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

Shannon Information and Entropy

Shannon information: The amount of information generated at the source X by the occurrence of the outcome x is defined as $I(x) := \log \frac{1}{p(x)}$. This is a measure of surprise:

- The less likely an outcome is, the more surprised we are.
- If an event is very probable, it is no surprise when that event happens as expected $\Longrightarrow$ transmission of such a message carries very little new information.
- Events that always occur do not communicate information.

Shannon entropy (a central role in information theory): A measure of uncertainty of a random variable, which is defined as the average Shannon information $E_X[I(X)]$.

- Entropy is an expression of the disorder or randomness of a system.
- Second law of thermodynamics: Entropy of an isolated system increases over time (associated with the concept of *arrow of time*; check out the movie "Tenet").
- 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).

Entropy and Its Units

Definition (Entropy)

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

By convention, $0\log 0 = 0$ (justified by the L'Hopital's rule).

- Gibbs entropy: $S = -\kappa_B \sum_i p_i \ln p_i$, where κ_B is the Boltzmann's constant; p_i is the probability that microstate i has the energy E_i during the system's fluctuations.
- The macrostate of a system is characterized by a distribution on the microstates .

Different units for difference bases of the logarithm:

- $\log_2 \rightarrow$ bits; $\log_e \rightarrow$ nats; $\log_{10} \rightarrow$ dits/bans/Harleys

Two properties of $H(X)$:

- $H(X) \geq 0$
- $H_b(X) = (\log_b a) H_a(X) \Leftarrow$ the change of base formula: $\log_b x = \frac{\log_a x}{\log_a b}$.

Entropy as a Measure of Average Information

Motivating question: 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; $\quad H(0) = H(1) = 0.$ ■

Q: What is the maximum value of $H(X)$?

A: $H'(X) = \frac{df(p)}{dp} = 0 \implies p = \frac{1}{2}$

$f''(p) < 0 \implies H(X)$ is a concave function (we'll learn convexity/concavity later)

Hence, we conclude that $H(X)$ achieves its maximum when $p = \frac{1}{2}$. That is, X has the greatest uncertainty when both outcomes are equally likely. In this case, $H(X) = 1$ bit.

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 encoding scheme:

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

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

$$A \text{ occurs with frequency } 0.70$$

$$B \text{ occurs with frequency } 0.26$$

$$C \text{ occurs with frequency } 0.02$$

$$D \text{ occurs with frequency } 0.02$$

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

$H(0.7, 0.26, 0.02, 0.02) = 1.09$ 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.

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 [$H(p_1, p_2, \ldots, p_m) = H(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$]: 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 $H(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i$ is the only function satisfying them.

Joint and Conditional Entropy

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

Joint entropy:

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

Conditional entropy:

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

- $H(Y|X=x)$ 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 $X \perp\!\!\!\perp Y \implies H(X|Y) = H(X)$
- $H(X|Y) \neq H(Y|X)$ 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 $\log p(X,Y) = \log p(X) + \log p(Y|X)$.

Corollary

- $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
- $H(f(X)) \leq H(X)$ *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	

Joint and Conditional Entropy (Cont'd)

Example (Cont'd)

$Y \setminus X$	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	

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

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

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

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

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:
  $D(p||q) \neq D(q||p), \qquad D(p||q) \nleq D(p||r) + D(r||q)$
- 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

- $D(p(x)||q(x)) = -\mathrm{E}_p(\log q(X)) + \mathrm{E}_p(\log p(X)) = \overbrace{H(p(x), q(x))}^{\text{cross entropy of } p \text{ and } q} - H(p(x))$.
- $H(p(x), q(x)) \neq H(q(x), p(x))$ 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.

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.

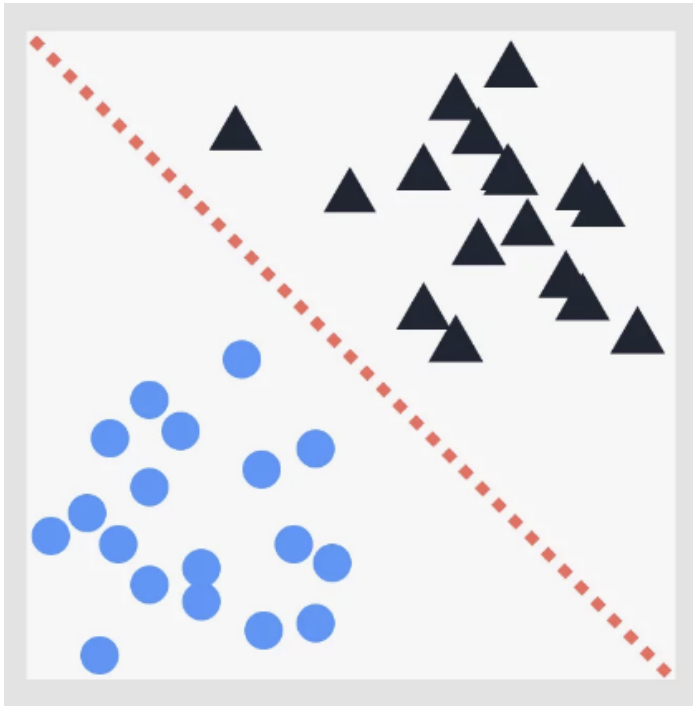


Figure: Binary classification.

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.

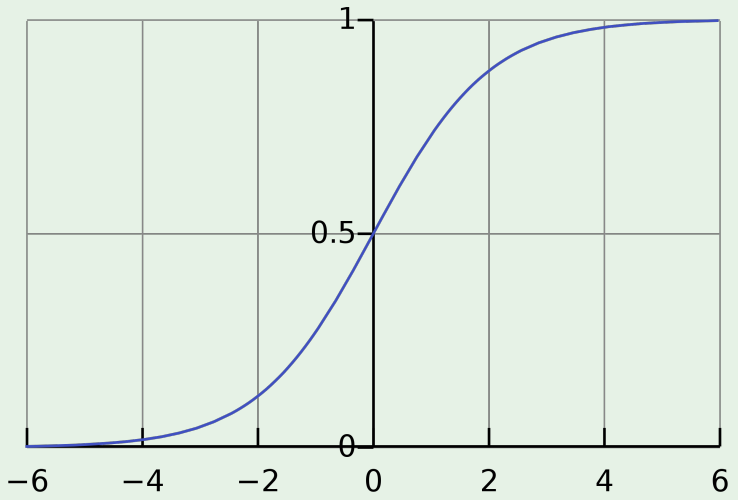


Figure: Logistic function

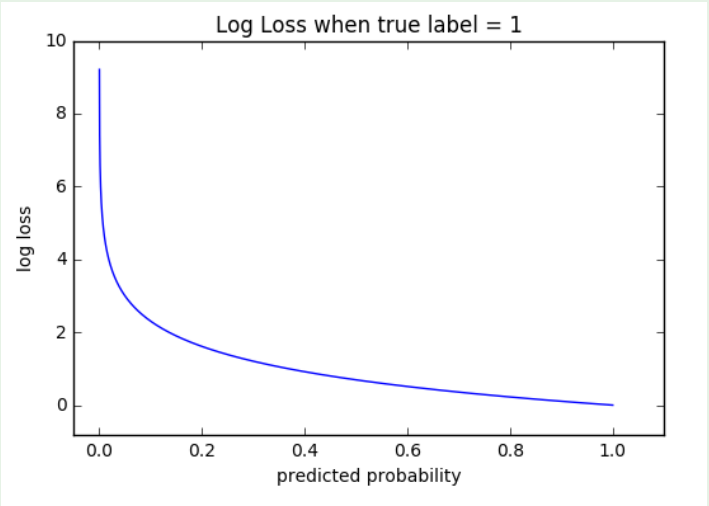


Figure: Cross entropy loss

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:

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

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 $p(x)p(y)$:

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

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

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

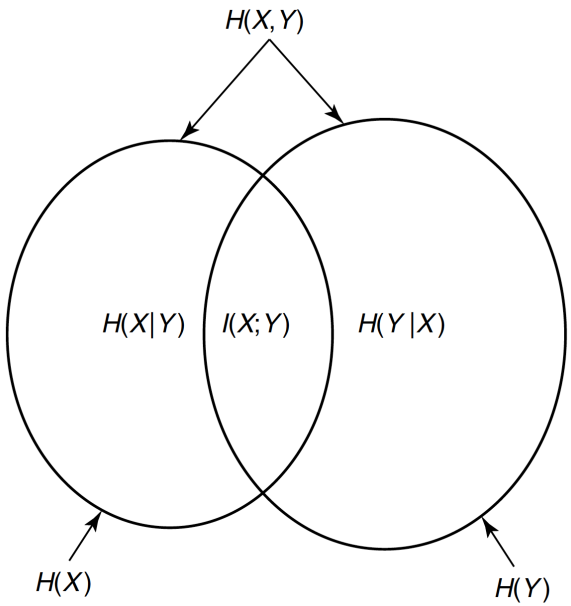
Uncertainty Reduction

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.



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.

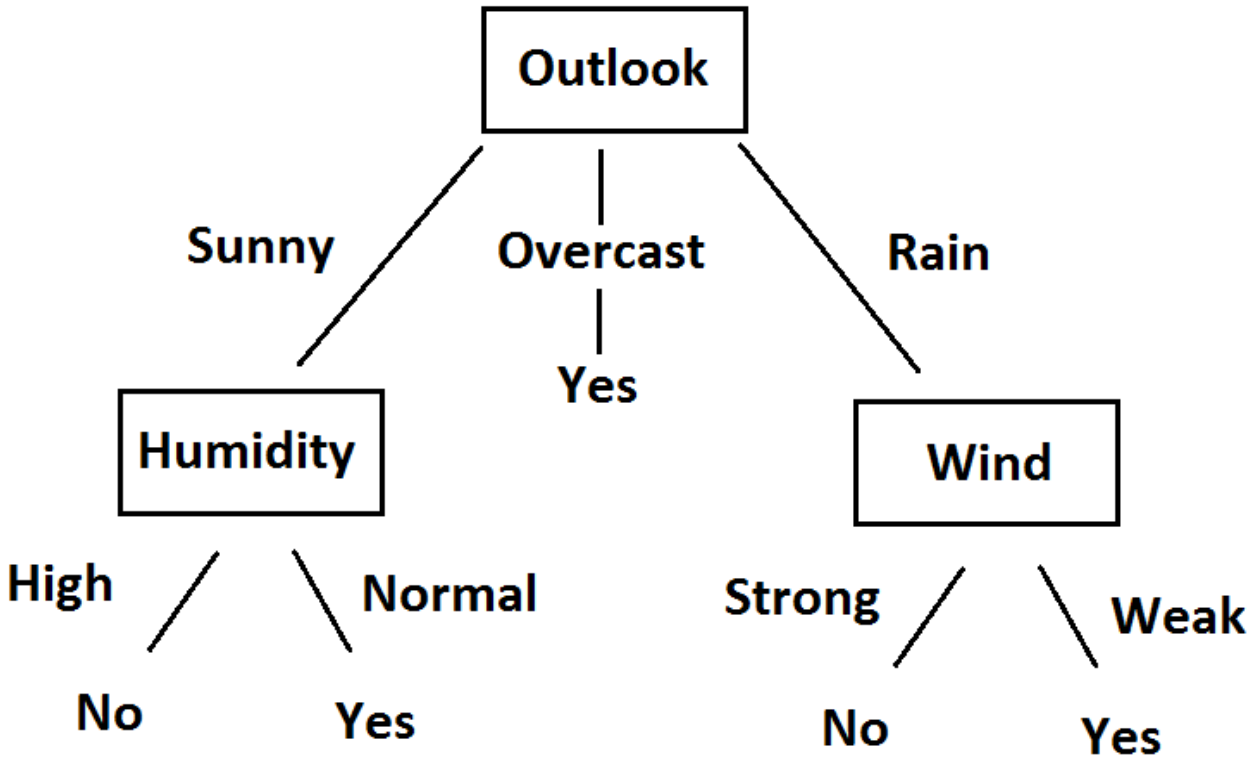


Figure: A decision tree to decide if we will play tennis.

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": $IG(\text{humidity}) = H(Y) - H(Y|\text{humidity})$.

$$H(Y|\text{humidity}) = \Pr(\text{humidity=high}) \times H(Y|\text{humidity=high}) + \Pr(\text{humidity=normal}) \times H(Y|\text{humidity=normal}) =$$

$$\frac{7}{14} \times H(\tfrac{3}{7}, \tfrac{4}{7}) + \frac{7}{14} \times H(\tfrac{1}{7}, \tfrac{6}{7}) \implies IG(\text{humidity}) = 0.152$$

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:

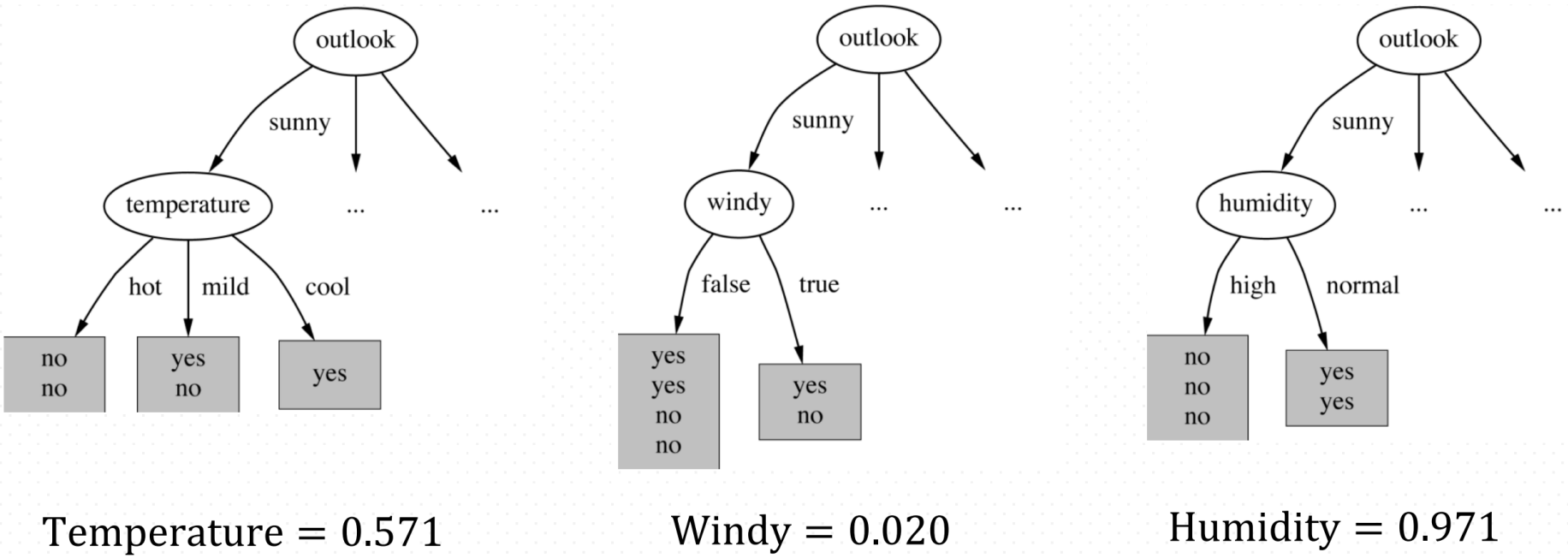


Figure: Humidity has the highest info gain at the second level when outlook = sunny.

Information Gain (Cont'd)

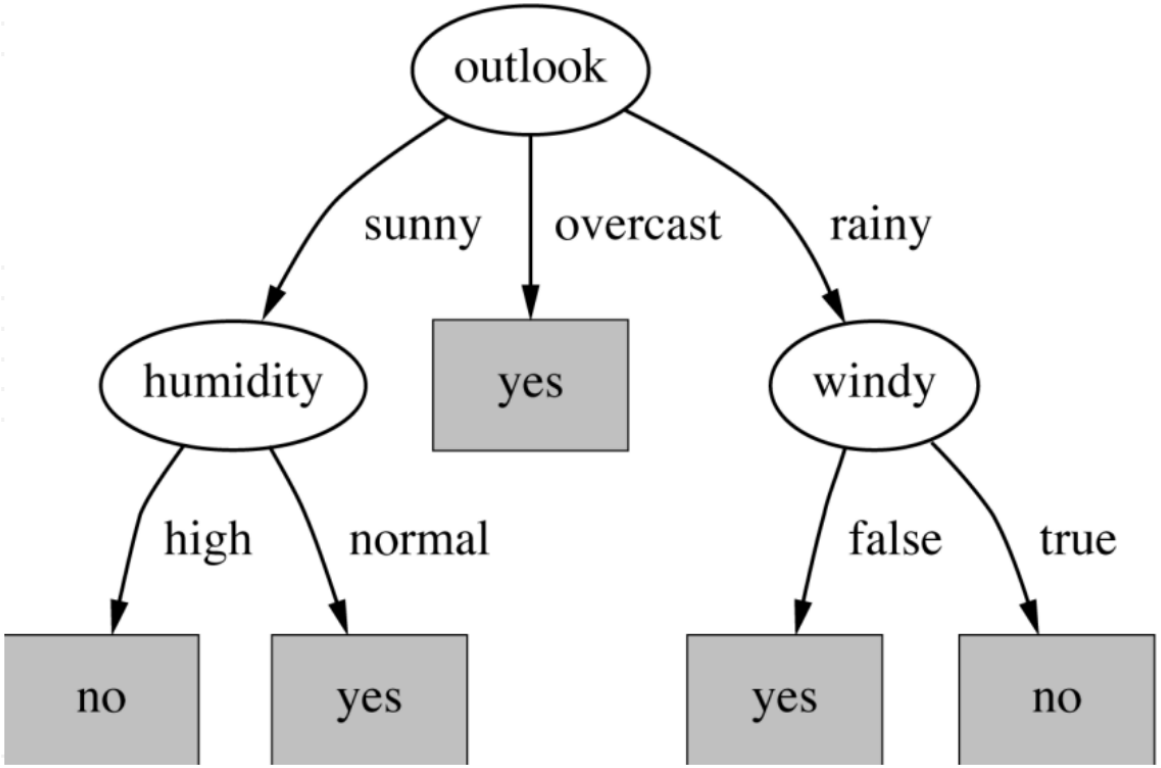


Figure: The final decision tree.

- 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)

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

Proof.

$$
\begin{aligned}
H(X_1, X_2) &= H(X_1) + H(X_2|X_1) \\
H(X_1, X_2, X_3) &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \\
&\vdots \\
H(X_1, X_2, ..., X_n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, ..., X_1) \\
&= \sum_{i=1}^{n} H(X_i|X_{i-1}, X_{i-2}, ..., X_1)
\end{aligned}
$$

□

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)

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

Proof.

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

□

Chain Rule of Relative Entropy

Definition (Conditional relative entropy)

Conditional relative entropy $D\left(p(y|x) \,||\, q(y|x)\right)$ 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).

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

Theorem (Chain rule for relative entropy)

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

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

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

Chain Rule of Relative Entropy (Cont'd)

Proof.

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

□

Thank You!

Email: <zhangy@ucsc.edu>

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