Topics in Probabilistic and Statistical Databases

Lecture 10: Sampling and Review

Dan Suciu
University of Washington

References

- Towards Estimation Error Guarantees for Distinct Values, Charikar, Chaudhuri, Motwani, Narasayya, PODS 2000
- Sampling-Based Estimation of the Number of Distinct Values of an Attribute, Haas, Naughton, Seshadri, Stokes, VLDB 1995

Distinct Values

- Problem definition:
- Data set with n tuples
- Column of interest has values{1,…,D}
- Let n_i = number of times value i occurs
- $n = \Sigma_{i=1,D}\ n_i$
- Goal: estimate D, denote the estimate Ď
- Error is Ď/D, or D/Ď, whichever is > 1

Negative Result

Theorem [Charikar'00] Consider any (possibly adaptive and randomized) estimator Ď for the number of distinct values D that examines at most r rows in a table with n rows. Then, for any $\gamma > \exp(-r)$, there exist a choice of the input data such that with probability at least γ:

$$\text{error}(\check{D}) \geq \text{sqrt}((n-r)/2r * \ln(1/\gamma))$$

Proof in class

Estimators

- Goodman's unbiased estimator
- Many specialized estimators from the statistics literature (won't discuss; see [Haas'95])
- GEE [Charikar'95]; will discuss because it matches the lower bound

Notations

- Select random sample of size r
- d=number of distinct values in the sample
- f_i=number of distinct values that occur exactly i times

- Thus: $d = \Sigma_{i=1,r} f_i$ $\quad r = \Sigma_{i=1,r} i*f_i$

Goodman's Unbiased Estimator

Goodman proved in 1949 that:

- If $r \geq \max(n_1, \ldots, n_D)$ then there exists only one unbiased estimator:

$$\widehat{D}_{\text{Good}} = d + \sum_{i=1}^{n} (-1)^{i+1} \frac{(N - r + i - 1)!\,(r - i)!}{(N - r - 1)!\,r!} f_i$$

- If $r < \max(n_1, \ldots, n_D)$ then there exists no unbiased estimator

Very unstable, with errors of 20,000%

The GEE Estimator

Definition The GEE is: $\check{D} = \text{sqrt}(n/r)\, f_1 + \Sigma_{i=2,r}\, f_i$

Theorem. Expected ratio error is $O(\text{sqrt}(n/r))$

Review of this Course

Three areas in Probabilistic and Statistical Databases

- Explicit probabilities
- Implicit probabilities
- Statistics

Explicit Probabilistic Data

- "Classical" probabilistic databases
- Each tuple has a probability value
  - "maybe-tuple"
  - "x-tuple"
- Possible worlds semantics

Explicit Probabilistic Data

- What are some key applications ?
- What is lineage and why is it important ?

Explicit Probabilistic Data

- Rule of thumb 1:
  – ProbDB = IncompleteDB + Probabilities
- Rule of thumb 2:
  – ProbDB = Disjoint/IndependentDB + Joins
- Rule of thumb 3:
  – GM Factorization = DB-normalization + prob-identities

Explicit Probabilistic Data

Query Evaluation is #P hard in general:

- General methods: Monte Carlo, OBDDs, …
- Safe queries and safe plans
- Top k query answering

Explicit Probabilistic Data

- Major Open Research Problems
  [IN CLASS]

Implicit Probabilistic Data

- All tuples have the same probability
- What are the major differences from explicit probabilistic data ?

Implicit Probabilistic Data

- Dense random graphs
  - Pr(t) = ½
- Fagin's 0/1 law for FO
  - For every sentence φ, lim Pr(φ) = 0 or =1
- "Theory of almost certain sentences" = ?
- "**THE** random graph" = ?

Implicit Probabilistic Data

- Material random graphs:
  - $Pr(t) = \beta / n^{arity(R)}$

- Every conjunctive query has an explicit asymptotic formula:
  - $Pr(q) = C(q) / n^{exp(q)} + O(n^{exp(q)+1})$

Implicit Probabilistic Data

- General Random Graphs: G(n,p)

  [WHAT IS THAT ?]

- Erdos and Renyi's theorem

- Random graphs $G(n, \beta/n^{\alpha})$:
  - Threshold values for α (no 0/1 laws): 2, 1+1/2, 1+1/3, …, 1+1/k, … 1, [rationals], 0
  - Everywhere else: 0/1 Law for FO

Implicit Probabilistic Data

- The major applications today:
  - ?

- … but great theory !

Implicit Probabilistic Data

- Research topics: [IN CLASS]

Data Statistics

- What is their main usage in database systems ?

Data Statistics

- Histograms
  - Eqwidth, eqdepth, V-optimal
- Sampling
  - Sequential sampling techniques
  - Join synopses

Data Statistics

- Limitations of how data statistics are used today: [IN CLASS]

- Major research topics in data statistics: [IN CLASS]

Final Thoughts

- Computer Science in the past:
  - Driven by better algorithms
- Computer Science today:
  - Driven by massive amounts of data
  - Processed with approximate methods
  - Data itself is often imprecise
- Computer Science tomorrow:
  - Probabilistic databases ☺