Chapter 3

機率論 Probability Theory

Statistical Machine Translation

Introduction to Chap. 3

- 本章介紹以下概念
  - 統計學 *Statistics*
  - 機率論 *Probability theory*
  - 資訊理論 *Information theory*
- 並非全面的介紹
- 只是介紹本書用到的一般的原理
- 更詳細的描述請見 *Cover and Thomas*-Ch2 Entropy, Relative Entropy and Mutual Information (https://web.cse.msu.edu/ cse842/Papers/ CoverThomas-Ch2.pdf)

3.1 估算機率分布 Probability Distributions

- 機率論 *Probability theory* 用數學的方式，研究「未確定性」 *uncertainty* 的性質 （多種可能 *possible* 的事件結果 outcomes

- 來看看氣象報告：*On Monday, there is a 20% chance of rain*

- 我們根據「知識」對「事實」（或意見）的陳述，而且包含了我們知識對該事實陳述的不確定性 *uncertainty about the facts*

- 機率事件 probabilistic events (rain) 我們預測後的行動 our action (帶雨傘) 有其風險
  - 風險 risk 有其代價 *unnecessary weight*：淋濕 *getting wet*

3.1.1 透過分析來估算機率 Estimation by Analysis

- 考慮最減的例子：投骰子 *cast of a dice*

- 估算事件結果*outcome event*的「機率」 *probability*
  - 取 600 樣本，計算事件的次數： (1:102, 2:100, 3:100, 4:98, 5:101, 6:99)
  - 可能的結果 outcomes, 1-6
  - 估算結果: 1-6 are (almost) equally likely, $p(i) = \frac{100}{600} = 0.167, i = 1, 6$
  - 這就是平均分布（不常見）

3.1.2 常見機率分布

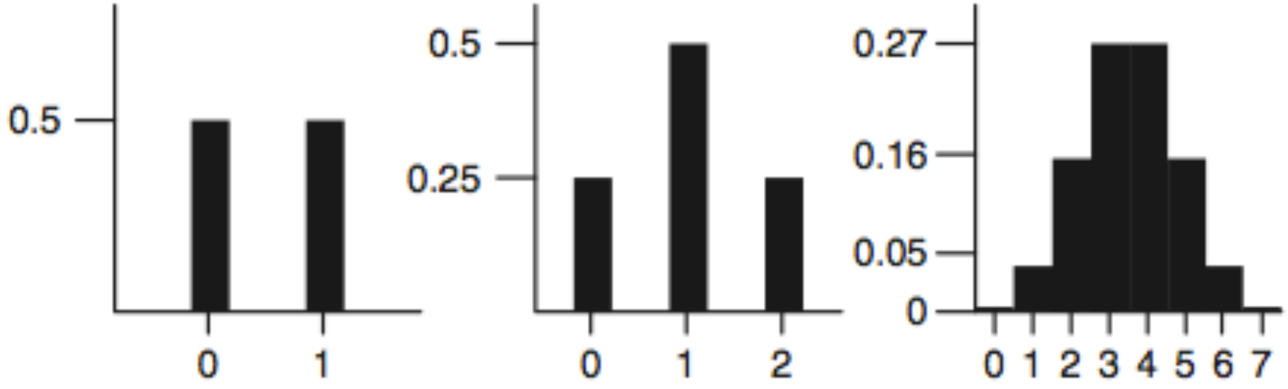
- 投骰子：平均分布 $uniform\ distribution$
- 投錢幣： 人頭 H (*Head* 或「尾巴」*Tail*) 也是平均分布 $uniform\ distribution$
- 投錢幣 n 次 H 的累積次數 b = 二項分布 binomial distribution

二項分布：人頭有幾次?

- 投錢幣 n 次，人頭 H 出現 k 次

$$
\begin{aligned}
b(n,k;p) &= \binom{n}{k} p^k \,(1-p)^{n-k} \\
&= \frac{n!}{(n-k)!\,k!}\, p^k \,(1-p)^{n-k}
\end{aligned}
\tag{3.1}
$$

- $n = 1$, 2, 7 (愈來愈接近常態分布)



意見調查 opinion poll、誤差範圍 margin of error、樣本數 sample size

- 透票給候選人 X 的機率 p
- 進行意見調查，詢問 n 個人，他們是否會投票給 X
- 調查結果 $= n$ 中有 k 回答 YES，機率值 $= b(n, k; p)$
- $b(n, k; p)$ 讓我們可以用來計算預測 p 的錯誤邊際 (is it $\pm 3\%$?)
- 機率論可以讓民調機構，計算誤差範圍，適當地調整樣本數降低誤差範圍
- 第 8 章將會使用這些現象來討論機器翻譯評估 MT evaluation

二項分布 Binomial Distribution

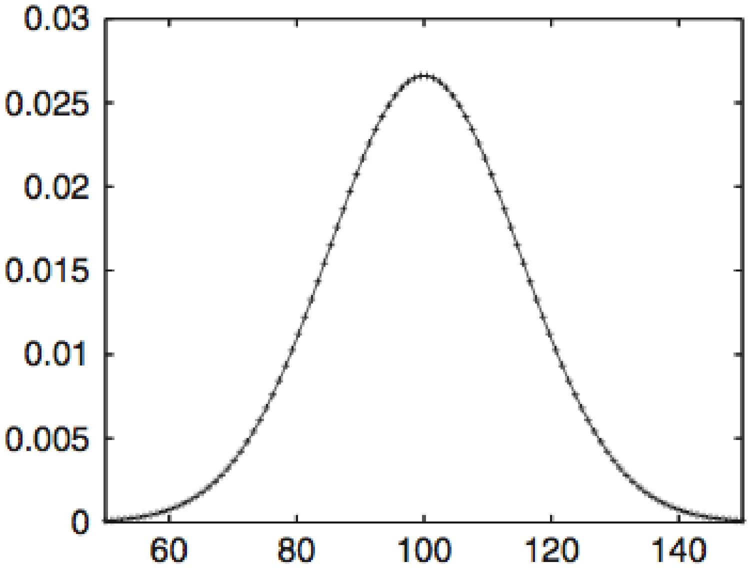
- 當 $n \to \infty$ 時，二項分布趨近於常態分布
- 許多自然現象 (如人口的智商、身高分布) 就是常態分布
- 又稱為鐘形曲線 *bell curve*
- 專家則愛用「高斯分布」 *Gaussian distribution* 的行話

常態分布數學式

- 公式：平均值 μ 變異數 σ^2 (σ 為標準差)

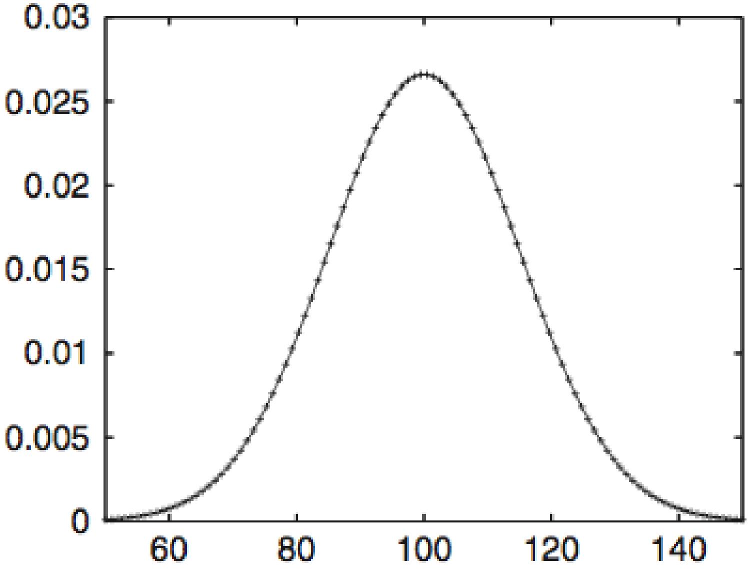
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3.2}$$

- 一般人都知道「智商」是常態分布（$\mu = 100$, $\sigma = 15$)



IQ 的常態分布與百分比

- 用常態分布的數學式，很容易計算某智商 IQ 值的人數百分比
  - IQ $\leq$ 115 (μ +1 σ) 的人數 *84%*
  - IQ $\leq$ 130 (μ +2 σ) 的人數 *96%*
  - IQ $\leq$ 145 (μ +3 σ) 的人數 *99.9%*



3.1.3 用統計來估算機率 Estimation from Statistics

- 蒐集實際的資料與統計來估算機率 (而不是坐著沙發上推理)

- （從一個資料的來源 population) 抽取樣本 sample data

1 然後加以統計樣本的平均值 *average* 和標準差 *standard deviation*

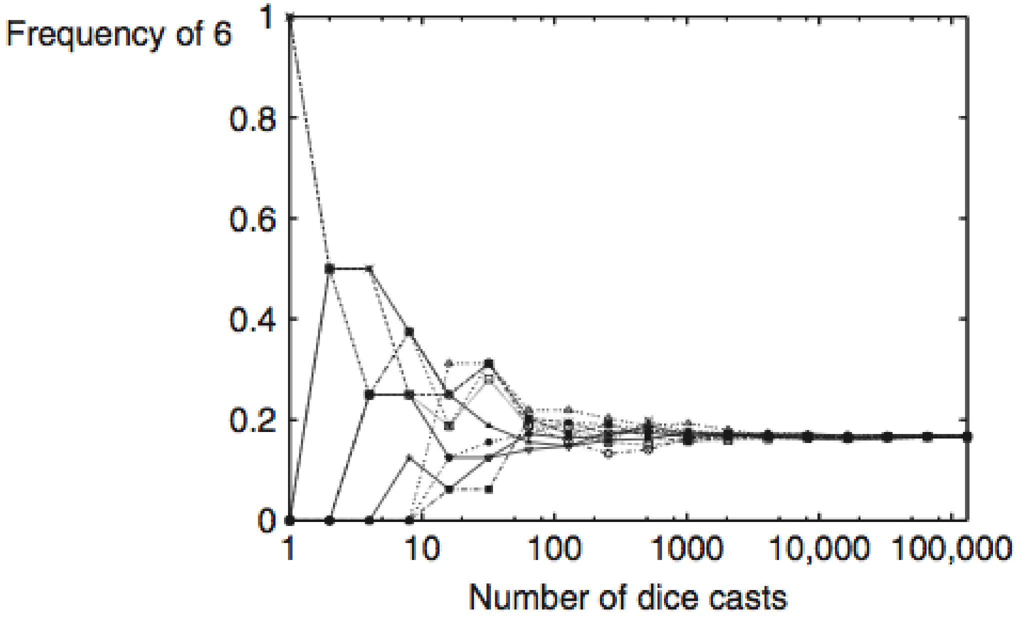
- 就可以用這兩個參數（母數）估算機率 *probabilities*

2 （若無參數）投骰子

- 1,000,000 次， 6 號出現 200,000 times, $p(6) = \frac{200,000}{1,000,000} = 0.2$

大數法則 Law of large numbers

- 當統計次數愈高（大數），機率估計值愈準確



預測天氣 Predicting Weather

- 考慮明天（和十年來以來一樣）的下雨機率
  - $P(rain) = 730/3,650 = 20\% = .20$
- 如果我們考慮附加的資訊
  - 今天是否下雨
  - 全季下雨天數
  - 鋒面是否將來（個別或三者加起來）

  $P(rain \mid \text{附加的資訊}) = ?$

- 要計算 $P(rain \mid \text{附加的資訊})$ 我們需要更加了解數學的機率論

3.2 計算機率分布

- Section 3.1 描述了隨機變數的機率分布
  - 使用標準的常用機率分布 (例如，平均、二項、常態)
  - 使用樣本的事件次數
- 機率論的數學讓我們估算更複雜狀況（聯合事件、條件事件）的機率分布

3.2.1 正式定義 Formal Definitions

- 隨機變數 *Random variable* X 代表不確定性的本質
- 不像固定值*fixed value* (e.g., a = 4) X 有不同的可能值x（機率＝$p(x)$）
- 機率函數有下列性質
  - $\forall x : 0 \leq p(x) \leq 1$
  - $\sum_x p(x) = 1$
- 機率分布的函數的定義域，可以有 2, 6 甚至無限多個事件

3.2. 聯合機率分布

- Two random variables, D (dice) and C (coin)

- Define *joint probability distribution*, $p(D = d, C = c)$, or for short $p(d, c)$

- How likely is it that we cast a *6* and get *head*?

- ANS: Total of 12 different combinations of both outcomes (equally likely), hence $p(6, heads) = 1$

- *independence* of random variables: value of one variable has no impact on the second, and vice versa

- Formally, X and Y are independent $\iff \forall x, y : p(x, y) = p(x)p(y)$

Examples of Joint Probability Distributions

- $p(6, heads) = p(6)p(heads) = 1/6$ x $1/2 = 1/12$

- rain on any given day is *0.20*

- raining days are *not independent*
  - $p(rain, rain) = \mathit{0.12}$
  - $p(rain)p(rain) = \mathit{0.2} \times \mathit{0.2} = \mathit{0.04} \neq 0.12$
  - $p(rain, rain) \neq p(rain)p(rain)$
  -

p(T, M)	rain tomorrow	no rain tomorrow
rain today	*0.12*	0.08
no rain today	0.08	0.72

3.2.3 Conditional Probability Distributions

- Consider "*If it rains today, how likely is it that it will rain tomorrow?*"

- We need conditional probability distribution for this
  - conditional prob: $p(y|x) = \frac{p(x,y)}{p(x)}$ $\quad (3.8)$
  - chain rule: $p(x,y) = p(x)p(y|x)$ (1st part x 2nd part cond. on 1st)
  - independency: $p(x,y) = p(x)p(y)$
  - E.g.,
  - $$p(rain|rain) = \frac{p(rain, rain)}{p(rain)} = \frac{0.12}{0.2} = 0.6$$
  - $$p(rain|no\_rain) = \frac{p(rain, no\_rain)}{p(rain)} = \frac{0.08}{0.8} = 0.1$$

markov assumption:

p(w1,w2,,,,wn)
= p(w1)*p(w2lw1)
*p(w3lw1,w2)
....

- *Conditional probability distributions* are also called *marginal distributions*

3.2.4 Bayes Rule

- Express a conditional probability distribution $p(x|y)$ in terms of
  - its inverse $p(y|x)$ ($posterior$)
  - $p(x)$ ($prior$) and p(y)

- Bayes Rule

language model

輸入 y 產生 x

希望輸出 x 機率高
輸入 y 固定

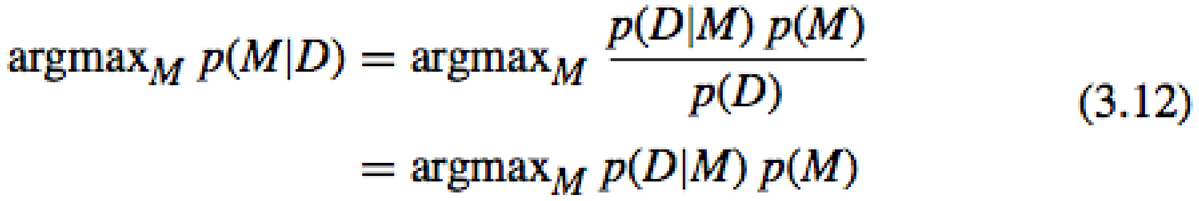
  - $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$ (3.11)

- Use $Bayes\ Rule$ in $Bayesian\ model\ estimation$

$$\begin{aligned} \text{argmax}_M\, p(M|D) &= \text{argmax}_M \frac{p(D|M)\, p(M)}{p(D)} \\ &= \text{argmax}_M\, p(D|M)\, p(M) \end{aligned} \qquad (3.12)$$

Use *Bayes Rule* in *Bayesian model estimation*

- best model M is selected by considering
  - $p(D|M)$: how well M explains the sample D,
  - $p(M)$: how likely M is a good model in general



$$\begin{aligned}\operatorname{argmax}_M p(M|D) &= \operatorname{argmax}_M \frac{p(D|M)\,p(M)}{p(D)} \\ &= \operatorname{argmax}_M p(D|M)\,p(M)\end{aligned} \tag{3.12}$$

3.2.5 Interpolation

避免機率為0

- $interpolation$ = combine 2 distributions p_1 and p_2 for one random variable X
  - $p(x) = \lambda p_1(x) + (1\text{-}\lambda)p_2(x)$ (3.13)
  - $p_1(x)$ and $p_2(x)$ arise from sampling in different conditions

- Example
  - predict tomorrow's weather M with prob. $p(m|t, d)$
  - based on today's weather T and current calendar day D
  - $p(m|t, d) = \lambda p(m|t, d) + (1\text{-}\lambda)p(m|t)$

- Why? $p(m|t, d)$ conditions on specific days may lead to small data sample. So, we also consider the more robust $p(m|t)$

- In $machine\ learning$, researchers combine results of several methods ($classifier\ combination$ or $ensemble\ learning$) based on $interpolation$

3.3 Properties of Probability Distributions

- Important, frequently-used concepts in properties of probability distributions
  - mean, variance
  - expectation
  - entropy
  - mutual information

3.3.1 Mean and Variance

- Outcomes of uncertain events can be represented with numerical values
- With numerical outcomes, we can compute mean $\bar{x}$ and variance σ
- $\bar{x} = \frac{1}{n} \sum x_i$ (3.15)
- $\sigma = \frac{1}{n} \sum (x_i - \bar{x})^2$ (3.17)
- Example: 10 dice casts = 5,6,4,2,1,3,4,3,2,4
- $\bar{x} = \frac{5+6+4+2+1+3+4+3+2+4}{10} = 3.4$
- $\sigma_2 = \frac{1}{10}((5-3.4)^2 + (6-3.4)^2 + (4-3.4)^2 + (2-3.4)^2 + (1-3.4)^2 + (3-3.4)^2 + (4-3.4)^2 + (3-3.4)^2 + (2-3.4)^{+}(4-3.4)^2) = 2.4$

3.3.2 Expectation and Variance

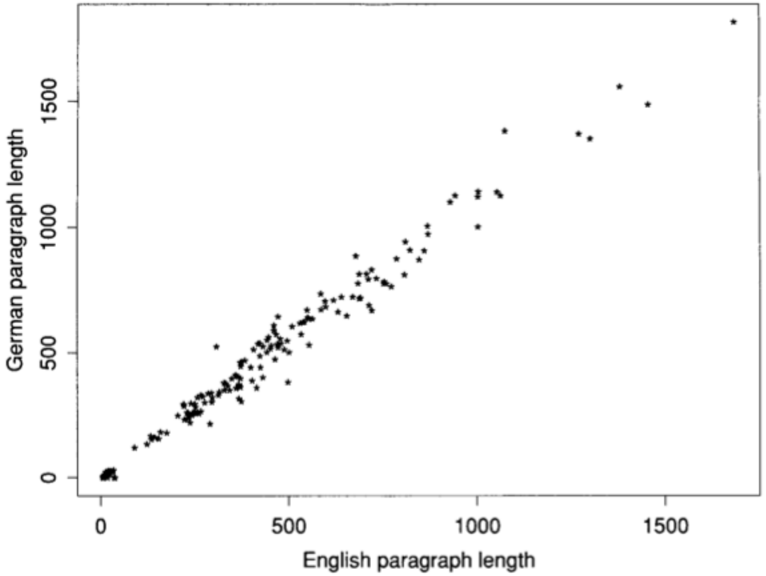
- *expectation* = mean of a probability distribution
  - $E[X] = \sum_{x \in X} x\ p(x)$ (3.20)
  - with possible values x_i of X weighted by their probability $p(x_i)$

- *variance* = expected squared difference between each x_i and expected value $E[X]$
  - $Var[X] = \sum_{x \in X} (x - E[x])^2\ p(x)$ (3.21)
  - $Var[X] = E[(X - E[X])^2]$ (3.22)

- Parametric statistics (assuming distributed according)
  - sample data from a population and compute μ and σ μ 決定平移位置 σ 定型
  - determine a normal distribution (completely fixed)

Distribution of length ratio in sentence alignment

- Paragraph lengths are highly correlated. The horizontal axis shows the length of English paragraphs, while the vertical scale shows the lengths of the corresponding German paragraphs. Note that the correlation is quite large (.991).

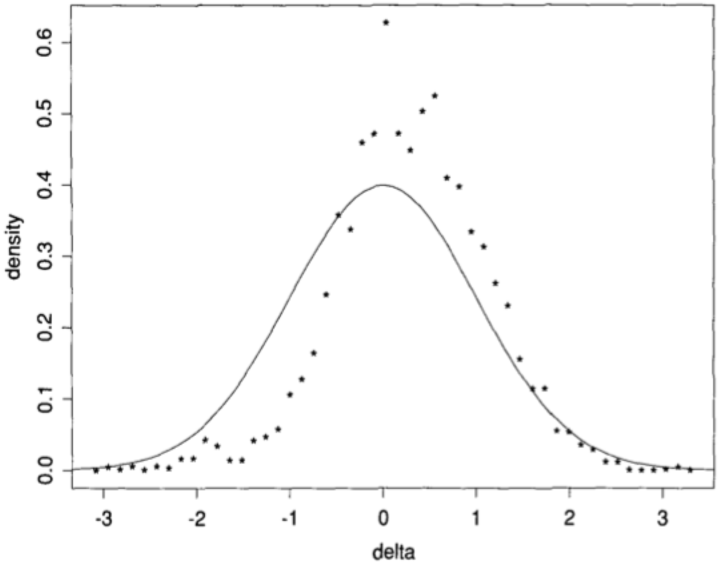


Distribution of length ratio in sentence alignment

- English-French mean length ratio $\mu = 72302/68450 \approx 1.06$ and $\sigma^2 = 5.6$

- $\mu =$ expected # characters in French / # characters in English

- $\sigma^2 =$ variance of # characters in French per # character in English



Determine the distribution

- Define $\delta = \frac{(len_2 - len_1 * \mu)}{\sqrt{len_1 \sigma^2}}$

- Variable δ has a normal distribution with mean zero and variance one

3.3.3 Entropy

- $entropy$ is widely used in computer science to measure *disorder*
- from the laws of thermodynamics, which state that unless energy is added to a system its entropy never decreases – Nature tends towards disorder
- In probability, theory $entropy$ measures $uncertainty$ of outcomes.
- When making predictions, we want to put in *effort* to $increase\ the\ certainty$ of our predictions (i.e. decrease the entropy)
  - what is the next word, following two preceding words
  - what is the best sentence as translation in MT

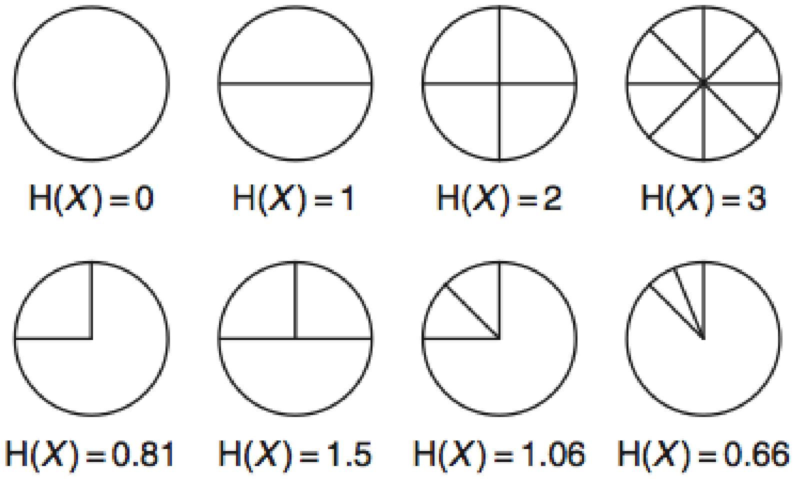
Define Entropy

- Definition: entropy of a random variable is
  - $H(X) = \sum_{x \in X} p(x) \log_2 p(x)$
- If a random variable X has one certain outcome, its entropy is 0
- If there are two equally likely outcomes, its entropy is 1
- If there are four equally likely out- comes, its entropy is 2, and so on
- Entropy is lower if the probability is distributed unevenly
- If two outcomes has 1/4 and 3/4 probabilities, the entropy is $0.81 < 1$

Entropy: Examples

- If possible events are equally likely, then entropy H(X) increases by 1 with each doubling of the number of events (top row of examples)

- Entropy is lower if the probabilities are less evenly distributed (lower row of examples).



example:

編碼長度：1,2,2
乘上機率：1*0.5+2*0.25*2*0.25 = 1.5

Explain Entropy as information

- *Entropy* has a very natural explanation in *information theory*

- When communicating events by encoding each with a binary code, entropy = lower bound of # bits needed / event

- With four equally likely events, we encode them with 00, 01, 10, 11, i.e., 2 bits / event

- With 3 events of .50, .25, .25 probabilities, we encode them with 0, 10, 11, and need on average (0.5×1+0.25×2+0.25×2) 1.5 bits / event

- *Entropy* correlates nicely with *news worthiness*: Dog bites man, a common event, receives scarcely a mention, while man bites dog, an uncommon event, is a news story.

3.3.4 Mutual Information

- Important to know about X impacts *certainty* over the outcome of Y

- Define *joint entropy* H(X, Y) as entropy when considering two random variables at the same time
  - $H(X, Y) = -\sum_{x,y \in X,Y} p(x, y) \log_2 p(x, y)$

- Define *conditional entropy* $H(X|Y)$ as uncertainty removed when one variables is known
  - $H(Y|X) = H(X, Y) - H(X)$
  - If we know X (weather one day), how much more certain we can predict Y (weather the next day)
  - $H(Y|X)$ is not necessary equal to $H(X|Y)$

3.3.4 Mutual Information

- Define a symmetric measure *mutual information* $I(X;Y)$ as

- Define *joint entropy* H(X, Y) as entropy when considering two random variables at the same time
  - $I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$

跨語言：
搭配詞：

- Two extreme cases:
  - X and Y are independent $p(x,y) = p(x)p(y)$
    - $I(X;Y) = 0$
  - X and Y are totally dependent $p(x,y) = p(x)$
    - $I(X;Y) = H(Y)$

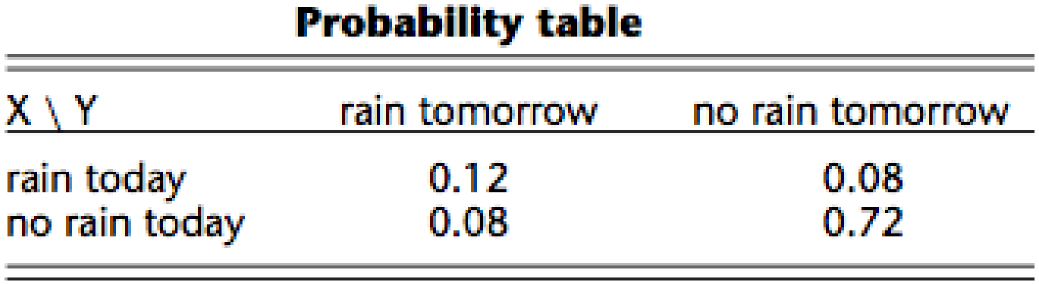
相互資訊和熵值的關係

- 相互資訊$I(X;Y)$ 可以表達為熵值 $H(X)$, $H(Y)$, $H(X,Y)$

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \qquad (3.27)$$

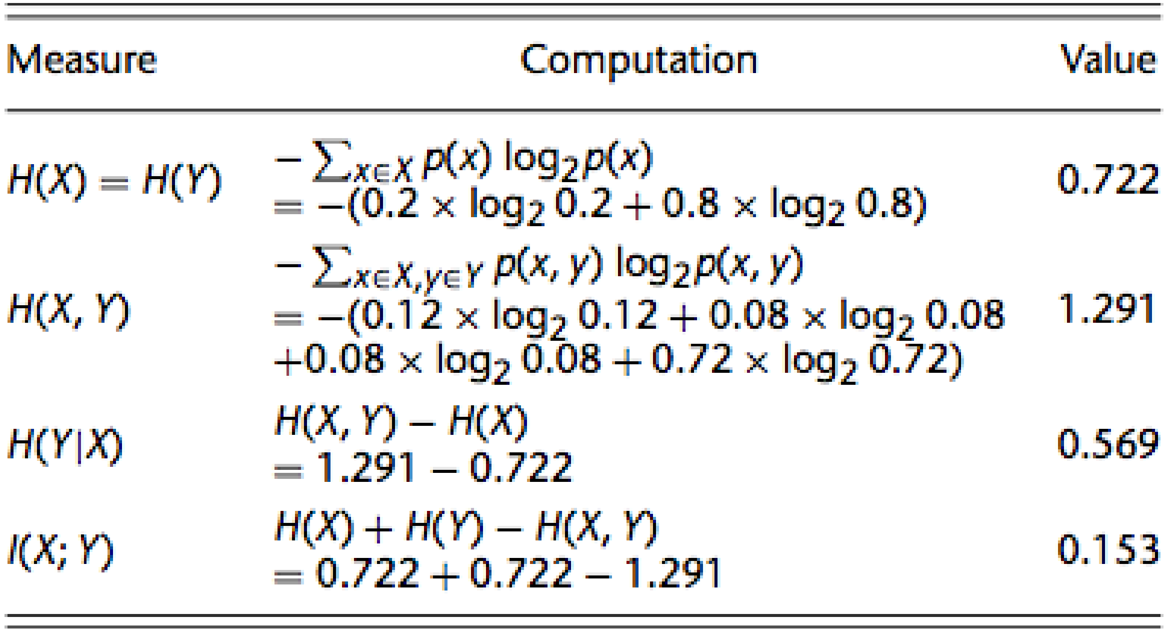
相互資訊舉例

- 今天下雨機率X、明天下雨機率 Y 先關程度＝相互資訊 $I(X;Y) = 0.153$



Probability table

X \ Y	rain tomorrow	no rain tomorrow
rain today	0.12	0.08
no rain today	0.08	0.72



Measure	Computation	Value
$H(X) = H(Y)$	$-\sum_{x\in X} p(x)\log_2 p(x)$ $= -(0.2 \times \log_2 0.2 + 0.8 \times \log_2 0.8)$	0.722
$H(X, Y)$	$-\sum_{x\in X, y\in Y} p(x,y)\log_2 p(x,y)$ $= -(0.12 \times \log_2 0.12 + 0.08 \times \log_2 0.08$ $+0.08 \times \log_2 0.08 + 0.72 \times \log_2 0.72)$	1.291
$H(Y\|X)$	$H(X,Y) - H(X)$ $= 1.291 - 0.722$	0.569
$I(X;Y)$	$H(X) + H(Y) - H(X,Y)$ $= 0.722 + 0.722 - 1.291$	0.153

3.4 摘要

- 我們會用適當的常見機率分布來描述特定事件
  - 平均分布
  - 二項分布
  - 常態分布
- 我們同時也需要蒐集樣本，用以估計機率分布
- 大數法則告訴我們，資料愈大，估計值愈接近真正機率分布
- 嚴謹的數學機率論，提供了豐富的工具，可以用「隨機變數」代表不確定事件，以及計算機率分布，事件的機率函數值

3.4 續

- 我們用聯合機率分布來描述多個隨機變數（例如，句子中一個個詞），如果他們彼此互相獨立
- 我們用連鎖規則 chain rule 來計算某事件，在另外事件發生後的條件機率分布 (又稱邊際分布 marginal distributions)
- 貝氏規則 Bayes rule 讓我們可以將條件機率分布，調整成兩部分：先驗機率、後驗機率
- 多個機率分布可以做成內插 interpolation 可以
  - 合併兩個分類器（機率函數）
  - 合併不同的機器學習方法

3.4 摘要

- 我們常常想知道資料樣本的一些性質：
  - 均值 mean (所有資料值的平均數)
  - 變異 d variance (偏離均質的幅度）
  - 期望值？ corresponding to expectation and variance of probability distributions
  - 熵值 entropy measures the degree of uncertainty of a probabilistic event and can be intuitively interpreted using information theory，allow us to analyze the relationship between different probability distributions
    * Measures such as joint entropy,
    * conditional entropy
    * mutual information

Coding Lab

- Obtain a text corpus that is tokenized and split into sentences, such as the Europarl corpus. Collect statistics on the length of sentences (in words) and the length of words (in characters).
  - Compute the mean and variance for these samples.
  - If you plot the samples, are they normally distributed?
- `http://www.erv-nsa.gov.tw/user/article.aspx?Lang=1&SNo=03000450`

網路語料範例

- corpus: 油綠桑田的青翠景致 a beautiful countryside with verdant, green mulberries
  - `http://www.erv-nsa.gov.tw/user/article.aspx?Lang=1&SNo=03000450`
  - `http://www.erv-nsa.gov.tw/user/article.aspx?Lang=2&SNo=03000450`

Web corpus

關山環鎮自行車道完工於民國86年，是全台第一條環鎮自行車道，不僅闢建最早，也享有最高的知名度。全長15.2公里，路寬約3~4公尺，全程為自行車專道，嚴禁非農耕之車輛進入；

The Guanshan Town Circle Bikeway was completed in 1997, and was the first bicycle path to circle any town in Taiwan. It is not only the earliest, but also the most famous of Taiwan's town-encircling bicycle paths. It is around 15.2 kilometers in length, and about 3 to 4 meters in width. The whole course is exclusively for cyclist use, and other vehicles are strictly prohibited with the exception of farming vehicles.

Web corpus

路面鋪水泥，路中央並以紅色磚塊排列成直線做為雙向車道的區隔；每隔一百公尺處以彩色地磚拼出花紋圖案並標示公里處和公尺數；水圳溪邊架設枕木護欄，各路口有明顯標示和安全告示；沿途並有涼亭和大理石桌椅方便休憩，整體規劃和安全考量都相當用心。

The surface is paved with cement, and is divided into a left and right lane by means of a red-brick line in the middle. At every hundred meters, there is a decorative design made of colorful bricks to mark the number of kilometers and meters already passed. Protective balustrades line the canal, and at every intersection, there are clear signs and safety notices. Along the way are pavilions with marble tables and chairs to provide rest. Much effort has gone into the entire planning and safety considerations.

a beautiful countryside with verdant, green mulberries

關山環鎮自行車道完工於民國86年，是全台第一條環鎮自行車道，

The Guanshan Town Circle Bikeway was completed in 1997, and was the first bicycle path to circle any town in Taiwan.

不僅闢建最早，也享有最高的知名度。

It is not only the earliest, but also the most famous of Taiwan's town-encircling bicycle paths.

全長15.2公里，路寬約3~4公尺，

It is around 15.2 kilometers in length, and about 3 to 4 meters in width.

全程為自行車專道，嚴禁非農耕之車輛進入；

The whole course is exclusively for cyclist use, and other vehicles are strictly prohibited with the exception of farming vehicles.