Information Theory And Machine Learning

Daniel Sadoc Menasche

2011

Notation

	probability mass function	expectation	expecation of function
X	$p(x) = P(X = x)$	$E[X] = \sum_{\forall x} p(x)x$	$E[f(X)] = \sum_{\forall x} p(x)f(x)$
Y	$q(x) = P(Y = x)$	$E[Y] = \sum_{\forall x} q(x)x$	$E[f(Y)] = \sum_{\forall x} q(x)f(x)$

Examples,

- to emphasize expectation over p, denote E by E_p,

$$E[q(X)] = E_p[q(X)] = \sum_{\forall x} p(x)q(x)$$

- entropy, $E[-\log p(X)] = -\sum_{\forall x} p(x) \log p(x)$

Entropy

Properties

- $H(X) > 0$
- let $H_b(X)$ be the entropy in base b: $H_b(X) = (\log_b a) H(X)$

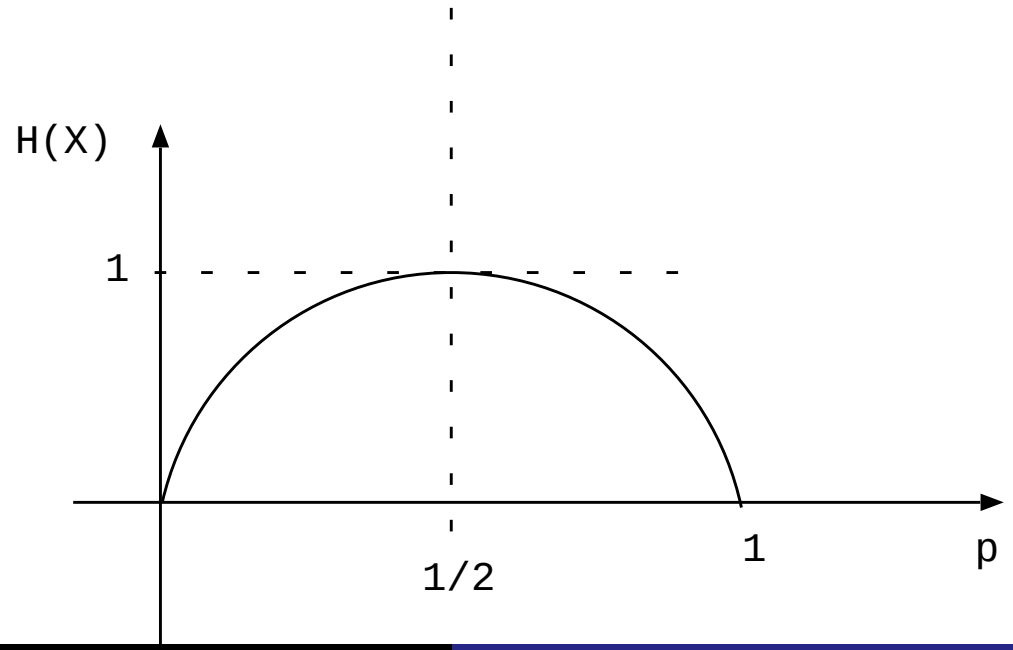
$$
\begin{aligned}
H_b(X) &= (\log_b a) H_a(X) && (1)\\
&= (\log_b a) \sum p(x) \log_a p(x) && (2)\\
&= (\log_b a) \sum p(x) \frac{\log_b p(x)}{\log_b a} && (3)
\end{aligned}
$$

Entropy of Indicator Variable

Let X be an indicator variable

$$X = \begin{cases} 1, & \text{probability } p \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

- $H(X) = p \log p + (1-p)\log(1-p)$
- maximized when $p = 1/2$
- is concave

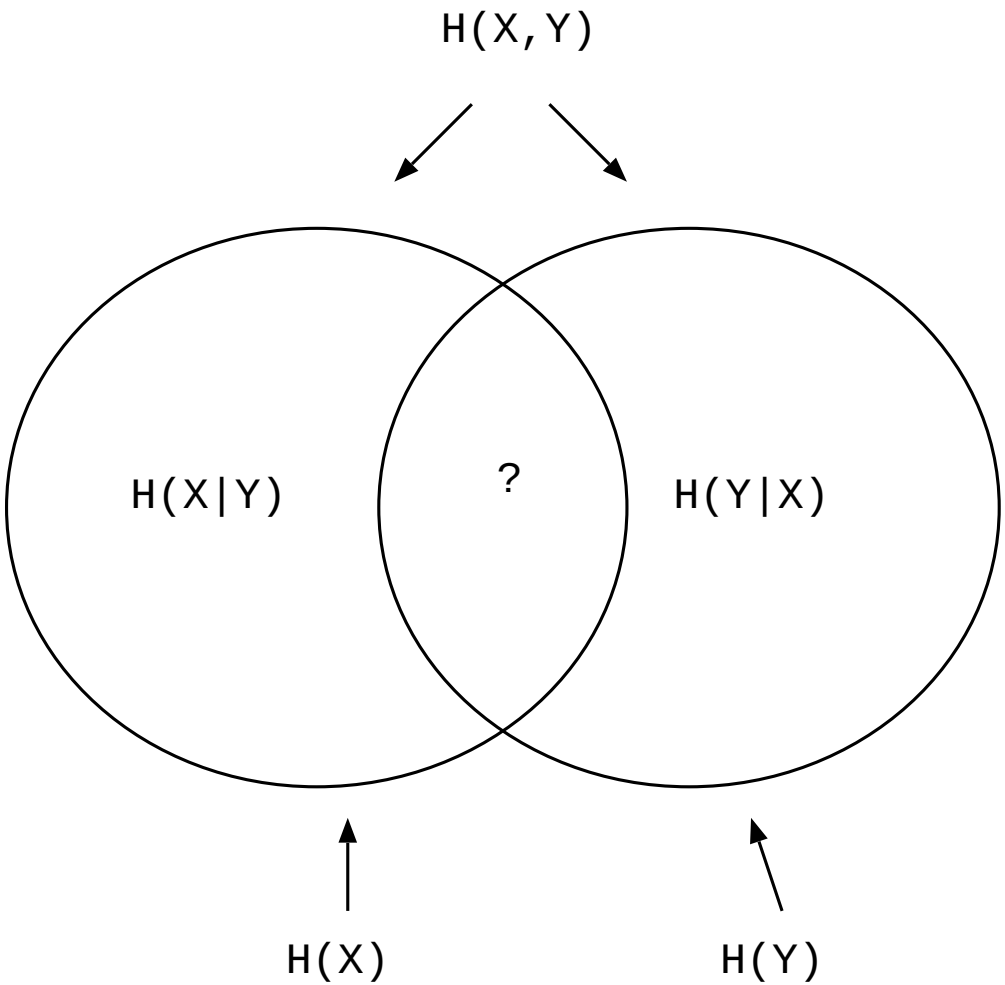


Another Example

Y \ X	1	2	3	4	
1	1/8	1/16	1/32	1/32	4/16=1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	4/16=1/4
4	1/4	0	0	0	1/4
	1/2	1/4	1/8	1/8	

- $H(X|Y) = \sum_y H(X|Y=y)P(Y=y) = 11/8$
- $H(Y|X) = \sum_x H(Y|X=x)P(X=x)$
- $H(X) = 7/4$ bits
- $H(Y) = 2$ bits
- $H(X,Y)$ (2 ways to compute)

Entropy and Conditional Entropy



KL Divergence

- Definition: KL Divergence between distributions p and q is

$$KL(p,q) = -\sum_{\forall x} P(X=x)\log\left(\frac{q(x)}{p(x)}\right)$$

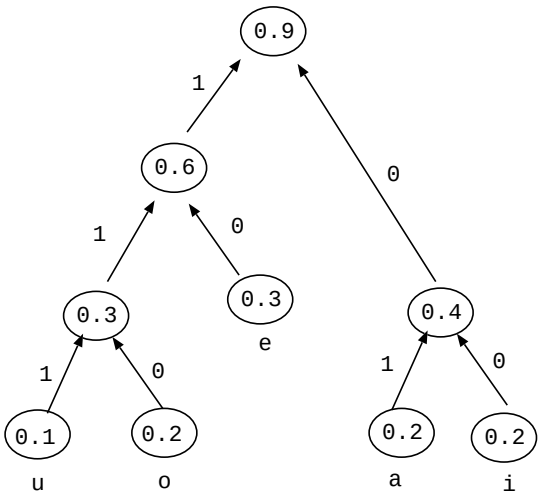
$$= E_p\left[-\log\left(\frac{q(X)}{p(X)}\right)\right]$$

- KL Divergence serves to
  - compare two distributions
  - find the number of extra bits that we need to encode function if we make wrong guess

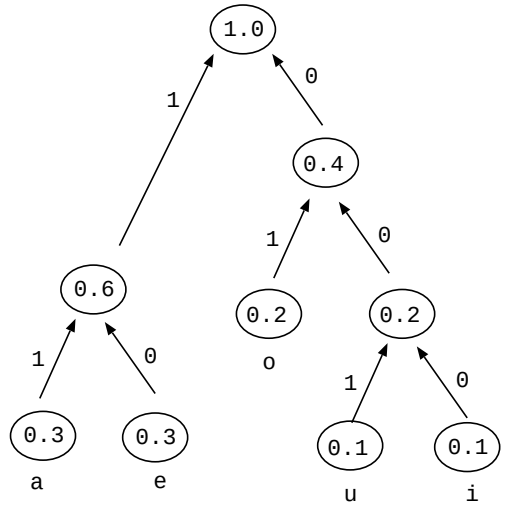
Huffman Code and KL Divergence Example

			a	e	i	o	u
p	X	Portuguese	0.3	0.3	0.1	0.2	0.1
		Portuguese code	11	10	000	01	001
q	Y	English	0.2	0.3	0.2	0.2	0.1
		English code	01	10	00	110	111

http://en.wikipedia.org/wiki/Huffman_coding



English



Portuguese

KL Divergence Example

			a	e	i	o	u
p	X	Portuguese	0.3	0.3	0.1	0.2	0.1
		Portuguese code	11	10	000	01	001
q	Y	English	0.2	0.3	0.2	0.2	0.1
		English code	01	10	00	110	111

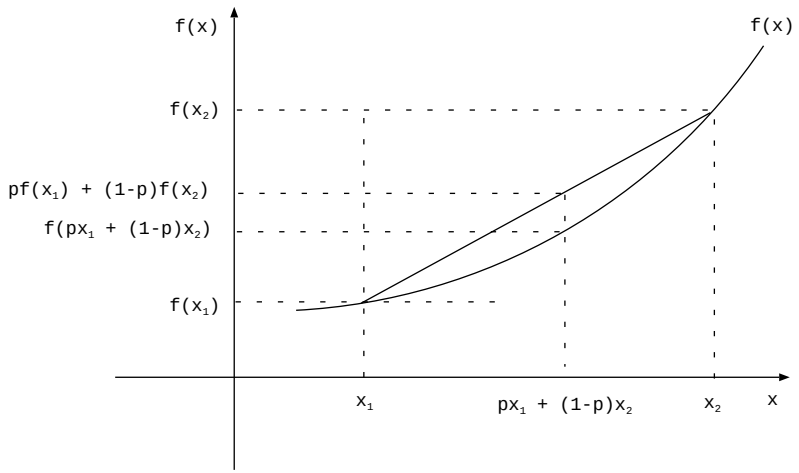
- $H(X) = 2.1710$ bits (minimum code length)
- $KL(p, q) = 0.0830$ bits (penalty using wrong distr.)
- expected code length Portuguese = L = 2.2 bits $\approx H(X)$
- expected code length Portuguese when using English code = M = 2.3 bits $\approx H(X) + KL(p, q)$
- in general,
  - $H(X) \leq L \leq L + 1$ and
  - $H(X) + KL(p, q) \leq M \leq L + KL(p, q) + 1$

KL Divergence Properties

- KL Divergence $KL(p, q)$,
  - $KL(p, q) \geq 0$
  - $KL(p, q) = 0$ iff $p = q$
  - $KL(p, q) \neq KL(q, p)$
- **Our goal:** prove first property

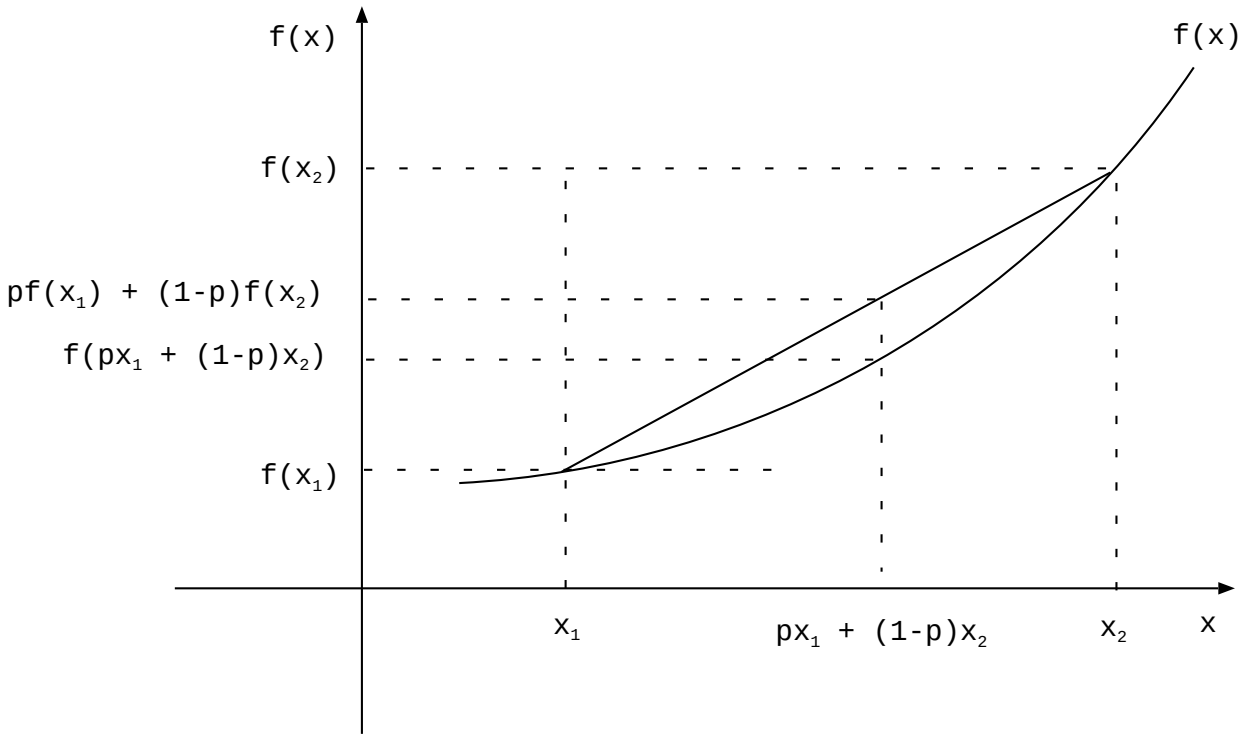
Convex and Concave Functions

- Convex function: “chord is above function”
- A function f is concave is $-f$ is convex
- A line is both concave and convex
- Example of convex function



Jensen Inequality

- Theorem: For any convex function $f(x)$, $E[f(X)] \geq f(E[X])$



KL Divergence is Always Positive

- Theorem: $KL(p,q) \geq 0$ (Gibb's Inequality)
- Proof:

$$
\begin{aligned}
KL(p,q) &= E_p\left[-\log\left(\frac{q(X)}{p(X)}\right)\right] = -\sum_{\forall x} P(X=x)\log\left(\frac{q(x)}{p(x)}\right) \\
&\geq -\log E_p\left[\frac{q(X)}{p(X)}\right] = \\
&= -\log\sum_{\forall x} P(X=x)\left(\frac{q(x)}{p(x)}\right) = \\
&= -\log\sum_{\forall x} p(x)\left(\frac{q(x)}{p(x)}\right) = \\
&= -\log\sum_{\forall x} q(x) = \\
&= -\log 1 = 0
\end{aligned}
$$

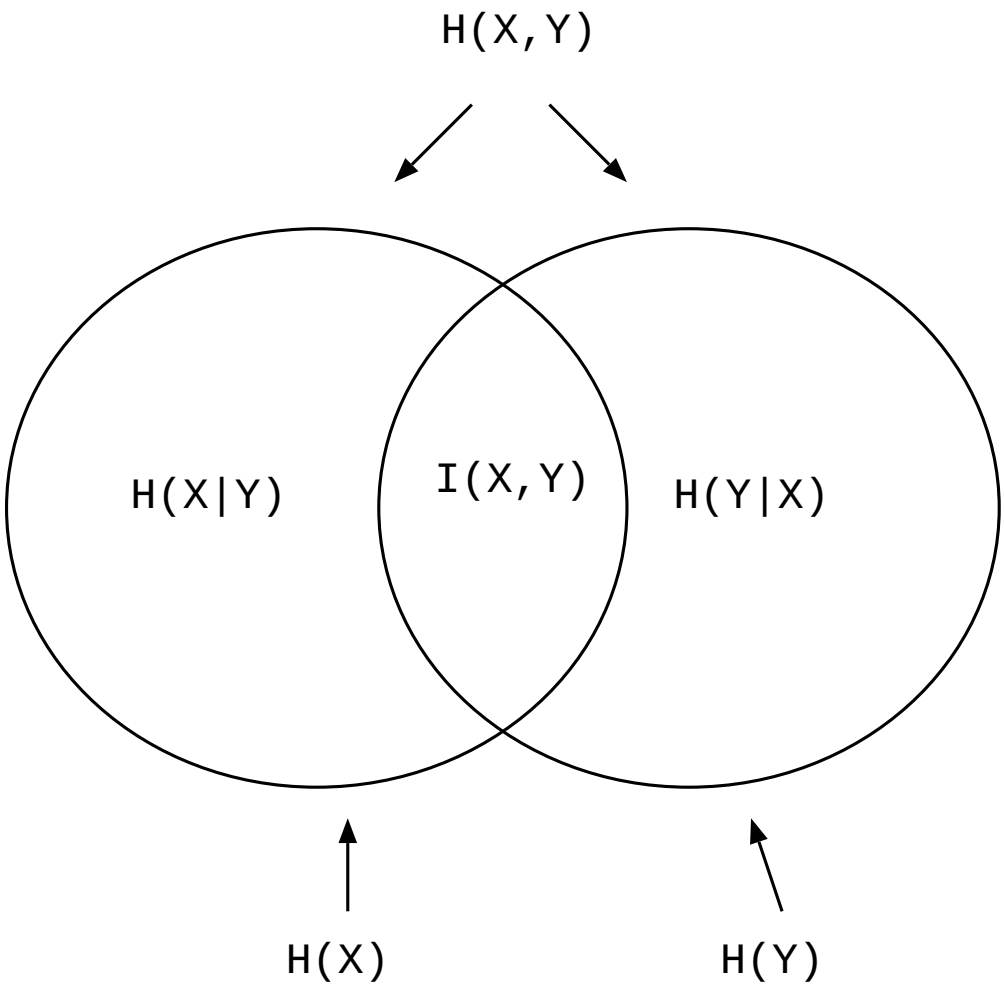
From KL Divergence to Mutual Information

- Mutual information between X and Y is amount of information that X gives about Y
  - recall $p(x) = P(X = x)$ and $q(x) = P(Y = x)$
  - let $j(x, y)$ be the joint distribution of X and Y, $j(x, y) = P(X = x, Y = y)$
- **Definition**: Mutual information between X and Y is

$$\begin{aligned} I(X, Y) &= KL(j, pq) = \\ &= \sum_{\forall x} \sum_{\forall y} p(X = x, Y = y) \log \left(\frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \right) \end{aligned}$$

  - if X and Y are independent, mutual information is zero
  - if $X = Y$, then $I(X, Y) = H(X)$
  - in general, $I(X, Y) = H(X) - H(X|Y)$

Entropy, Conditional Entropy and Mutual Information



Conditioning Reduces Entropy

- **Theorem:** $H(X|Y) \leq H(X)$
- **Proof:** $H(X) = H(X|Y) + I(X,Y)$ then $0 \leq I(X,Y) = H(X) - H(X|Y)$
  - $I(X,Y)$, information that Y brings about X
  - $H(X|Y)$, uncertainty that rests in X after *seeing* Y
  - $H(X)$, uncertainty in X
- **Theorem:** $H(X,Y) \leq H(X) + H(Y)$
- **Proof:** $H(X,Y) = H(X|Y) + H(Y)$
  - if X and Y are independent, $H(X,Y) = H(X) + H(Y)$
  - the result follows from Theorem above

Four Applications of Information Theory (IT) to Machine Learning

- IT interpretation to maximum likelihood
- IT interpreation to Bayes theorem
- Using mutual information for feature selection
- Using entropy for medical image alignment

Maximum Likelihood Is Maximum KL Divergence

- Given
  - empirical distribution $q(x)$
  - model distribution $p(x|\theta)$
- Find θ that maximizes likelihood of data **is equivalent** to minimize KL divergence between $q(x)$ and $p(x;\theta)$

Maximum Likelihood Is Maximum KL Divergence

- Example
  - Given set of points (1, 3, 4, 3, 5)
  - Empirical distribution is $P(Y = 1) = 1/5$, $P(Y = 3) = 2/5$, $P(Y = 4) = 1/5$, $P(Y = 5) = 1/5$
- Empirical distribution
  - Given set of points $x_1, x_2, x_3, \ldots, x_N$
  - $q(x) = P(Y = x) = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}_{x=x_t}$

Maximum Likelihood $\equiv$ Minimum KL Divergence

$$
\begin{aligned}
KL(q(x), p(x|\theta)) &= \sum_x q(x) \log \frac{q(x)}{p(x|\theta)} = \\
&= \sum_x q(x) \log q(x) - \sum_x q(x) \log(p(x|\theta))
\end{aligned}
$$

Maximum Likelihood $\equiv$ Minimum KL Divergence

$$
\begin{aligned}
\min_\theta KL(q(x), p(x|\theta)) &= \min_\theta -\sum_x q(x)\log(p(x|\theta)) = \\
&= \max_\theta \sum_x q(x)\log(p(x|\theta)) \\
&= \max_\theta \sum_x \left(\frac{1}{N}\sum_{t=1}^{N} \mathbf{1}_{x=x_t}\right)\log(p(x|\theta)) \\
&= \max_\theta \sum_{t=1}^{N}\sum_x \mathbf{1}_{x=x_t}\log(p(x|\theta)) \\
&= \max_\theta \sum_{t=1}^{N}\log(p(x_t|\theta))
\end{aligned}
\tag{5}
$$

Four Applications of Information Theory (IT) to Machine Learning

- IT interpretation to maximum likelihood
- IT interpreation to Bayes theorem
- Using mutual information for feature selection
- Using entropy for medical image alignment

Interpreation to Bayes Theorem

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{6}$$

- ingredients,
  - x, data
  - $p(\theta|x)$, posterior (e.g., probability of class given data)
  - $p(x|\theta)$, likelihood
  - $p(\theta)$, prior
- conditioning reduces entropy, therefore

$$\begin{aligned} H(\theta|X) &\leq H(\theta) \\ I(\theta, X) &= H(\theta) - H(\theta|X) \end{aligned}$$

- using Bayes theorem, we reduce uncertainty about class
- $I(\theta, X)$ = information that data brings about class

Four Applications of Information Theory (IT) to Machine Learning

- IT interpretation to maximum likelihood
- IT interpreation to Bayes theorem
- Using mutual information for feature selection
- Using entropy for medical image alignment

Feature Selection For Information Retrieval

Feauture selection: e.g., text classification in categories

- (in training set) select terms occuring in text
- (in test set/real text classification) use these terms to classify text
- e.g., the word *britain* is a good feature to classify texts into *UK*
- e.g., the word *viagra* is a good feature to classify texts into *spam*

Feature Selection For Information Retrieval

Good feature

- tells a lot about the class (informative)
- does not overlap with other features (non-redundant)

Noisy feature

- increases classification error
- e.g., overfitting (e.g., all texts in test set about China had word “hand”)

Feature Selection For Information Retrieval

http://nlp.stanford.edu/IR-book/html/htmledition/feature-selection-1.html

Four Applications of Information Theory (IT) to Machine Learning

- IT interpretation to maximum likelihood
- IT interpreation to Bayes theorem
- Using mutual information for feature selection
- Using entropy for medical image alignment

Medical Image Alignment

- Data Driven Image Models through Continuous Joint Alignment, Erik G. Learned-Miller