

ECE253/CSE208 Introduction to Information Theory

Lecture 3: Shannon Information and Entropy

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 2 of *Elements of Information Theory (2nd Edition)* by Thomas Cover & Joy Thomas
- *Introduction to Learning & Decision Trees*, CMU

Information is Surprise¹

- Shannon thought of Many ways to communicate (phone calls, msg, email, Morse code, etc) as ways of producing strings of symbols.
- How to measure the amount of information contained in such a string?
- How about counting the number of words in a message? Then, “The Sun will rise tomorrow” has the same information value as “The world will end tomorrow”.
- Shannon’s idea: **information is related to what’s new; related to surprise.**
- Think about reading news, words like “the” or “a” are pretty unsurprising (redundant since we understand the message without them). The real essence of the message lies in words that aren’t as common, such as “alien” or “invasion”.
- This suggests a measure of surprise for each word: the frequency with which it appears in the English language.
- More generally, **the information of an outcome is inversely proportional to its probability.**

¹Source: <https://plus.maths.org/content/information-surprise>

Shannon Information: A Measure of Surprise

The amount of information generated at the source X by the occurrence of the outcome x is defined as

$$I(x) := \log \frac{1}{p(x)}.$$

The less probable an event is, the more surprising it is and has more info.

Why logarithm:

- It is a monotonically increasing transformation
- It converts the value of info into number of bits
- It makes the information additive: $I(xy) = \log \frac{1}{p(x)p(y)} = I(x) + I(y)$
- Must-happen thing carries no information: $p(x) = 1 \implies I(x) = 0$

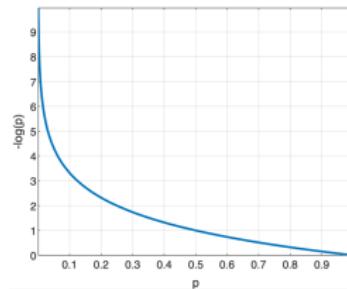


Figure: The function plot of Shannon Information.

Entropy: A Central Role in Information Theory

Entropy is a measure of uncertainty of a random variable, which is defined as the average Shannon information $\text{E}_X[I(X)]$.

Definition (Entropy)

$$H(X) := \text{E}_X[I(X)] = \text{E}_X \left[\log \frac{1}{p(X)} \right] = - \sum_{x \in \mathcal{X}} p(x) \log p(x).$$

- $p(x) = 0 \implies 0 \log 0 = 0$ (by convention or justified by the L'Hopital's rule).
- Entropy is always nonnegative: $H(X) \geq 0$.
- Units for different log bases: $\log_2 \rightarrow$ bits; $\log_e \rightarrow$ nats; $\log_{10} \rightarrow$ harleys/dits/bans
- $H_b(X) = (\log_b a) H_a(X) \Leftarrow$ the change of base formula: $\log_b x = \frac{\log_a x}{\log_a b}$.

Different Bases → Different Units

units

log basis	units
2	bit
e	nat
10	ban or hartley
256	byte
2^{1000}	kilobit
2^{8000}	kilobyte
2^{10^6}	megabit
$2^{2^{95^6}}$	megabyte
2^{10^9}	gigabit
$2^{2^{12}5^9}$	gigabyte

Axiomatic Definition of Entropy

Entropy should satisfy the following properties:

Continuity [$H(p, 1 - p)$ is continuous in p]: The amount of info associated with an outcome should vary continuously as the probability of that outcome changes.

Maximum [$H(0.5, 0.5) = 1$]: The amount of info associated with a set of outcomes cannot be increased if those outcomes are equally probable.

Grouping [$H(p_1, p_2, \dots, p_m) = H(p_1 + p_2, p_3, \dots, p_m) + (p_1 + p_2)H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$]: The amount of info associated with a set of outcomes is obtained by adding the info of each outcome.

Question 2.46 (page 53) of Cover-Thomas's book: Start from these properties and prove that $H(p_1, p_2, \dots, p_m) = -\sum_{i=1}^m p_i \log p_i$ is the only function satisfying them.

See more properties of $H(X)$: <http://www.mtm.ufsc.br/~taneja/book/node6.html>

Entropy: A Measure of the Disorder or Randomness of a System

- The term entropy was coined in 1865 by the German physicist Rudolf Clausius from Greek word “en-” = in + “trope” = change or transformation.
- Second law of thermodynamics: Entropy of an isolated system increases over time
- Statistical mechanics: Entropy (Gibbs entropy or Boltzmann's entropy) is a thermodynamic property of a system, which is a bridge between the microscopic world (positions and momenta of all the atoms) and the macroscopic (e.g., total energy E , volume V , pressure P , temperature T , etc).
- Gibbs entropy: the entropy of the system is $S = -\kappa_B \sum_i p_i \ln p_i$, where $\kappa_B = 1.380649 \times 10^{-23} \text{ J/K}$ is the Boltzmann's constant; p_i is the probability that microstate i has the energy E_i during the system's fluctuations. Therefore, the macrostate of a system is characterized by a distribution on the microstates.
- Entropy is determined by the number of random microstates. It's related to the amount of additional information needed to specify the exact physical state.
- Entropy and Energy: Low entropy = energy concentrated; High entropy = energy spreads out²

²https://www.youtube.com/watch?v=YM-uykVfq_E&ab_channel=TED-Ed

Why the Metric is Called Entropy

Why Call This Metric Entropy?

“My greatest concern was what to call it. I thought of calling it ‘information’, but the word was overly used, so I decided to call it ‘uncertainty’. When I discussed it with John von Neumann, he had a better idea.”

- Claude Shannon



“You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name. In the second place, and more importantly, no one knows what entropy really is, so in a debate you will always have the advantage.”

- John von Neumann



Entropy as a Measure of Average Information

Given a random variable X , what is the average information of it?

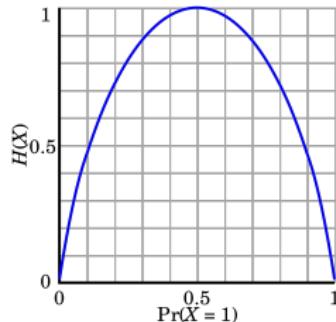
Example (Bernoulli distribution)

Consider a Bernoulli random variable $X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1 - p \end{cases}$

$$H(X) = -p \log p - (1 - p) \log(1 - p) =: f(p)$$

Special cases: $H(0.5) = 1$ bit; $H(0) = H(1) = 0$. ■

$H(X)$ achieves its maximum when $p = \frac{1}{2}$. That is, X has the greatest uncertainty when both outcomes are equally likely.



Implication 1: Entropy for Equally Probable Outcomes

We can use 2 bits ($\{00, 01, 10, 11\}$) to fully describe a system of 4 equally probable states. In general, $\log_2(m)$ bits are needed for m equally probable states.

- For uniform distribution random variable $X \in \mathcal{X} \implies H(X) = \log(|\mathcal{X}|)$.
- Uniform probability yields maximum uncertainty and therefore maximum entropy.
- For equiprobable events, the entropy increases with the number of outcomes \implies Hence, casting a die has higher entropy than tossing a coin.

Example (An unfair coin)

Suppose we have a two-sided coin where $P(X_h) = 0.9, P(X_t) = 0.1$.

$$H(X) = 0.469 \text{ bits} \implies m = 2^{H(X)} = 1.38 \text{ equally probable outcomes}$$

A coin with $H(X) = 0.469$ bits has the same entropy as a dice with 1.38 sides. ■

Insight:

A random variable with entropy $H(X)$ bits provides enough Shannon information to choose $m = 2^{H(X)}$ equally probable outcomes.

Implication 2: Entropy for Average Description Length

Example (Encoding a message)

Bob will send a message composed of letters from the alphabet $\{A, B, C, D\}$ to Alice. He wants to encode a string from this alphabet as a sequence of bits. If each letter occurs with equal probability, consider the following ASCII encoding scheme:

$$A \rightarrow 00; \quad B \rightarrow 01; \quad C \rightarrow 10; \quad D \rightarrow 11 \implies \text{average length} = 2 \text{ bits per symbol.}$$

Now, if those four letter occur with different probabilities and are coded as:

A occurs with frequency 0.70

B occurs with frequency 0.26

C occurs with frequency 0.02

D occurs with frequency 0.02

Consider the coding scheme: $A \rightarrow 0; \quad B \rightarrow 10; \quad C \rightarrow 110; \quad D \rightarrow 111 \implies$
average length = $0.7 \times 1 + 0.26 \times 2 + 0.02 \times 3 + 0.02 \times 3 = 1.34$ bits
while $H(0.7, 0.26, 0.02, 0.02) = 1.09$ bits.



Implication 2: Entropy for Average Description Length (cont'd)

The above example reflects the key idea of Huffman encoding: More likelihood symbols should be assigned with shorter length of codes to minimize the average length of the description.

Insight:

Entropy $H(X)$ provides a lower bound on the average length of the shortest description of the random variable X . That is, entropy can be viewed as the cost of encoding the specific distribution X .

Theorem (Shannon's Source Coding Theorem)

Given a categorical random variable X over a finite source alphabet \mathcal{X} and a code alphabet \mathcal{A} , then for all uniquely decodable $C : \mathcal{X} \rightarrow \mathcal{A}^$, it holds that*

$$\mathbb{E}[|C(X)|] \geq H(X).$$

Joint and Conditional Entropy

Consider a pair of random variables (X, Y) with joint distribution $p(x, y)$.

Joint entropy:

$$H(X, Y) = \text{E}_{(X, Y)} \left(\log \frac{1}{p(X, Y)} \right) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y).$$

Conditional entropy:

$$\begin{aligned} H(Y|X) &= \text{E}_{(X, Y)} \left(\log \frac{1}{p(Y|X)} \right) = - \sum_{x, y} p(x, y) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \end{aligned}$$

- $H(Y|X = x)$ is the specific conditional entropy, i.e., the entropy of Y knowing that X takes a specific value x . $H(Y|X)$ is the expectation of $H(Y|X = x)$.
- $H(f(X)|X) = 0$ for any function $f(\cdot) \Leftarrow$ no randomness of $f(X)$ if X is given.
- If $X \perp\!\!\!\perp Y \implies H(X|Y) = H(X), \quad H(Y|X) = H(Y)$
- $H(X|Y) \neq H(Y|X)$ in general.

Joint and Conditional Entropy (Cont'd)

Chain rule: $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$

Proof: Take expectation on both sides of $\log p(X, Y) = \log p(X) + \log p(Y|X).$

Interpretation:

- $H(Y|X)$ measures how much information, on average, X encodes about Y .
- Now if we first learn the value of X , we have gained $H(X)$ bits of information. Once X is known, we only need $H(Y|X) = H(X, Y) - H(X)$ additional bits to describe the state of the whole system.
- Joint uncertainty of (X, Y) is equal to uncertainty of X plus uncertainty of Y after knowing X .

Corollary

- *Bayes' rule for conditional entropy:* $H(Y|X) = H(X|Y) - H(X) + H(Y)$
- $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$
- $H(f(X)) \leq H(X)$ for any function $f(\cdot) \rightarrow$ post-processing reduces entropy

Example

Consider two random variables X and Y, whose joint distribution is given below. Find $H(X)$, $H(Y)$, $H(X|Y)$, $H(Y|X)$, and $H(X,Y)$.

$Y \setminus X$	1	2	3	4	$p(y)$
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
$p(x)$	1/2	1/4	1/8	1/8	

$$H(X) = H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) = -\left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4} + 2 \times \frac{1}{8} \log \frac{1}{8}\right) = \frac{7}{4} \text{ bits}$$

$$H(Y) = H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = 2 \text{ bits}$$

$$\begin{aligned} H(X|Y) &= \sum_y p(y) H(X|Y=y) \\ &= \frac{1}{4} \times \left(H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + H(1, 0, 0, 0) \right) \\ &= \frac{11}{8} \text{ bits.} \end{aligned}$$

Similarly, we can get $H(Y|X) = \frac{13}{8}$ bits, $H(X, Y) = \frac{27}{8}$ bits.

Relative Entropy (Kullback-Leibler Divergence)

Definition (Relative Entropy)

The relative entropy (a.k.a. KL divergence) between two probability massive functions $p(x)$ and $q(x)$ over **the same sample space (underlying set of events)** is defined as:

$$D(p||q) = E_p \left(\log \frac{p(X)}{q(X)} \right) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- $0 \times \log(\frac{0}{0}) = 0$; $D(p||q) = \infty$, $\forall x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$.
- $\log \frac{p(x)}{q(x)}$ is the log likelihood ratio (LLR). Hence, KL divergence is the expected LLR.
- Distribution p represents the data/observations/true distribution.
Distribution q represents a theory/model/estimated distribution.
- In Bayesian inference, $D(p||q)$ is a measure of the information gained when one revises the belief from the prior distribution q to the posterior distribution p . In other words, it's the amount of info lost when q is used to approximate p .

KL Divergence (Cont'd)

$D(p||q)$ is the average number of extra bits needed to encode the data, because we used distribution q to encode the data instead of the true distribution p .

Example (Extra bits for approximate distribution)

	a	b	c
Distribution p	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Distribution q	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

	a	b	c
Distribution p	0	01	11
Distribution q	01	0	11

Table: Left: PMF of two distributions. Right: The Huffman codes.

Hence, we get the average code lengths and the extra bits as follows:

$$\mathbb{E}[\ell(p)|p] = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = \frac{3}{2}$$

$$\mathbb{E}[\ell(q)|p] = \frac{1}{2} \times 2 + \frac{1}{4} \times 1 + \frac{1}{4} \times 2 = \frac{7}{4}$$

$$\mathbb{E}[\ell(q) - \ell(p)|p] = \frac{7}{4} - \frac{3}{2} = \frac{1}{4}$$

$$D(p||q) = \frac{1}{2} \times \log(2) + \frac{1}{4} \times \log\left(\frac{1}{2}\right) = \frac{1}{4}$$

KL Divergence (Cont'd)

- The KL distance is NOT a metric of distance since it generally does not satisfy the properties of symmetry and the triangle inequality:
$$D(p||q) \neq D(q||p), \quad D(p||q) \nleq D(p||r) + D(r||q)$$
- Other popular divergence: Jensen-Shannon divergence, Rényi divergence, Hellinger divergence, and Wasserstein distance (optimal transport problem).

Example (KL divergence is not symmetric)

Consider two distribution functions $p(x)$ and $q(x)$, where $x \in \mathcal{X} = \{0, 1\}$. Let $p(0) = 1 - r$ and $p(1) = r$ while $q(0) = 1 - s$ and $q(1) = s$. Then,

$$D(p || q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s}$$

$$D(q || p) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}$$

If $r = \frac{1}{2}$ and $s = \frac{1}{4}$, then $D(p || q) = 0.2075$ bits and $D(q || p) = 0.1887$ bits.

A special case: if $r = s \Rightarrow p(x) = q(x) \Rightarrow D(p || q) = D(q || p) = 0$.



Cross Entropy

- $D(p(x)||q(x)) = -\text{E}_p(\log q(X)) + \text{E}_p(\log p(X)) = \overbrace{H(p(x), q(x))}^{\text{cross entropy of } p \text{ and } q} - H(p(x)).$
- $H(p(x), q(x)) \neq H(q(x), p(x))$ in general.
- For fixed $p(x)$, $\arg \min_q H(p(x), q(x)) = \arg \min_q D(p(x)||q(x))$
- Cross entropy $H(p(x), q(x))$ measures the average number of bits needed to identify an event drawn from the estimated distribution q instead of the true distribution p .
- In other words, the average length of a message from distribution p by using the optimal code designed for distribution q .

Supervised Learning: Classification

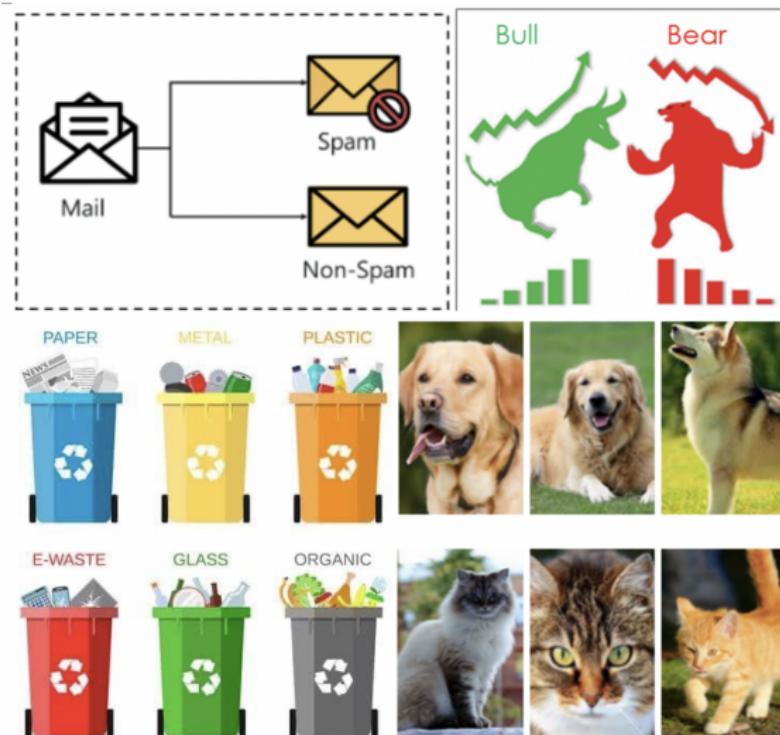


Figure: Examples of classification.

Supervised Learning: Classification

Classification task: Identify to which of a set of classes a new observation belongs, on the basis of a training set of data containing observations whose class membership is known.

For the training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, \mathbf{x}_i is the feature vector (a.k.a. input or explanatory variables, regressors, covariates, predictors), and y_i is the true label of sample i .

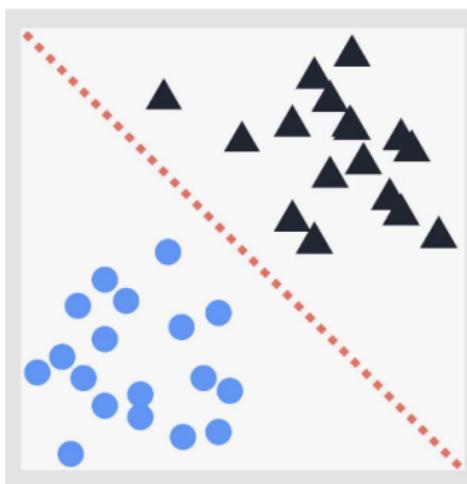


Figure: Binary classification.

Cross Entropy (Cont'd)

Example (Cross-entropy loss function for binary classification)

Let the true probability p_i be the label y_i (hence no uncertainty). A model outputs the predicted probability q_i for classifying sample i by using the logistic function $\ell(\cdot) \implies$

$$\begin{cases} \Pr(y_i = 1 \mid \mathbf{x}_i) = \ell(\mathbf{w} \cdot \mathbf{x}_i) = \frac{1}{1+e^{-\mathbf{w} \cdot \mathbf{x}_i}} =: \hat{y}_i \\ \Pr(y_i = 0 \mid \mathbf{x}_i) = 1 - \hat{y}_i \end{cases}$$

Now, for the distributions $p \in \{y, 1-y\}$ and $q \in \{\hat{y}, 1-\hat{y}\}$, we can use $H(p, q) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$ to measure the dissimilarity.

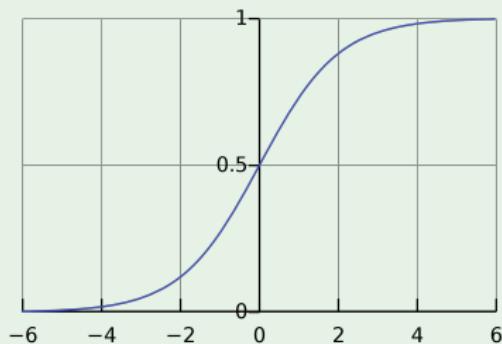


Figure: Logistic function

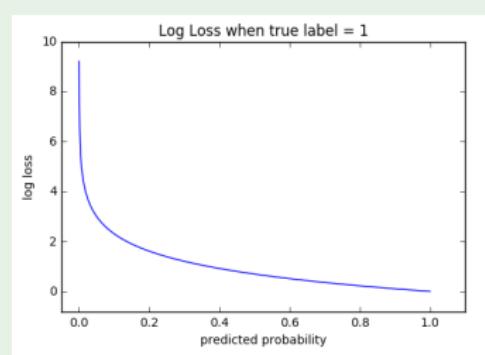


Figure: Cross entropy loss

Cross Entropy (Cont'd)

Example (Cross-entropy loss function for binary classification)

The machine learning training phase is to find the optimal \mathbf{w} , which can be obtained (e.g., via gradient descent) by minimizing the average cross entropy loss:

$$J(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N H(p_n, q_n) = -\frac{1}{N} \sum_{n=1}^N [y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)].$$

Note that the likelihood function of those N samples (assuming i.i.d.) is given as

$$\mathcal{L}\left(\{\hat{y}_i\}_{i=1}^N \mid \{\mathbf{x}_i, y_i\}_{i=1}^N\right) = \prod_{i=1}^N \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i} \implies \boxed{-\frac{1}{N} \log \mathcal{L}(\cdot) = J(\mathbf{w})}$$

Thus, minimizing the cross-entropy is the same as maximizing the log-likelihood.

Note that maximum likelihood estimation (MLE) aims at finding a set of parameters (weights in \mathbf{w}) that best explain the observed data.

“... using the cross-entropy error function instead of the sum-of-squares for a classification problem leads to faster training as well as improved generalization.”

— *Pattern Recognition and Machine Learning (2006), Page 235.*

Mutual Information

Definition (Mutual Information)

Consider two random variables X and Y . The mutual information $I(X; Y)$ is the relative entropy between the joint distribution $p(x, y)$ and the product distribution $p(x)p(y)$:

$$\begin{aligned} I(X; Y) &= D(p(x, y) \parallel p(x)p(y)) \\ &= \text{E}_{p(x, y)} \left(\log \frac{p(X, Y)}{p(X)p(Y)} \right) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \end{aligned}$$

Let us rewrite $I(X; Y)$ as:

$$\begin{aligned} I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x|y)}{p(x)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) \\ &= H(X) - H(X|Y) \end{aligned}$$

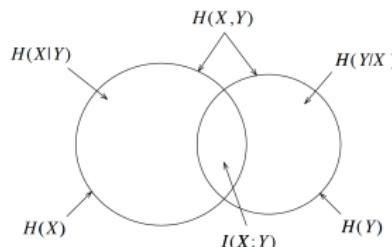
Uncertainty Reduction

Insight:

$I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y .

Properties of mutual information:

- $I(X;Y) = I(Y;X) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, which means that X says as much about Y as Y says about X .
- $I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.
- $I(X;X) = H(X) - H(X|X) = H(X) \rightarrow$ Entropy is self-information.
- Variation of information: $VI(X;Y) = H(X) + H(Y) - 2I(X,Y) = 2H(X,Y) - H(X) - H(Y) = H(X|Y) + H(Y|X)$.



An Application of Mutual Information: Decision Tree

A tree structure: each *internal node* denotes a test on an attribute, each *branch* represents an outcome of the test, and each *leaf* represents a class (or class distribution).

Q: Which attribute to choose for splitting?

A: Choose the most relevant attribute for classification.

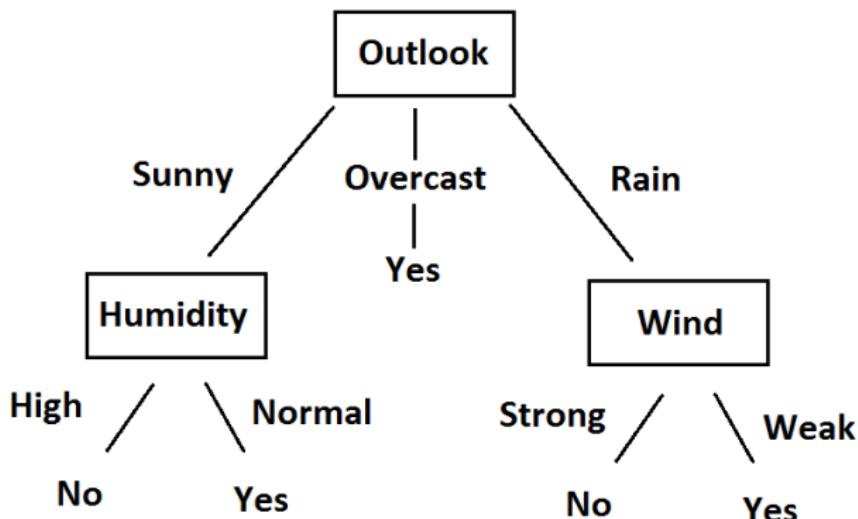


Figure: A decision tree to decide if we will play tennis.

Information Gain

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Figure: Play or not play.

Without any splitting: $H(Y) = -\sum_{i=1}^K p_i \log_2 p_i = -\frac{5}{14} \log_2 \frac{5}{14} - \frac{9}{14} \log_2 \frac{9}{14} = 0.94$ bits.

Information Gain (Cont'd)

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Figure: The Information Gain of by splitting the attribute “Humidity”:
 $IG(\text{humidity}) = H(Y) - H(Y|\text{humidity})$.

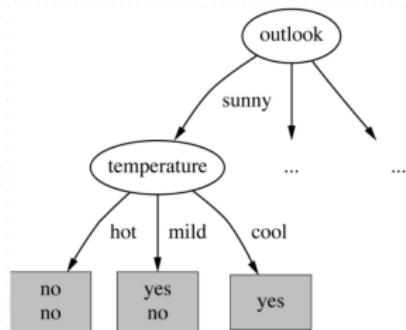
$$H(Y|\text{humidity}) = \Pr(\text{humidity}=\text{high}) \times H(Y|\text{humidity}=\text{high}) + \\ \Pr(\text{humidity}=\text{normal}) \times H(Y|\text{humidity}=\text{normal}) = \\ \frac{7}{14} \times H\left(\frac{3}{7}, \frac{4}{7}\right) + \frac{7}{14} \times H\left(\frac{1}{7}, \frac{6}{7}\right) \implies IG(\text{humidity}) = 0.152$$

Information Gain (Cont'd)

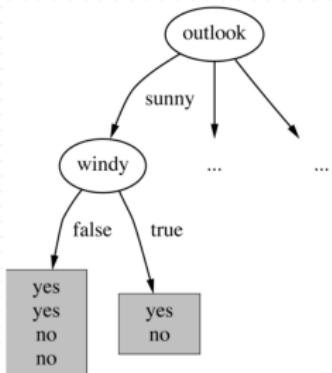
Similarly, $IG(\text{outlook}) = 0.247$, $IG(\text{temperature}) = 0.029$, $IG(\text{windy}) = 0.048$

Hence, the initial split is on **outlook** that has the highest info gain.

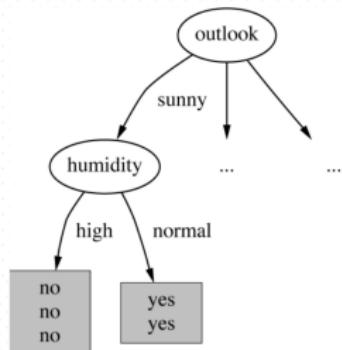
Now search for the best split at the next level:



$$\text{Temperature} = 0.571$$



$$\text{Windy} = 0.020$$



$$\text{Humidity} = 0.971$$

Figure: Humidity has the highest info gain at the second level when **outlook = sunny**.

Information Gain (Cont'd)

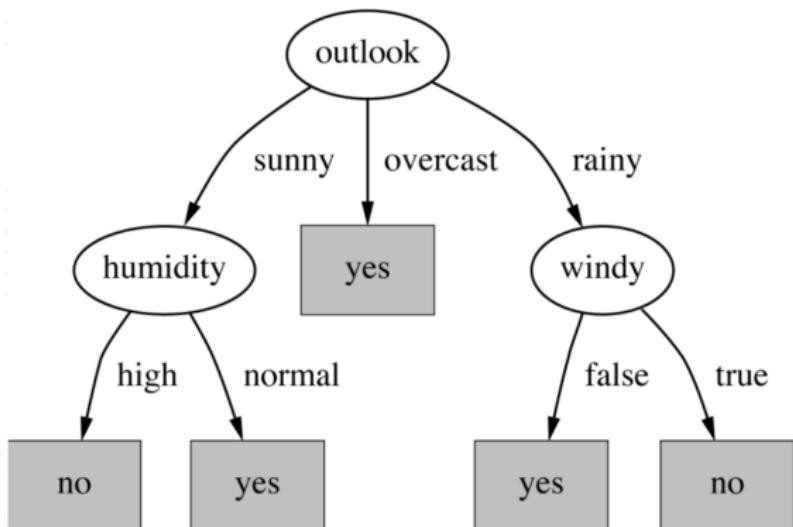


Figure: The final decision tree.

- Not all leaves need to be pure.
- Sometimes similar (even identical) instances have different classes.
- Splitting stops when data cannot be split any further.

Chain Rule of Entropy

Theorem (Chain rule for entropy)

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, X_{i-2}, \dots, X_1).$$

Proof.

$$H(X_1, X_2) = H(X_1) + H(X_2 | X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1)$$

⋮

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_{n-1}, \dots, X_1)$$

$$= \sum_{i=1}^n H(X_i | X_{i-1}, X_{i-2}, \dots, X_1)$$



Chain Rule of Mutual Information

Definition (Conditional mutual information)

The conditional mutual information of random variables X and Y given Z is defined by:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = E_{p(x,y,z)} \left(\log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \right).$$

Theorem (Chain rule for mutual information)

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1).$$

Proof.

$$I(X_1, X_2, \dots, X_n; Y) = H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n | Y)$$

$$\begin{aligned} &= \sum_{i=1}^n H(X_i | X_{i-1}, X_{i-2}, \dots, X_1) - \sum_{i=1}^n H(X_i | X_{i-1}, X_{i-2}, \dots, X_1, Y) \\ &= \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1) \end{aligned}$$



Chain Rule of Relative Entropy

Definition (Conditional relative entropy)

Conditional relative entropy $D(p(y|x) \parallel q(y|x))$ for joint PMFs $p(x,y)$ and $q(x,y)$ is the average of the relative entropies between the conditional PMFs $p(y|x)$ and $q(y|x)$ averaged over the PMF $p(x)$.

$$D(p(y|x) \parallel q(y|x)) = E_{p(x,y)} \left(\log \frac{p(Y|X)}{q(Y|X)} \right) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)}.$$

Theorem (Chain rule for relative entropy)

$$D(p(x,y) \parallel q(x,y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x))$$

Corollary (KL divergence is additive for independent distributions)

If p_1, p_2 are independent distributions, with the joint distribution $p(x,y) = p_1(x)p_2(y)$ and $q(x,y), q_1(x), q_2(y)$ likewise, then

$$D(p(x,y) \parallel q(x,y)) = D(p_1(x) \parallel q_1(x)) + D(p_2(y) \parallel q_2(y))$$

Chain Rule of Relative Entropy (Cont'd)

Proof.

$$\begin{aligned} D(p(x,y) \parallel q(x,y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)) \end{aligned}$$



Thank You!

Email: <zhangy@ucsc.edu>

Homepage: <https://people.ucsc.edu/~yzhan419/>