187 - Day 10

Wednesday, April 25, 2018 9:34 AM

Day 10 – Breaking Rectangular Transposition

Conditional Probability

Definition 1. *The **conditional probability** of an event B is the probability that this event will occur, given the knowledge that another event A has already occurred.*

$$\boxed{\mathbb{P}(B|A) = \frac{\mathbb{P}(B \text{ and } A)}{\mathbb{P}(A)},}$$

assuming that $\mathbb{P}(A) > 0$.

b/c A has occured, it will block some of the outcomes

Example. Two women state the following:

- A: "I have two children, the eldest is a girl."
- B: "I also have two children, and one of them is a girl."

Which of them is more likely to have two girls?

2 children : Sample space = set of all outcomes
$\{BB, BG, GB, GG\}$.

<u>A</u>

New sample space $\{GB, GG\}$

$$\mathbb{P}(2G\text{'s} \mid 1\text{st} = G) = \frac{1}{2}.$$

$$= \frac{\mathbb{P}(2G\text{'s and } 1^{st} = G) \rightarrow \mathbb{P}(GG)}{\mathbb{P}(1^{st} = G)}$$

$$= \frac{1/4}{1/2} = \frac{1}{2}$$

<u>B</u>

New sample space = $\{BG, GB, GG\}$

$$\mathbb{P}(2G\text{'s} \mid \text{atleast } 1 \text{ is a } G) = \frac{1}{3}$$

Example.

- The Z-virus is a rare disease that hits about **5% of the population**.
- Meanwhile, Umbrella Corporation claim that they have a test that can detect Z-virus with **95% accuracy**. → 95% it will get the correct result.

Suppose that you are **tested positive** for having Z-virus (by Umbrella's test), what is the probability that you actually carry the deadly disease?

Take a population 10,000

has disease | no disease.

500 | 9,500

+ | − | + | −

500×0.95 | 500×0.05 | 9500×0.05 | 9500×0.95

$\underbrace{500 \times 0.95}_{475}$ | $\underbrace{9500 \times 0.05}_{475}$

- 9,500 ppl w/no disease
- Test is correct 95% time
  ↳ 5% time test is wrong

$$P(\text{you have virus} \mid \text{test } +) = \frac{P(\text{have virus AND test } +)}{P(\text{test } +)}$$

$$= \frac{475}{475 + 475} = \frac{1}{2}$$

$\sim\sim\ \underline{\mu}\ \underline{\lambda}\ \sim\sim$

Example. Given an English text. What is the probability that a randomly chosen letter λ is A?

$$\mathbb{P}(\lambda = \text{"A"}) = p_A = 0.08399$$ (from table on Day 09)

Now suppose that we also know about the letter μ that is immediately to the left of λ. What is the probability that $\lambda =$ "A" given that

only letter after Q is U

- $\mu =$ "Q"?

$\sim\ \underline{Q}\ \underline{A}^{?}\sim\sim$

$$\mathbb{P}(\lambda = A \mid \mu = Q) = \frac{\mathbb{P}(\lambda = A \text{ and } \mu = Q)}{\mathbb{P}(\mu = Q)} = \frac{0}{\mathbb{P}(\mu = Q)} = 0.$$

$\sim\ \underline{E}\ \underline{A}^{?}\sim$

- $\mu =$ "E"?

$$\mathbb{P}(\lambda = A \mid \mu = E) = \frac{\mathbb{P}(\mu\lambda = EA)}{\mathbb{P}(\mu = E)} = \frac{110/10{,}000}{1237/10{,}000} = \frac{110}{1237}$$

110: # in cell EA

1237: Σ of E-row.

- $\mu =$ "L"?

$$\mathbb{P}(\lambda = A \mid \mu = L) = \frac{\mathbb{P}(\mu\lambda = LA)}{\mathbb{P}(\mu = L)} = \frac{40}{391}$$

(Now suppose that μ and λ are far apart. What is the probability that $\mu =$ "L" and $\lambda =$ "A"?

$\_\ \underline{L}\ \sim\sim\ \underline{A}\ \sim$

If the letters are far apart.

↳ they're independant.

$$\mathbb{P}(\lambda = A \text{ and } \mu = L) = \mathbb{P}(\mu = L)\cdot\mathbb{P}(\lambda = A)$$
$$= p_L \cdot p_A$$

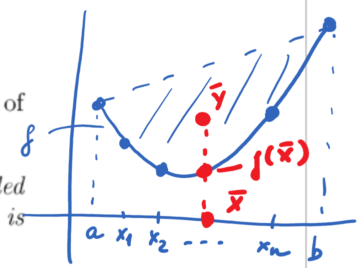
from Day 09 Table.

3

Breaking Rectangular Transposition

Some remarks:

1. Single letter frequencies are useless here
2. Not all pairs of adjacent English letters are equally probable
3. Table of bi-letter frequencies should reveal which pairs of letters of ciphertext were adjacent in the plaintext



Definition 2. *A function $y = f(x)$ is **convex** on an interval $[a, b]$ provided that $f''(x) \geq 0$ for all $a \leq x \leq b$. In particular, the first derivative $f'(x)$ is increasing on the interval $[a, b]$.*

Theorem 1. *Let $x_1, x_2, \dots, x_n \in [a, b]$ and let $\overbrace{p_1, p_2, \dots, p_n}^{\text{weights}}$ be real numbers such that $p_1 + \cdots + p_n = 1$. If f is **convex** on $[a, b]$ then*

$f(\bar{x}) \leq \bar{y}$

$$f(\underbrace{p_1x_1 + p_2x_2 + \cdots + p_nx_n}_{\bar{x}}) \leq \underbrace{p_1f(x_1) + p_2f(x_2) + \cdots + p_nf(x_n)}_{\text{average for } f(x_i) = \bar{y}}.$$

Here, equality occurs if and only if $x_1 = x_2 = \cdots = x_n$.

Corollary 1.1. *Let $f(x) = \log\left(\frac{1}{x}\right)$ in the above theorem to obtain*

$$\log\left(\frac{1}{p_1x_1 + p_2x_2 + \cdots + p_nx_n}\right) \leq p_1 \log\left(\frac{1}{x_1}\right) + \cdots + p_n \log\left(\frac{1}{x_n}\right)$$

Theorem 2. *Let $p_1, \dots, p_n$ be probabilities with $p_1 + \cdots + p_n = 1$. Then for any set of probabilities $q_1, \dots, q_n$ such that $q_1 + \cdots + q_n = 1$, we have*

$$\sum_{i=1}^{n} p_i \log(q_i) \leq \sum_{i=1}^{n} p_i \log(p_i)$$

key to break rec. trans.

To prove this, use Corollary 1.1 with $x_i = q_i/p_i$.

The steps for breaking rectangular transposition:

1. Guess a length for the decrypting permutation, says k.
2. Arrange the ciphertext into k columns and let N be the height (i.e. number of rows) of the resulting rectangle.
3. For each pair $1 \le i \ne j \le k$, extract the columns i and j and count the number of occurrence of the pair of letters $\alpha\beta$ and call this $n_{\alpha\beta}^{(ij)}$.

do this for all pairs $\alpha\beta$ that appear in these columns

4. For each pair $\alpha\beta$, let $p_{\alpha\beta}$ be the probability of the pair $\alpha\beta$ in the English language (obtain from the table of frequency for letter pairs). Compute

$$\boxed{C_{ij} = \sum_{\alpha,\beta} p_{\alpha\beta} \log(n_{\alpha\beta}^{(ij)}).}$$

now repeat for all values i, j in $1, 2, \ldots, k$ with $i \neq j$

$k = 10,\ N = 23,\ i = 3,\ j = 7$

$k = 10$

$N = 23$

$$\begin{pmatrix}
E & C & T & I & H & N & O & H & G & I \\
O & K & R & O & B & C & A & O & H & F \\
E & I & N & S & G & N & N & S & A & A \\
E & T & C & N & I & I & E & C & N & H \\
O & A & S & R & E & E & H & C & T & L \\
H & S & A & A & T & E & I & B & N & E \\
S & F & N & E & U & C & N & O & E & R \\
R & E & T & I & U & S & S & S & A & A \\
R & E & O & C & U & W & S & O & I & F \\
M & N & D & A & O & D & I & D & V & A \\
T & E & C & H & E & X & O & T & T & E \\
H & O & F & E & T & C & E & R & L & A \\
I & I & A & T & S & O & E & S & M & S \\
M & S & T & E & I & O & N & K & W & N \\
N & I & C & S & O & S & F & S & O & T \\
X & Y & S & T & I & U & H & F & R & O \\
A & R & E & G & X & S & A & A & E & M \\
S & M & C & Y & H & L & Z & B & I & O \\
B & A & E & Y & D & R & I & P & T & A \\
L & R & C & A & U & R & N & A & A & R \\
M & N & G & E & E & F & I & T & S & O \\
T & A & X & R & S & H & A & I & T & G \\
B & O & N & R & D & N & I & K & L & E
\end{pmatrix}
\Longrightarrow
\begin{matrix}
T & O \\
R & A \\
N & N \\
C & E \\
S & H \\
A & I \\
N & N \\
T & S \\
O & S \\
D & I \\
C & O \\
F & E \\
A & E \\
T & N \\
C & F \\
S & H \\
E & A \\
C & Z \\
E & I \\
C & N \\
G & I \\
X & A \\
N & I
\end{matrix}$$

$i = 3$ $\qquad$ $j = 7$

count the pairs

1 TO, 1 RA

2 NN, ...

- $n_{TO}^{(3,7)} = 1$
- $n_{RA}^{(3,7)} = 1$
- $n_{NN}^{(3,7)} = 2$

$\vdots$

$$C_{37} = p_{TO} \cdot \cancel{\log\left(\overset{1}{n_{TO}^{(3,7)}}\right)}^{\,0} + p_{RA} \cdot \log\left(\overset{1}{n_{RA}^{(3,7)}}\right) + p_{NN} \cdot \log\left(n_{NN}^{(3,7)}\right) + \dots$$

$$\begin{array}{c|c}
 & 10 \\
\hline
10 & \begin{bmatrix} & & \\ \text{———} & C_{37} & \text{—} \\ & & \end{bmatrix}
\end{array}$$

6

Compute all $C_{i;j}$

If col j not follow col i then C_{ij} small.

Define $f^{(ij)}_{\alpha\beta} = \frac{n^{(ij)}_{\alpha\beta}}{N}$. When two columns i and j **were not consecutive in the plaintext**, then

$$\begin{aligned}
C_{ij} &= \sum_{\alpha,\beta} p_{\alpha\beta} \log(N \cdot f^{(ij)}_{\alpha\beta}) \\
&= \log(N) + \sum_{\alpha,\beta} p_{\alpha\beta} \log(f^{(ij)}_{\alpha\beta}) \\
&\le \sum_{\alpha,\beta} p_{\alpha\beta} \log(p_{\alpha\beta})
\end{aligned}$$

$$\begin{bmatrix} \ddots & & \\ s \ s \ 0 & & s \ s \ B \ s \\ & & \ddots \end{bmatrix}$$

so C_{ij} is much smaller comparing to when two columns i and j were consecutive in the plaintext.

So **if** we guessed the correct period then the matrix $[C_{ij}]_{1\le i\ne j\le k}$ will have a substantially bigger number in each row, **except one.**

- If C_{ij} is the substantially big number on row i then j follows i in the decryption permutation.
- If row k is the only row with no substantially big entry, then k is the first entry in the decryption permutation.

$$\begin{bmatrix} 0 & s & s & B & s \\ s & 0 & B & s & s \\ s & s & 0 & s & s \\ s & s & s & 0 & B \\ s & B & s & s & 0 \end{bmatrix}$$

B's cannot be on the same row/col.