(阴/晴)
假设	一位朋友，每天告诉你他的活动
问题	在他告诉你每天所做事情基础上，你要猜他那边的天气

2. 两种空间：观测$\xrightarrow{\text{预测}}$状态

空间	符号	示例
观测空间	$O=\{o_1,o_2,\dots,o_T\}$	朋友的活动
状态空间	$X=\{x_1,x_2,\dots,x_T\}$	朋友那边的天气

3. 符号：三种矩阵$\mathbf{A}/\mathbf{B}/\boldsymbol{\pi}$的项

符号	含义	示例
初始状态概率π_{x_i}	初始状态($t=0$时状态)为x_i的概率	第一天的天气
状态转移概率$a_{x_ix_j}$	从一状态x_i转化为下一状态x_j的概率	今天晴天明天阴天的概率
符号发射概率$b_{x_io_j}$	从状态x_j从而导向某种观测o_j的概率	晴天时朋友散步的概率

1 问题1：给定$\mu=(A,B,\pi)$求解观察序列$O=\{o_1,o_2,\dots,o_T\}$发生的概率$P(O|\mu)$

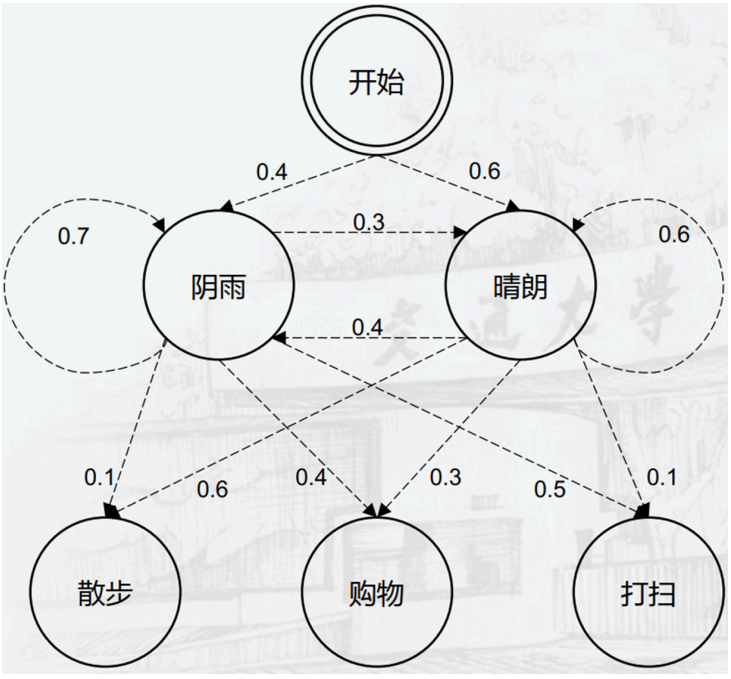
1. 分解：$P(O|\mu)\xrightarrow{\text{边缘化}}\sum_X P((O,X)|\mu)\xrightarrow{\text{分解}}\sum_X P(O|(X,\mu))\times P(X|\mu)$

2. 代入：

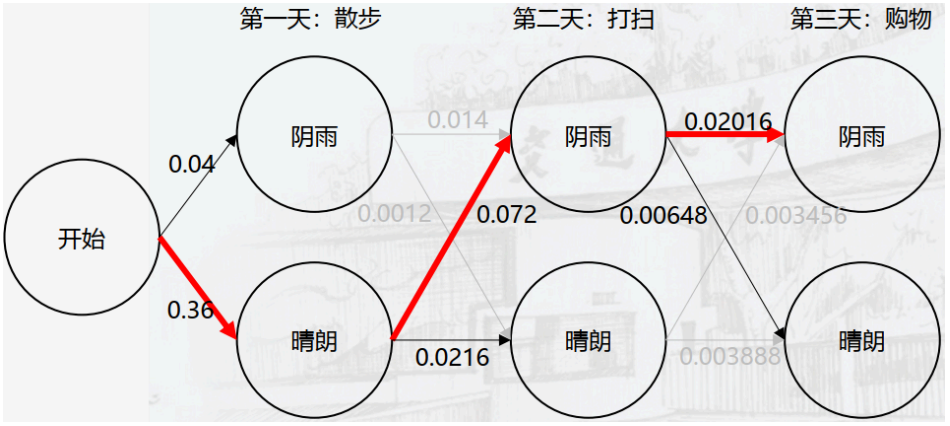
$$\mathrm{P}(O|\mu)\xrightarrow[\mathrm{P}(O|(X,\mu))=b_{x_1o_1}\times b_{x_2o_2}\times\ldots\times b_{x_To_T}]{\mathrm{P}(X|\mu)=\pi_{x_1}\times a_{x_1x_2}\times a_{x_2x_3}\times\ldots\times a_{x_{T-1}x_T}}\sum_{x_1\to x_T}\left(\pi_{x_1}\times b_{x_1o_1}\times\prod_{t=1}^{T-1}a_{x_tx_{t+1}}\times b_{x_{t+1}o_{t+1}}\right)$$

2 问题2：给定参数$\mu=(A,B,\pi)$与观察所得$O=\{o_1,o_2,\dots,o_T\}$，最大化$P(X|(O,\mu))$

1. 含义：找到最可能的状态序列$X=\{x_1,x_2,\dots,x_T\}$，以解释观测
2. 求解：Viterbi算法，计算每个时刻t的最路径概率，最终找出整体最优路径；示例如下
   - 转换概率：状态↔状态/状态→观测



   - Viterbi算法路径：



3 问题3：观测到O，如何调整$\mu=(A,B,\pi)$最大化$P(O|\mu)$

1. 期望值最大化算法(EM)
2. 最大似然估计(MLE).....

1.2. 贝叶斯模型

1 Bayes定律

1. 不展开形式：$P(A|B)=\dfrac{P(A)P(B|A)}{P(B)} \Leftarrow \begin{cases} P(A|B)=\dfrac{P(AB)}{P(B)} \\ P(B|A)=\dfrac{P(AB)}{P(A)} \end{cases}$

参数	含义	示例
$P(A)/P(B)$	先验概率，即对A/B发生的经验推断(臆测)	发病率历史数据
$P(A\|\|B)/P(B\|\|A)$	后验概率，即已知B/A后A/B发生的概率	检测后个体的患病概率

2. 展开后：

$$P(A|B)=\frac{P(A)P(B|A)}{\sum_j P(B|A_j)P(A_j)} \Leftarrow P(B)=\sum_j P(BA_j)=\sum_j P(B|A_j)P(A_j)$$

2 朴素贝叶斯分类器

1. 模型描述：
   - 假设：决定各分类的属性之间是相互独立(简单粗暴/但也损失了分类精度)
   - 前提：样本$X(a_1,a_2,\dots,a_n)$有n个属性$\{A_1,A_2,\dots,A_n\}$且有m个类$\{C_1,C_2,\dots,C_m\}$
   - 分类：将$X(a_1,a_2,\dots,a_n)$归类为$C_i \overset{\text{等价}}{\Longleftrightarrow} P(C_i|X)>P(C_j|X)$，其中$C_j$为除$C_i$任一类
2. 模型分析：
   - $P(C_i|X)=\dfrac{P(X|C_i)P(C_i)}{P(X)} \xrightarrow{P(X)\text{为常数}}$最大化$P(X|C_i)P(C_i)$以分类
   - $P(X|C_i)P(C_i)=P(A_1=a_1,A_2=a_2,\cdots,A_n=a_n|C_i)P(C_i) \xrightarrow{\text{各属性独立}} \prod_{k=1}^{n} P(a_k|C_i)P(C_i)$
3. 模流程：

阶段	操作
准备	确定样本的属性，获取相应的样本
训练	对每个类别计算$P(C_i)$→对每个属性计算$P(a_k\|C_i)$
应用	对新样本Λ计算其对每个类别的$P(\Lambda\|C_i)P(C_i)$→找到使之最大的C_i以归类之

4. 示例：判断学历为大学，年薪30-40，薪水20000-30000的员工的性别

样本	性别	工作内容	学历	年龄	薪水
1	女	送货	大学	$20-30$	$20000-30000$
2	男	包装	大学	>40	>40000
3	男	烘烤	大学	$30-40$	$20000-30000$
4	男	包装	高中	$30-40$	$20000-30000$
5	男	送华	大学	>40	$30000-40000$
6	女	烘烤	高中	$20-30$	$20000-30000$
7	男	烘烤	大学	$20-30$	<20000
8	女	包装	大学	$30-40$	$20000-30000$
9	男	烘烤	大学	>40	$20000-30000$
10	男	包装	大学	$20-30$	<20000

   - $\begin{cases} P(\text{包装}\|\text{女})\times P(\text{大学}\|\text{女})\times P(\text{30-40}\|\text{女})\times P(\text{20000-30000}\|\text{女})\times P(\text{女})=0.0222 \\ P(\text{包装}\|\text{男})\times P(\text{大学}\|\text{男})\times P(\text{30-40}\|\text{男})\times P(\text{20000-30000}\|\text{男})\times P(\text{男})=0.0315 \end{cases}$
   - 根据给定条件，应归类为男性

1.3. 平滑技术

1 基本概念

1. 稀疏问题：某些变量在训练集未出现(但可能在测试集出现)→概率被估计为0→为后续计算造成麻烦
2. 平滑概念：为所有分配一个非零概率→解决稀疏问题→提高模型泛化能力

2 Additive平滑

1. 方法：普通(频率估计的)概率$p_i=\frac{x_i}{N}\xrightarrow{\text{Additive平滑}}p_{i,\alpha\text{-smoothed}}=\frac{x_i+\alpha}{N+\alpha}$
2. 参数：d是可能变量的总数，α为平滑参数($\alpha=1$时即变为Laplace变换)

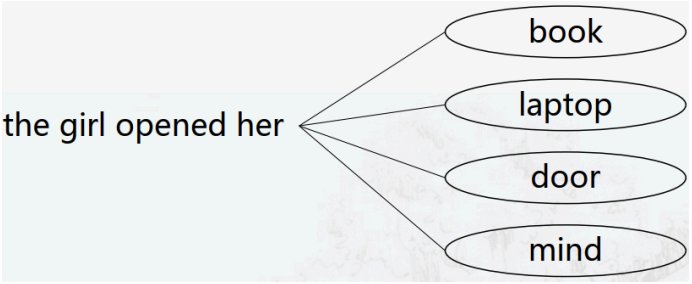
3 图灵(Good-Turing)估计

1. 方法：原有计数$r\xrightarrow{\text{调整}}r^*=(r+1)\frac{n_{r+1}}{n_r}$，其中$n_r$表示计数为$r$的变量个数
2. 示例：若鱼类及数目为3 perch/2 white/1 trout/1 salmon/1 eel→trout数目估计为$(1+1)\frac{1}{3}$

2. 语言模型概述

2.1. 基本概念

1️⃣语言模型含义：计算一个句子的概率的概率模型，如已知一个句子已有的词预测下一个词



2️⃣数学描述：对句子$\{x^{(1)}, x^{(2)}, ..., x^{(t)}\}$

1. 对单词：计算下一词出现的概率分布$P\left(x^{(t+1)}|x^{(t)}, ..., x^{(1)}\right)$
2. 对句子：计算句子出现的概率分布$P\left(x^{(1)}, ..., x^{(T)}\right)=\prod_{t=1}^{T} P\left(x^{(t)}|x^{(t-1)}, ..., x^{(1)}\right)$

2.2. N-gram概率模型

1️⃣N-gram含义：以The girl opened her book为例

N-gram	含义	示例
Unigrams	单个词	The, girl, opened, her, book
Bigrams	两个连续的词	The girl, girl opened, opened her, her book
Trigrams	三个连续的词	The girl opened, girl opened her, opened her book
4-grams	四个连续的词	The girl opened her, girl opened her book

2️⃣N-gram概率计算

1. 假设：$P\left(x^{(t+1)}|x^{(t)}, x^{(t-1)}, \ldots, x^{(1)}\right)\approx P\left(x^{(t+1)}|x^{(t)}, \ldots, x^{(t-n+2)}\right)$即当前词仅依赖前$n$-1词
2. 概率：$P\left(x^{(t+1)}|x^{(t)}, ..., x^{(t-n+2)}\right)\approx\frac{\text{count}\left(x^{(t+1)}, x^{(t)}, ..., x^{(t-n+2)}\right)}{\text{count}\left(x^{(t)}, ..., x^{(t-n+2)}\right)}$即
   $\frac{\text{N-gram数量}}{\text{(N-1)-gram数量}}$
3. 示例：$P(\text{book}|\text{girl opened her})=\frac{\text{count(girl opened herbook)}}{\text{count(girl opened her)}}$

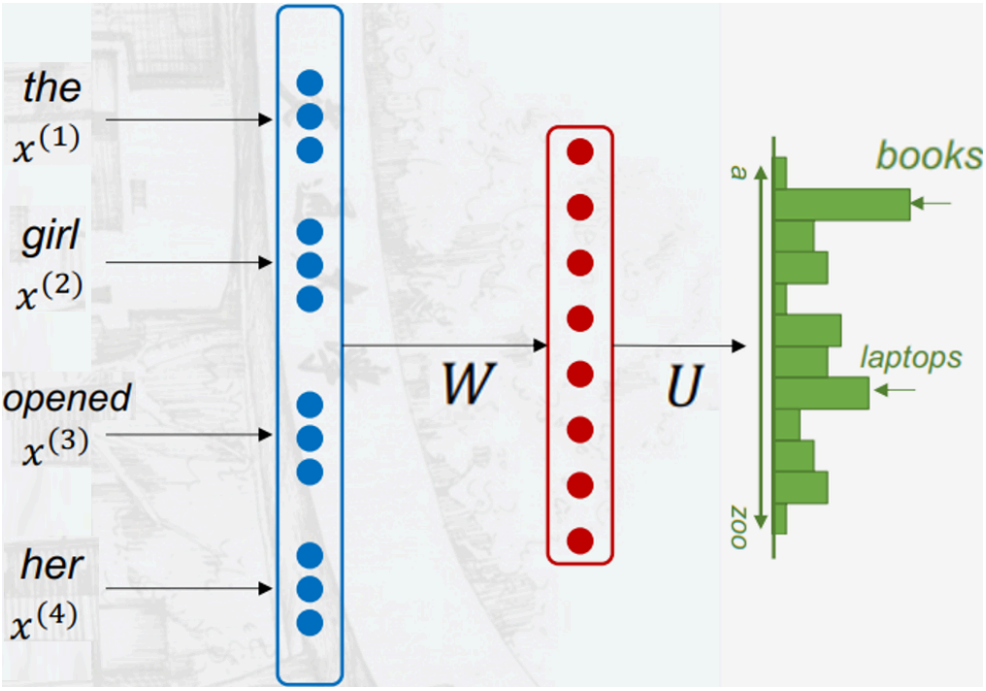
3️⃣稀疏性问题：

处理	示例	处理
平滑化	w_1, w_2, w_3, w_X不在语料库	$P(w_X|w_1, w_2, w_3)=\frac{\text{count}(w_1, w_2, w_3, w_X)+1}{\text{count}(w_1, w_2, w_3)+V}$

处理	示例	处理
回退	w_1, w_2, w_3不在语料库	用$w_1, w_2,$ __的3-gram代替w_1, w_2, w_3的3-gram

2.3. 神经网络语言模型简述

1️⃣结构&流程



1. 词向量：通过one-hot编码/分布式表示等，得到$\{x^{(1)}, x^{(2)}, ..., x^{(t)}\}$
    - 在基于窗口的神经网络中词向量通常为定长(窗口)，而RNN正是为了处理任意长度句子而生
2. 词嵌入：$\{x^{(1)}, x^{(2)}, ..., x^{(t)}\} \xrightarrow{\text{embedding}} \{e^{(1)}, e^{(2)}, e^{(3)}, ...\}$
3. 隐藏层：获得$h=f(We+b_1)$
4. 输出层：获得$\hat{y}=\text{Softmax}\left(Uh+b_2\right) \in \mathbb{R}^{|V|}$

2️⃣相比于N-gram：没有稀疏性问题，无需存储所有观察到的N-gram