Some Useful Concepts from Information Theory

COMP9418 — Advanced Topics in Statistical Machine Learning

Edwin V. Bonilla

School of Computer Science and Engineering
UNSW Sydney

Last Update: Wednesday 16th August, 2017 at 09:35

Acknowledgments

- [Cover & Thomas, EoIT, 2012] Elements of Information Theory. Thomas M. Cover and Joy A. Thomas. John Wiley & Sons, 2012.
- [Mackay, ITILA, 2003] Information Theory, Inference, and Learning Algorithms . David J. C. Mackay. Cambridge University Press. 2003.
- [Bishop, PRML, 2006] Pattern Recognition and Machine Learning, Christopher Bishop, 2006
- [Murphy, MLaPP, 2012] Machine Learning: A Probabilistic Perspective, Kevin P. Murphy, 2012

Information Content

Information as:

- Amount of unexpected data
- message that its uncertain to receivers
  - If we are told that a very likely event has happened vs an unlikely event

Information Content

Information as:

- Amount of unexpected data
- message that its uncertain to receivers
  - If we are told that a very likely event has happened vs an unlikely event

How can we *measure* information content?:

- Information content of an outcome must depend on its probability
- Information content of a random variable must depend on its probability distribution

Information Content

Information as:

- Amount of unexpected data
- message that its uncertain to receivers
  - If we are told that a very likely event has happened vs an unlikely event

How can we *measure* information content?:

- Information content of an outcome must depend on its probability
- Information content of a random variable must depend on its probability distribution

Entropy (or information content) of an outcome x:

$$h(x) = \log_2 \frac{1}{p(x)}$$

- Choice of logarithm basis is arbitrary
- If we use $\log_2$ we measure information in *bits*

Entropy of a Random Variable

Let X be a discrete r.v. with values drawn from alphabet $\mathcal{X}$ from a total number of states $|\mathcal{X}|$.

Entropy of a Random Variable

Let X be a discrete r.v. with values drawn from alphabet $\mathcal{X}$ from a total number of states $|\mathcal{X}|$.

If we want to transmit the value of X, the average amount of information is given by:

$$H(X) = \sum_{x} p(x) \log_2 \frac{1}{p(x)}$$

- The expectation of the entropy of each outcome wrt $p(x)$

We call this the entropy of the r.v. X (Note dependency on distribution)

Entropy of a Random Variable

Let X be a discrete r.v. with values drawn from alphabet $\mathcal{X}$ from a total number of states $|\mathcal{X}|$.

If we want to transmit the value of X, the average amount of information is given by:

$$H(X) = \sum_{x} p(x) \log_2 \frac{1}{p(x)}$$

- The expectation of the entropy of each outcome wrt $p(x)$

We call this the entropy of the r.v. X (Note dependency on distribution)

Note that we can write:

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

and define: $0 \log 0 \equiv 0$, as $\lim_{p \to 0} p \log p = 0$.

- Non-negativity:

$$0 \leq p(x) \leq 1 \rightarrow \log \frac{1}{p(x)} \geq 0$$

$$\sum_x p(x) \log \frac{1}{p(x)} \geq 0$$

$$H(X) \geq 0$$

- Non-negativity:

$$0 \le p(x) \le 1 \rightarrow \log \frac{1}{p(x)} \ge 0$$
$$\sum_x p(x) \log \frac{1}{p(x)} \ge 0$$
$$H(X) \ge 0$$

- Change of base:

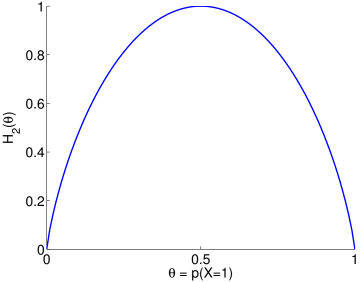
$$H_b(X) = -\sum_x p(x) \log_b p(x)$$
$$= \sum_x p(x) \log_a p(x) \log_b a$$
$$H_b(X) = \log_b a \; H_a(X)$$

  - If we use natural logarithm the units are call *nats*

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$:

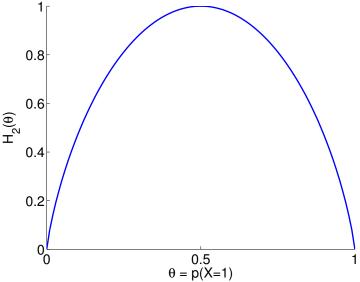
$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$

Why?



Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$:

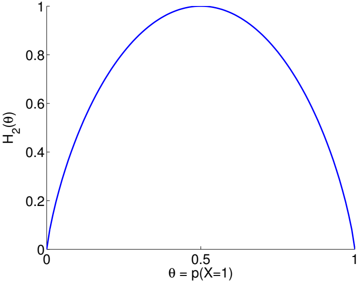
$$H(X) = -\theta \log \theta - (1-\theta) \log(1-\theta)$$ Why?



- *Concave* function of the distribution

Let $X \in \{0,1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X=1)$:

$$H(X) = -\theta \log \theta - (1-\theta)\log(1-\theta)$$ Why?



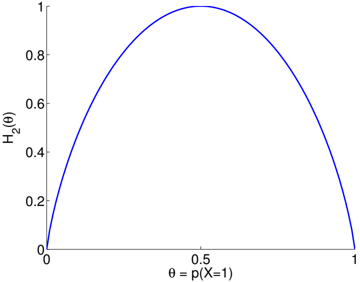
- *Concave* function of the distribution
- Minimum when there is no uncertainty $\theta = 1$ or $\theta = 0$

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$:

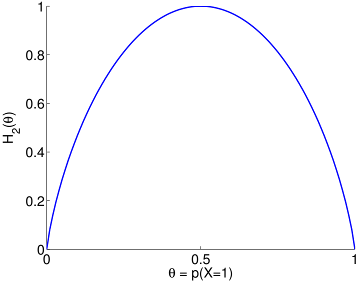
$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$ Why?



- *Concave* function of the distribution
- Minimum when there is no uncertainty $\theta = 1$ or $\theta = 0$
- Maximum when there is maximum uncertainty $\theta = 0.5$

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$:

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$ Why?



- *Concave* function of the distribution
- Minimum when there is no uncertainty $\theta = 1$ or $\theta = 0$
- Maximum when there is maximum uncertainty $\theta = 0.5$
- For $\theta = 0.5$ (e.g. a fair coin) $H_2(X) = 1$ bit.

Consider a random variable X with uniform distribution over 32 outcomes:

- To identify an outcome we need a label that that takes on 32 different values

Consider a random variable X with uniform distribution over 32 outcomes:

- To identify an outcome we need a label that that takes on 32 different values
- Hence 5-bit strings suffice as labels

Entropy of a Random Variable
Example 2

Consider a random variable X with uniform distribution over 32 outcomes:

- To identify an outcome we need a label that that takes on 32 different values
- Hence 5-bit strings suffice as labels

The entropy of this rv is given by:

$$H(X) = -\sum_{i=1}^{32} p(i) \log_2 p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log_2 \frac{1}{32} = \log_2 32 = 5 \text{ bits.}$$

Consider a random variable X with uniform distribution over 32 outcomes:

- To identify an outcome we need a label that that takes on 32 different values
- Hence 5-bit strings suffice as labels

The entropy of this rv is given by:

$$H(X) = -\sum_{i=1}^{32} p(i) \log_2 p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log_2 \frac{1}{32} = \log_2 32 = 5 \text{ bits.}$$

- This agrees with the number of bits needed to describe X
- In this case all the outcomes have representations of the same length

Categorical distributions with 30 different states:

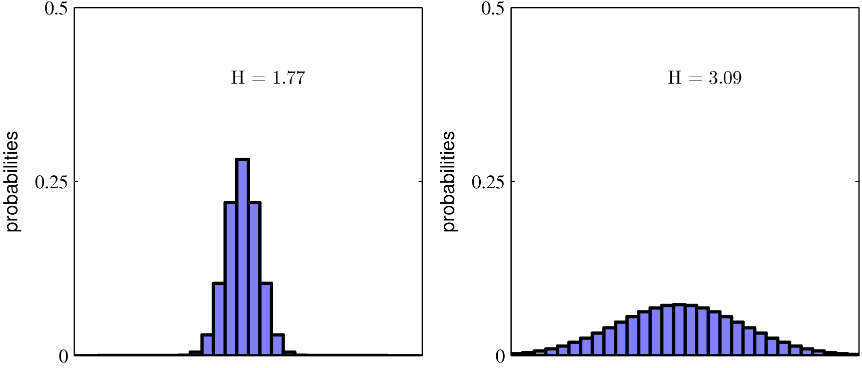


Figure from Bishop, PRML, 2006)

Categorical distributions with 30 different states:

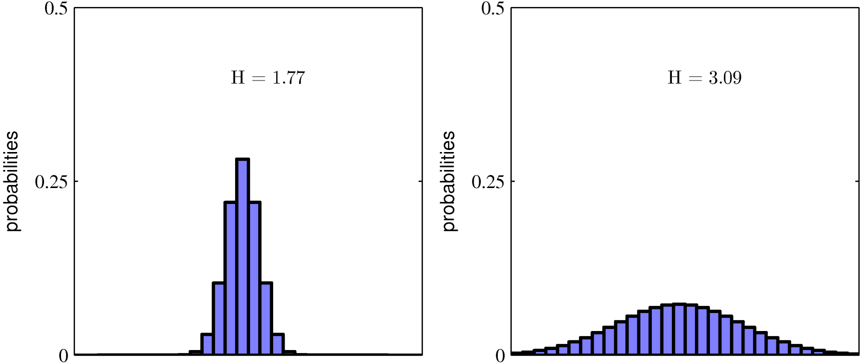


Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

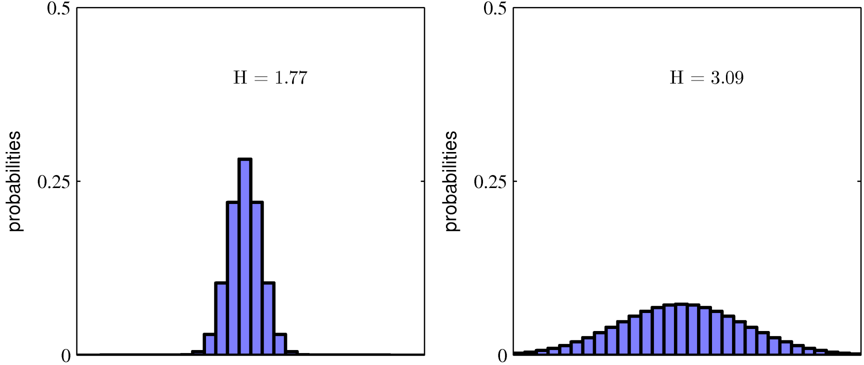


Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy
- The more evenly spread the higher the entropy

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

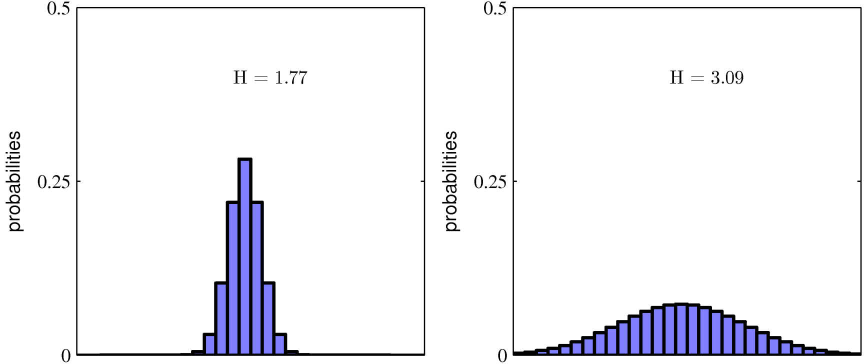


Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy
- The more evenly spread the higher the entropy
- Maximum for *uniform* distribution: $H(X) = -\log \frac{1}{30} \approx 3.40$ nats
  - When will the entropy be minimum?

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \dots, |\mathcal{X}|$
- Denote the vector of probabilities with $\mathbf{p}$

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$
- Denote the vector of probabilities with $\mathbf{p}$

The entropy is maximized if $\mathbf{p}$ is uniform:

$$H(X) \leq \log|\mathcal{X}|$$

With equality iff $p_i = \frac{1}{|\mathcal{X}|}$ for all i

Joint Entropy

The joint entropy $H(X, Y)$ of a pair of discrete random variables with joint distribution $p(X, Y)$ is given by:

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

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

$$H(Y|X = x) = \sum_{y \in \mathcal{Y}} p(y|X = x) \log \frac{1}{p(y|X = x)}$$

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

$$H(Y|X = x) = \sum_{y \in \mathcal{Y}} p(y|X = x) \log \frac{1}{p(y|X = x)}$$

The conditional entropy of Y given X, is the average over X of the conditional entropy of Y given $X = x$:

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

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

$$H(Y|X = x) = \sum_{y \in \mathcal{Y}} p(y|X = x) \log \frac{1}{p(y|X = x)}$$

The conditional entropy of Y given X, is the average over X of the conditional entropy of Y given $X = x$:

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

This measures the average uncertainty that remains about Y when X is known.

- $D(p||q)$: Distance/divergence between two distributions

- $D(p||q)$: Distance/divergence between two distributions
- Machine Learning: Approximating a posterior $p(X)$ with $q(X)$

Relative Entropy

Introduction

- $D(p||q)$: Distance/divergence between two distributions
- Machine Learning: Approximating a posterior $p(X)$ with $q(X)$
- Information Theory: Inefficiency of assuming q when the true distribution is p
    - $p(X)$: Can construct a code with average description length $H(p)$
    - $q(X)$: Would need $H(p) + D(p\|q)$ bits on average to describe the r.v.

- $D(p||q)$: Distance/divergence between two distributions
- Machine Learning: Approximating a posterior $p(X)$ with $q(X)$
- Information Theory: Inefficiency of assuming q when the true distribution is p
  - $p(X)$: Can construct a code with average description length $H(p)$
  - $q(X)$: Would need $H(p) + D(p\|q)$ bits on average to describe the r.v.

A measure of divergence between two distributions is the *relative entropy* or *Kullback-Leibler divergence*

Relative Entropy

Definition

The relative entropy or Kullback-Leibler (KL) divergence between two probability distributions $p(X)$ and $q(X)$ is defined as:

$$\mathrm{KL}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{p(X)} \left[\log \frac{p(X)}{q(X)} \right].$$

Relative Entropy

Definition

The relative entropy or Kullback-Leibler (KL) divergence between two probability distributions $p(X)$ and $q(X)$ is defined as:

$$\mathrm{KL}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{p(X)}\left[\log \frac{p(X)}{q(X)}\right].$$

- Note:
  - Both $p(X)$ and $q(X)$ are defined over the same alphabet $\mathcal{X}$
  - KL in statistics
- Conventions:

$$0 \log \frac{0}{0} \stackrel{\text{def}}{=} 0 \qquad 0 \log \frac{0}{q} \stackrel{\text{def}}{=} 0 \qquad p \log \frac{p}{0} \stackrel{\text{def}}{=} \infty$$

Relative Entropy
Properties

- $\mathrm{KL}(p \| q) \geq 0$

Relative Entropy

Properties

- $\mathrm{KL}(p\|q) \geq 0$
- $\mathrm{KL}(p\|q) = 0 \Leftrightarrow p = q$

Relative Entropy
Properties

- $\mathrm{KL}(p\|q) \geq 0$
- $\mathrm{KL}(p\|q) = 0 \Leftrightarrow p = q$
- $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$

Relative Entropy
Properties

- $\mathrm{KL}(p\|q) \geq 0$
- $\mathrm{KL}(p\|q) = 0 \Leftrightarrow p = q$
- $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$
- Not a true distance since is not symmetric and does not satisfy the triangle inequality

- $\text{KL}(p \| q) \geq 0$
- $\text{KL}(p \| q) = 0 \Leftrightarrow p = q$
- $\text{KL}(p \| q) \neq \text{KL}(q \| p)$
- Not a true distance since is not symmetric and does not satisfy the triangle inequality
- Very important in machine learning and information theory

Relative Entropy

Example (from Cover & Thomas, 2006)

Let $X \in \{0, 1\}$ and consider the distributions $p(X)$ and $q(X)$ such that:

$$\begin{aligned} p(X = 1) &= \theta_p \qquad & p(X = 0) &= 1 - \theta_p \\ q(X = 1) &= \theta_q \qquad & q(X = 0) &= 1 - \theta_q \end{aligned}$$

What distributions are these?

Let $X \in \{0, 1\}$ and consider the distributions $p(X)$ and $q(X)$ such that:

$$p(X = 1) = \theta_p \qquad p(X = 0) = 1 - \theta_p$$
$$q(X = 1) = \theta_q \qquad q(X = 0) = 1 - \theta_q$$

What distributions are these?

Compute $\mathrm{KL}(p\|q)$ and $\mathrm{KL}(q\|p)$ with $\theta_p = \frac{1}{2}$ and $\theta_q = \frac{1}{4}$

Relative Entropy

Example (from Cover & Thomas, 2006) — Cont'd

$$\mathrm{KL}(p\|q) = \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p) \log \frac{1-\theta_p}{1-\theta_q}$$

Relative Entropy

Example (from Cover & Thomas, 2006) — Cont'd

$$
\begin{aligned}
\mathrm{KL}(p\|q) &= \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p)\log\frac{1-\theta_p}{1-\theta_q} \\
&= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} = 1 - \frac{1}{2}\log 3 \approx 0.2075 \text{ bits}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{KL}(p\|q) &= \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p)\log\frac{1-\theta_p}{1-\theta_q} \\
&= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} = 1 - \frac{1}{2}\log 3 \approx 0.2075 \text{ bits}
\end{aligned}
$$

$$
\mathrm{KL}(q\|p) = \theta_q \log \frac{\theta_q}{\theta_p} + (1-\theta_q)\log\frac{1-\theta_q}{1-\theta_p}
$$

Relative Entropy

Example (from Cover & Thomas, 2006) — Cont'd

$$
\begin{aligned}
\mathrm{KL}(p\|q) &= \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p) \log \frac{1-\theta_p}{1-\theta_q} \\
&= \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}} = 1 - \frac{1}{2} \log 3 \approx 0.2075 \text{ bits}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{KL}(q\|p) &= \theta_q \log \frac{\theta_q}{\theta_p} + (1-\theta_q) \log \frac{1-\theta_q}{1-\theta_p} \\
&= \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} + \frac{3}{4} \log \frac{\frac{3}{4}}{\frac{1}{2}} = -1 + \frac{3}{4} \log 3 \approx 0.1887 \text{ bits}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{KL}(p\|q) &= \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p)\log\frac{1-\theta_p}{1-\theta_q} \\
&= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} = 1 - \frac{1}{2}\log 3 \approx 0.2075 \text{ bits}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{KL}(q\|p) &= \theta_q \log \frac{\theta_q}{\theta_p} + (1-\theta_q)\log\frac{1-\theta_q}{1-\theta_p} \\
&= \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} + \frac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} = -1 + \frac{3}{4}\log 3 \approx 0.1887 \text{ bits}
\end{aligned}
$$

When will $\mathrm{KL}(p\|q) = \mathrm{KL}(q\|p)$?

Mutual Information

Definition

Let X, Y be two r.v. with joint distribution $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

Let X, Y be two r.v. with joint distribution $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

Definition

The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

Let X, Y be two r.v. with joint distribution $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

Definition

The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

$$I(X; Y) = \mathrm{KL}\left(p(X, Y) \| p(X)p(Y)\right)$$

Let X, Y be two r.v. with joint distribution $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

Definition

The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

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

Let X, Y be two r.v. with joint distribution $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

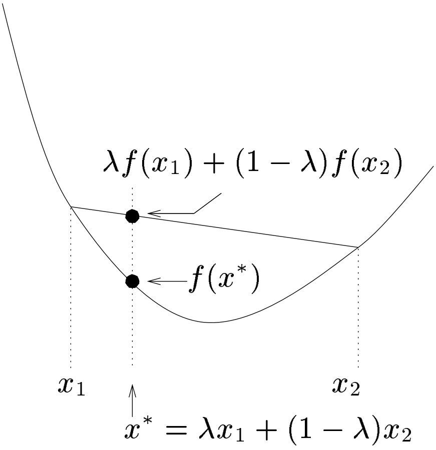
Definition

The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

$$I(X; Y) = \mathrm{KL}\left(p(X, Y) \| p(X)p(Y)\right)$$
$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Intuitively, how much information, on average, X conveys about Y (or vice versa).

Convex Functions:
Introduction



$0 \leq \lambda \leq 1$ (Figure from Mackay, 2003)

A function is convex $\smile$ if every cord of the function lies above the function

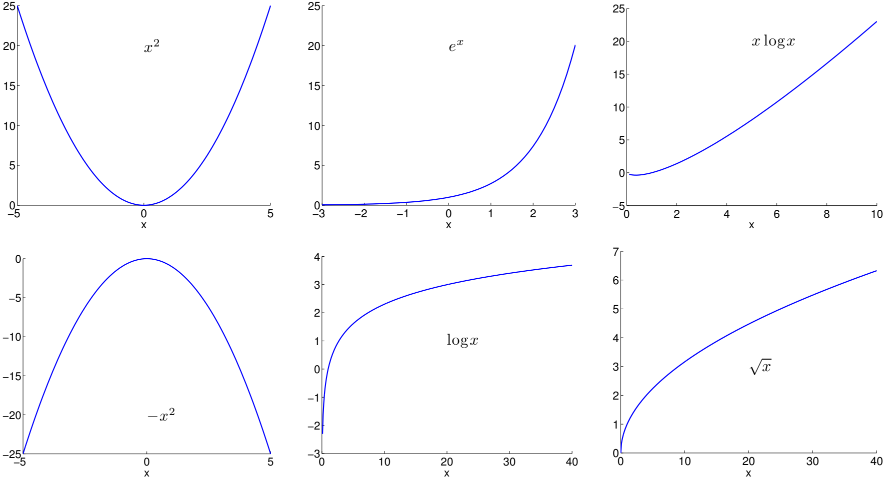
Definition

A function $f(x)$ is convex $\smile$ over (a, b) if for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

We say f is strictly convex $\smile$ if for all $x_1, x_2 \in (a, b)$ the equality holds only for $\lambda = 0$ and $\lambda = 1$.

Similarly, a function f is concave $\frown$ if $-f$ is convex $\smile$, i.e. if every cord of the function lies below the function.


x^2
e^x
$x \log x$
$-x^2$
$\log x$
$\sqrt{x}$
x

Verifying Convexity

Theorem (Cover & Thomas, Th 2.6.1)

If a function f has a second derivative that is non-negative (positive) over an interval, the function is convex $\smile$(strictly convex $\smile$) over that interval.

This allows us to verify convexity or concavity.

Examples:

- x^2: $\dfrac{d}{dx}\left(\dfrac{d}{dx}(x^2)\right) = \dfrac{d}{dx}(2x) = 2$
- e^x: $\dfrac{d}{dx}\left(\dfrac{d}{dx}(e^x)\right) = \dfrac{d}{dx}(e^x) = e^x$
- $\sqrt{x}$, $x > 0$: $\dfrac{d}{dx}\left(\dfrac{d}{dx}(\sqrt{x})\right) = \dfrac{1}{2}\dfrac{d}{dx}\left(\dfrac{1}{\sqrt{x}}\right) = -\dfrac{1}{4}\dfrac{1}{\sqrt{x^3}}$

Convexity, Concavity and Optimization

if $f(x)$ is concave $\frown$ and there exists a point at which

$$\frac{df}{dx} = 0,$$

then $f(x)$ has a maximum at that point.

if $f(x)$ is concave $\frown$ and there exists a point at which

$$\frac{df}{dx} = 0,$$

then $f(x)$ has a maximum at that point.

Note: the converse does not hold: if a concave $\frown$ $f(x)$ is maximized at some x, it is not necessarily true that the derivative is zero there.

- $f(x) = -|x|$: is maximized at $x = 0$ where its derivative is undefined
- $f(p) = \log p$ with $0 \leq p \leq 1$, is maximized at $p = 1$ where $\frac{df}{dp} = 1$

Convexity, Concavity and Optimization

if $f(x)$ is concave $\frown$ and there exists a point at which

$$\frac{df}{dx} = 0,$$

then $f(x)$ has a maximum at that point.

Note: the converse does not hold: if a concave $\frown$ $f(x)$ is maximized at some x, it is not necessarily true that the derivative is zero there.

- $f(x) = -|x|$: is maximized at $x = 0$ where its derivative is undefined
- $f(p) = \log p$ with $0 \leq p \leq 1$, is maximized at $p = 1$ where $\frac{df}{dp} = 1$

- Generalization to multivariate functions.
- Similarly for convex functions (and minimization)
- Can use second derivative and first derivative together

Jensen's Inequality for Convex Functions

Theorem: Jensen's Inequality

If f is a convex $\smile$ function and X is a random variable then:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Moreover, if f is strictly convex $\smile$, the equality implies that $X = \mathbb{E}[X]$ with probability 1, i.e X is a constant.

Similarly for a concave $\frown$ function: $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$.

There squares have average area $\overline{A} = 100\ \text{m}^2$. The average of the lengths of their sides is $\bar{\ell} = 10$ m. What can be said about the largest of the three squares?

There squares have average area $\overline{A} = 100 \text{ m}^2$. The average of the lengths of their sides is $\bar{\ell} = 10$ m. What can be said about the largest of the three squares?

Solution:

Let $X \in \{\ell_1, \ell_2, \ell_3\}$ denote the length of the side of a square with $\mathbf{p} = (1/3, 1/3, 1/3)$.

There squares have average area $\overline{A} = 100\ \text{m}^2$. The average of the lengths of their sides is $\bar{\ell} = 10$ m. What can be said about the largest of the three squares?

Solution:

Let $X \in \{\ell_1, \ell_2, \ell_3\}$ denote the length of the side of a square with $\mathbf{p} = (1/3, 1/3, 1/3)$.

We are given:

$$\mathbb{E}[X] = 10 \qquad \mathbb{E}[f(X)] = 100,$$

Jensen's Inequality

Example (from Mackay, 2003)

There squares have average area $\overline{A} = 100$ m^2. The average of the lengths of their sides is $\overline{\ell} = 10$ m. What can be said about the largest of the three squares?

Solution:

Let $X \in \{\ell_1, \ell_2, \ell_3\}$ denote the length of the side of a square with $\mathbf{p} = (1/3, 1/3, 1/3)$.

We are given:

$$\mathbb{E}[X] = 10 \qquad \mathbb{E}[f(X)] = 100,$$

where $f(x) = x^2$, which is a strictly convex $\smile$ function.

There squares have average area $\overline{A} = 100$ m^2. The average of the lengths of their sides is $\overline{\ell} = 10$ m. What can be said about the largest of the three squares?

Solution:

Let $X \in \{\ell_1, \ell_2, \ell_3\}$ denote the length of the side of a square with $\mathbf{p} = (1/3, 1/3, 1/3)$.

We are given:

$$\mathbb{E}[X] = 10 \qquad \mathbb{E}[f(X)] = 100,$$

where $f(x) = x^2$, which is a strictly convex $\smile$ function.

Therefore $f(\mathbb{E}[X]) = \mathbb{E}[f(X)]$, implying that X is a constant and the three lengths $\ell_1 = \ell_2 = \ell_3 = 10$.

Gibbs' Inequality

Theorem

The relative entropy (or KL divergence) between two distributions $p(X)$ and $q(X)$ with $X \in \mathcal{X}$ is non-negative:

$$\mathrm{KL}(p\|q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Proof Using Jensen's inequality.

- Differential Entropy: It is possible to define:

$$H(x) = \mathbb{E}_{p(x)}[-\log p(x)] = -\int p(x) \log p(x) dx$$

However, it does not satisfy all properties, e.g. it can be negative.

- KL Divergence: Similarly, we have seen:

$$\mathrm{KL}(p \| q) \stackrel{\text{def}}{=} \mathbb{E}_{p(x)}\left[\log \frac{p(x)}{q(x)}\right] = \int p(x) \log \frac{p(x)}{q(x)} dx,$$

which always satisfies Gibbs' inequality.

Summary & Conclusions

- Entropy as a measure of information content
- Computation of entropy of discrete random variables
- Joint and conditional entropies
- Relative entropy
- Convex Functions, Jensen's inequality, Gibbs' inequality
- Reading:
  * Mackay (ITILA, 2003): Sec. 1.2 – 1.5, 2.5, 2.6 – 2.10, 8.1
  * Bishop (PRML, 2006): Sec. 1.6
  * Murphy (MLaPP, 2012): Sec. 2.8