CMU SCS

15-826: Multimedia Databases and Data Mining

Data Mining - Statistics reminders

C. Faloutsos

CMU SCS

Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
- Data Mining

CMU SCS

Data Mining - Detailed outline

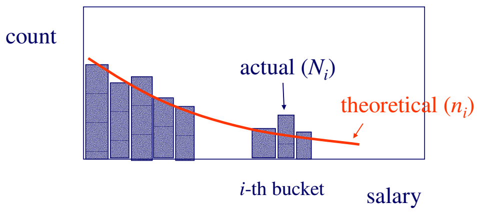
- Statistics
  - Hypothesis testing
  - Entropy
  - Contingency analysis
  - Linear correlation
  - Conclusions
- AI - decision trees
- DB

CMU SCS

Hypothesis testing (Chi-square)

Problem: is it true that the salary distribution is Zipf?

'Null Hypothesis H_0': It is Zipf



CMU SCS

Hypothesis testing (Chi-square)

Approach: very intuitive: Accept hypothesis if the theoretical values are 'close enough' to the actual ones

Formally:

Step 1: bucketize (how many? how wide?)

Step 2: Compute deviation of theoretical from actual

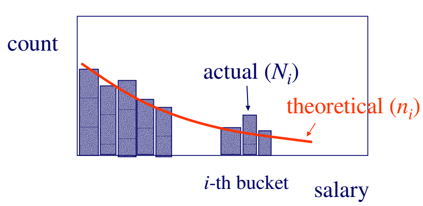
Step 3: Accept H_0 if the deviation is 'small'

CMU SCS

Hypothesis testing (Chi-square)

Step 2: Compute deviation of theoretical from actual

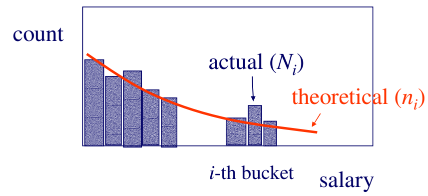
How?



CMU SCS

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$



15-826 Copyright: C. Faloutsos (2007) 7

CMU SCS

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

B: Number of buckets (~ 'degrees of freedom')

Step 3: Accept H_0 if the deviation is 'small'
Q: How small is 'small'?

15-826 Copyright: C. Faloutsos (2007) 8

CMU SCS

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

A:
- find the PDF of the χ^2 variable
- decide on a confidence level (say, 95%)
- figure out the χ^2 that is exceeded with 5% probability

Pictorially:

15-826 Copyright: C. Faloutsos (2007) 9

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

PDF

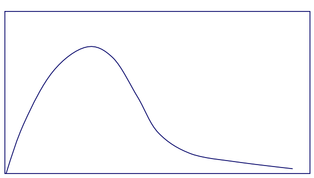
Assumption: sum of B squared Gaussians

0 χ^2 value

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

PDF

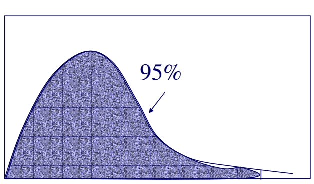


0 χ^2 value

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

PDF



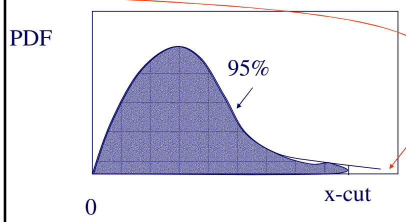
x-cut

0 χ^2 value

CMU SCS

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

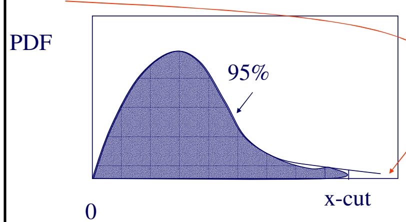


Accept H_0?

CMU SCS

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

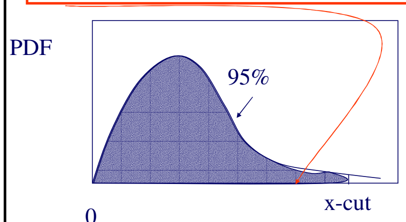


Accept H_0?
NO! (Why?)

CMU SCS

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$

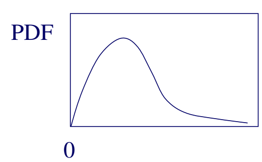


Accept H_0?

CMU SCS

Hypothesis testing (Chi-square)

$\chi^2 = \text{sum}_i \; \{ (N_i - n_i)^2 / n_i \} \quad (i = 1 \ldots B)$



Drill: if this is the PDF with B degrees of freedom, how is the PDF for B+1?

CMU SCS

Hypothesis testing (Chi-square)

Final touch:

(Degrees of freedom) =

(number of buckets)

-

(number of parameters we computed from the data)

CMU SCS

Data Mining - Detailed outline

- Statistics
  - Hypothesis testing
  - Entropy
  - Contingency analysis
  - Linear correlation
  - Conclusions
- AI - decision trees
- DB

CMU SCS

Entropy

- (useful in compression; classification; ...)
- Informally: Entropy of a (categorical) variable *X*
  - ~ min. number of yes/no questions to discover it
  - ~ min. # bits to encode it

Eg#1: fair coin { $p_H = 1/2 = p_T$ }

CMU SCS

Entropy

Eg#1: fair coin { $p_H = 1/2 = p_T$ }
entropy: 1 bit

Eg#2: fair tetrahedral dice
({p1 = p2 = p3 = p4 = 1/4}
entropy =2 (why?)

CMU SCS

Entropy

Formula for entropy H()

$$\boxed{H(\{ p_1, p_2, \ldots , p_N \}) =_{def} - \Sigma^N_{i=1} (p_i \log p_i)}$$

log: typically, base 2 (to give 'bits')

Sanity checks:

- fair coin: H(1/2, 1/2) = ?
- tetrahedral dice: H(1/4, ... 1/4) = ?

CMU SCS

Entropy

$\boxed{H(\{ p_1, p_2, \ldots, p_N \}) = - \Sigma^N_{i=1} (p_i \log p_i)}$

Fact1: for N equi-probable outcomes

$H(1/N, \ldots 1/N) = \log N$

Fact2: that's the **maximum**, for a r.v. with N outcomes:

$H(p_1, \ldots p_N) <= \log N$

15-826 Copyright: C. Faloutsos (2007) 22

CMU SCS

Entropy

Conditional entropy: H(X / Y)

Intuitively: min# of questions to recover X, when somebody tells us the value of Y

Eg: R: fair, red dice (1/6, 1/6 ... 1/6)

G: fair, green dice

S: sum of outcomes

15-826 Copyright: C. Faloutsos (2007) 23

CMU SCS

Entropy

Q1: H(R) = ?

Q2: H(R/G) = ?

Q3: H(R/S) = ?

15-826 Copyright: C. Faloutsos (2007) 24

CMU SCS

Entropy

Q1: $H(R) = \log_2 6$

Q2: $H(R/G) = \log_2 6$

Q3: $H(R/S) < \log_2 6$

(Fact3:

$$H(X/Y) <= H(X)$$

)

CMU SCS

Entropy

Formula for H(X/Y)

$$H(X/Y) = -\Sigma_i \Sigma_j [\, p_{ij} \log (p_{ij} / p_{*,j}) \,]$$

where p_{ij}, $p_{*,j}$ are defined as follows:

CMU SCS

Entropy

N outcomes for Y

M outcomes for X

				marginals
	$p_{1,1}$	…	$p_{1,N}$	$p_{1,*}$
M outcomes for X	…		…	…
	$p_{M,1}$	…	$p_{M,N}$	$p_{M,*}$
marginals	$p_{*,1}$	…	$p_{*,N}$	

CMU SCS

Entropy

- Proof of the above formula?

CMU SCS

Entropy

- Symmetrically, for H(Y/X)
- 'Joint information' I(X,Y)

$$I(X,Y) =_{def} H(X) - H(X/Y)$$

Intuitively: I(X,Y) = the bits of info that X and Y have in common.

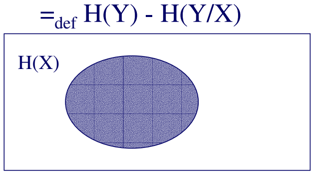
CMU SCS

Entropy

- also holds:

$$I(X,Y) =_{def} H(X) - H(X/Y)$$
$$=_{def} H(Y) - H(Y/X)$$

Pictorially:

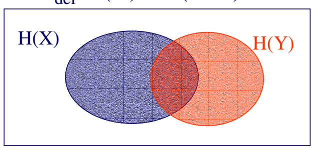


Entropy

- also holds:

$$I(X,Y) =_{def} H(X) - H(X/Y)$$
$$=_{def} H(Y) - H(Y/X)$$

Pictorially:

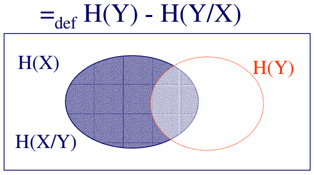


Entropy

- also holds:

$$I(X,Y) =_{def} H(X) - H(X/Y)$$
$$=_{def} H(Y) - H(Y/X)$$

Pictorially:

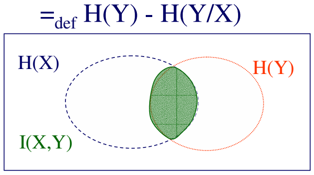


Entropy

- also holds:

$$I(X,Y) =_{def} H(X) - H(X/Y)$$
$$=_{def} H(Y) - H(Y/X)$$

Pictorially:



Entropy

- Fact4:

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

- Fact5: if X,Y are independent then:

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

Entropy

- Entropy H(X): impossible to compress outcomes of X with less bits than that (unless...)

(and compression = data mining!)

Entropy: useful for contingency analysis - 'is attribute X independent of Y'?

Data Mining - Detailed outline

- Statistics
  - Hypothesis testing
  - Entropy
  - Contingency analysis
  - Linear correlation
  - Conclusions
- AI - decision trees
- DB

CMU SCS

Contingency analysis

- Problem: we have insects, with attributes X='gender' and Y='color'.
- Q: is color and gender correlated?

(We are given a population of insects, as follows:)

CMU SCS

Contingency analysis

N outcomes for Y (color)

M outcomes for X (gender)

	red	green	…	marginals
m	$C_{1,1}$	…	$C_{1,N}$	$C_{1,*}$
f	…		…	…
..	$C_{M,1}$	…	$C_{M,N}$	$C_{M,*}$
marginals	$C_{*,1}$	…	$C_{*,N}$	

CMU SCS

Contingency analysis

- how to test for independence?

Contingency analysis

- how to test for independence?
- A: two tests:
  - T1: statistical significance with Chi-square (χ^2)
  - T2: strength, with 'joint information'
- Q1: how to set up χ^2
- Q2: how to compute strength

Contingency analysis

- Q1: Chi-square - assume independent. Then:
- theoretical $C'_{i,j}$ = ??
- degrees of freedom = ??

Contingency analysis

- Q1: Chi-square - assume independent. Then:
- theoretical $C'_{i,j} = C_{i,*} * C_{*,j} / (C_{*,*})$
- degrees of freedom = ??

Contingency analysis

- Q1: Chi-square - assume independent. Then:
- theoretical $C'_{i,j} = C_{i,*} * C_{*,j} / (C_{*,*})$
- degrees of freedom = M*N - M - N + 1

Contingency analysis

- Q2: Strength of dependency?
- why not I(X,Y)?

Contingency analysis

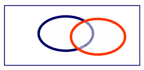
- Q2: Strength of dependency?
- why not I(X,Y)?
- why not I(X,Y) / [H(X) + H(Y)]

CMU SCS

Contingency analysis

- Q2: Strength of dependency?
- why not I(X,Y)?
- why not I(X,Y) / [H(X) + H(Y)]
- final answer

$$\boxed{2^* I(X,Y) / [H(X) + H(Y)]}$$



CMU SCS

Data Mining - Detailed outline

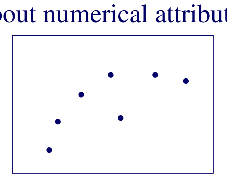
- Statistics
  - Hypothesis testing
  - Entropy
  - Contingency analysis
  - Linear correlation
  - Conclusions
- AI - decision trees
- DB

CMU SCS

Linear correlation

- just saw how to detect correlations between categorical attributes (‘gender’/ ‘color’)
- How about numerical attributes? Eg.:

salary



age

Linear correlation

- Answer: Pearson's r coefficient (== correlation coefficient)

Definition:

$$\boxed{r = \Sigma_i \{(x_i - x_{avg})\,(y_i - y_{avg})\} \,/\, (\sigma_x\, \sigma_y)}$$

where σ_x, σ_y the standard deviation of x, y

- Observation: r = cosine similarity of normalized vectors $(x1, \ldots xN)$, $(y1, \ldots yN)$

Linear correlation

$$\boxed{r = \Sigma_i \{(x_i - x_{avg})\,(y_i - y_{avg})\} \,/\, (\sigma_x\, \sigma_y)}$$

- Q: what is the max value of r?
- Q: when does this happen?
- Q: what is the min. value of r?

Linear correlation

$$\boxed{r = \Sigma_i \{(x_i - x_{avg})\,(y_i - y_{avg})\} \,/\, (\sigma_x\, \sigma_y)}$$

- Q: what is the max value of r? +1
- Q: when does this happen? y = a x+b
- Q: what is the min. value of r? -1

CMU SCS

Linear correlation

$r = \Sigma_i \{(x_i - x_{avg}) (y_i - y_{avg})\} / (\sigma_x \sigma_y)$

- Q: what is the value of *r* when x, y are independent?

CMU SCS

Linear correlation

$r = \Sigma_i \{(x_i - x_{avg}) (y_i - y_{avg})\} / (\sigma_x \sigma_y)$

- Q: what is the value of *r* when x, y are independent?
- A: 0

CMU SCS

Linear correlation

$r = \Sigma_i \{(x_i - x_{avg}) (y_i - y_{avg})\} / (\sigma_x \sigma_y)$

- NOTICE: *r* is a good measure of strength, if the correlation has been checked to be statistically significant (how?)

CMU SCS

Linear correlation

$$\boxed{r = \Sigma_i \{(x_i - x_{avg}) (y_i - y_{avg})\} / (\sigma_x \sigma_y)}$$

- NOTICE: *r* is a good measure of strength, if the correlation has been checked to be statistically significant (how?)
  - A: Chi-square

15-826 Copyright: C. Faloutsos (2007) 55

CMU SCS

Conclusions

- Chi-square test for **stat. significance**
- for **strength** of correlation:
  - entropy (for categorical attributes)
  - correlation coefficient, for numerical attributes.

15-826 Copyright: C. Faloutsos (2007) 56

CMU SCS

References

- Numerical Recipes in C.

15-826 Copyright: C. Faloutsos (2007) 57