NLU 的基础概率模型

李辰

2024年9月

下面三次课

课程提纲

一、隐马尔可夫模型

二、贝叶斯模型

三、平滑技术

课程提纲

一、隐马尔可夫模型

二、贝叶斯模型

三、平滑技术

隐马尔可夫模型

为什么是“隐”？

什么是“马尔可夫模型”？

马尔可夫模型

- 马尔可夫模型（Markov Model）是一种统计模型，广泛应用在词性自动标注，概率文法，语音识别，音字转换等多个自然语言处理等应用领域。
- 隐性马尔可夫模型（HMM）在现代人工智能系统中仍有广泛应用。
- 为了区别传统的 HMM，一般把显性马尔可夫模型（VMM）称为马尔可夫模型。

定义

- 如果一个系统有 N 个状态 $S_1, S_2, \dots S_N$。用 q_t 表示系统在 t 时间的状态变量，那么 t 时刻的状态取值为 S_j $(1 < j < N)$。
- 假设：系统在 t 时刻的状态只与其在 $t-1$ 时刻的状态有关，则该系统构成一个离散的一阶马尔可夫链

$$\mathrm{P}(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \dots) = \mathrm{P}(q_t = S_j | q_{t-1} = S_i)$$

- 如果只考虑上述公式独立于时间 t 的随机过程（不动性假设），状态与时间无关，那么：

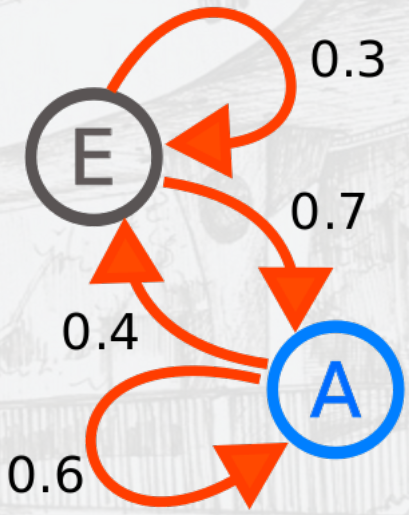
$$\mathrm{P}(q_t = S_j | q_{t-1} = S_i) = a_{ij}, 1 \le i, j \le N$$

该随机过程称为马尔可夫模型。

- 在马尔可夫模型中，状态转移概率 a_{ij} 必须满足下列条件：

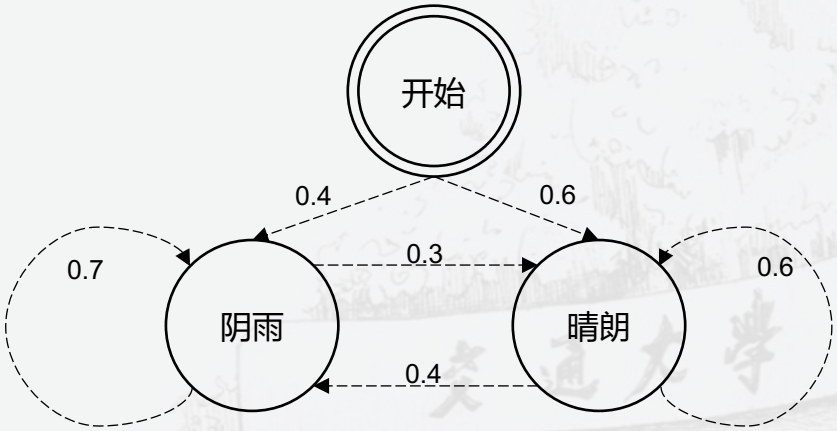
$$a_{ij} \ge 0$$

$$\sum_{j=1}^{N} a_{ij} = 1$$



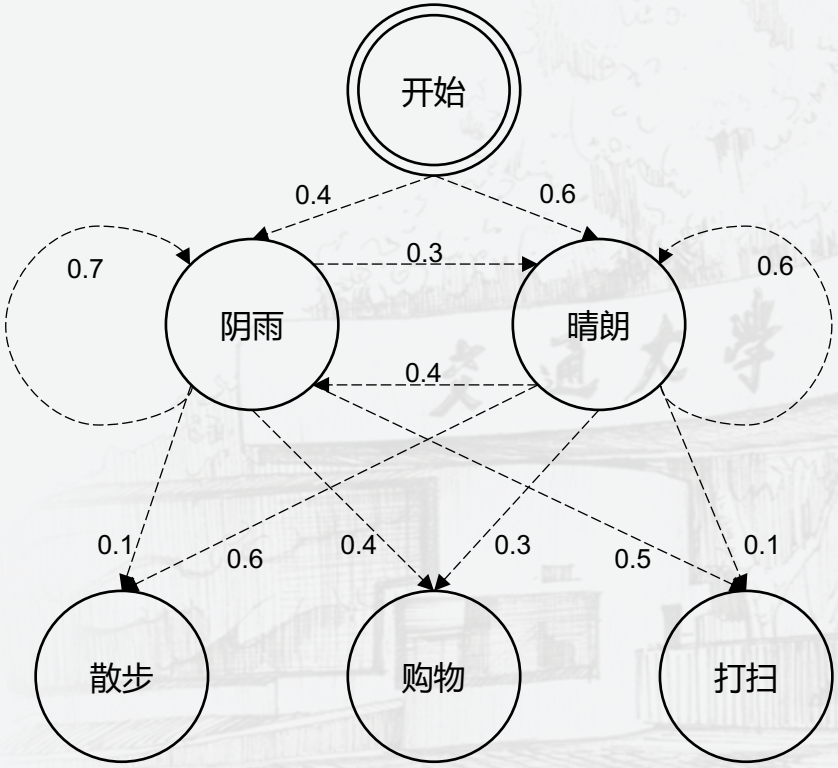
举例

- 假设：有一个住得很远的朋友，他每天跟你打电话告诉你他那里的天气。
- 条件：天气只有阴雨和晴朗两种情况。天气的转换只与前一天的天气有关。
- 问题：在他告诉你每天天气的基础上，你想要计算他那里天气情况的转换概率。



隐马尔可夫模型

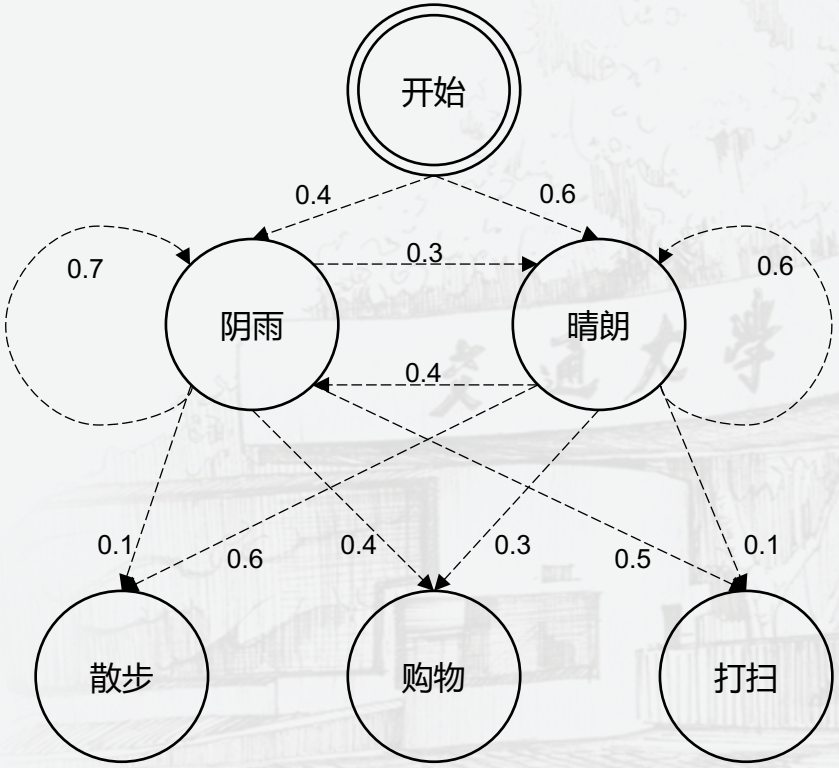
- 假设：有一个住得很远的朋友，他每天跟你打电话告诉你他那天做了什么。
- 条件：他仅对三种活动感兴趣：散步，购物以及打扫房间。他决定做什么事情只凭天气。
- 问题：在他告诉你每天所做的事情基础上，你想要猜测他所在地的天气情况。



Ref: https://en.wikipedia.org/wiki/Hidden_Markov_model

隐马尔可夫模型

- 状态数 $N = 2$ —— 天气种类
- 每个状态可能输出的不同符号数 $M = 3$ —— 活动种类
- 状态转移概率矩阵 $\mathrm{A} = a_{ij}$ —— 从一种天气 S_i 转向另一种天气 S_j 的概率
- 符号发射概率矩阵 $\mathrm{B} = \mathrm{b}_j(k)$ —— 从第 j 中天气从事第 k 种活动的概率
- 初始状态概率分布 π。



Ref: https://en.wikipedia.org/wiki/Hidden_Markov_model

隐马尔可夫模型

三个基本问题

1. 在给定模型 $\mu = (A, B, \pi)$，怎样计算某个观察序列发生的概率，即 $P(O|\mu)$?

$$P(O|X,\mu) = b_{x_1 o_1} b_{x_2 o_2} \ \dots \ b_{x_T o_T}$$

$$P(X|\mu) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} \ \dots \ a_{x_{T-1} x_T}$$

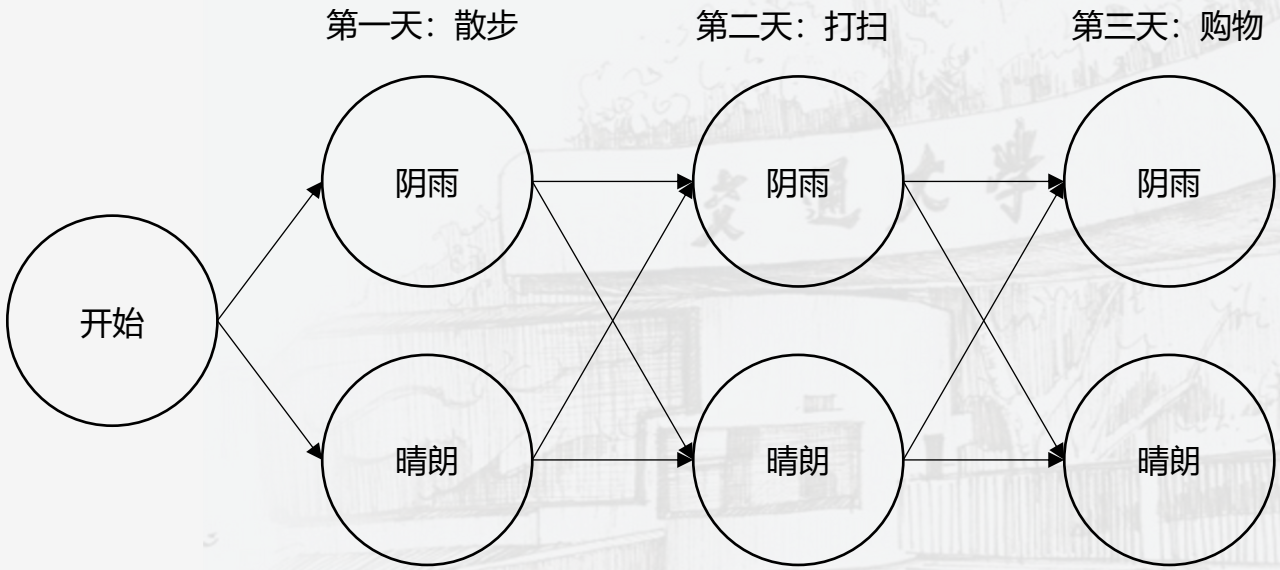
$$P(O,X|\mu) = P(O|X,\mu)P(X|\mu)$$

$$P(O|\mu) = \sum_{\{x_1,,,x_T\}} \pi_{x_1} b_{x_1 o_1} \prod_{t=1}^{T-1} a_{x_t x_{t+1}} \ b_{x_{t+1} o_{t+1}}$$

隐马尔可夫模型

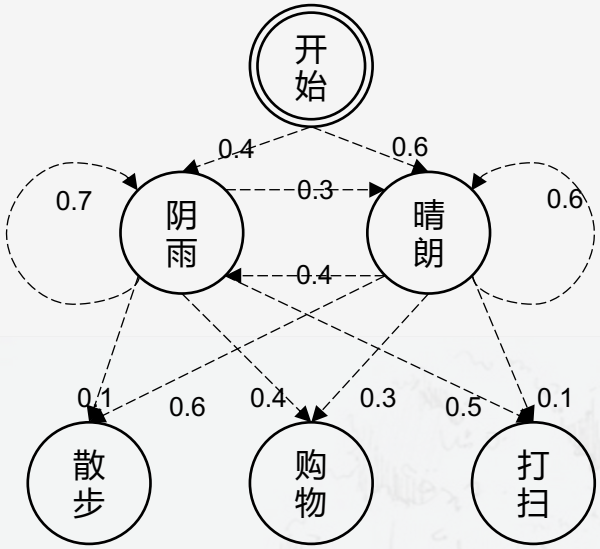
三个基本问题

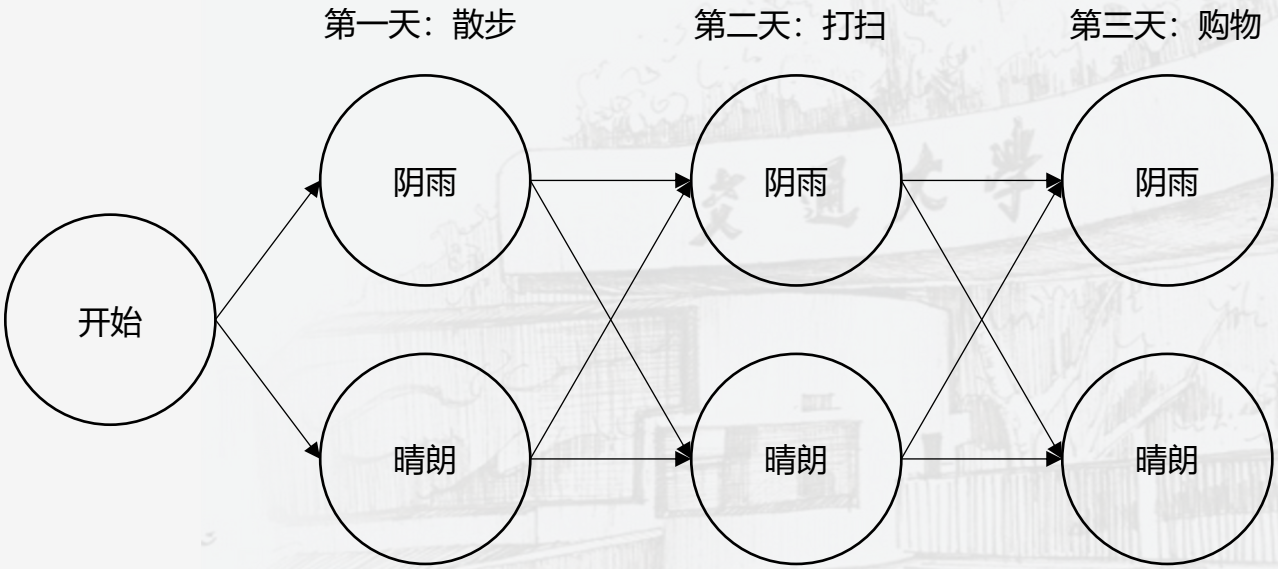
1. 在给定模型 $\mu = (A, B, \pi)$，怎样计算某个观察序列发生的概率，即 $P(O|\mu)$?
2. 给出观测序列 O 和模型 μ，怎样选择一个状态序列 $(X_1, \ldots, X_{T+1})$ 来最好的解释观测序列?
   - 有超过一种状态序列来解释观察序列。
   - 选择原则：对于 t，需要找到 X_t，使 $P(X_t|O, \mu)$ 最大。
   - Viterbi 算法



隐马尔可夫模型

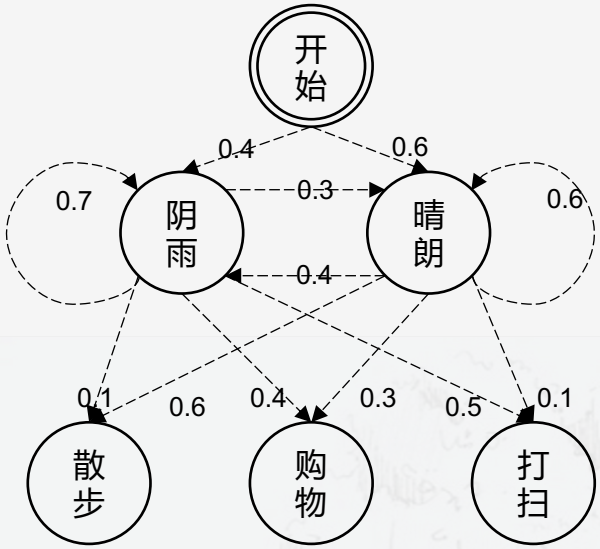
Viterbi算法

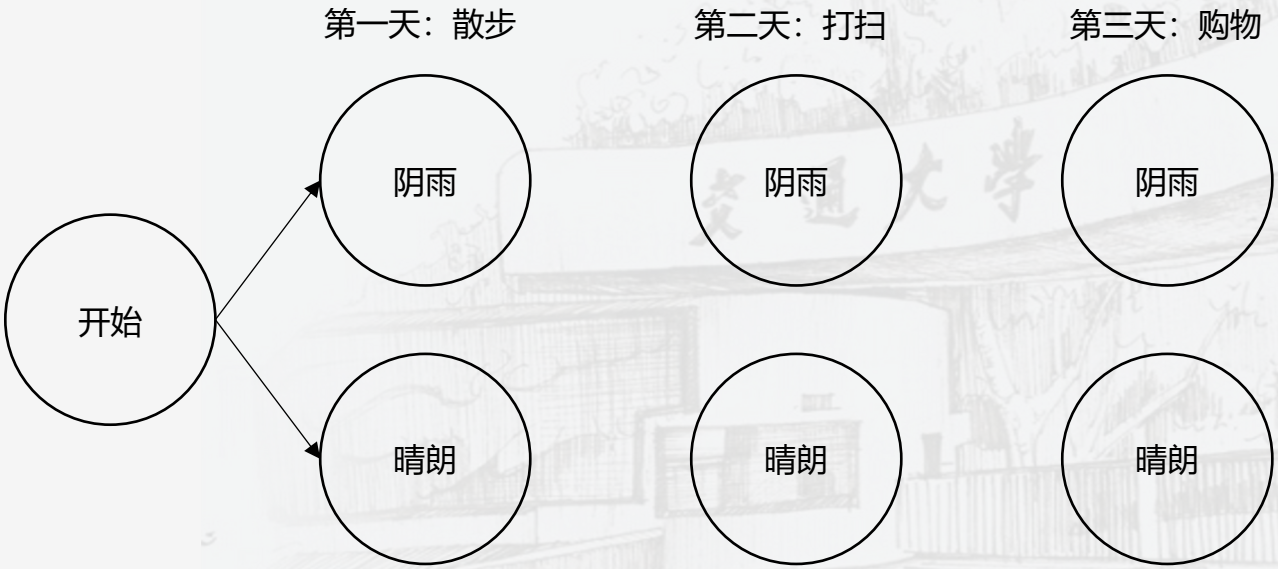




隐马尔可夫模型

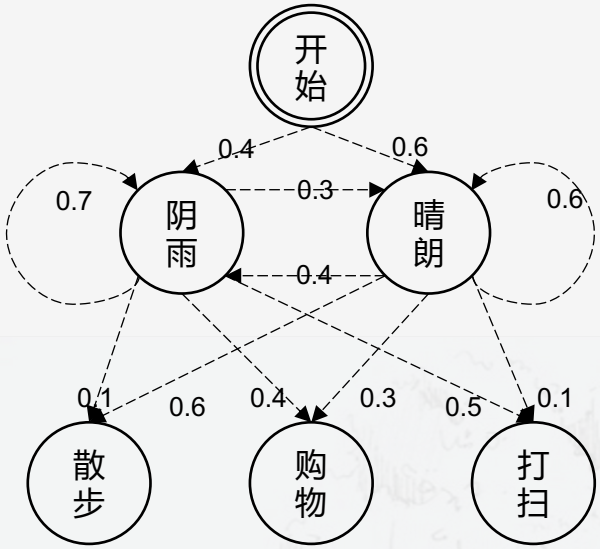
Viterbi算法

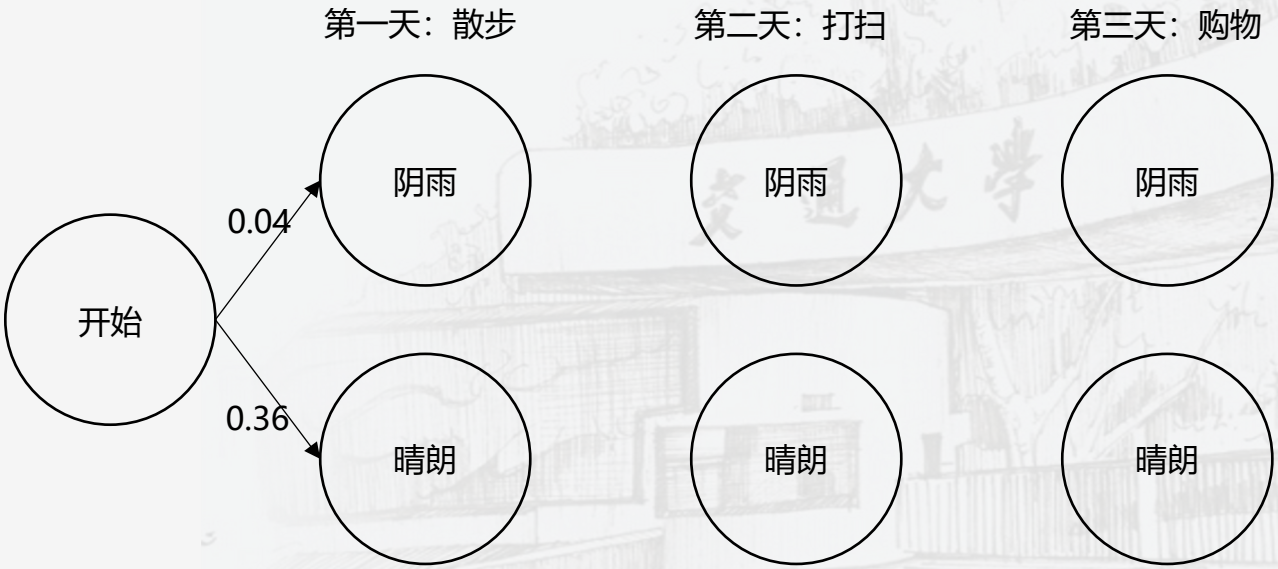




隐马尔可夫模型

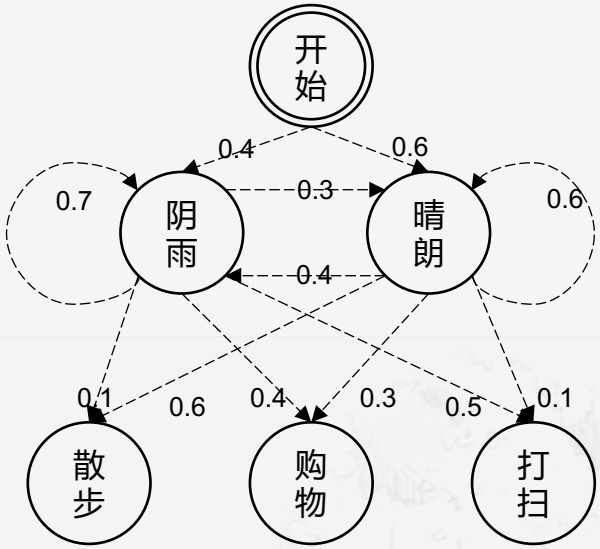
Viterbi算法

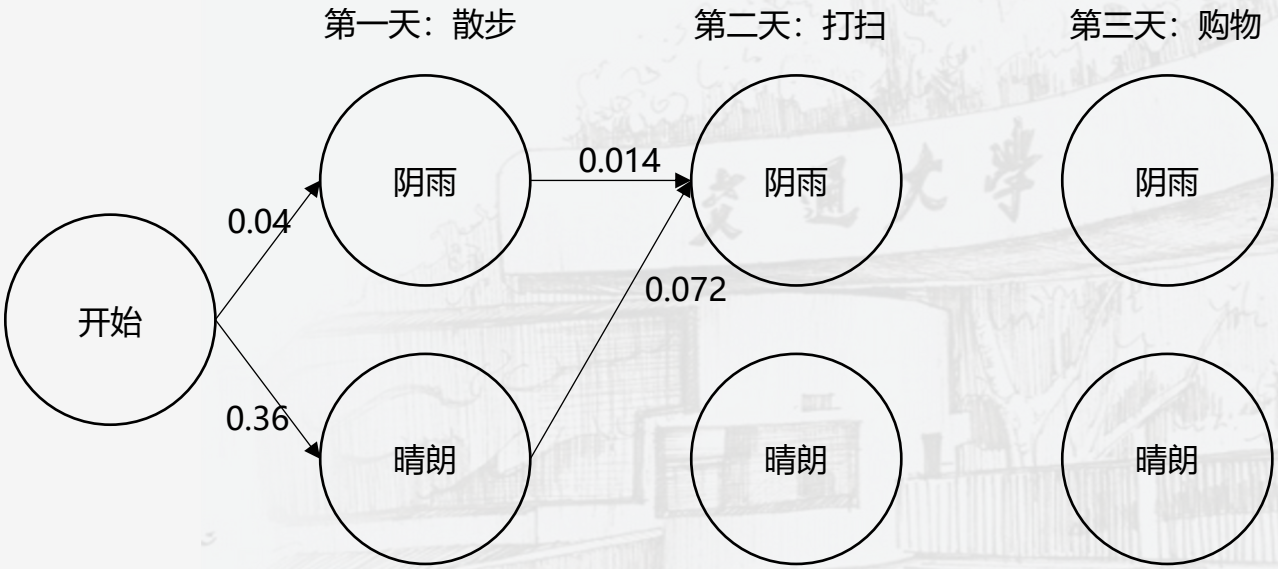




隐马尔可夫模型

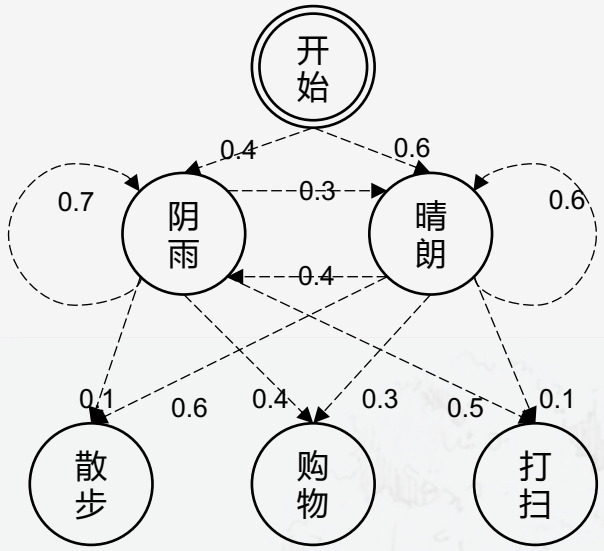
Viterbi算法

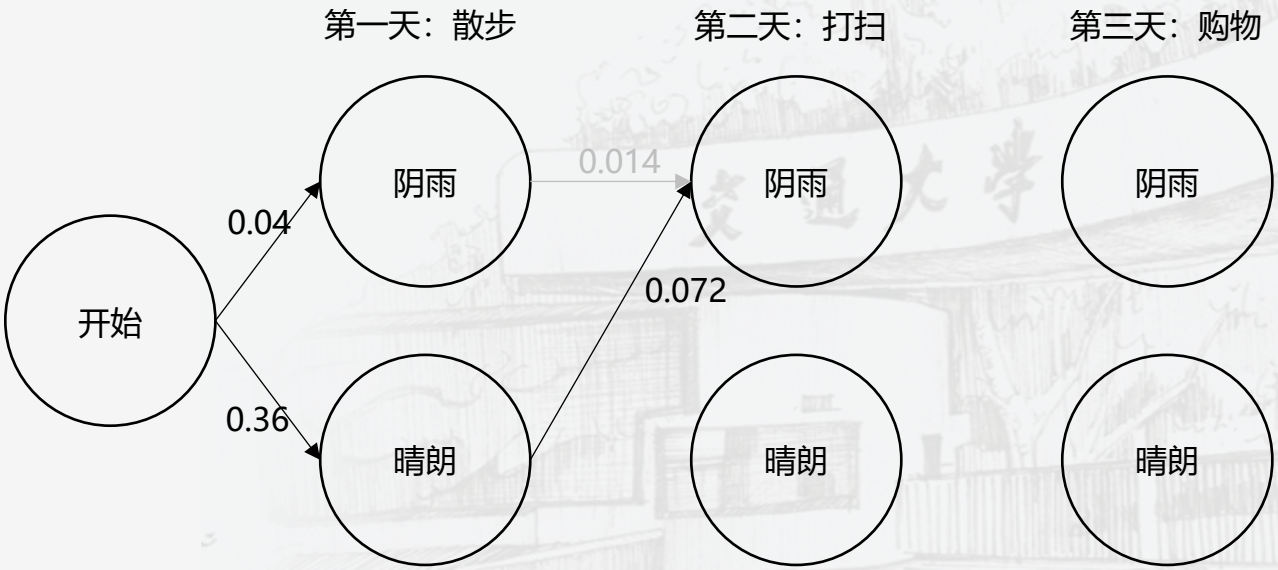




隐马尔可夫模型

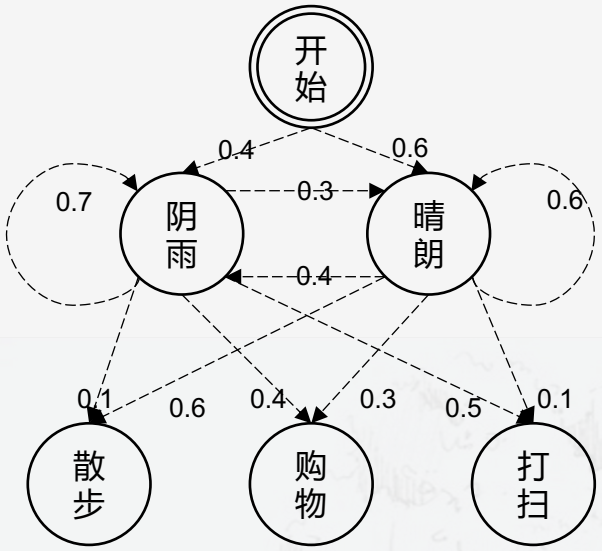
Viterbi算法

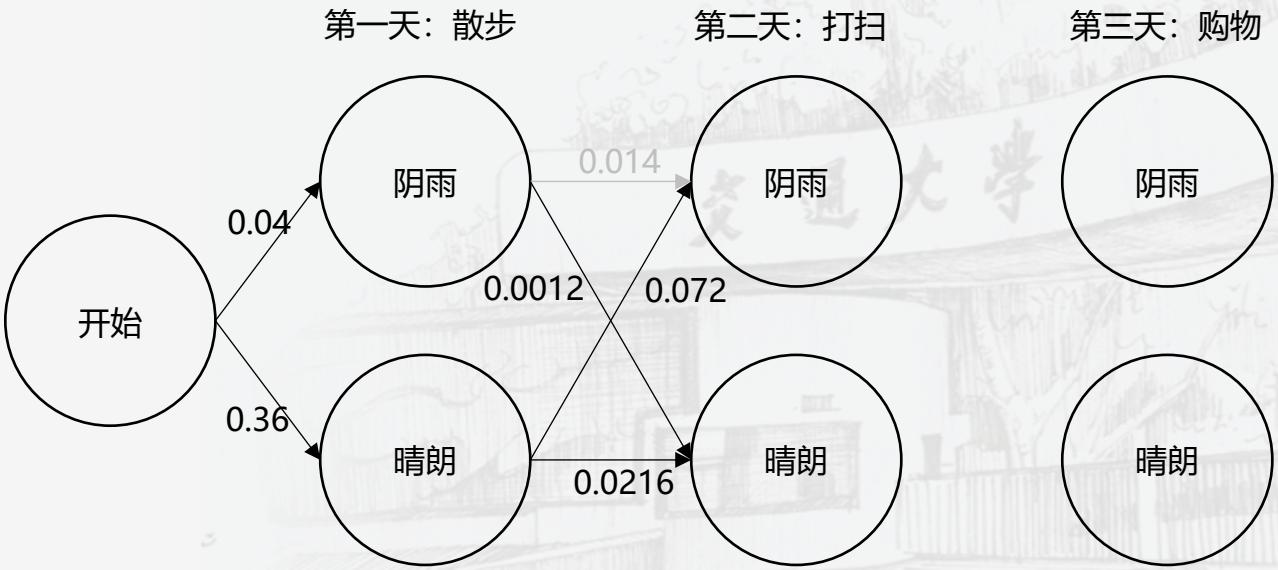




隐马尔可夫模型

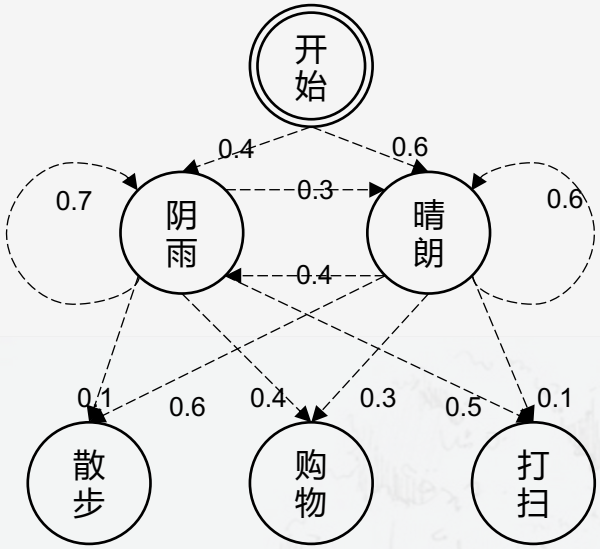
Viterbi算法

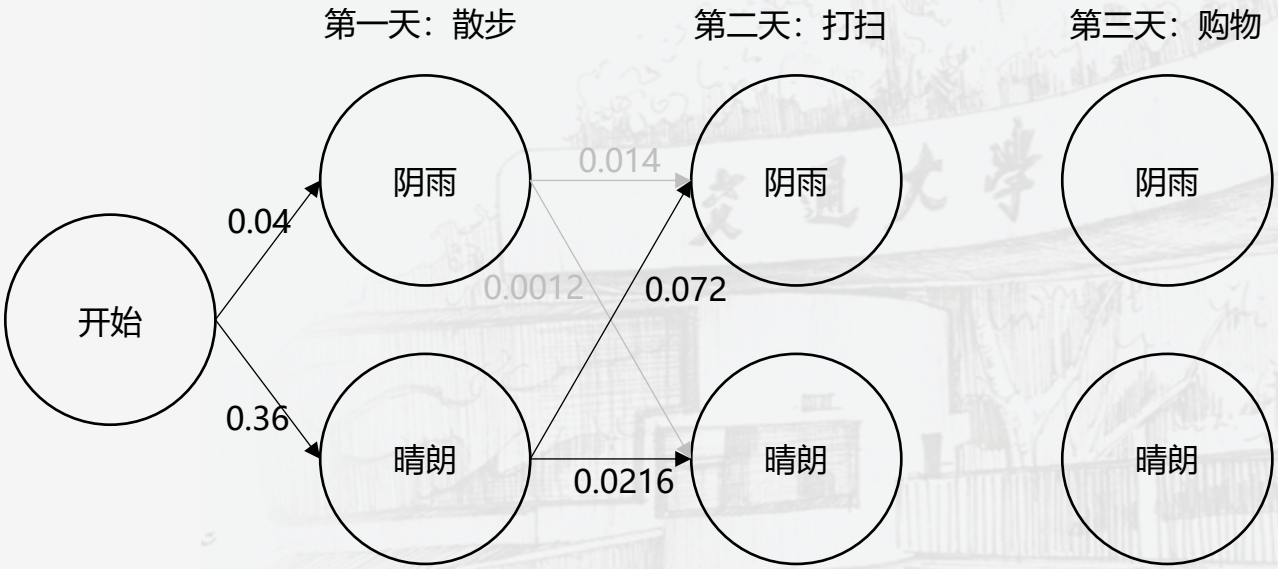




隐马尔可夫模型

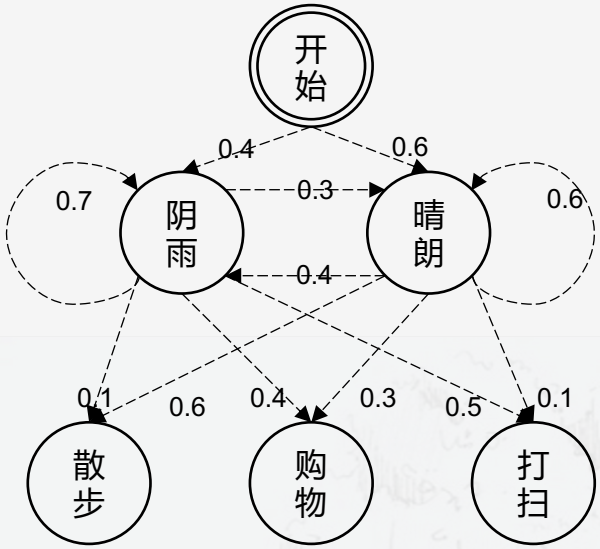
Viterbi算法

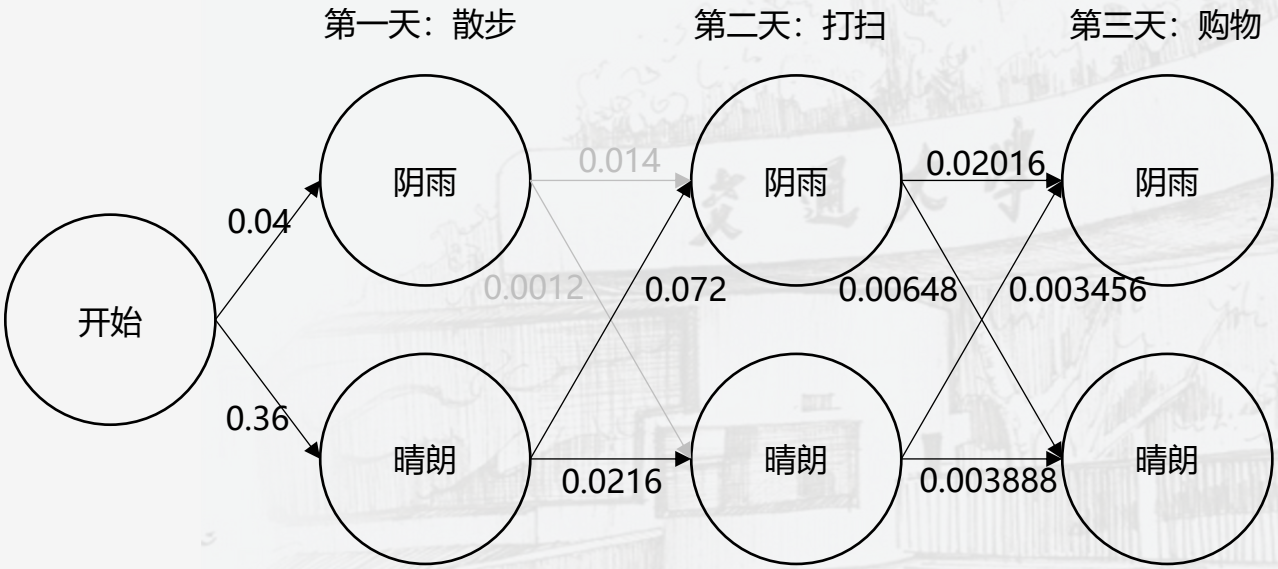




隐马尔可夫模型

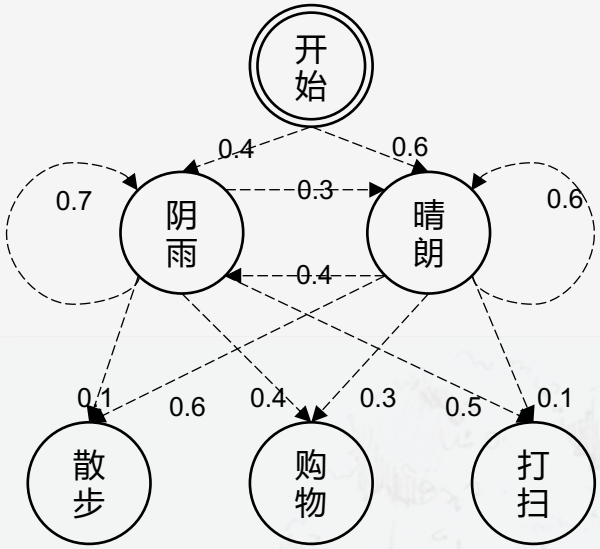
Viterbi算法

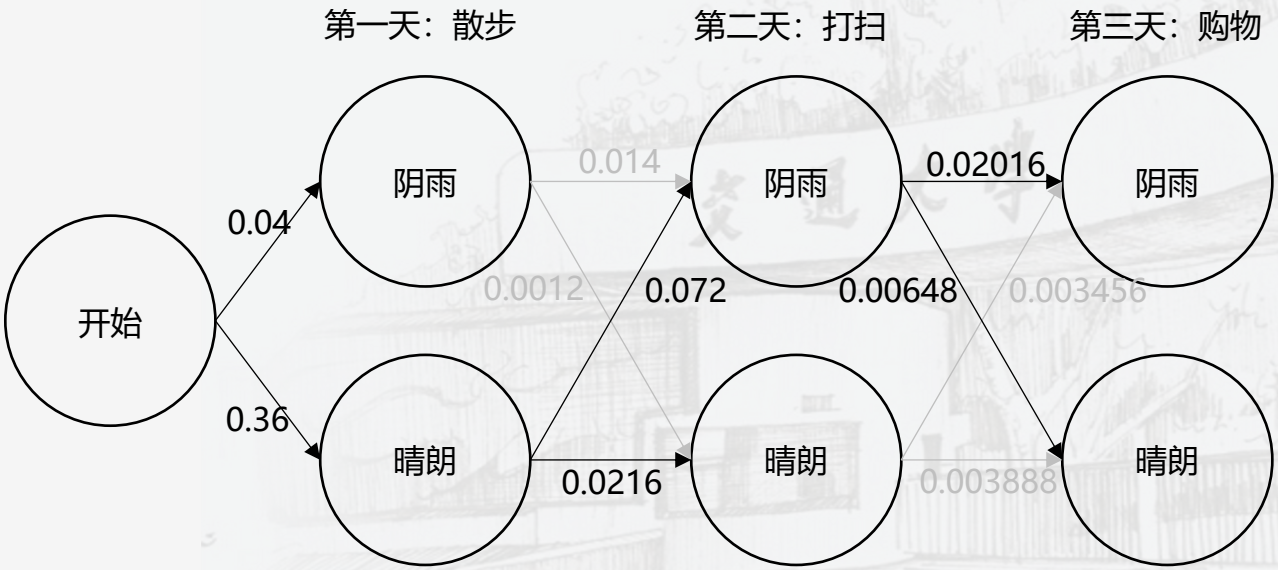




隐马尔可夫模型

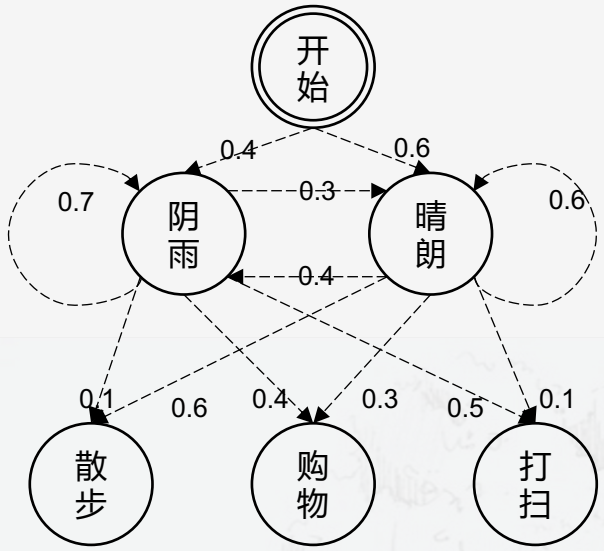
Viterbi算法

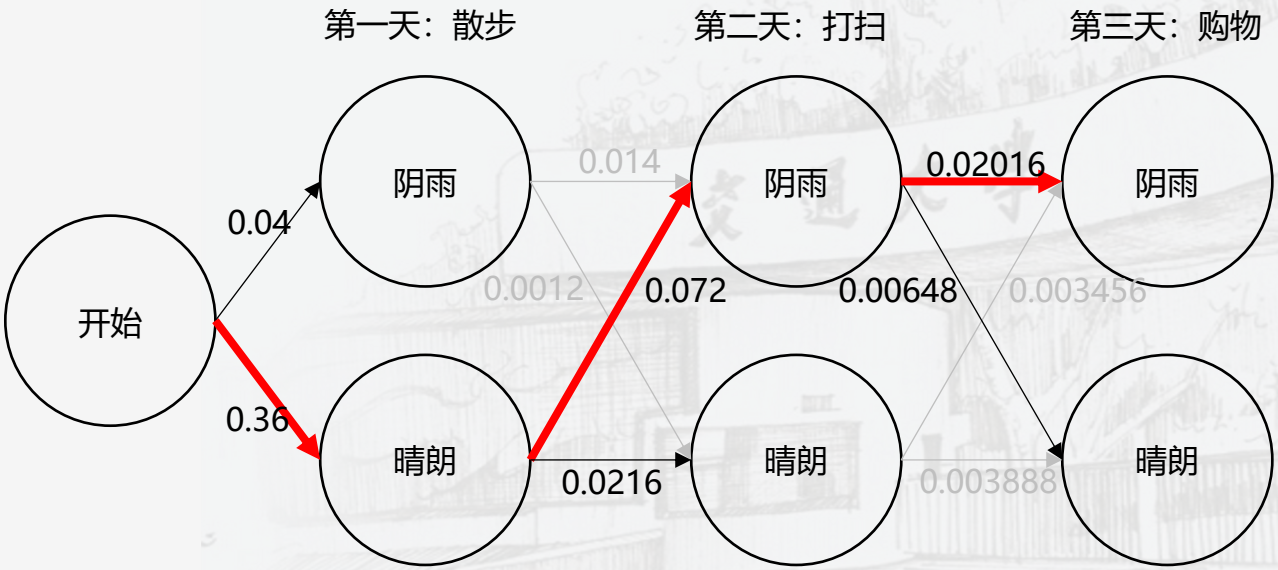




隐马尔可夫模型

Viterbi算法





隐马尔可夫模型

三个基本问题

1. 在给定模型 $\mu = (A, B, \pi)$，怎样计算某个观察序列发生的概率，即 $P(O|\mu)$?
2. 给出观测序列 O 和模型 μ，怎样选择一个状态序列 $(X_1, \ldots, X_{T+1})$ 来最好的解释观测序列?
3. 给定观测序列 O，如何调节模型 $\mu = (A, B, \pi)$ 参数使 $P(O|\mu)$ 最大?
   - 期望值最大化算法（Expectation-maximization algorithm，EM）
   - 最大似然估计（Maximum likelihood estimation，MLE）

最大似然估计

定义

- 给定一个概率分布 D；
- 已知其概率密度函数（连续分布）或概率质量函数（离散分布）为 f_D，以及一个分布参数 θ；
- 从这个分布中抽出一个具有 n 个值的采样 $X_1, X_2, \ldots, X_n$；
- 利用 f_D 计算出其似然函数：

$$\mathrm{L}(\theta \mid x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n).$$

- 若 D 是离散分布，f_θ 即是在参数为 θ 时观测到这一采样的概率。
- 若 D 是连续分布，f_θ 则为 $X_1, X_2, \ldots, X_n$ 联合分布的概率密度函数在观测值处的取值。

最大似然估计

抛硬币的例子

- 假设一个硬币正面跟反面轻重不同。
- 我们把这个硬币抛80次（获取一个采样 $x_1 = \mathrm{H}, x_2 = \mathrm{T}, \ldots, x_{80} = \mathrm{T}$ 并把正面的次数记下来，正面记为 H，反面记为 T）。
- 得到正面的概率 P，则反面的概率为 $1 - P$ （这里的 P 相当于上边的 θ）。
- 假设我们抛出了49个正面，31个反面。
- 假设这个硬币是我们从一个装了三个硬币的盒子里头取出的。这三个硬币抛出正面的概率分别为 $P = \frac{1}{3}$， $P = \frac{1}{2}$， $P = \frac{2}{3}$。这些硬币没有标记，所以我们无法知道哪个是哪个。

最大似然估计

抛硬币的例子

- 使用最大似然估计，基于二项分布中的概率质量函数公式，通过这些试验数据（即采样数据），我们可以计算出哪个硬币的可能性最大。这个似然函数取以下三个值中的一个：

$$\mathbb{L}(p=1/3 \mid \mathrm{H}{=}49, \mathrm{T}{=}31) = \mathbb{P}(\mathrm{H}{=}49, \mathrm{T}{=}31 \mid p=1/3) = \binom{80}{49}(1/3)^{49}(1-1/3)^{31} \approx 0.000$$

$$\mathbb{L}(p=1/2 \mid \mathrm{H}{=}49, \mathrm{T}{=}31) = \mathbb{P}(\mathrm{H}{=}49, \mathrm{T}{=}31 \mid p=1/2) = \binom{80}{49}(1/2)^{49}(1-1/2)^{31} \approx 0.012$$

$$\mathbb{L}(p=2/3 \mid \mathrm{H}{=}49, \mathrm{T}{=}31) = \mathbb{P}(\mathrm{H}{=}49, \mathrm{T}{=}31 \mid p=2/3) = \binom{80}{49}(2/3)^{49}(1-2/3)^{31} \approx 0.054$$

- 当 $\hat{p} = \frac{2}{3}$ 时，似然函数取得最大值。这就是 p 的最大似然估计。

课程提纲

一、隐马尔可夫模型

二、贝叶斯模型

三、平滑技术

贝叶斯定理

- 贝叶斯 (Thomas Bayes, 1701—1761) 英国牧师、业余数学家。在《论机会学说中一个问题的求解》中给出了贝叶斯定理。
- 并列于数据挖掘十大经典算法。
- 它解决了两个事件条件概率的转换问题。

贝叶斯定理

- 贝叶斯 (Thomas Bayes, 1701—1761) 英国牧师、业余数学家。在《论机会学说中一个问题的求解》中给出了贝叶斯定理。
- 并列于数据挖掘十大经典算法。
- 它解决了两个事件条件概率的转换问题。

贝叶斯定理

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

- P(A)是A的先验概率或边沿概率，之所以称为先验，是因为它不考虑任何B方面的因素
- P(A|B)是已知B发生后A的条件概率，也由于得自B的取值而被称为A的后验概率
- P(B|A)是已知A发生后B的条件概率，也由于得自B的取值而被称为B的后验概率
- P(B)是B的先验概率或边沿概率,之所以称为先验，是因为它不考虑任何A方面的因素

贝叶斯定理

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

- 先验概率：由以往的数据分析得到的概率
- 后验概率：得到“结果”的信息后重新修正的概率
- 简单地说，贝叶斯定理是基于假设的先验概率、给定假设下观察到不同数据的概率，提供了一种计算后验概率的方法
- 在人工智能领域，贝叶斯方法是一种非常具有代表性的不确定性知识表示和推理方法

贝叶斯定理

条件概率

- P(A|B)表示事件B已经发生的前提下，事件A发生的概率，叫做事件B发生下事件A的条件概率。其基本求解公式：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

贝叶斯公式

- P(B|A)是根据A判断其属于类别B的概率，称为后验概率。P(B)是直接判断某个样本属于B的概率，称为先验概率。P(A|B)是在类别B中观测到A的概率，P(A)是在数据库中观测到A的概率

$$(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

贝叶斯定理

条件概率

- P(A|B)表示事件B已经发生的前提下，事件A发生的概率，叫做事件B发生下事件A的条件概率。其基本求解公式：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

贝叶斯公式

- P(B|A)是根据A判断其属于类别B的概率，称为后验概率。P(B)是直接判断某个样本属于B的概率，称为先验概率。P(A|B)是在类别B中观测到A的概率，P(A)是在数据库中观测到A的概率

$$(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B) = P(A,B) + P(A^C,B) = P(B|A)P(A) + P(B|A^C)P(A^C)$$

$$= \frac{P(B|A)\,P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

贝叶斯定理

例子

假设一个常规的检测结果的灵敏度和特异度均为 99%，即吸毒者每次检测呈阳性（+）的概率为 99%。而不吸毒者每次检测呈阴性（-）的概率为 99%。假设某公司对全体雇员进行吸毒检测，已知 0.5% 的雇员吸毒。请问每位检测结果呈阳性的雇员吸毒的概率有多高?

- 令“D”为雇员吸毒事件，“N”为雇员不吸毒事件，“+”为检测呈阳性事件
- P(D) 代表雇员吸毒的概率，不考虑其他情况，该值为 0.005。因为公司的预先统计表明该公司的雇员中有 0.5%的人吸食毒品，所以这个值就是D的先验概率。
- P(N) 代表雇员不吸毒的概率，显然，该值为 0.995，也就是 1-P(D)。
- P(+|D) 代表吸毒者阳性检出率，这是一个条件概率，由于阳性检测准确性是 99%，因此该值为 0.99。
- P(+|N) 代表不吸毒者阳性检出率，也就是出错检测的概率，该值为 0.01，因为对于不吸毒者，其检测为阴性的概率为 99%，因此，其被误检测成阳性的概率为 1 - 0.99 = 0.01。

贝叶斯定理

例子

- P(+)代表不考虑其他因素影响的雇员阳性检出率。用公式表示为：

$$P(+) = P(+ \cap D) + P(+ \cap N) = P(+ \mid D)P(D) + P(+ \mid N)P(N)$$

- 雇员吸毒者阳性检出率　(0.5% x 99% = 0.495%) +
  雇员不吸毒者阳性检出率（99.5% x 1%= 0.995%)。
  P(+)=0.0149 是检测呈阳性的先验概率。

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+)}$$

$$= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid N)P(N)}$$

$$= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995}$$

$$= 0.3322$$

贝叶斯定理

结论

- 尽管吸毒检测的准确率高达99%，但贝叶斯定理告诉我们：

  如果某人检测呈阳性，其吸毒的概率只有大约33%，不吸毒的可能性比较大。

  假阳性高，则检测的结果不可靠。

朴素贝叶斯分类

- Naive Bayes classifier
- 一种构建分类器的简单方法
- 基于贝叶斯定理与特征条件独立假设。
  - 结合样本输入输出的联合概率分布和输出的概率分布
  - 对于给定的输入 x，利用贝叶斯定理求解后验概率的过程。

朴素贝叶斯分类

朴素贝叶斯分类

医学诊断

新闻分类

人脸识别

天气预测

朴素贝叶斯分类

例子

- 问题描述：通过测量的特征，判定一个人是男性还是女性。

性别	身高(英尺)	体重(磅)	脚的尺寸(英寸)
男	6	180	12
男	5.92 (5'11")	190	11
男	5.58 (5'7")	170	12
男	5.92 (5'11")	165	10
女	5	100	6
女	5.5 (5'6")	150	8
女	5.42 (5'5")	130	7
女	5.75 (5'9")	150	9

性别	均值(身高)	方差(身高)	均值(体重)	方差(体重)	均值(脚的尺寸)	方差(脚的尺寸)
男性	5.855	3.50E-02	176.25	1.23E+02	11.25	9.17E-01
女性	5.4175	9.72E-02	132.5	5.58E+02	7.5	1.67E+00

朴素贝叶斯分类

例子

- 问题描述：通过测量的特征，判定一个人是男性还是女性。

性别	身高	体重	脚的尺寸
?	6	130	8

基本思想：对于给定的待分类项 x，求解在此样本出现的条件下各个类别出现的概率 P，后验概率高的即为预测值。

$$posterior(male) = \frac{P(male)\,p(height|male)\,p(weight|male)\,p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\,p(height|female)\,p(weight|female)\,p(footsize|female)}{evidence}$$

$$evidence = P(male)\,p(height|male)\,p(weight|male)\,p(footsize|male) + P(female)\,p(height|female)\,p(weight|female)\,p(footsize|female)$$

$$posteriornumerator(male) = 6.1984e^{-09}$$

$$posteriornumerator(female) = 5.3778e^{-04}$$

西安交通大学
XI'AN JIAOTONG UNIVERSITY

Q & A