文本数据管理与分析

复旦大學
FUDAN UNIVERSITY

统计语言模型

邱锡鹏
复旦大学
http://nlp.fudan.edu.cn/xpqiu

基本概念

基本概念

- 样本空间
  - 一个实验或随机试验所有可能结果的集合，而随机试验中的每个可能结果称为样本点。
- 随机事件
  - 一个被赋与机率的事物集合，也就是样本空间中的一个子集。
- 概率
  - 表示对一个随机事件发生的可能性大小，为0到1之间的一个非负实数。

基本概念

- 随机变量
  - 是随着试验结果的不同而变化的，是样本点的一个函数。
  - 例子
    - 随机掷一个骰子，得到的点数就可以看成一个随机变量X，X的取值为{1,2,3,4,5,6}。
    - 如果随机掷两个骰子，构造两个随机变量
      - 随机变量X（获得的两个骰子的点数和）
      - 随机变量Y （获得的两个骰子的点数差）
      - 随机变量X可以有11个整数值，而随机变量Y只有6个。

基本概念

- 条件概率
  - 事件A在另外一个事件B已经发生条件下的发生概率。条件概率表示为P（A|B），读作“在B条件下A的概率”。

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

基本概念

- 贝叶斯定理

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

- 词条的独立假设
  - P(AB)= P(A) P(B) 当且仅当 A与B相互独立

- 对一篇文档而言，若文档中的各个索引词相互独立，则有
  - $P(dj)=P(k_1)\cdots P(k_t)$

统计语言模型

语言模型

- 给句子赋予概率，表示句子的可能性/合理性
  - ！在报那猫告做只
  - 那只猫在作报告！
  - 那个人在作报告！

$$P(x_1, x_2, \cdots, x_n)$$
$$= \prod_i P(x_i | x_{i-1}, \cdots, x_1)$$
$$\approx \prod_i P(x_i | x_{i-1}, \cdots, x_{i-n+1})$$

N元语言模型

谢 谢

如果您有任何意见、评论以及建议，请通过GitHub的Issues页面进行反馈。