Elementi di Teoria dell'Informazione

R. Basili

Corso di *Web Mining* e *Retrieval*
a.a. 2008-9

March 23, 2010

Outline

Outline

- Information Theory
- Entropy
- Joint-Entropy and Conditional entropy
- Mutual Information
- Cross-Entropy and Norms

Information Theory

Let ξ be a discrete stochastic variable with a finite range $\Omega_\xi = \{x_1, \ldots, x_M\}$ and let $p_i = p(x_i)$ be the corresponding probabilities.

How much information is there in knowing the outcome of ξ?

Information Theory

Let ξ be a discrete stochastic variable with a finite range $\Omega_\xi = \{x_1, \ldots, x_M\}$ and let $p_i = p(x_i)$ be the corresponding probabilities.

> How much information is there in knowing the outcome of ξ?

Or equivalently:

> How much uncertainty arises if the outcome ξ is unknown?

Information Theory

Let ξ be a discrete stochastic variable with a finite range $\Omega_\xi = \{x_1, \ldots, x_M\}$ and let $p_i = p(x_i)$ be the corresponding probabilities.

How much information is there in knowing the outcome of ξ?

Or equivalently:

How much uncertainty arises if the outcome ξ is unknown?

This is the information needed to specify which of the x_i has occurred. The problem is writing ξ.

Information Theory

Let ξ be a discrete stochastic variable with a finite range $\Omega_\xi = \{x_1, \ldots, x_M\}$ and let $p_i = p(x_i)$ be the corresponding probabilities.

How much information is there in knowing the outcome of ξ?

Or equivalently:

How much uncertainty arises if the outcome ξ is unknown?

This is the information needed to specify which of the x_i has occurred. The problem is writing ξ.
Let us assume further that we only have a small set of symbols $A = \{a_k : k = 1, \ldots D\}$, that is a *coding alphabet*.

Information Theory

Thus each x_i will be represented by a string over A.
Let us assume that ξ is *uniformly distributed*, i.e.

$$p_i = \frac{1}{M} \qquad \forall i = 1, \ldots, M,$$

and that the coding alphabet is exactly $A = \{0, 1\}$.

Information Theory

Thus each x_i will be represented by a string over A.
Let us assume that ξ is *uniformly distributed*, i.e.

$$p_i = \frac{1}{M} \qquad \forall i = 1, \ldots, M,$$

and that the coding alphabet is exactly $A = \{0, 1\}$.
Thus, each x_i will be represented by a binary number. To use N binary digits to specify which x_i actually occurred means:

$$N : 2^{N-1} < M \leq 2^N$$

Thus we need $N = \lceil \log_2 M \rceil$ digits.
So what if the distribution is *nonuniform*, i.e., if the p_is are not all equal?

Information Theory

How much uncertainty does a possible outcome with probability introduce?

Information Theory

How much uncertainty does a possible outcome with probability introduce?
The basic assumption is that p_i will introduce equally much uncertainty regardless of the rest of the probabilities p_j with $j \neq i$.

Information Theory

How much uncertainty does a possible outcome with probability introduce?
The basic assumption is that p_i will introduce equally much uncertainty regardless of the rest of the probabilities p_j with $j \neq i$.
We can thus reduce the problem to the case where all outcomes have probability p_i. In this case, there are $\frac{1}{p_i} = M_{p_i}$ possible outcomes.

Information Theory

How much uncertainty does a possible outcome with probability introduce?
The basic assumption is that p_i will introduce equally much uncertainty regardless of the rest of the probabilities p_j with $j \neq i$.
We can thus reduce the problem to the case where all outcomes have probability p_i. In this case, there are $\frac{1}{p_i} = M_{p_i}$ possible outcomes.
Example: if $p_i \approx 1$ then $M_{p_i} \approx 1$.

Information Theory

How much uncertainty does a possible outcome with probability introduce?
We can thus reduce the problem to the case where all outcomes have probability p_i. In this case, there are $\frac{1}{p_i} = M_{p_i}$ possible outcomes.

Information Theory

How much uncertainty does a possible outcome with probability introduce?

We can thus reduce the problem to the case where all outcomes have probability p_i. In this case, there are $\frac{1}{p_i} = M_{p_i}$ possible outcomes.

For a binary coding alphabet, we thus need

$$\log_2 M_{p_i} = \log_2 \frac{1}{p_i} = -\log_2 p_i$$

binary digits to specify that the outcome was x_i.

Thus, the uncertainty introduced by p_i is in the general case

$$-\log_2 p_i$$

Entropy

Uncertainty of ξ

The uncertainty introduced by the random variable ξ will be taken to be the *expectation value of the number of digits required to specify its outcome*.

Entropy

Uncertainty of ξ

The uncertainty introduced by the random variable ξ will be taken to be the *expectation value of the number of digits required to specify its outcome*.
This is the expectation value of $-\log_2 P(\xi)$, i.e.

$$E[-\log_2 P(\xi)] = \sum_i -p_i \log_2 p_i$$

Entropy

Entropy

The entropy $H[\xi]$ of ξ is precisely the amount of uncertainty introduced by the random variable ξ and it is more often referred to a natural logarithm $\ln(.)$, so that

$$H[\xi] = E[-\ln p(\xi)] = \sum_{x_i \in \Omega_\xi} -p(x_i) \ln p(x_i) = \sum_{i}^{M} -p_i \ln p_i$$

Entropy

Example 1: Dado

In the Dado example, $\forall i = 1, \ldots, 6$, it follows that $p_i = \frac{1}{6}$.

$$H[\xi] = E[-\ln p(\xi)] = \sum_{x_i \in \Omega_\xi} -p(x_i)\ln p(x_i) = 6 \cdot \frac{1}{6} \ln 6 = 1,792$$

Entropy

Example 1: Dado

In the Dado example, $\forall i = 1, \ldots, 6$, it follows that $p_i = \frac{1}{6}$.

$$H[\xi] = E[-\ln p(\xi)] = \sum_{x_i \in \Omega_\xi} -p(x_i) \ln p(x_i) = 6 \cdot \frac{1}{6} \ln 6 = 1,792$$

Example 2: Dado Perdente

A loosing Die: $p_1 = 1.00$, and $\forall i = 2, \ldots, 6$, $p_i = 0$.

$$H[\xi] = E[-\ln p(\xi)] = \sum_{x_i \in \Omega_\xi} -p(x_i) \ln p(x_i) = 1 \ln 1 = 0$$

Entropy

Consequence

Given a distribution $p_i \quad (i = 1, , \ldots, M)$ for a discrete random variable ξ then for any other distribution $q_i \quad (i = 1, , \ldots, M)$ over the same sample space Ω_ξ it follows that:

$$H[\xi] = -\sum_i^M p_i \ln p_i \leq -\sum_i^M p_i \ln q_i$$

where equality holds **iff** the two distribution are the same, i.e. $\forall i = 1, \ldots, M \qquad p_i = q_i$

Joint-Entropy

Given two random variable ξ and η:

Joint-Entropy

the *joint entropy* of ξ and η is defined as:

$$H[\xi,\eta] = -\sum_{i=1}^{M}\sum_{j=1}^{L} p(x_i,y_j)\ln p(x_i,y_j)$$

Joint-Entropy

Given two random variable ξ and η:

Joint-Entropy

the *joint entropy* of ξ and η is defined as:

$$H[\xi,\eta] = -\sum_{i=1}^{M}\sum_{j=1}^{L} p(x_i,y_j)\ln p(x_i,y_j) = H[\eta,\xi]$$

Conditional-entropy

Conditional Entropy

the *conditional entropy* $H[\xi|\eta]$ of ξ and η is defined as:

$$
\begin{aligned}
H[\xi|\eta] &= -\sum_{j=1}^{L} p(y_j) \sum_{i=1}^{M} p(x_i|y_j) \ln p(x_i|y_j) = \\
&= -\sum_{j=1}^{L} \sum_{i=1}^{M} p(x_i, y_j) \ln p(x_i|y_j)
\end{aligned}
$$

Conditional and joint entropy

Conditional and Joint Entropy

The conditional and joint entropies are related just like the conditional and joint probabilities:

$$H[\xi,\eta] = H[\eta] + H[\xi|\eta]$$

Conditional and joint entropy

Conditional and Joint Entropy

The conditional and joint entropies are related just like the conditional and joint probabilities:

$$H[\xi,\eta] = H[\eta] + H[\xi|\eta]$$

Conveyed Information

The *information conveyed* by η, denoted $I[\xi|\eta]$, is the reduction in entropy of ξ by finding out the outcome of η. This is defined by:

$$I[\xi|\eta] = H[\xi] - H[\xi|\eta]$$

Conditional and joint entropy

Conditional and Joint Entropy

$$H[\xi,\eta] = H[\eta] + H[\xi|\eta]$$
$$I[\xi|\eta] = H[\eta] - H[\xi|\eta]$$

Conditional and joint entropy

Conditional and Joint Entropy

$$H[\xi,\eta] = H[\eta] + H[\xi|\eta]$$
$$I[\xi|\eta] = H[\eta] - H[\xi|\eta]$$

Consequences

Note that:

$$\begin{aligned} I[\xi|\eta] &= H[\xi] - H[\xi|\eta] = H[\xi] - (H[\xi,\eta] - H[\eta]) = \\ &= H[\xi] + H[\eta] - H[\xi,\eta] = H[\xi] + H[\eta] - H[\eta,\xi] = \\ &= H[\eta] + H[\xi] - H[\eta,\xi] = H[\eta] - H[\eta|\xi] = \\ &= I[\eta|\xi] \end{aligned}$$

Mutual Information

Given two random variable ξ and η:

Mutual Information

the *mutual information* between ξ and η is defined as:

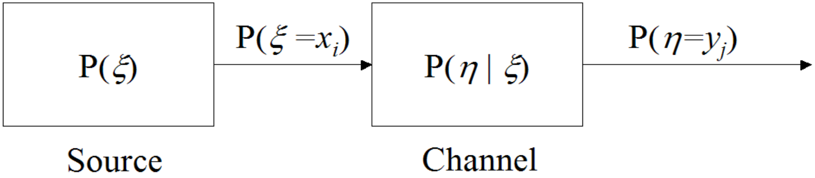
$$
\begin{aligned}
MI[\xi,\eta] &= E[\ln \frac{P(\xi,\eta)}{P(\xi)\cdot P(\eta)}] = \\
&= \sum_{(x,y)\in\Omega_{(\xi,\eta)}} f_{(\xi,\eta)}(x,y) \ln \frac{f_{(\xi,\eta)}(x,y)}{f_\xi(x)\cdot f_\eta(y)}
\end{aligned}
$$

Mutual Information

Mutual Information measures the amount of information about a random variable ξ an observer receives when the outcome of a random variable η is available.

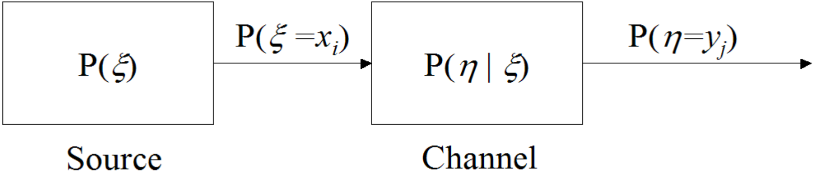
Mutual Information

Mutual Information measures the amount of information about a random variable ξ an observer receives when the outcome of a random variable η is available.



Mutual Information

Mutual Information measures the amount of information about a random variable ξ an observer receives when the outcome of a random variable η is available.



How much information about the source output x_i does an observer gain by knowing the channel output y_j?

Mutual Information

Mutual Information measures the amount of information about a random variable ξ an observer receives when the outcome of a random variable η is known, in fact:

Mutual Information

$$
\begin{aligned}
MI[\xi,\eta] &= H[\xi] - H[\xi|\eta] = \\
&= \sum_{(x,y)\in\Omega_{(\xi,\eta)}} f_{(\xi,\eta)}(x,y) \ln \frac{f_{(\xi,\eta)}(x,y)}{f_\xi(x)\cdot f_\eta(y)}
\end{aligned}
$$

Mutual Information

Mutual Information measures the amount of information about a random variable ξ an observer receives when the outcome of a random variable η is known, in fact:

Mutual Information

$$\begin{aligned} MI[\xi,\eta] &= H[\xi] - H[\xi|\eta] = \\ &= \sum_{(x,y)\in\Omega_{(\xi,\eta)}} f_{(\xi,\eta)}(x,y) \ln \frac{f_{(\xi,\eta)}(x,y)}{f_\xi(x)\cdot f_\eta(y)} \end{aligned}$$

Mutual Information

MI and H

$$MI[\xi,\eta] = H[\xi] - H[\xi|\eta]$$

Mutual Information

MI and H

$$MI[\xi,\eta] = H[\xi] - H[\xi|\eta]$$
$$H[\xi,\eta] = H[\eta,\xi]$$
$$H[\xi,\eta] = H[\eta] + H[\xi|\eta],$$

Mutual Information

MI and H

$$MI[\xi,\eta] = H[\xi] - H[\xi|\eta]$$
$$H[\xi,\eta] = H[\eta,\xi]$$
$$H[\xi,\eta] = H[\eta] + H[\xi|\eta], \qquad H[\xi|\eta] = H[\xi,\eta] - H[\eta]$$

Mutual Information

MI and H

$MI[\xi,\eta] = H[\xi] - H[\xi|\eta]$
$H[\xi,\eta] = H[\eta,\xi]$
$H[\xi,\eta] = H[\eta] + H[\xi|\eta], \qquad H[\xi|\eta] = H[\xi,\eta] - H[\eta]$

Symmetry

Note that mutual information is symmetric in ξ and η, that is $MI[\xi,\eta] = MI[\eta,\xi]$, as

$$H[\xi] - H[\xi|\eta] = H[\xi] + H[\eta] - H[\xi,\eta] = H[\eta] - H[\eta|\xi]$$

Pointwise Mutual Information

Another way to look to mutual information is about the individual values (i.e. outcomes) $\xi = x_i$ and $\eta = y_j$.

Pointwise Mutual Information

Another way to look to mutual information is about the individual values (i.e. outcomes) $\xi = x_i$ and $\eta = y_j$.

Pointwise Mutual Information

Given the two random variable ξ and η: the *pointwise mutual information* between $\xi = x_i$ and $\eta = y_j$ is defined as:

$$MI[x_i, y_j] = f_{(\xi,\eta)}(x_i, y_j) \ln \frac{f_{(\xi,\eta)}(x_i, y_j)}{f_\xi(x_i) \cdot f_\eta(y_j)}$$

Pointwise Mutual Information

Another way to look to mutual information is about the individual values (i.e. outcomes) $\xi = x_i$ and $\eta = y_j$.

Pointwise Mutual Information

Given the two random variable ξ and η: the *pointwise mutual information* between $\xi = x_i$ and $\eta = y_j$ is defined as:

$$MI[x_i, y_j] = f_{(\xi,\eta)}(x_i, y_j) \ln \frac{f_{(\xi,\eta)}(x_i, y_j)}{f_\xi(x_i) \cdot f_\eta(y_j)} = P(x_i, y_j) \ln \frac{P(x_i, y_j)}{P(x_i) \cdot P(y_j)}$$

Pointwise Mutual Information

Pointwise Mutual Information (pmi)

$$MI[x_i, y_j] = P(x_i, y_j) \ln \frac{P(x_i, y_j)}{P(x_i) \cdot P(y_j)}$$

Pointwise Mutual Information

Pointwise Mutual Information (pmi)

$$MI[x_i, y_j] = P(x_i, y_j) \ln \frac{P(x_i, y_j)}{P(x_i) \cdot P(y_j)}$$

Use of the pmi

If $MI[x_i, y_j] >> 0$, there is a strong correlation between x_i and y_j
If $MI[x_i, y_j] << 0$, there is a strong negative correlation.
When $MI[x_i, y_j] \approx 0$ the two outcomes are almost independent.

Perplexity

Perplexity

The *perplexity* of a random variable ξ is the exponential of its entropy, i.e.

$$Perp[\xi] = e^{H[\xi]}$$

Perplexity

Perplexity

The *perplexity* of a random variable ξ is the exponential of its entropy, i.e.

$$Perp[\xi] = e^{H[\xi]}$$

Example

Predicting the next w of a sequence of n words $w_k \in Dict$:

$$P(\xi_n = w | \xi_{n-1} = w_{n-1}, \xi_{n-2} = w_{n-2}, \ldots, \xi_1 = w_1)$$

What is $Perp[(\xi_n, \ldots, \xi_1)]$?
OSS: In case of a uniform distribution $P(\xi_n = w|\ldots) = \frac{1}{|Dict|}$ ...

Cross-entropy

Cross-entropy

If we have two distributions (collections of probabilities) $p(x)$ and $q(x)$ on Ω_ξ, then the *cross entropy* of p with respect to q is given by:

$$H_p[q] = -\sum_{x \in \Omega_\xi} p(x) \ln q(x)$$

Cross-entropy

Cross-entropy

If we have two distributions (collections of probabilities) $p(x)$ and $q(x)$ on Ω_ξ, then the *cross entropy* of p with respect to q is given by:

$$H_p[q] = -\sum_{x\in\Omega_\xi} p(x)\ln q(x)$$

Minimality

$$H_p[q] = -\sum_{x\in\Omega_\xi} p(x)\ln q(x) \geq -\sum_{x\in\Omega_\xi} p(x)\ln p(x) \quad \forall q$$

implies that the cross entropy of a distribution q w.r.t. another distribution p is **minimal** when q is identical to p.

Cross-entropy as a Norm

Cross-entropy

$$H_p[q] = -\sum_{x \in \Omega_\xi} p(x) \ln q(x)$$

Cross-entropy as a Norm

Cross-entropy

$$H_p[q] = -\sum_{x \in \Omega_\xi} p(x) \ln q(x)$$

Relative Entropy (or Kullback-Leibler distance)

$$D[p||q] = \sum_{x \in \Omega_\xi} p(x) \ln \frac{p(x)}{q(x)} = H_p[q] - H[p]$$

Cross-entropy and Norms

Relative Entropy (or Kullback-Leibler distance)

$$D[p||q] = \sum_{x \in \Omega_\xi} p(x) \ln \frac{p(x)}{q(x)} = H_p[q] - H[p]$$

KL distance: properties

$$D[p||q] \geq 0 \quad \forall q$$

Cross-entropy and Norms

Relative Entropy (or Kullback-Leibler distance)

$$D[p||q] = \sum_{x \in \Omega_\xi} p(x) \ln \frac{p(x)}{q(x)} = H_p[q] - H[p]$$

KL distance: properties

$$D[p||q] \geq 0 \quad \forall q$$

$$D[p||q] = 0 \qquad \textbf{iff } q = p$$

Cross-entropy and Norms

Relative Entropy (or Kullback-Leibler distance)

$$D[p||q] = \sum_{x \in \Omega_\xi} p(x) \ln \frac{p(x)}{q(x)} = H_p[q] - H[p]$$

Cross-entropy and Norms

Relative Entropy (or Kullback-Leibler distance)

$$D[p||q] = \sum_{x \in \Omega_\xi} p(x) \ln \frac{p(x)}{q(x)} = H_p[q] - H[p]$$

KL distance as a norm?

Unfortunately, as

$$D[p||q] \neq D[q||p]$$

the KL distance is *not* a valid metric in the classical terms. It is a *measure of the dissimilarity* between p and q.

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- During a learning process we need to figure out the circumstances (i.e. the state of affairs of the world) under which a certain concept/class/property manifest.

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- During a learning process we need to figure out the circumstances (i.e. the state of affairs of the world) under which a certain concept/class/property manifest.
- This make a direct reference to the probability of some (stochastic) event. Stochastic events are used to describe circumstances and properties.

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- During a learning process we need to figure out the circumstances (i.e. the state of affairs of the world) under which a certain concept/class/property manifest.
- This make a direct reference to the probability of some (stochastic) event. Stochastic events are used to describe circumstances and properties.
- Moreover, learning proceeds from experience, i.e. known facts or previous classified examples, to rules, i.e. probability joint distributions over *decisions* and *circumstances*

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- During a learning process we need to figure out the circumstances (i.e. the state of affairs of the world) under which a certain concept/class/property manifest.
- This make a direct reference to the probability of some (stochastic) event. Stochastic events are used to describe circumstances and properties.
- Moreover, learning proceeds from experience, i.e. known facts or previous classified examples, to rules, i.e. probability joint distributions over *decisions* and *circumstances*
- Learning in general means **to induce the proper probability distributions from the known examples**. There are several many ways to do it!!!

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- During a learning process we need to figure out the circumstances (i.e. the state of affairs of the world) under which a certain concept/class/property manifest.
- This make a direct reference to the probability of some (stochastic) event. Stochastic events are used to describe circumstances and properties.
- Moreover, learning proceeds from experience, i.e. known facts or previous classified examples, to rules, i.e. probability joint distributions over *decisions* and *circumstances*
- Learning in general means **to induce the proper probability distributions from the known examples**. There are several many ways to do it!!!

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- **Consequences.** In general, we need to compare different inductive hypothesis (*IH*), that are different probability distributions q_i of the same decision,

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- **Consequences.** In general, we need to compare different inductive hypothesis (*IH*), that are different probability distributions q_i of the same decision,
- In order to do it, we measure the agreement of our hypothesis with the observations (i.e. a pool of annotated data kept aside, the *held out*, to validate the different q_i)

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- **Consequences.** In general, we need to compare different inductive hypothesis (*IH*), that are different probability distributions q_i of the same decision,
- In order to do it, we measure the agreement of our hypothesis with the observations (i.e. a pool of annotated data kept aside, the *held out*, to validate the different q_i)
- The result is an estimate of the similarity between the probability q_i induced at the i-th learning stage with the probability p characterizing the known examples.

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- **Consequences.** In general, we need to compare different inductive hypothesis (*IH*), that are different probability distributions q_i of the same decision,
- In order to do it, we measure the agreement of our hypothesis with the observations (i.e. a pool of annotated data kept aside, the *held out*, to validate the different q_i)
- The result is an estimate of the similarity between the probability q_i induced at the i-th learning stage with the probability p characterizing the known examples.
- The KL divergence $D[p||q] = H_p(q) - H(p)$ can be the suitable dissimilarity function.

Norms, Similarity and Learning

Why ranking probability distributions is necessary?

- **Consequences.** In general, we need to compare different inductive hypothesis (*IH*), that are different probability distributions q_i of the same decision,
- In order to do it, we measure the agreement of our hypothesis with the observations (i.e. a pool of annotated data kept aside, the *held out*, to validate the different q_i)
- The result is an estimate of the similarity between the probability q_i induced at the i-th learning stage with the probability p characterizing the known examples.
- The KL divergence $D[p||q] = H_p(q) - H(p)$ can be the suitable dissimilarity function.
- The probability $\hat{q}$ such that $\hat{q} = argmax_i D[p||q_i]$ is returned.

Norm

What makes a function a norm?

Norm

What makes a function a norm? Any binary mapping m between a set of objects $D \times D$ and the real numbes is a norm **iff**:

Axioms

- (*Positive*) $m(X,Y) \geq 0 \quad \forall X, Y \in D$ whereas $m(X,Y) = 0 \rightarrow X = Y$.

Norm

What makes a function a norm? Any binary mapping m between a set of objects $D \times D$ and the real numbes is a norm **iff**:

Axioms

- (*Positive*) $m(X,Y) \geq 0 \qquad \forall X,Y \in D$ whereas $m(X,Y) = 0 \rightarrow X = Y$.
- (*Simmetry*) $m(X,Y) = m(Y,X) \qquad \forall X,Y \in D$

Norm

What makes a function a norm? Any binary mapping m between a set of objects $D \times D$ and the real numbes is a norm **iff**:

Axioms

- (*Positive*) $m(X,Y) \geq 0 \qquad \forall X,Y \in D$ whereas $m(X,Y) = 0 \rightarrow X = Y$.
- (*Simmetry*) $m(X,Y) = m(Y,X) \qquad \forall X,Y \in D$
- (*Triangle inequality*) $m(X,Y) \leq m(X,Z) + m(Z,Y) \qquad \forall X,Y,Z \in D$

Norm

What makes a function a norm? Any binary mapping m between a set of objects $D \times D$ and the real numbes is a norm **iff**:

Axioms

- (*Positive*) $m(X,Y) \geq 0 \qquad \forall X, Y \in D$ whereas $m(X,Y) = 0 \rightarrow X = Y$.
- (*Simmetry*) $m(X,Y) = m(Y,X) \qquad \forall X, Y \in D$
- (*Triangle inequality*) $m(X,Y) \leq m(X,Z) + m(Z,Y) \qquad \forall X, Y, Z \in D$

Euclidean Norm

$$\sqrt[2]{\sum_{x \in \Omega(\xi)} (p(x) - q(x))^2}$$

References

Elementary Information Theory

- in (Krenn & Samuelsson, 1997), Brigitte Krenn, Christer Samuelsson, *The Linguist's Guide to Statistics Don't Panic*, Univ. of Saarlandes, 1997.
  URL:
  `http://nlp.stanford.edu/fsnlp/dontpanic.pdf`