Shannon's Entropy

Kullback-Leibler Divergence

香农熵 KL散度

东川路第一可爱猫猫虫

Shannon's Entropy

Kullback-Leibler Divergence

香农熵 KL散度

东川路第一可爱猫猫虫

感谢粉丝大佬
秋山凌
的充电支持

主要内容

- 什么是信息
- 如何度量信息
- 惊喜度与香农熵
- 香农熵的代码
- 如何度量两个变量之间的差异
- KL散度（相对熵）
- KL散度的代码实战

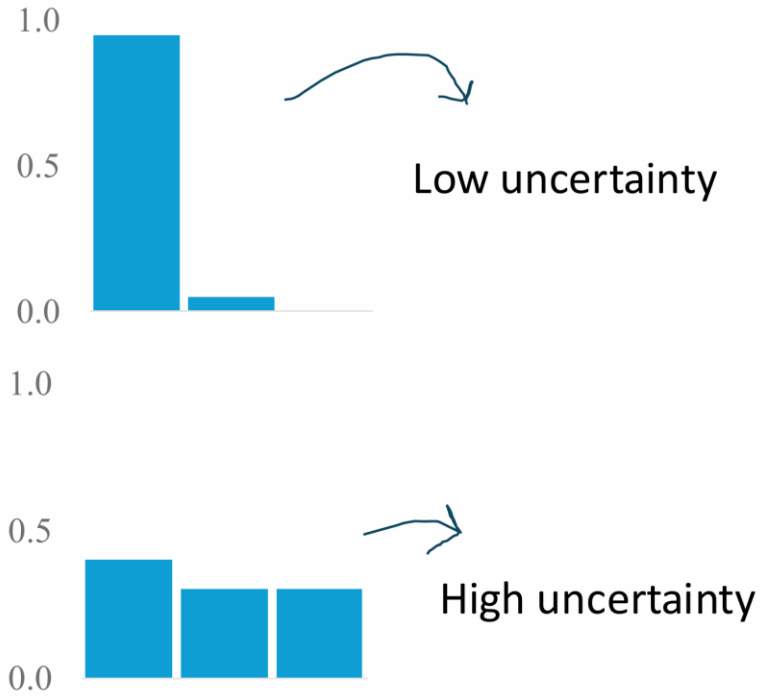
什么是信息

- 变量
  - 变化的量
  - 有不同取值（不同事件）
  - 在不同时间根据某种**过程**取多个特定值中的一个
    - 每个取值（事件）有其概率
- 变量是信息的容器
  - 开关
  - 骰子
  - 温度计
- 一个变量可以看作一个存储单元 用来存储信息

举个例子

- 图灵机
  - 想象一个双向无限长的一维磁带
  - 被划分为一个个大小相等的小方格
  - 每个小方格可以存放一个符号
  - 每个小方格就构成了一个用来存储信息的存储单元
- 现代计算机
  - 数据的最小存储单位是一个比特
  - 这个变量取值为0或1
  - 每个比特的信息就蕴含在它的取值里
  - 多个比特组合在一起可以表示更复杂的信息

我们该如何度量信息



- 文章

  如果文章里的内容毫无新意　信息量小

  如果文章爆出惊天大猛料　信息量大

- 抽奖　一等奖 二等奖 三等奖

  奖池里只有一等奖　毫无惊喜 信息量为0

  一等奖概率0.1 二0.2 三0.7　中等惊喜 信息量中等

  奖池里一二三等奖概率相同　很难猜中 信息量大

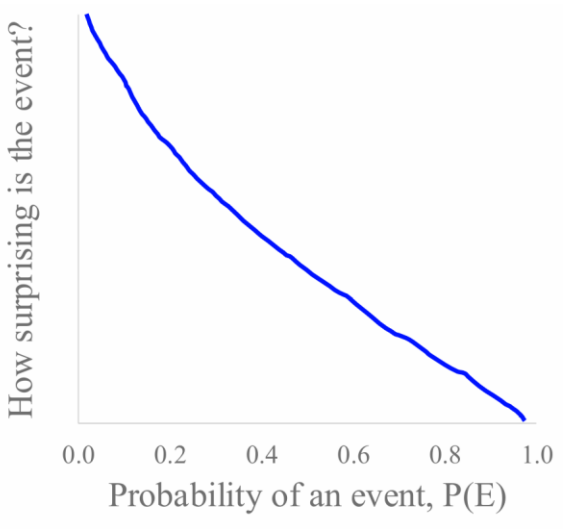
- 信息=惊喜

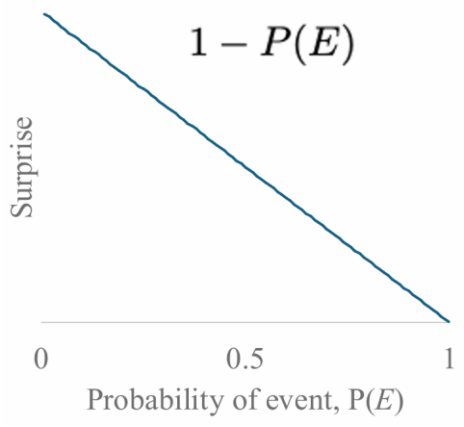
  信息量取决于惊喜度

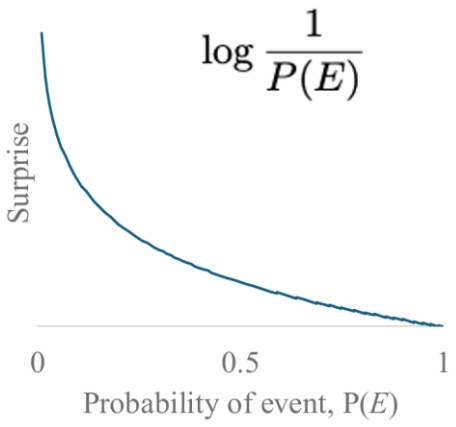
https://web.stanford.edu/class/cs109/lectures/18-InformationTheory/18-InformationTheory.pdf

我们该如何度量惊喜度

- 对单个事件而言
  - 发生概率越低
  - 惊喜度越高
- 如何量化地定义**惊喜度**







https://web.stanford.edu/class/cs109/lectures/18-InformationTheory/18-InformationTheory.pdf

单个事件的惊喜度

- 事件x的惊喜度I

$$I(X = x) = \log_2 \frac{1}{P(X=x)}$$

- 为什么选对数?

  可加性

  两个独立事件的联合惊喜度 = 各自惊喜度之和

  概率每减半，惊喜度+1

- 惊喜度即自信息

  self-information

Probability of Event $P(X = x)$	Surprise of Event $I(X = x)$
1	0
½	1
1/4	2
1/8	3
1/16	4
1/32	5
1/64	6

https://web.stanford.edu/class/cs109/lectures/18-InformationTheory/18-InformationTheory.pdf

平均惊喜度

- 我们知道了单个事件x的惊喜度

  随机变量X的平均惊喜度

  观测X时得到的期望惊喜度

$$H(X) = E\big(I(X)\big) = \Sigma_x \log_2 \frac{1}{P(X=x)} \cdot P(X = x)$$

$$= \Sigma_x - \log_2 P(X = x) \cdot P(X = x)$$

- $\Sigma_x - \log_2 P(X = x) \cdot P(X = x)$

"H(X) Is called Shannon Entropy to scare students and impress Physics people"

$H(X)$就是**香农熵 Shannon Entropy**

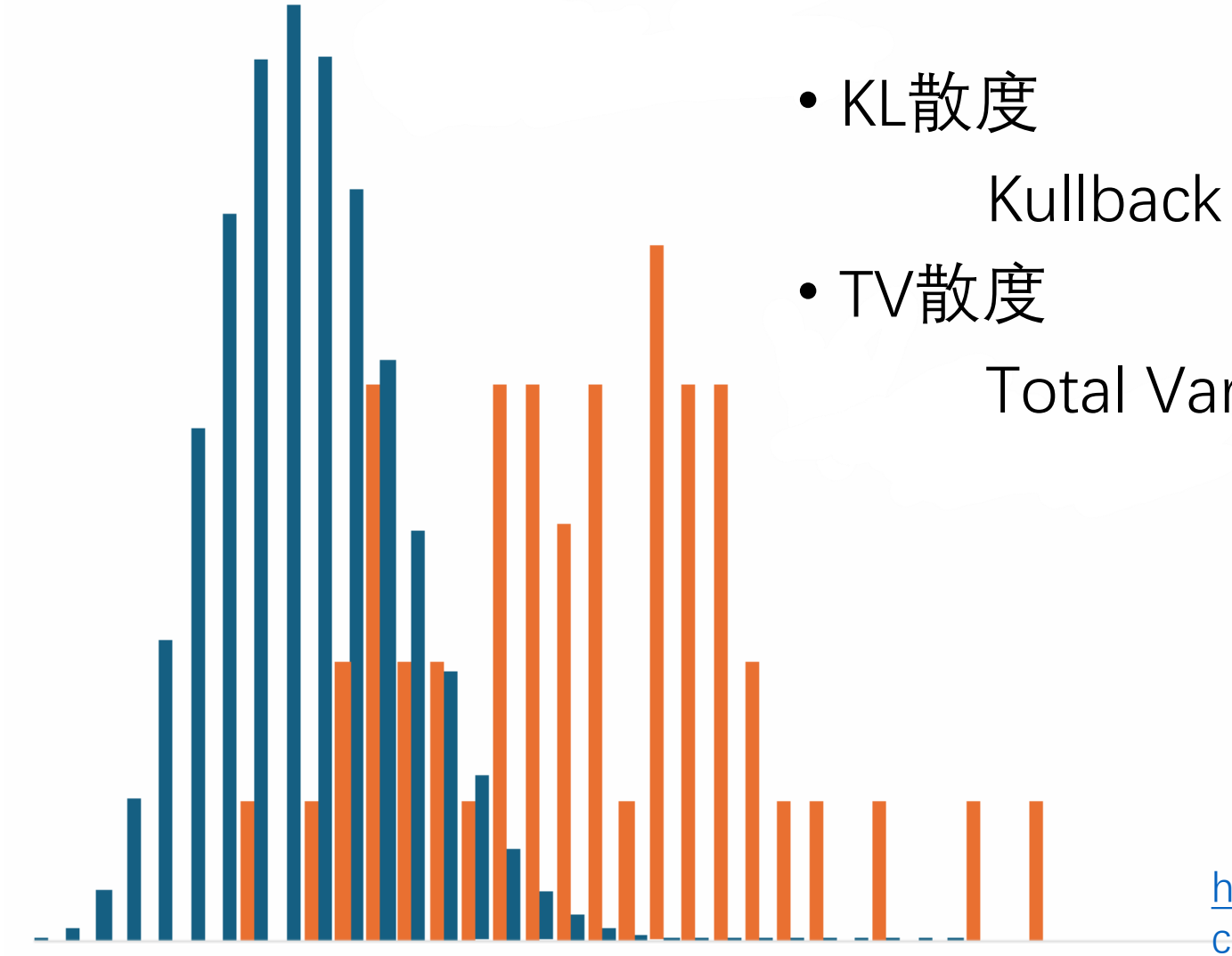
香农熵直接度量了变量里信息量的大小

香农熵代表存储一个变量的信息所需的比特数

$$H(X) = \Sigma_x \log_2 \frac{1}{P(X = x)} \cdot P(X = x)$$

```
def calc_uncertainty(pmf):
    uncertainty = 0
    for x in pmf:
        p_x = pmf[x]
        if p_x == 0:
            continue
        surprise_x = np.log2(1/p_x)
        uncertainty += surprise_x * p_x
    return uncertainty
```

我们该如何度量两个变量之间的差异



- KL散度

  Kullback Leibler Divergence

- TV散度

  Total Variation Distance

https://web.stanford.edu/class/cs109/lectures/19-KL/19-KL.pdf

TV散度

- 对两个变量
  - 遍历所有可能的取值
  - 计算绝对差值
  - 再将这些绝对差值求和

$$TV(X,Y) = \sum_i |P(X=i) - P(Y=i)|$$

KL散度就是期望额外惊喜度

- 当真实分布是X时，用Y作为模型来近似X所带来的期望额外惊喜度
- 期望额外惊喜度

  对于事件x，在分布Y下的期望惊喜度减去在分布X下的期望惊喜度

  KL 散度是额外惊喜度关于分布X取期望

$$\log_2 \frac{1}{P(Y=x)} - \log_2 \frac{1}{P(X=x)}$$

$$KL(X,Y) = \sum_{x \in X} \text{ExcessSurprise}(x) \cdot P(X=x)$$

$$= \sum_{x \in X} \log \frac{P(X=x)}{P(Y=x)} \cdot P(X=x)$$

KL散度（相对熵）

- 从香农信息论的角度来衡量变量的差异

  对于离散随机变量

$$KL(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

  对于连续随机变量

$$KL(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

飓风次数预测

- 目标是预测每年发生的飓风次数
- 气象团队提出假设：飓风次数服从**泊松分布**

  $X \sim \text{Poisson}(\lambda)$
- 我们需要评估不同λ值的预测效果
- 解决方案：用KL散度衡量预测分布与观测分布的差异

```
def kl_divergence_poisson(predicted_lambda, observed_pmf,
max_val=None, epsilon=1e-10):
    if max_val is None or max_val > len(observed_pmf):
        max_val = len(observed_pmf)
    X = stats.poisson(predicted_lambda)
    divergence = 0.0
    observed_pmf_smooth = observed_pmf + epsilon
    observed_pmf_smooth = observed_pmf_smooth /
observed_pmf_smooth.sum()
    for i in range(0, max_val):
        pr_X_i = X.pmf(i)
        pr_Y_i = observed_pmf_smooth[i]
        if pr_X_i == 0:
            continue
        divergence += pr_X_i * np.log(pr_X_i / pr_Y_i)
    return divergence
```

预测lambda=1: KL散度 = 0.8675
预测lambda=2: KL散度 = 0.1745
预测lambda=3: KL散度 = 0.0180 <-- 最小值
预测lambda=4: KL散度 = 0.2831
预测lambda=5: KL散度 = 1.0440
预测lambda=6: KL散度 = 2.4092