PROCESSING SOCIAL DATA

Devavrat Shah
Laboratory for Information and Decision Systems
Department of EECS
Massachusetts Institute of Technology
SOCIAL DATA

- Data generated by *us*
  - Purchases or transactions
  - Movie or restaurant ratings or reviews
  - Tweets
  - Mechanical Turk
  - Facebook posts, blogs, web-crawl history, ....
OPPORTUNITY

- Tremendous amount of social data
  - Can be useful for all sorts of decision making

- For example,
  - Better business practices (a la analytics)
    - Revenue management, pricing, advertising, ...
  - Improved social living
    - Policy making, consensus building, recommendations, ...
  - Uplifting of societies (?)
    - Arab spring, Crowd computation, CNN i report, ...
CHALLENGES

Want to make decisions using data that is

- Unstructured, noisy  ➔  Statistical challenge
- A lot of it  ➔  Computational challenge
Challenges: Towards Resolution

- Solution = Statistical Inference + Message-Passing
  - Find meaningful statistical model
    - Relating “decision variables” with “data”
  - Identify *optimal* inference algorithms
    - To estimate “decision variables”
  - Develop it’s approximation: Data to Decision
    - Message-passing: fast, distributed and performance preserving
We have developed this program successfully

- Crowd-sourcing
  - Reliable task using unreliable *Turks*
- Group decision making and recommendation
  - Understanding *people’s choice*
- Advertising and Searching in Twitter
  - Identifying influential agents using Tweets/ReTweets
CHALLENGES: TOWARDS RESOLUTION

- We have developed this program successfully
  - Crowd-sourcing
    - With S. Oh + D. Karger
  - Group decision making and recommendation
    - With A. Ammar + V. Farias + S. Jagabathula + S. Negahban + S. Oh
  - Advertising and Searching in Twitter
    - With T. Zaman
We have developed this program successfully

- Crowd-sourcing
  - Reliable task using unreliable *Turks*
- Group decision making and recommendation
  - Understanding *people’s choice*
- Advertising and Searching in Twitter
  - Identifying influential agents using Tweets/ReTweets
CROWD SOURCING
CROWD SOURCING

CrowdFlower
livework
Data Discoveries
SmartSheet
YAHOO! ANSWERS
Aardvark
Chaordix
KICKSTARTER
IdeaScale
Distributed innovation
Crowdfunding
Crowdfunding aggregators
Innovation prizes
Content markets
intrade
IPO
Image labeling
Data entry
Transcription

$30 million to land on moon

$0.05 for
Image Labeling
Data Entry
Transcription
Micro-task Crowdsourcing

Undergrad Intern: 300 image/hr, cost: $15/hr  
Mturk (single label): 3000 image/hr, cost: $15/hr  
Mturk (mult. labels): 600 image/hr, cost: $15/hr

Reliability

95%  
65%  
95%
AN EXAMPLE
An Example
AN EXAMPLE
Goal: reliable estimate the tasks with min’l cost

Key operational questions:
- Task assignment
- Inferring the “answers”
PREVIEW OF RESULTS

Graph showing the relationship between labels per image and probability of error. The graph compares different methods:
- Majority Voting
- EM Algorithm
- Our Approach
- Lower Bound

Legend:
- 1 label/image: 35% Error
- 15 labels/image: 1% Error

Axes:
- Labels per image
- Probability of Error
PREVIEW OF RESULTS

Target Accuracy $\rightarrow$ 0.1

Probability of Error

Lower Bound

Majority Voting

EM Algorithm

Our Approach

Labels per image

Our Approach: 8
Existing Algorithm: 12
Majority Voting: 17
Task Assignment

- Random $(\ell; r)$-regular bipartite graphs
- Locally Tree-like
- Good expander
- Sharp analysis
- High Signal to Noise Ratio
Binary tasks: \( t_i \in \{+1, -1\} \)

Worker reliability: \( p_j \in [0, 1] \)

\[
A_{ij} = \begin{cases} 
  t_i & \text{with probability } p_j \\
  -t_i & \text{with probability } 1 - p_j 
\end{cases}
\]

\( A_{ij} \)'s are independent conditioned on \( t_i \) and \( p_i \)

Necessary assumption: we know \( \frac{1}{n} \sum_{j=1}^{n} p_j > 0.5 \) or not
**Inference Problem**

- **Majority:**
  \[ \hat{t}_i = \text{sign} \left( \sum_j A_{ij} \right) \]

- **Oracle:**
  \[ \hat{t}_i = \text{sign} \left( \sum_j \log \left( \frac{p_j}{1 - p_j} \right) A_{ij} \right) \]

![Graph showing the relationship between cost and probability of error]

- **Majority Voting**
- **Oracle Estimator**

Cost = Number of assignments per task

Probability of error vs. Cost
Inference Problem

- **Majority:**
  \[ \hat{t}_i = \text{sign} \left( \sum_j A_{ij} \right) \]

- **Oracle:**
  \[ \hat{t}_i = \text{sign} \left( \sum_j \log \left( \frac{p_j}{1 - p_j} \right) A_{ij} \right) \]

- **Our Approach:**
  \[ \hat{t}_i = \text{sign} \left( \sum_j W_{ij} \sum_j A_{ij} \right) \]

---

**Graph:**

- Probability of error vs. Cost (Number of assignments per task)
- Majority Voting
- Iterative Algorithm
- Oracle Estimator
Iteratively learn $W_{ij}$'s:
\[
\hat{t}_i = \text{sign}\left( \sum_j W_{ij} A_{ij} \right)
\]
- Message-passing
  - $O(\# \text{ edges})$ operations
- Approximation of
  - Maximum Likelihood

Task-likelihood update

$$T_{ij} = \sum_{j' \neq j} W_{ij'} A_{ij'}$$

A task is likely to be ‘+’ if reliable workers agree that it is ‘+’

Worker-reliability update

$$W_{ij} = \sum_{i' \neq i} T_{i'j} A_{i'j}$$

A worker is reliable if the worker agreed with our belief on other tasks
Iterative Inference

Iteratively learn $W_{ij}$’s:

$$\hat{t}_i = \text{sign} \left( \sum_j W_{ij} A_{ij} \right)$$

- Message-passing
  - $O(\# \text{edges})$ operations
- Approximation of
  - Maximum Likelihood

Task-likelihood update

Worker-reliability update

A task is likely to be ‘+’ if reliable workers agree that it is ‘+’

A worker is reliable if the worker agreed with our belief on other tasks
EXPERIMENTS: AMAZON MTURK

- Learning similarities
  - Recommendations
  - Searching, ...

You wanted to get the tie on the top, but it was not available. Which one would be a better substitute?

- [ ] the tie the left
- [x] the tie on the right
EXPERIMENTS: AMAZON MTURK

- Learning similarities
  - Recommendations
  - Searching, ...

Which tie matches my shirt better?
- [ ] the tie on the left
- [x] the tie on the right
EXPERIMENTS: AMAZON MTURK

Which colors...

- *left*
- *right*

- EM Algorithm
- Majority voting
- Iterative Algorithm
Task Assignment: Why Random Graph

Average probability of error

Number of responses per task

Graphs with small spectral gap

EM Algorithm

Iterative Algorithm
Quality of Crowd

Prob of err determined by Crowd Quality Parameter

\[ q \equiv \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2 \]

Phase transition at \( q^2(\ell - 1)(r - 1) = 1 \)
- \( q^2(\ell - 1)(r - 1) < 1 \): Iteration Hurts
- \( q^2(\ell - 1)(r - 1) > 1 \): Iteration Helps

Majority Voting
EM Algorithm
Iterative Algorithm
Oracle Estimator
**Quality of Crowd**

Prob of err determined by Crowd Quality Parameter

\[ q = \frac{1}{n} \sum_{j=1}^{n} (2p_j - 1)^2 \]

**Theorem. [Karger, Oh, Shah ’11]**

In the large system limit, and for \((\ell - 1)(r - 1)q^2 > 1\), using a random graph and \(k\) iterations of the iterative algorithm achieves

\[ P_{\text{error}} \leq \exp \left\{ -\frac{q\ell}{2\sigma_k^2} \right\}, \]

for \(\sigma_k^2 \equiv \left(3 + \frac{1}{qr}\right)\frac{q^2(\ell-1)(r-1)}{q^2(\ell-1)(r-1)-1} + \frac{2}{q}(q^2(\ell-1)(r-1))^{-k}.\)

\[ q^2(\ell - 1)(r - 1) > 1: \text{Iteration Helps} \]
How Good Is This?

Iterative algorithm ($\ell \leq r$, $q(\ell - 1) > 1$, $k \to \infty$):

$$P_{\text{error}} \leq e^{-\frac{1}{8}(q\ell - 1)}$$

Matching minimax lower bound:

$$\inf_{\text{Alg}, G(\ell)} \sup_{\{t_i\}, \{p_j\} \in \mathcal{F}(q)} P_{\text{error}} \gtrsim \frac{1}{2} e^{-(q\ell + O(q^2\ell))}$$

Majority Voting:

$$\inf_{G(\ell)} \sup_{\{t_i\}, \{p_j\} \in \mathcal{F}(q)} P_{\text{error}} \gtrsim e^{-C(q^2\ell + 1)}$$

![Graph showing error rates for different algorithms](chart.png)

- Majority Voting
- EM Algorithm
- Iterative Algorithm
- Oracle Estimator
Cost of Reliability

\[ P_{\text{error}} \leq e^{-\frac{1}{8}(q\ell-1)} \]

How much do we need to spend to achieve \( P_{\text{error}} \leq \epsilon \)?

- **Sufficient** to choose \( \ell = O\left(\frac{1}{q} \log\left(\frac{1}{\epsilon}\right)\right) \)
- **Necessary** to have \( \ell = \Omega\left(\frac{1}{q} \log\left(\frac{1}{\epsilon}\right)\right) \)

Since Oracle knows reliability of every worker

- There is no significant gain in using side-information:
  (Worker reputation, golden questions with known answers)
ADAPTIVE TASK ASSIGNMENT: DOES IT HELP?

Surprise: No gain in adaptive scheme!
Adaptive Task Assignment: Does it Help?

Theorem. [Karger, Oh, Shah ’11]

Let $\Delta$ be the budget necessary to achieve a target accuracy $P_{\text{error}} \leq \epsilon$. Then, for any $\text{Alg}$ that can adaptively assign tasks,

$$\inf_{\text{Alg} \{t_i\}, \{p_j\} \in \mathcal{F}(q)} \sup \mathbb{E}[\Delta] = \Omega\left(\frac{1}{q \log \left(\frac{1}{\epsilon}\right)}\right)$$

Surprise: No gain in adaptive scheme!
Which Crowd To Employ

Cost

\[ c_1 = \$0.04 \]

Worker Quality

\[ P_1 \]

\[ c_2 = \$0.05 \]

\[ P_2 \]
Which Crowd To Employ

Cost

$c_1 = \$0.04$

$c_2 = \$0.05$

Worker Quality

$q_1 = \mathbb{E}[(2P_1 - 1)^2]$  
$q_2 = \mathbb{E}[(2P_2 - 1)^2]$

Invest all resources on the crowd that gives best value/cost $= \frac{q_k}{c_k}$
SUMMARY

- Social Data
  - All sorts, every where, and a lot of it
  - Opportunity: processing it for better
    - Business, social living, policy, and societal uplifting
  - Challenge:
    - Statistical and computational

- Our approach:
  - Statistical inference + message-passing algorithm
  - This talk
    - Crowd-sourcing
  - Other successful examples
    - Twitter search/influential agents/advertisement
    - Recommendation and group decision making