Machine Learning

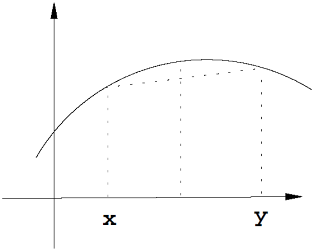
Lecture 01-2: Basics of Information Theory

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

Outline

Concave functions



- A function f is **concave** on interval I if for any $x, y \in I$,

$$\lambda f(x) + (1-\lambda) f(y) \leq f(\lambda x + (1-\lambda) y) \text{ for any} \lambda \in [0, 1]$$

Weighted average of function is upper bounded by function of weighted average.

It is **strictly concave** if the equality holds only when $x{=}y$.

Jensen's Inequality

Theorem (1.1)

Suppose function f is concave on interval I. Then

- *For any $p_i \in [0,1], \sum_{i=1}^{n} p_i = 1$ and $x_i \in I$.*

$$\sum_{i=1}^{n} p_i f(x_i) \leq f(\sum_{i=1}^{n} p_i x_i)$$

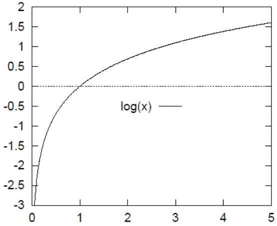
  Weighted average of function is upper bounded by function of weighted average.

- *If f is strictly CONCAVE, the equality holds iff $p_i \times p_j \neq 0$ implies $x_i = x_j$.*

Exercise: Prove this (using induction).

Logarithmic function

- The logarithmic function is concave in the interval $(0, \infty)$:



- Hence

$$\sum_{i=1}^{n} p_i log(x_i) \leq log(\sum_{i=1}^{n} p_i x_i) \qquad\qquad 0 \leq x_i$$

- In words, exchanging $\sum_i p_i$ with *log* increases quantity. Or, swapping expectation and logarithm increases quantity:

$$E[\log x] \leq \log E[x].$$

Outline

Entropy

- The **entropy** of a random variable X:

$$H(X) = \sum_X P(X) \log \frac{1}{P(X)} = -E_P[\log P(X)]$$

with convention that $0 \log(1/0) = 0$.

  - Base of logarithm is 2, unit is bit.
  - Sometimes, also called the entropy of the distribution, $H(P)$.
  - $H(X)$ **measures the amount of uncertainty about** X**.**

For real-valued variable, replace $\sum_X \ldots$ with $\int \ldots dx$.

Entropy

Example:

- X — result of coin tossing
- Y — result of dice throw
- Z — result of randomly pick a card from a deck of 54
- Which one has the highest uncertainty?
- Entropy:

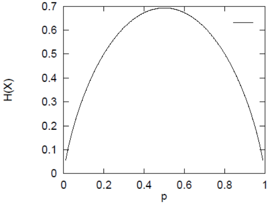
$$
\begin{aligned}
H(X) &= \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1(\log 2) \\
H(Y) &= \frac{1}{6}\log 6 + \ldots + \frac{1}{6}\log 6 = \log 6 \\
H(Z) &= \frac{1}{54}\log 54 + \ldots + \frac{1}{54}\log 54 = \log 54
\end{aligned}
$$

Indeed we have:

$$H(X) < H(Y) < H(Z).$$

Entropy

- X binary. The chart on the right shows $H(X)$ as a function of $p{=}P(X{=}1)$.
- The higher $H(X)$ is, the more uncertainty about the value of X



Entropy

Proposition (1.2)

- $H(X) \geq 0$
- $H(X) = 0$ *iff* $P(X{=}x) = 1$ *for some* $x \in \Omega_X$. *i.e. iff no uncertainty.*
- $H(X) \leq log(|X|)$ *with equality iff* $P(X{=}x){=}1/|X|$.
  Uncertainty is the highest in the case of uniform distribution.

Proof: Because *log* is concave, by Jensen's inequality:

$$
\begin{aligned}
H(X) &= \sum_X P(X) log \frac{1}{P(X)} \\
&\leq log \sum_X P(X) \frac{1}{P(X)} = log|X|
\end{aligned}
$$

Conditional entropy

- The **conditional entropy** of Y given event $X{=}x$:
  - Entropy of the conditional distribution $P(Y|X=x)$, i.e.

$$H(Y|X{=}x) = \sum_{Y} P(Y|X{=}x) log \frac{1}{P(Y|X{=}x)}$$

The uncertainty that remains about Y when X is known to be y.

- It is possible that $H(Y|X{=}x) > H(Y)$
  - Intuitively $X{=}x$ might contradicts our prior knowledge about Y and increase our uncertainty about Y
  - Exercise: Give example.

Conditional Entropy

- The **conditional entropy** of Y given variable X:

$$
\begin{aligned}
H(Y|X) &= \sum_{x} P(X = x) H(Y|X{=}x) \\
&= \sum_{X} P(X) \sum_{Y} P(Y|X) log \frac{1}{P(Y|X)} \\
&= \sum_{X,Y} P(X, Y) log \frac{1}{P(Y|X)} \\
&= -E[log P(Y|X)]
\end{aligned}
$$

The average uncertainty that remains about X when Y is known.

Outline

Kullback-Leibler divergence

- **Relative entropy** or **Kullback-Leibler divergence**
  - Measures how much a distribution $Q(X)$ differs from a "true" probability distribution $P(X)$.
  - **K-L divergence** of Q from P is defined as follows:

$$KL(P||Q) = \sum_{X} P(X) log \frac{P(X)}{Q(X)}$$

$0 log \frac{0}{0} = 0$ and $p log \frac{p}{0} = \infty$ if $p \neq 0$

Kullback-Leibler divergence

Theorem (1.2)

*(**Gibbs' inequality**)*

$$KL(P, Q) \geq 0$$

with equality holds iff P is identical to Q

Proof:

$$\begin{aligned}
\sum_X P(X) log \frac{P(X)}{Q(X)} &= -\sum_X P(X) log \frac{Q(X)}{P(X)} \\
&\geq -log \sum_X P(X) \frac{Q(X)}{P(X)} \qquad \text{Jensen's inequality} \\
&= -log \sum_X Q(X) = 0.
\end{aligned}$$

KL divergence between P and Q is larger than 0 unless P and Q are identical.

Cross Entropy

- Entropy: $H(P) = \sum_X P(X) log \frac{1}{P(X)} = -E[\log P(x)]$
- **Cross entropy**:

$$H(P, Q) = \sum_X P(X) log \frac{1}{Q(X)} = -E_P[log Q(X)]$$

- Relationship with KL:

$$\begin{aligned} KL(P||Q) &= \sum_X P(X) log \frac{P(X)}{Q(X)} = E_P[log P(X)] - E_P[log Q(X)] \\ &= H(P, Q) - H(P) \end{aligned}$$

Or,

$$H(P, Q) = KL(P||Q) + H(P)$$

A corollary

Corollary (1.1)

(Gibbs Inequality)

$$\begin{aligned} H(P, Q) &\geq H(P), \textit{ or} \\ \sum_X P(X) \log Q(X) &\leq \sum_X P(X) \log P(X) \end{aligned}$$

In general, let $f(X)$ be a non-negative function. Then

$$\sum_X f(X) \log Q(X) \leq \sum_X f(X) \log P^*(X)$$

where $P^*(X) = f(X) / \sum_X f(X)$.

Unsupervised Learning

- Unknown true distribution $P(\mathbf{x})$.

$$P(\mathbf{x}) \xrightarrow{\text{sampling}} \mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N} \xrightarrow{\text{learning}} Q(\mathbf{x})$$

- Objective:
  - **Minimizing KL**: $KL(P||Q)$
  - Same as **minimizing cross entropy**: $H(P, Q)$
  - Approximating the cross entropy using data:

$$\begin{aligned} H(P, Q) &= -\int P(\mathbf{x}) \log Q(\mathbf{x}) d\mathbf{x} \\ &\approx -\frac{1}{N} \sum_{i=1}^{N} \log Q(\mathbf{x}_i) \\ &= -\frac{1}{N} \log Q(\mathcal{D}) \end{aligned}$$

  - Same as **maximizing likelihood**: $\log Q(\mathcal{D})$.

Supervised Learning

- Unknown true distribution $P(\mathbf{x}, y)$, where y is **label** of input $\mathbf{x}$.

$$P(\mathbf{x}, y) \xrightarrow{\text{sampling}} \mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N \xrightarrow{\text{learning}} Q(y|\mathbf{x})$$

- Objective:
  - **Minimizing cross (conditional) entropy**:

$$\begin{aligned} H(P, Q) &= -\int P(\mathbf{x}, y) \log Q(y|\mathbf{x}) d\mathbf{x} dy \\ &\approx -\frac{1}{N} \sum_{i=1}^{N} \log Q(y_i|\mathbf{x}_i) \end{aligned}$$

  - Same as **maximizing loglikelihood**: $\sum_{i=1}^N \log Q(y_i|\mathbf{x}_i)$,
  - Or **minimizing the negative loglikelihood (NLL)**: $-\sum_{i=1}^N \log Q(y_i|\mathbf{x}_i)$

Jensen-Shannon divergence

- KL is not symmetric: $KL(P||Q)$ usually is not equal to reverse KL $KL(Q||P)$.
- **Jensen-Shannon divergence** is one symmetrized version of KL:

$$JS(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$$

  where $M = \frac{P+Q}{2}$
- Properties:
  - $0 \leq JS(P||Q) \leq \log 2$
  - $JS(P||Q) = 0$ if $P = Q$
  - $JS(P||Q) = \log 2$ if P and Q has disjoint support.

Outline

Mutual information

- The **mutual information** of X and Y:

$$I(X;Y) = H(X) - H(X|Y)$$

- Average reduction in uncertainty about X from learning the value of Y, or
- Average amount of information Y conveys about X.

Mutual information and KL Divergence

- Note that:

$$
\begin{aligned}
I(X;Y) &= \sum_{X} P(X) log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) log \frac{1}{P(X|Y)} \\
&= \sum_{X,Y} P(X,Y) log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) log \frac{1}{P(X|Y)} \\
&= \sum_{X,Y} P(X,Y) log \frac{P(X|Y)}{P(X)} \\
&= \sum_{X,Y} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)} \qquad \text{equivalent definition} \\
&= KL(P(X,Y)||P(X)P(Y))
\end{aligned}
$$

- Due to equivalent definition:

$$I(X;Y) = H(X) - H(X|Y) = I(Y;X) = H(Y) - H(Y|X)$$

Property of Mutual information

Theorem (1.3)

$$I(X;Y) \geq 0$$

with equality holds iff $X \perp Y$.

Interpretation: X and Y are independent iff X contains no information about Y and vice versa.

Proof: Follows from previous slide and Theorem 1.2.

Conditional Entropy Revisited

Theorem (1.4)

$H(X|Y) \leq H(X)$ with equality holds iff $X \perp Y$

Observation reduces uncertainty in average except for the case of independence.

Proof: Follows from Theorem 1.3.

Mutual information and Entropy

- From definition of mutual information

$$I(X;Y) = H(X) - H(X|Y)$$

  and the chain rule,

$$H(X,Y) = H(Y) + H(X|Y)$$

  we get

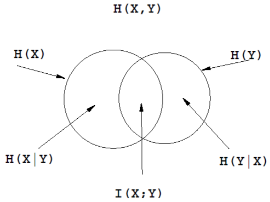
$$H(X) + H(Y) = H(X,Y) + I(X;Y)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- Consequently
  - $H(X,Y) \leq H(X) + H(Y)$ with equality holds iff $X \perp Y$.

Mutual information and entropy

Venn Diagram: Relationships among joint entropy, conditional entropy, and mutual information



$$H(X) + H(Y) = H(X, Y) + I(X; Y)$$

$$I(X; Y) = H(X) - H(X|Y)$$

$$I(Y; X) = H(Y) - H(Y|X)$$

Conditional Mutual information

- The **conditional mutual information** of X and Y given Z:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

- Average amount of information Y conveys about X given Z.

Conditional mutual information and KL Divergence

Note:

$$
\begin{aligned}
I(X;Y|Z) &= \sum_{X,Z} P(X,Z) log \frac{1}{P(X|Z)} - \sum_{X,Y,Z} P(X,Y,Z) log \frac{1}{P(X|Y,Z)} \\
&= \sum_{X,Y,Z} P(X,Y,Z) log \frac{1}{P(X|Z)} - \sum_{X,Y,Z} P(X,Y,Z) log \frac{1}{P(X|Y,Z)} \\
&= \sum_{X,Y,Z} P(X,Y,Z) log \frac{P(X|Y,Z)}{P(X|Z)} \qquad \text{equivalent definition} \\
&= \sum_{Z} P(Z) \sum_{X,Y} P(X,Y|Z) log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)} \\
&= \sum_{Z} P(Z) KL(P(X,Y|Z), P(X|Z)P(Y|Z)) \geq 0.
\end{aligned}
$$

Property of conditional mutual information

Theorem (1.5)

$$I(X;Y|Z) \geq 0$$

$$H(X|Z) \geq H(X|Y,Z)$$

with equality hold iff $X \perp Y|Z$.

Interpretation:

- More observations reduce uncertainty on average except for the case of conditional independence.
- X and Y are independently given Z iff X contain no information about Y given Z and vice versa:

  $$X \perp Y|Z \equiv I(X;Y|Z) = 0.$$

  Another characterization of conditional independence.