Data Mining

#2 – Information theory

מציאת תבניות אינפורמטיביות

- אנו זקוקים לקריטריון על מנת לבחור תבניות אינפורמטיביות מתוך כלל התבניות הקיימות במסד הנתונים.
- תורת האינפורמציה מספקת מסגרת מתמטית פורמלית שמאפשרת לנו לזהות תבניות אינפורמטיביות, ולהעריך את מידת האינפורמטיביות שלהן ביחס לתבניות אחרות.

תורת האינפורמציה

- **<u>אי-וודאות</u>**: המידע המוגבל שיש לנו לגבי תוצאה של אירוע כלשהו, בד"כ אירוע עתידי
- **<u>אנטרופיה</u>**: המטרה של מדד זה היא להעריך את אי-הוודאות של משתנה מקרי כלשהו (X)

Motivating Entropy

- Let's assume that we have a device that emits one symbol (A). We have no uncertainty as to what we will see: uncertainty is *zero*.
- A device that emits two symbols (A, B). We have one choice, either A or B: our uncertainty is *one*, because we could use one bit (0 or 1) to encode the outcome.
- A device that emits four symbols (A, B, C, D). We would need *two* bits (00, 01, 10, 11) to encode the outcome.
- What we are describing is a $\log_2 M$, where M is the number of symbols.
- Entropy = average number of bits required to transmit the signal.

Adopted from: www9.georgetown.edu/faculty/mad87/07/420/

Entropy explained

- Let's consider a set of possible events with equal probability (a fair dice with values from *1* to *n*). The *uncertainty* for such set of outcomes is defined by $u = \log_2 n$

https://www.miniwebtool.com/log-base-2-calculator/

- The logarithm is used to provide the **additive** characteristic for independent uncertainty.
  - The uncertainty of playing with two dice (n*m possible outcomes) is obtained by adding the uncertainty of the second dice to the uncertainty of the first dice:
  - $u = \log_2(n*m) = \log_2 n + \log_2 m$.

Entropy explained, cont.

- Now return to the playing with one dice only (the first one); since the probability of each event is *1/n*, we can write

  $$u = \log_2(1/p(x_i)) = -\log_2(p(x_i)),\ \forall\ 1 \leq i \leq n$$

- In the case of a non-uniform probability mass function (or distribution in the case of continuous random variable), we let $u_i = -\log_2(p(x_i))$, (the lower the probability, the higher the uncertainty or the surprise)

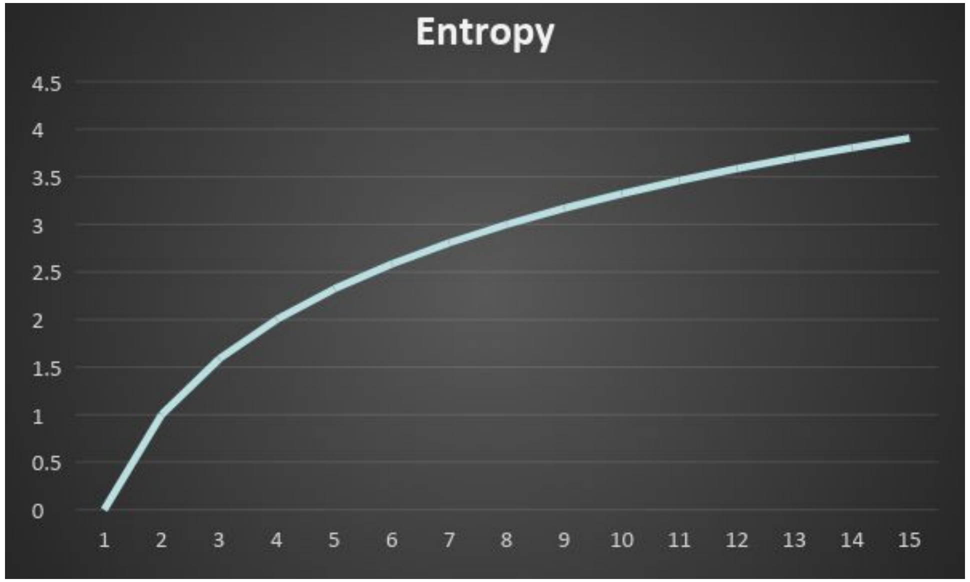
- The average uncertainty is obtained by $\Sigma\, p(x_i) \cdot u_i = -\Sigma\, p(x_i) \cdot \log_2 p(x_i)$ and is used as the definition of the information entropy

Entropy

- **Entropy** $H(X) = -\sum p(x) \cdot \log_2 p(x)$

  Where:
  - X - a discrete random variable
  - x - value of X
  - $p(x)$ - probability of x
- **Interpretation**: measure of uncertainty of X.

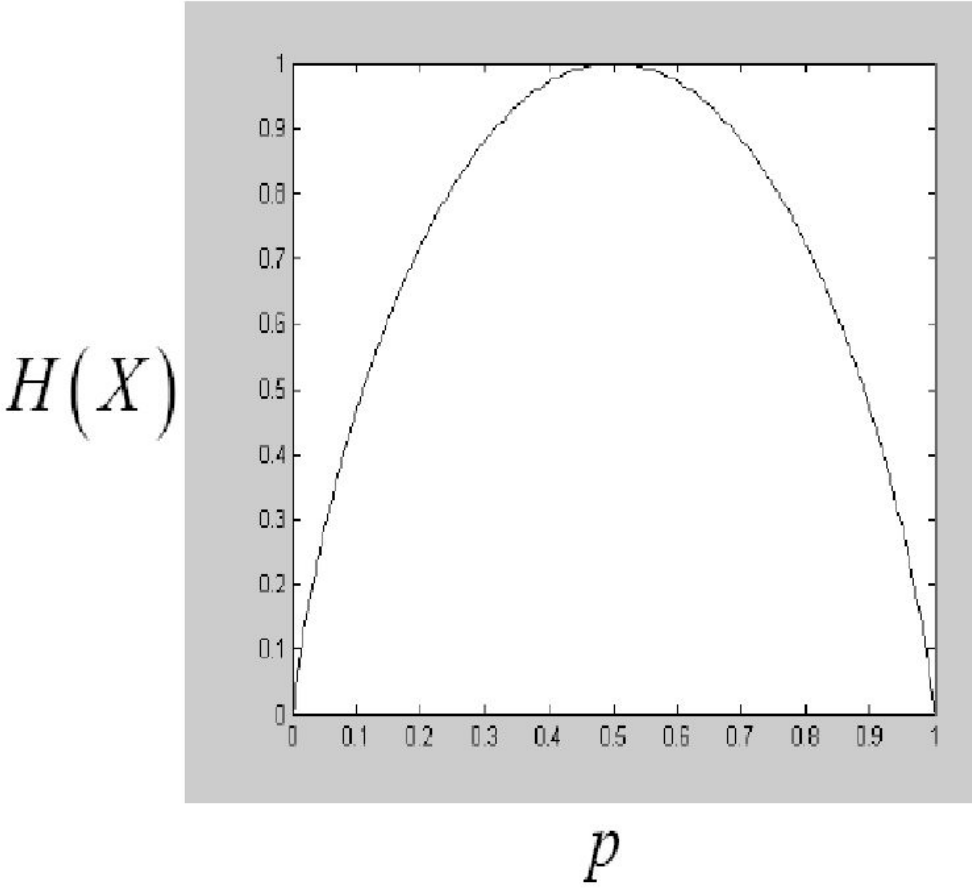
Entropy



אם לכל ערכי המשתנה X יש את אותה הסתברות, האנטרופיה היא פונקציה מונוטונית עולה.

Entropy - with two states

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1\text{-}p \end{cases}$$



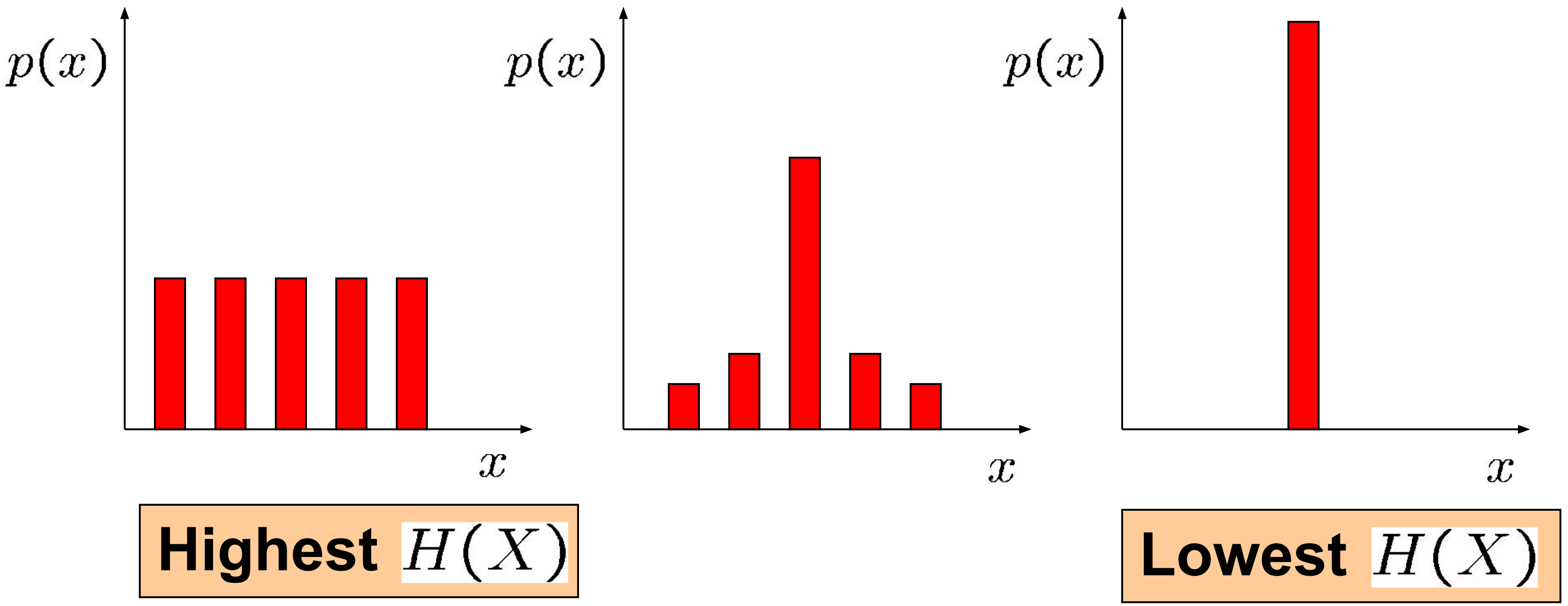
$$H(X) = -\sum_i \Pr(X = x_i) \log_2 \Pr(X = x_i)$$

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p)$$

In general:
Entropy is maximal and equal to $\log_2 n$ **if all** ***n*** **states have the same probability.**

Adopted from:
http://www.c3.chalmers.se/undergraduate/courses0405/css155/lectures/lecture2.pdf

Entropy - Illustration



Adopted from:
www.wisdom.weizmann.ac.il/~deniss/vision_spring04/files/infomation_bottleneck/InformationBottleneck.ppt

Entropy – Small Examples

$$A_n \in \{0,1\}$$

$$p_0 = \Pr(A_n = 0)$$

$$p_1 = \Pr(A_n = 1) = 1 - p_0$$

Example: $p_0 = 0.2$, $p_1 = 0.8$

$-0.2 * \log_2(0.2) - 0.8 * \log_2(0.8) =$ **0.72**

1,1,1,0,1,1,0,1,1,1

Source

Example 2.1 (Weather). As a specific weather example, suppose that weather in California is either sunny or cloudy with probabilities 7/8 and 1/8, respectively. The entropy of this source of information is the average information of sunny and cloudy days. Hence

$$H = -(7/8)\log(7/8) - (1/8)\log(1/8) = .54 \text{ bits.}$$

From: http://press.princeton.edu/chapters/s02_8214.pdf
http://www.s2.chalmers.se/undergraduate/courses0405/ess155/lectures/lecture2.pdf

Entropy – Examples

	x=A	x=B	x=C	x=D	Total
Data	1	2	30	5	**38**
p(x)	1/38	2/38	30/38	5/38	
-log(p(x))	5.25	4.25	0.34	2.93	
-p(x)*log(p(x))	0.14	0.22	0.27	0.38	$\sum$ **1.016** bit

H(X)

	y=A	y=B	y=C	y=D	Total
Data	19	21	22	18	**80**
p(x)	19/80	21/80	22/80	18/80	
-log(p(x))	2.07	1.93	1.86	2.15	
-p(x)*log(p(x))	0.49	0.51	0.51	0.48	$\sum$ **1.995** bit

H(Y)

Conditional Entropy

- $H(Y|X) = -\sum p(x,y) \cdot \log p(y|x)$
- Where
- X, Y – discrete random variables
- $p(x,y)$ – joint probability of x and y
- $p(y|x)$ – conditional probability of y given x
- **Interpretation**: measure of uncertainty of Y, when X is given.

Conditional Entropy

- **If Y(X) $\longrightarrow$ H(Y|X)=0**
- **If Y not dependent on X $\longrightarrow$ H(Y|X)=H(Y)**

Conditional Entropy – Example

Data	y=A	y=B	y=C	y=D	Total
x=A	6	8	2	5	**21**
x=B	9	10	3	11	**33**
x=C	15	1	2	1	**19**
x=D	14	5	6	2	**27**
Total	**44**	**24**	**13**	**19**	**100**

H(Y)=1.85

P(x,y)	y=A	y=B	y=C	y=D	**Total**
x=A	6/100	8/100	2/100	5/100	**21/100**
x=B	9/100	10/100	3/100	11/100	**33/100**
x=C	15/100	1/100	2/100	1/100	**19/100**
x=D	14/100	5/100	6/100	2/100	**27/100**
Total	**44/100**	**24/100**	**13/100**	**19/100**	**100/100**

P(y|x)	y=A	y=B	y=C	y=D	**Total**
x=A	6/21	8/21	2/21	5/21	**21/21**
x=B	9/33	10/33	3/33	11/33	**33/33**
x=C	15/19	1/19	2/19	1/19	**19/19**
x=D	14/27	5/27	6/27	2/27	**27/27**

-P(x,y)*log(P(y|x))	y=A	y=B	y=C	y=D	**Total**
x=A	0.11	0.11	0.07	0.10	**0.39**
x=B	0.17	0.17	0.10	0.17	**0.62**
x=C	0.05	0.04	0.06	0.04	**0.20**
x=D	0.13	0.12	0.13	0.08	**0.46**
Total	**0.46**	**0.45**	**0.37**	**0.40**	**1.67**

H(Y|X)

Conditional Entropy – Example

Data	y=A	y=B	y=C	y=D	Total
x=A	6	8	2	5	**21**
x=B	9	10	3	11	**33**
x=C	15	1	2	1	**19**
x=D	14	5	6	2	**27**
Total	**44**	**24**	**13**	**19**	**100**

H(X)=1.96

P(x,y)	y=A	y=B	y=C	y=D	**Total**
x=A	6/100	8/100	2/100	5/100	**21/100**
x=B	9/100	10/100	3/100	11/100	**33/100**
x=C	15/100	1/100	2/100	1/100	**19/100**
x=D	14/100	5/100	6/100	2/100	**27/100**
Total	**44/100**	**24/100**	**13/100**	**19/100**	**100/100**

P(x|y)	y=A	y=B	y=C	y=D
x=A	6/44	8/24	2/13	5/19
x=B	9/44	10/24	3/13	11/19
x=C	15/44	1/24	2/13	1/19
x=D	14/44	5/24	6/13	2/19
Total	**44/44**	**24/24**	**13/13**	**19/19**

-P(x,y)*log(P(x|y))	y=A	y=B	y=C	y=D	Total
x=A	0.17	0.13	0.05	0.10	**0.45**
x=B	0.21	0.13	0.06	0.09	**0.48**
x=C	0.23	0.05	0.05	0.04	**0.38**
x=D	0.23	0.11	0.07	0.06	**0.48**
Total	**0.84**	**0.41**	**0.24**	**0.29**	**1.78**

H(X|Y)

Mutual Information

- **Mutual Information** (of variables *X* and *Y*)

I(X;Y) = H(Y) - H(Y/X) = $$\sum_{x,y} p(x,y) \bullet \log \frac{p(y/x)}{p(y)}$$

- **Interpretation**: the reduction in the uncertainty of *Y* as a result of knowing *X.*

Mutual Information - Symmetry

I(X;Y) = H(Y) - H(Y|X) = **I(Y;X)** = H(X) – H(X|Y)

- *I(X;Y) =* $$\sum_{x,y} p(x,y) \bullet \log \frac{p(y/x)}{p(y)}$$
- $$p(y \mid x) = \frac{p(x,y)}{p(x)}$$
- *I(X;Y) =* $$\sum_{x,y} p(x,y) \bullet \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual information

- **Symmetry $\rightarrow$ I(X;Y)= I(Y;x)**
- **MI always positive or zero**
- **Max(MI) $\rightarrow$ Y function of X**
- **Min(MI) $\rightarrow$ No connection between Y and X**

Mutual Information - Example

Data	y=A	y=B	y=C	y=D	Total
x=A	6	8	2	5	**21**
x=B	9	10	3	11	**33**
x=C	15	1	2	1	**19**
x=D	14	5	6	2	**27**
Total	**44**	**24**	**13**	**19**	**100**

$$\sum_{x,y} p(x,y) \bullet \log \frac{p(y/x)}{p(y)}$$

P(x,y)	y=A	y=B	y=C	y=D	**Total**
x=A	6/100	8/100	2/100	5/100	**21/100**
x=B	9/100	10/100	3/100	11/100	**33/100**
x=C	15/100	1/100	2/100	1/100	**19/100**
x=D	14/100	5/100	6/100	2/100	**27/100**
Total	**44/100**	**24/100**	**13/100**	**19/100**	**100/100**

	y=A	y=B	y=C	y=D	**Total**
P(y)	44/100	24/100	13/100	19/100	**100/100**

P(y|x)	y=A	y=B	y=C	y=D	**Total**
x=A	6/21	8/21	2/21	5/21	**21/21**
x=B	9/33	10/33	3/33	11/33	**33/33**
x=C	15/19	1/19	2/19	1/19	**19/19**
x=D	14/27	5/27	6/27	2/27	**27/27**

P(x,y)*log(P(y|x) / P(y))	y=A	y=B	y=C	y=D	**Total**
x=A	-0.04	0.05	-0.01	0.02	**0.02**
x=B	-0.06	0.03	-0.02	0.09	**0.05**
x=C	0.13	-0.02	-0.01	-0.02	**0.08**
x=D	0.03	-0.02	0.05	-0.03	**0.03**
Total	**0.06**	**0.05**	**0.02**	**0.06**	**0.18**

I(X;Y)

Mutual Information – Example

Another way to calculate

	y=A	y=B	y=C	y=D	**Total**
P(y)	44/100	24/100	13/100	19/100	**100/100**

↓

	y=A	y=B	y=C	y=D	**Total**
H(Y)=-P(y)*log(P(y))	0.52	0.49	0.38	0.46	1.85

↓

$$I(X;Y) = H(Y)-H(Y|X) = 1.85 - 1.67 = \mathbf{0.18}$$

Mutual Information – Example

	x=A	x=B	x=C	o=D	**Total**
P(x)	21/100	33/100	19/100	27/100	**100/100**

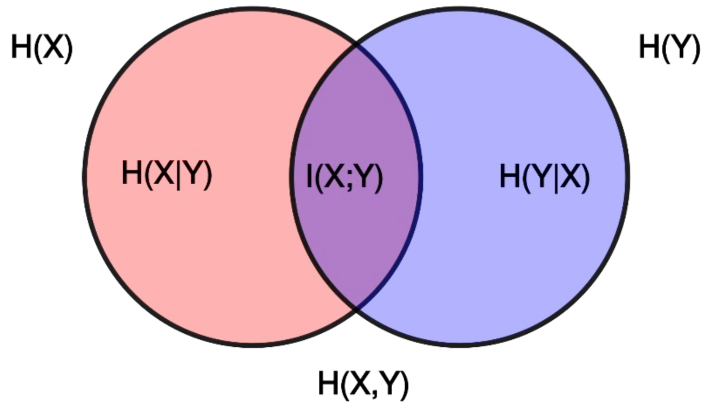
	y=A	y=B	y=C	y=D	**Total**
H(Y)=-P(y)*log(P(y))	0.47	0.53	0.46	0.51	1.96

I(Y;X) = H(X)-H(X|Y) = 1.96 - 1.78 = **0.18**

And via chain rule

- Chain rule: $H(X,Y) = H(X) + H(Y \mid X)$

- $\mathrm{I(X;Y) = H(Y) - H(Y \mid X) = H(Y) - (H(X,Y) - H(X)) = H(Y) + H(X) - H(X,Y)}$



https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Entropy-mutual-information-relative-entropy-relation-diagram.svg/2000px-Entropy-mutual-information-relative-entropy-relation-diagram.svg.png