COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

University of Houston

Entropy and Information Grain

Probability of winning using independents events

- What is the probability to draw the same sequence
- Set1=AAAA 1 1 1 1

  → P(winning)1*1*1*1=1
- Set2=AAAB 0.75 0.75 0.75 0.25

  → P(winning)0.75*0.75*0.75*0.25=0.105
- Set3=AABB 0.5 0.5 0.5 0.5
- → P(winning)0.5*0.5*0.5*0.5=0.0625

	P(Winning)	$-\log_2 p$(Winning)	Entropy= average
AAAA	1*1*1*1	0+0+0+0	0
AAAB	0.75*0.75*0.75*0.25	0.415+0.415+0.415+0.2	0.81
AABB	0.5*0.5*0.5*0.5	1+1+1+1	1

AAAAABBB ➔Entropy formula?

$$=-5/8\log_2(5/8) - 3/8\log_2(3/8)$$

What if we have more than 2 ?

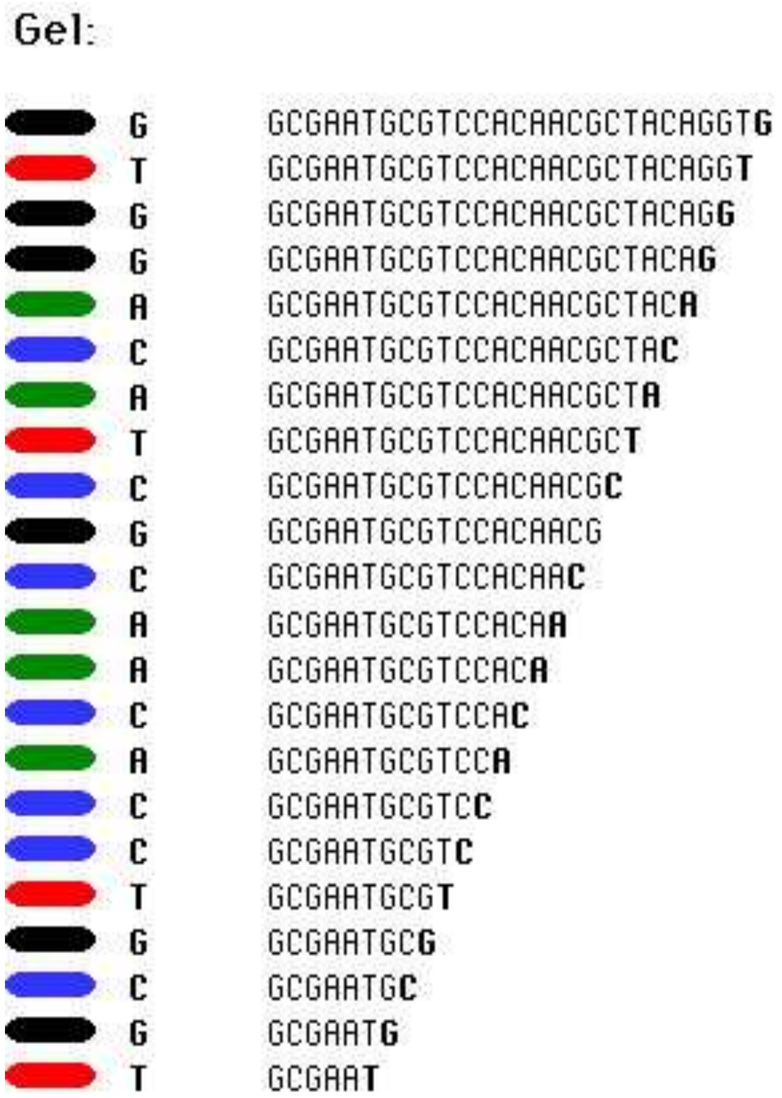
Probability of winning using MORE classes

	P(Winning)	$-\log_2 p$(Winning)	E
AAAAAAAA	1*1*1*1*1*1*1*1	0+0+0+0+0+0+0+0	0
AAAABBCD	0.5*0.5*0.5*0.5*0.25*0.25*0.125 *0.125	-1/2log0.5-1/4log0.25-1/8log0.125-1/8log0.125	1.75
AABBCCDD	0.25*0.25*0.25*0.25*0.25*0.25* 0.25*0.25	-8/8log0.25	2

Shannon ➔Entropy is the average number of questions needed to get an answer (Bits)

The entropy concept in information theory first time coined by Claude Shannon (1850).

DNA sequencing



Entropy & Bits

- You are watching a set of independent random sample of X
- X has 4 possible values:

$$P(X=A)=1/4, P(X=B)=1/4, P(X=C)=1/4, P(X=D)=1/4$$

- You get a string of symbols ACBABBCDADDC…
- To transmit the data over binary link you can encode each symbol with bits (A=00, B=01, C=10, D=11)
- You need 2 bits per symbol

Fewer Bits – example 1

- Now someone tells you the probabilities are not equal

$P(X=A)=1/2, P(X=B)=1/4, P(X=C)=1/8, P(X=D)=1/8$

- Now, it is possible to find coding that uses only 1.75 bits on the average. How?

Fewer bits – example 2

- Suppose there are three equally likely values

  $P(X=A)=1/3, P(X=B)=1/3, P(X=C)=1/3$

- Naïve coding: A = 00, B = 01, C=10
- Uses 2 bits per symbol
- Can you find coding that uses 1.6 bits per symbol?

- In theory it can be done with 1.58496 bits

Entropy – General Case

- Suppose X takes *n* values, $V_1, V_2, \ldots V_n$, and

$$P(X=V_1)=p_1,\ P(X=V_2)=p_2,\ \ldots\ P(X=V_n)=p_n$$

- What is the smallest number of bits, on average, per symbol, needed to transmit the symbols drawn from distribution of X? It's

$$H(X) = p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots\ p_n \log_2 p_n$$

$$= -\sum_{i=1}^{n} p_i \log_2 (p_i)$$

- $H(X)$ = the entropy of X

High, Low Entropy

- "High Entropy"
  - X is from a uniform like distribution
  - Flat histogram
  - Values sampled from it are less predictable
- "Low Entropy"
  - X is from a varied (peaks and valleys) distribution
  - Histogram has many lows and highs
  - Values sampled from it are more predictable

Specific Conditional Entropy, H(Y|X=v)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- I have input X and want to predict Y
- From data we estimate probabilities

  P(LikeG = Yes) = 0.5

  P(Major=Math & LikeG=No) = 0.25

  P(Major=Math) = 0.5

  P(Major=History & LikeG=Yes) = 0

- Note

  H(X) = 1.5

  H(Y) = 1

Specific Conditional Entropy, H(Y|X=v)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- Definition of Specific Conditional Entropy
- $H(Y|X=v)$ = **entropy of *Y* among only those records in which *X* has value *v***
- Example:

$H(Y|X=Math) = 1$

$H(Y|X=History) = 0$

$H(Y|X=CS) = 0$

Conditional Entropy, H(Y|X)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- Definition of Conditional Entropy

$H(Y|X)$ = **the average conditional entropy of** ***Y***

$$= \Sigma_i \, P(X=v_i) \, H(Y|X=v_i)$$

- Example:

v_i	$P(X=v_i)$	$H(Y\|X=v_i)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$$H(Y|X) = 0.5*1+0.25*0+0.25*0 = 0.5$$

Information Gain

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- Definition of Information Gain
- $IG(Y|X)$ = **I must transmit** *Y.*

  How many bits on average would it save me if both ends of the line knew X?

$$IG(Y|X) = H(Y) - H(Y|X)$$

- Example:

  $H(Y) = 1$

  $H(Y|X) = 0.5$

  Thus:

  $IG(Y|X) = 1 - 0.5 = 0.5$

Example what IG tells us about the target

- Predict whether someone is going to live more than 80 years
- From consensus data ➔
  Ig(Longlife|haircolor)=0.01 (tells nothing!)
  IG (Longlife|Smoker)=0.2
  IG (Longlife|Gender)=0.25
  IG (Longlife|last4digitsSSn)=0.00001
- If IG(y|x) ➔x is a good attribute to spilt on