

Information Dependencies

Outline

Introduction

InD Measure

InD Inequalities

InD Constraints

InDs and Traditional Dependencies

Applications

Database Context

We are still dealing with the surprise factor of information

- not the surprise of *receiving* a message
- but the surprise of *finding* a value in a relation

Assumptions:

- based on relation instance(s)
- instance is a “multiset” (or “bag”) where one tuple may occur multiple times

Notations & Conventions

1. r is a relation instance with attributes R
2. $X \subseteq R$, with values $x_1, \cdot \cdot \cdot, x_l$
similarly for Y, Z
3. XY means $X \cup Y$, *etc.*
4. Assume all tuples equally likely and
define $p_{X=x_i}$ (also written p_{x_i}) as

$$\frac{(\text{SELECT COUNT} (*) \text{ FROM } r \text{ WHERE } X = x_i)}{(\text{SELECT COUNT} (*) \text{ FROM } r)}$$
5. Because $\lim_{p \rightarrow 0} (p \times \log(1/p)) = 0$,
assume $0 \times \log(1/0) = 0$

Probability & Entropy from Counts

Define the entropy of X in \mathbf{r} as

$$H_X(\mathbf{r}) = \sum_{i=1}^l p_{x_i} \times \log(1/p_{x_i})$$

(omit \mathbf{r} when only one instance is being discussed)

(This is the same formula as $H(\mathbf{P})$.)

H increases with X . That is, $H_X \leq H_{XY}$

If \mathbf{r} is a set instance (no duplicates) with k tuples,
then $H_R = \log(k)$

InD Definition

The *Information Dependency measure* answers:

“what information do we need to determine Y provided we already know X ?”

Abbreviate by “InD measure” and denote by $H_{X \rightarrow Y}$

Define by evaluating H_Y in each partition on x_i , that is in

```
SELECT *
FROM r
WHERE X = xi
```

and weight by $p_{X=x_i}$

Equivalently: $H_{X \rightarrow Y} = H_{XY} - H_X$

Example

Instance and entropies:

r				
<i>X</i>	<i>Y</i>			
a	e		H_X	= 7/4
a	f		H_Y	= 3/2
a	e		H_{XY}	= 9/4
a	f			
b	g		$H_{X \rightarrow Y}$	= 1/2
b	g			
c	g		$H_{Y \rightarrow X}$	= 3/4
d	g			

Encodings:

	<i>X</i>	<i>XY</i>	<i>Y</i>	
a	0	00 01	00 01	f e
b	10	10		
c	110	110	1	g
d	111	111		

Inequality Facts

The following hold in all instances r :

1. $H_{XY \rightarrow X} = 0$

2. $H_{X \rightarrow Y} + H_{X \rightarrow Z} \geq H_{X \rightarrow YZ}$

3. $H_{XZ \rightarrow YZ} = H_{XZ \rightarrow Y}$

4. $H_{XY \rightarrow Z} \leq H_{X \rightarrow Z}$

5. $H_{XZ \rightarrow YZ} \leq H_{X \rightarrow Y}$

6. $H_{X \rightarrow Y} + H_{Y \rightarrow Z} \geq H_{X \rightarrow Z}$

Constraints and Observations

InD inequalities hold for any \mathbf{r} , but we can also stipulate other numeric relationships.

⇒ InD constraint

e.g. $H_{X \rightarrow Y} \leq 0.5$

Similar to “foreign key constraint”:

☆ an InD constraint only holds because we have chosen to enforce it

Also, for some particular instance \mathbf{r} , we might observe (that is, calculate this H in \mathbf{r})

$$H_{X \rightarrow Y}(\mathbf{r}) = 0.463 \leq 0.5$$

Functional Dependencies

Theorem:

$$X \rightarrow Y \text{ IFF } H_{X \rightarrow Y} = 0$$

Moreover, Armstrong's Axioms are just special cases of InD inequalities:

- reflexivity:
 $X \rightarrow Y$ if $Y \subseteq X$ (ineq. 1)
- monotonicity:
 $X \rightarrow Y \Rightarrow XZ \rightarrow YZ$ (ineq. 5)
- transitivity:
 $X \rightarrow Y \ \& \ Y \rightarrow Z \Rightarrow X \rightarrow Z$ (ineq. 6)

Multivalued Dependencies

A new characterization of MVDs:

- X , Y , and Z partition R

Definition: $X \twoheadrightarrow Y|Z$ in r IFF there exists r' and r'' , over XY and XZ respectively, such that

$$r = r' \bowtie r''$$

When r is a set (no duplicates), this is a standard MVD characterization:

- $r' = \pi_{XY}(r)$
- $r'' = \pi_{XZ}(r)$

Theorem:

$$X \twoheadrightarrow Y|Z \text{ in } r \quad \text{IFF} \quad H_{X \rightarrow Y} + H_{X \rightarrow Z} = H_{X \rightarrow YZ}$$

- Alternative definition of MVDs works for both sets and multisets

Example Decomposition

Numbers to the right are counts

r			
X	Y	Z	
a	b	d	8
a	b	e	4
a	c	d	4
a	c	e	2

$$H_X = 0$$

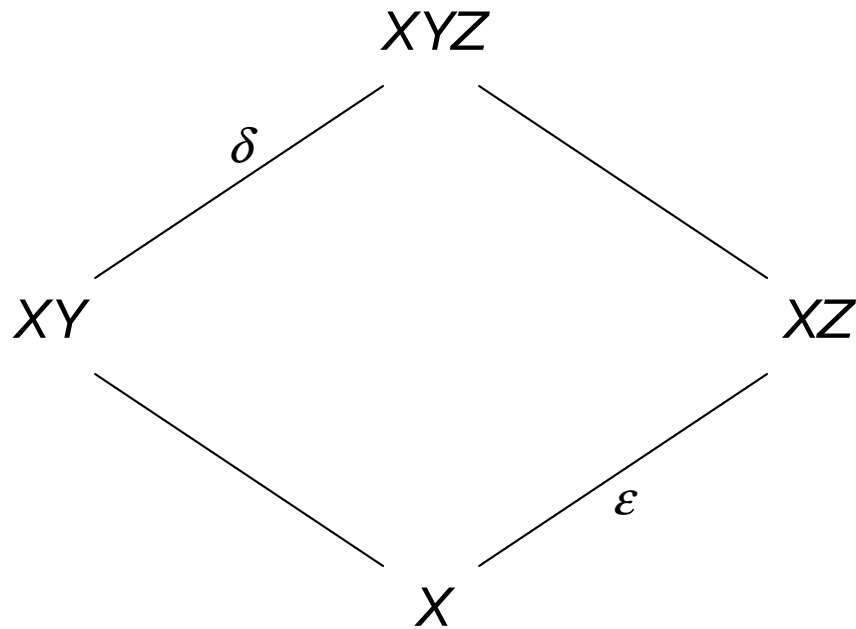
$$H_{X \rightarrow Y} = H_Y = H_Z = H_{X \rightarrow Z} = 0.918$$

$$H_{X \rightarrow YZ} = H_{YZ} = 1.837$$

r'			r''		
X	Y		X	Z	
a	b	4	a	d	2
a	c	2	a	e	1

Relating FDs and MVDs

Consider H over



Bounds on δ :

- $\delta \geq 0$
- $\delta \leq \epsilon$

Explaining “Lossless” Joins

Explaining why $X \twoheadrightarrow Y|Z$ fails, decompose r into

$$r' = \Pi_{XY}(r) \text{ and } r'' = \Pi_{XZ}(r)$$

This decomposition is *lossless* if $r' \bowtie r'' = r$ and *lossy* otherwise

Example of lossy decomposition:

r		
X	Y	Z
a	b	d
a	b	e
a	c	d

r'	
X	Y
a	b
a	c

r''	
X	Z
a	d
a	e

Q: $r' \bowtie r'' \supseteq r$, so what is **lost**?

A: $H_{X \rightarrow Z} - H_{XZ \rightarrow Y}$ ($\varepsilon - \delta$ in previous diagram)

- known as “joint information between Y and Z given X ”

Computing H

Datacube technique computes the counts needed to compute H

- standard datamining operation
- compute counts on all subsets of R
- huge output
 - ⇒ costly (unnecessarily so?)
 - ⇒ heuristic disregards small cases
- optimization + heuristics
 - ◇ currently used heuristics may not apply
 - heuristics with small datacube errors
 - may have big impact on entropy
 - ◇ new ones might

Approximate FDs and MVDs

Given r , what are all X, Y such that

$$H_{X \rightarrow Y} \leq 0.10?$$

or any other constant instead of 0.10

Given r , for what partitions XYZ of R is

$$H_{X \rightarrow Y} + H_{X \rightarrow Z} - H_{X \rightarrow YZ} \leq .10?$$

Comparing Functional Dependency Approximations

Kivinen and Mannila define three measures for approximate functional dependencies:

g_1 fraction of violating pairs

g_2 fraction of violating tuples

g_3 minimum fraction of tuples removed to eliminate all violations

where s and t are violating if
 $s.X = t.X$ & $s.Y \neq t.Y$

InD's can make distinctions that the g measures cannot

Approximate Functional Dependency Example

X	Y
d	1
d	1
a	1
a	2
a	3
a	4
a	5
a	6
b	1
c	1
c	2

Diagram illustrating the relationship between sets r , s , and t and the data rows in the table above:

- Set r is indicated by a bracket on the left, covering the first 7 rows (X values: d, d, a, a, a, a, a).
- Set s is indicated by a bracket on the right, covering the last 6 rows (X values: a, a, a, a, b, c).
- Set t is indicated by a bracket on the right, covering the last 5 rows (X values: a, a, a, a, b).

	H_X	$H_{X \rightarrow Y}$	g_1	g_2	g_3
r	1.52	.80	.16	.8	.4
s	1.37	.95	.36	.8	.4
t	1.57	1.55	.36	.8	.4

Data Warehousing

Data Warehouse:

Archived transaction data organized to support subsequent “decision support” processing

Use of InD measures:

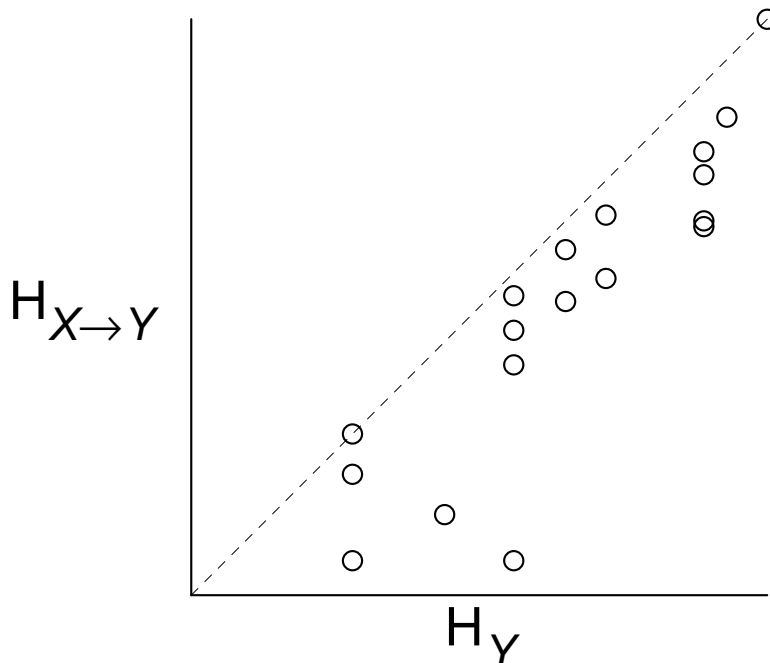
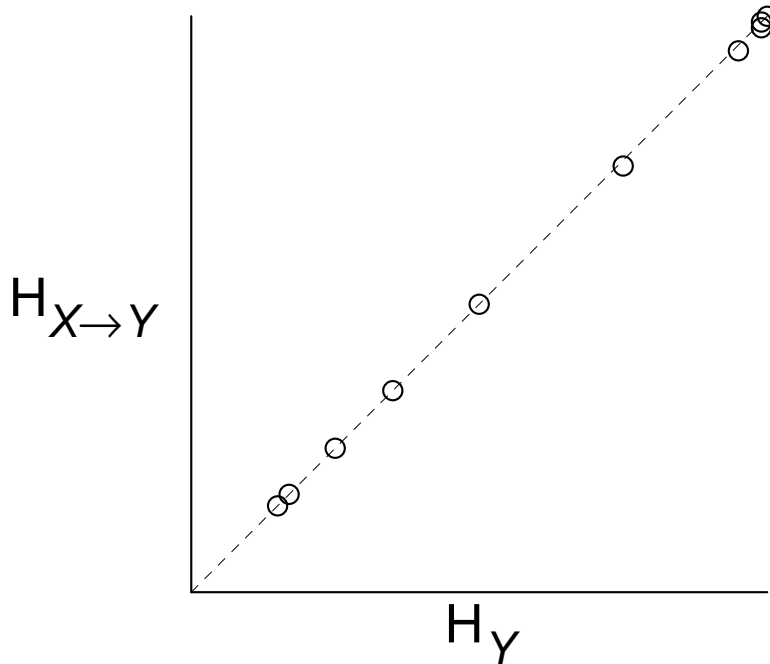
indicate what data and relationships may be significant

e.g. “partial” functional dependencies may indicate space savings

Visualizing Entropy Properties

Information maps of two database benchmarks
each point corresponds to one attribute

Which benchmark is natural and which is synthetic?



Query Optimization

- Given r over X , Y , and Z , where $H_{X \rightarrow Y} < \varepsilon$
 - ◊ (so $X \rightarrow Y$ holds approximately)
- Decompose r by:
 - ◊ subtract “correction” r^C from r so FD holds in remainder
 - ◊ do FD decomposition of remainder into r' (over XZ) and r'' (over XY)

$$\text{Thus } r = r' \bowtie r'' \cup r^C$$

- Improves performance of certain queries over a wide range of ε
provided that query optimizer doesn't interfere
- More difficult with MVDs

Mining for AFD's

Dissertation of Jeremy Engle

Search space similar to many other datamining problems

set of subsets of $\{A_1, \cdot \cdot \cdot A_n\}$

New global algorithmic approach: attribute at a time
each iteration adds one more attribute to exploration

order of picking attributes important

Framework for experimentation with local search tactics

Opportunities to incorporate machine learning, visualization, $\cdot \cdot \cdot$

Other Future Work

Extend to multiple relations:

adding probabilities facilitates set-based relational algebra

Consider InD constraint systems as a logical formalism

linear algebra does a lot

More work with information theory and statistics

e.g. the best hash function maximizes entropy