

Estimation of the Click Volume by Large Scale Regression Analysis

Yury Lifshits^{1,2} and Dirk Nowotka³

¹Steklov Institute of Mathematics at St.Petersburg,

²California Institute of Technology,

³University of Stuttgart



General area:

Fast web-scale algorithms for machine learning

Least Squares: Find α minimizing $\|\alpha M - y\|_2$

Assume M has $k \ll mn$ nonzero elements

Can we solve least squares in $\mathcal{O}(k)$ time?

Our contribution:

- 1 Solving least squares on sparse matrices
- 2 Application to on-line advertisement

Part I

Fast Algorithm for Solving Least Squares on Sparse Matrices

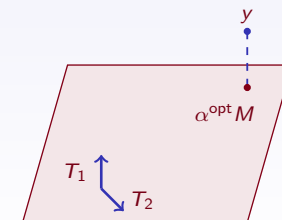
Geometric View on Least Squares

Find α minimizing $\|\alpha M - y\|_2$

Let V be a linear hull of M 's rows T_1, \dots, T_m

Interpretation: every vector α represents coordinates of αM point in V in T_1, \dots, T_m basis

Thus $\alpha^{\text{opt}} M$ is just the projection of y to V



Complexity of One Global Round

Global Round:

Sequentially do one update for every j from 1 to m

Theorem

Componentwise Iterations algorithm uses only $\mathcal{O}(k)$ time for global round of updates. Recall, k is the number of nonzero entries in M .

Algorithm: Discussion

- A vector $\alpha^{(k)}$ can be safely updated in two components j_1 and j_2 in parallel if we have $T_{j_1} \perp T_{j_2}$
- In the case of orthogonal vectors T_1, \dots, T_m one global round is sufficient for reaching α^{opt}
- Joint optimal update of k components requires inverting $k \times k$ matrix

Part II

Application to On-line Advertisements

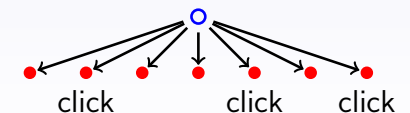
On-line Ads: Terminology

Opportunity: all available information about ad request to advertising system. E.g. website, search query, user id, ...

Click-through rate $CTR(a, o)$: a probability of clicking ad a at opportunity o

Click volume for collection O of all opportunities for some time interval: $CV(a) = \sum_{o \in O} CTR(a, o)$

The same ad
to all ad requests



Importance of Click Volume

- Measuring **effectiveness of advertising system** by a ration between sold clicks and click volume
- Measuring **preciseness of targeting** by ration between click volume in target group and overall click volume
- Defining **coverage/agressiveness tradeoff**: How many users should we contact in order to get a given fraction of click volume?
- Helpful for **setting prices** for on-line ads

Estimation of Click Volume: Methodology

- **Indexing**: mapping advertising system logs to well-defined data structure
- **Regression Analysis**: deriving formula for click-through rate prediction
- **Estimation**: applying resulting formula to a given ad and monthly collection of opportunities

Indexing: History Table

Logs of advertising system:

Set of triples (a_i, o_i, b_i) : ad, opportunity, click

Events indexing:

Mapping a pair (a, o) to internal representation $E(a, o)$

Techniques: clustering, dimensionality reduction

Resulting reduced table:

Set of pairs (E_i, CTR_i)

Quick Recall: Regression Analysis

Input:

Training collection of n documents

Document i : m -dimensional vector D_i

and additional parameter y_i

Regression Problem:

Find function f such that $f(D_i) \approx y_i$

Linear Regression Problem (least squares):

find m -dimensional vector α such that the sum of squared prediction errors $\sum |\alpha D_i - y_i|^2$ is minimized

Click Volume via Regression

- 1 Building history table from logs of advertising system: matrix $M = \{E_1, \dots, E_n\}$ and CTR vector
- 2 Solving least squares
- 3 Computing click volume by formula:

$$CV(a_{\text{new}}) = \sum_{1 \leq i \leq n} \alpha \cdot E(a_{\text{new}}, o_i)$$

Directions for Further Work

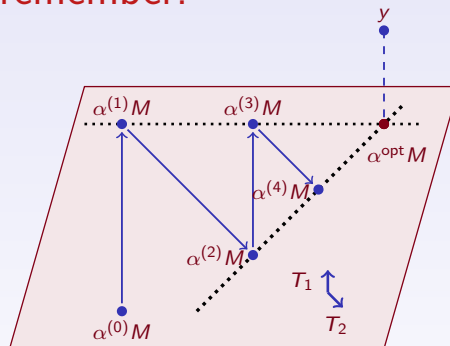
Theoretic problems:

- Prove upper bounds for **speed of convergence** of our algorithm
- Can we compute **click volume for all ads** in the system faster than doing it separately for every ad?

Experimental problems:

- Apply our algorithm for some real data set. Measure the **empirical precision** of CTR prediction
- Study effects of **heuristic ingredients** for algorithm: indexing, dimensionality reduction, update order

A picture to remember:



Thanks for your attention! Questions?

Yury Lifshits

<http://yury.name>

Dirk Nowotka

[google://dirk nowotka](google://dirk%20nowotka)

Some related work:



Y. Lifshits, D. Nowotka

Estimation of the Click Volume by Large Scale Regression Analysis. CSR'07.

<http://yury.name/papers/lifshits2007click.pdf>



B. Hoffmann, Y. Lifshits, D. Nowotka

Maximal Intersection Queries in Randomized Graph Models. CSR'07.

<http://yury.name/papers/hoffmann2007maximal.pdf>



N. Goyal, Y. Lifshits, H. Schütze

Disorder Inequality: A Combinatorial Approach to Nearest Neighbor Search. Submitted.

<http://yury.name/papers/goyal2008disorder.pdf>