Evaluating Crowdsourcing Participants in the Absence of Ground-Truth

Ram Subramanian
rsubramanian@coe.neu.edu

Advised by Prof. Jennifer Dy

Department of Electrical & Computer Engineering,
Northeastern University, Boston



Publications Resulted

- Subramanian, R., Rosales, R., Fung, G., and Dy, J. *"Evaluating Crowdsourcing Participants in the Absence of Ground-Truth,"* in NIPS 2012, Lake Tahoe, NV. 7 Dec 2012.



Summary Problem Definition

- Supervised/semi-supervised learning where:
  - Ground-truth exist but is not available or is very expensive
  - Multiple annotation sources (participants) are available and label some data points
- Problem:
  - How to evaluate participants
- Example use:
  - Identify/eliminate unhelpful or adversarial participants

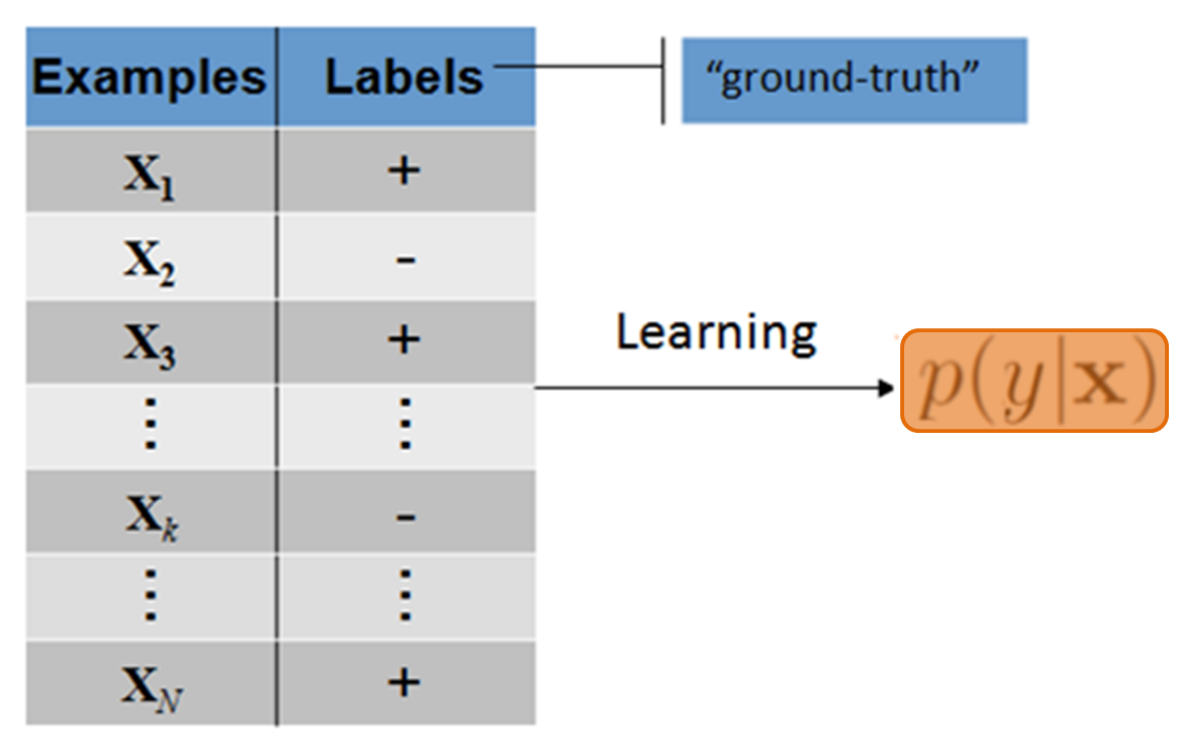


Learning with a Single Participant/Annotator

- Predict a binary event y given (related) observations about examples $\mathbf{x}$

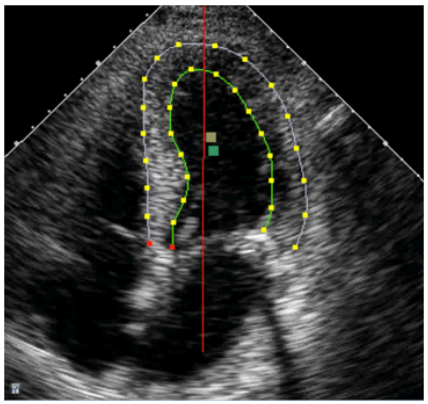
Training data

$$\left\{ \begin{array}{lrcl} \text{Examples:} & X & = & [\mathbf{x}_1^T, \mathbf{x}_2^T, ..., \mathbf{x}_N^T] \\ \text{Labels:} & Y & = & [y_1, y_2, ..., y_N]^T \end{array} \right\} \begin{array}{l} \mathbf{x}_i \in \mathbb{R}^D \\ y \in \{+, -\} \end{array}$$



Fundamental Limitations

- Example scenario: Coronary artery disease diagnosis
  - CAD can be diagnosed by measuring and scoring regional heart-wall motion in echocardiography images.



Cardiologists in agreement Heart-level (A/N)	% of cases N=240
5	48.75%
4 or more	62.09%
3 or more	70.42%

  - No objective ground-truth
  - Quality of diagnosis highly dependent on skill and training of reader. High variability in quality of interpretation

Fundamental Limitations

- Increasingly common scenario
  - Product quality from user reviews, conference paper rating from multiple reviews, quality of on-line contributions from user history, ...
- What is common in these problems?
  - Ground-truth exists but is unverifiable/subjective
    - Or it is verifiable but at a high cost
  - Multiple annotator participants exist
    - Multiple inexpensive labels may be available, but will depend on labeler idiosyncrasies.
- We could address these problems using the standard supervised/unsupervised approach
  - Simply aggregate training data
  - Majority voting (But not ideal)



Approaches to Estimating the Ground Truth

- Assign the majority vote among the participants as the true label
  - Assumes all the participants are equally skilled while in reality some participants can be more reliable than others.
- Weight the participants' opinion based their error rate in labeling the past scans.
  - Assumes participants' uniform level of skill across the entire domain of the data while in reality some participants are more reliable for certain instances of the data.
  - A better approach, but again, suboptimal.
- Model the expertise of participant as a function of the data instance observed.
  - Our Approach. Models the probability that the participant labels the scan correctly.



Try to Learn by Discovering the Expertise in Everyone

- Basic idea:
  - Face the unavailability of ground-truth
  - Learn different aspects from different participants/annotators
  - Employ labeler identity $\rightarrow$ learn to discriminate
    - Create participant/annotator-specific models

Examples	L1	L2	L3	L4	L5	Ground-truth
X_1	+	+	+	-	+	X
X_2	-	+	-	-	+	X
X_3	-	+	+	-	-	X
X_k	-	+	-	-	-	X
X_N	+	+	+	-	+	X

Challenges when Modeling Multiple Participants

- Some participants/annotators may:
  - be more reliable than others
  - be more reliable for certain types of data
  - be malicious
  - be correlated with others
  - **have varying accuracy depending on the data instance observed**
- Questions (e.g.; medical setting)
  - How should the patients be diagnosed when radiologists disagree?
  - How do we evaluate the radiologists' diagnoses?
  - How do we evaluate radiologists?



Modeling Multiple Participants

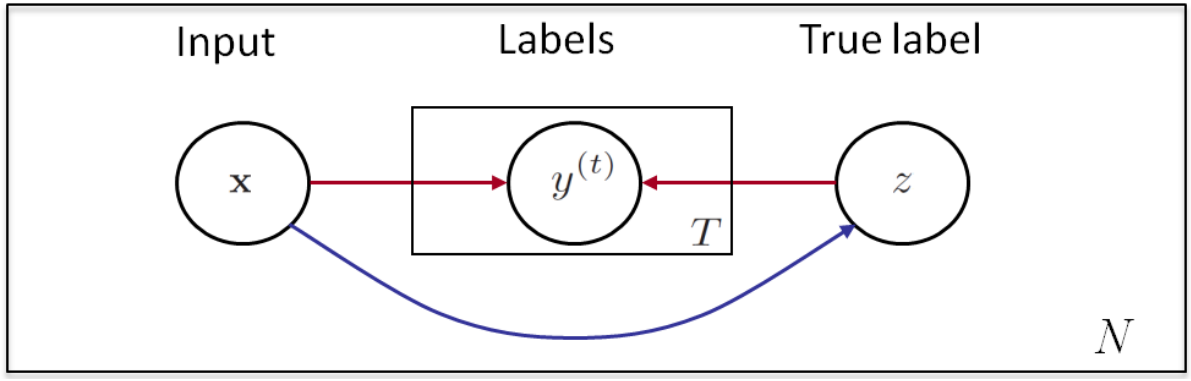
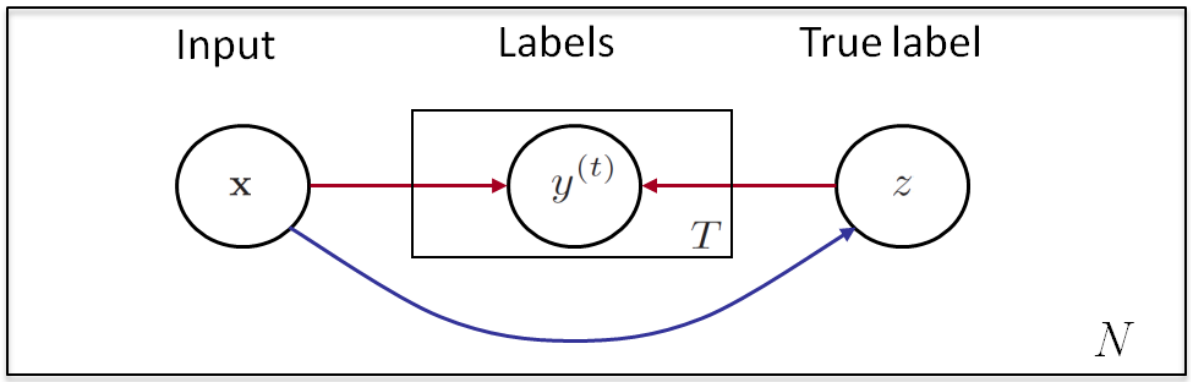


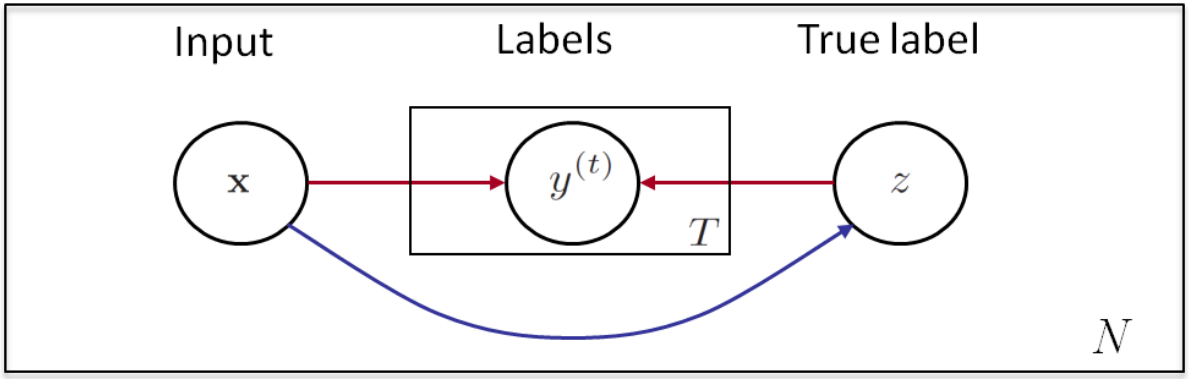
Figure: Graphical Model for **x**, **y**, and z respectively inputs, annotator-specific labels, and ground truth label (for simplicity α, β, $\{\gamma_t\}$, and $\{\mathbf{w}_t\}$, with $t \in \{1, .., T\}$, are excluded)

Modeling Multiple Participants



$$p(Y, Z|X) = \prod_i p(z_i|\mathbf{x}_i) \prod_i p(y_i^{(t)}|\mathbf{x}_i, z_i)$$

Modeling Multiple Participants

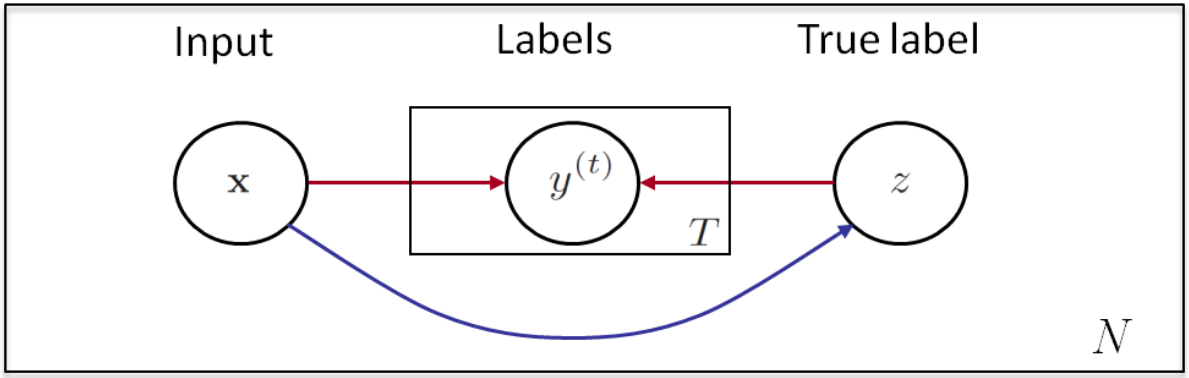


$$p(Y, Z|X) = \prod_i \boxed{p(z_i|\mathbf{x}_i)} \prod_i p(y_i^{(t)}|\mathbf{x}_i, z_i)$$

Classification Model: Logistic regression

$$p(z_i = 1|\mathbf{x}_i) = (1 + exp(-\alpha^T \mathbf{x}_i - \beta))^{-1}$$

Modeling Multiple Participants



$$p(Y, Z|X) = \prod_i p(z_i|\mathbf{x}_i) \prod_i \boxed{p(y_i^{(t)}|\mathbf{x}_i, z_i)}$$

Multi-Labeler Model: Bernoulli || Gaussian

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = (1 - \eta_t(\mathbf{x}_i))^{|y_i^{(t)} - z_i|} \; \eta_t(\mathbf{x}_i)^{1-|y_i^{(t)} - z_i|}$$

$\eta_t(\mathbf{x})$: Probability of labeler t to be correct (given input)

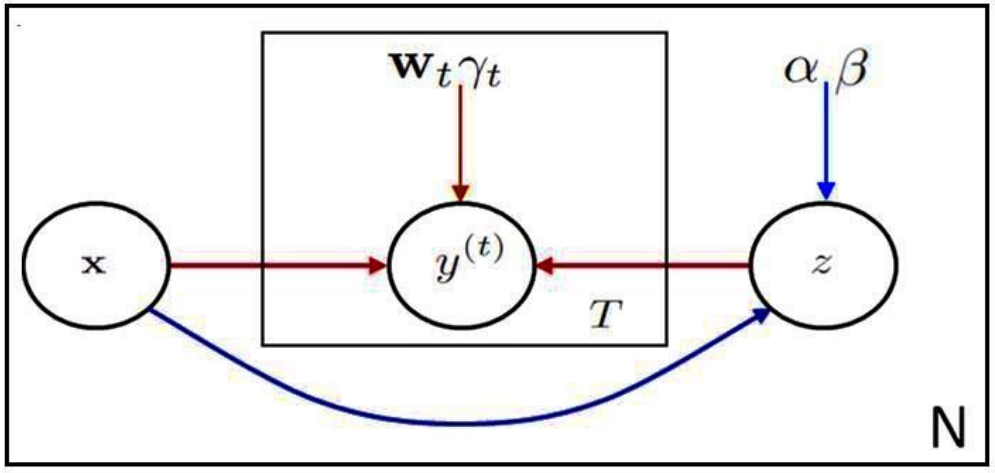
$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = N(y_i^{(t)}; z_i, \sigma_t(\mathbf{x}_i))$$

$\sigma_t(\mathbf{x})$: How much labeler t deviates from the true label z (given input)

Modeling Multiple Participants

- Bernoulli/Gaussian model parameters, as a function of the input

$$\eta_t(\mathbf{x}) \quad = \quad (1 + exp(-\mathbf{w}_t^T \mathbf{x} - \gamma_t))^{-1}$$



- Participant/annotator specific distributions

	$p(y_i^{(t)} \mid \mathbf{x}_i, A)$	$p(y_i^{(t)} \mid \mathbf{x}_i, z_i)$
L1:	{N,A,A,N,A,N,A,A,N,A,A}	{N,N,N,N,N,N,A,N,N,A,N}
L2:	{A,A,A,N,A,A,A,A,N,A,A}	{N,N,A,N,N,N,N,N,N,N,N}
L3:	{A,N,A,N,A,N,A,N,A,N,A}	{A,N,A,N,A,N,A,N,A,N,A}

Formally Evaluating Participants

- Known ground-truth (GT):
  - Relatively simple solution (can compare against GT)
- Unknown ground-truth
  - No ground-truth but we could rely on label redundancy
  - Improve performance by annotator expertise modeling
  - Underlying idea (for a single point): evaluate participant/annotator k by determining how easy/difficult it is to predict its label from the labels given by other annotators

$$p(y^{(k)}|\{y^{(t\setminus k)}\},\mathbf{x})$$



Evaluating Participants - Formulation Details

- Partially annotated data distribution given by model

$$p(Y_O|X) = \prod_i \sum_{z_i} p(z_i|\mathbf{x}_i) \prod_{t|t\epsilon\mathcal{T}_i} p(y_i^{(t)}|\mathbf{x}_i, z_i)$$

- Using model independence assumptions

$$p(y^{(k)}|\{y^{(t\backslash k)}\}, \mathbf{x}) = \frac{p(\{y^{(t)}\}|\mathbf{x})}{p(\{y^{(t\backslash k)}\}|\mathbf{x})}$$

all annot. GT estimate

$$= \frac{\sum_z p(\{y^{(t)}\}|z, \mathbf{x}) p(z|\mathbf{x})}{\sum_z p(\{y^{(t\backslash k)}\}|z, \mathbf{x}) p(z|\mathbf{x})}$$

w/o k-th annot.

- In the limit of known ground-truth we have that annotators are mutually independent:

$$p(y^{(k)}|\{y^{(t\backslash k)}\}, \mathbf{x}) = p(y^{(k)}|\mathbf{x})$$



Experimental Setup

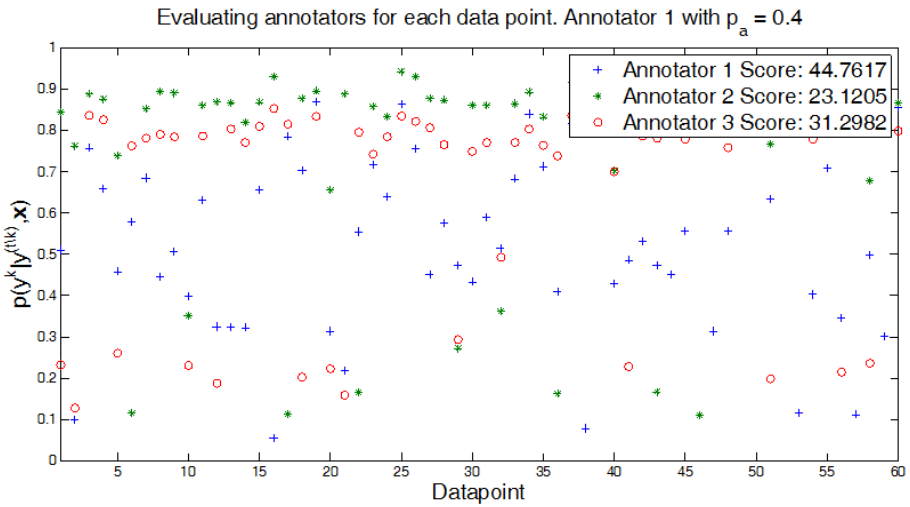
- Train multi-annotator model on 70% of the data
  - Employs real annotators and (when applicable) simulated adversaries
- Test on remaining 30% of the data
- Results show averages and deviations on 10 repeated experiments
- Annotator score (sum): log probability of annotator k label assignments given remaining annotator label assignments

$$S_k = -\sum_i \log p(y_i^{(k)} | y_i^{(t \setminus k)}, \mathbf{x}_i)$$

- Helpful/adversarial participants added by generating a random label flip with varying probability p_a



Conditional Label Probability Values - Breast Cancer



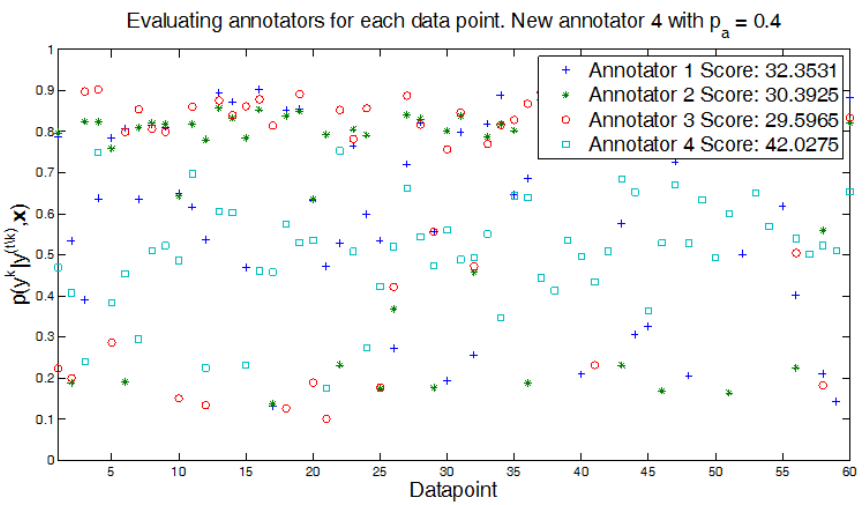


Figure: Breast cancer diagnosis from X-ray images [N=75, d=8, A=3]

Experiments [score/pa/A] - Multiple Participant Data

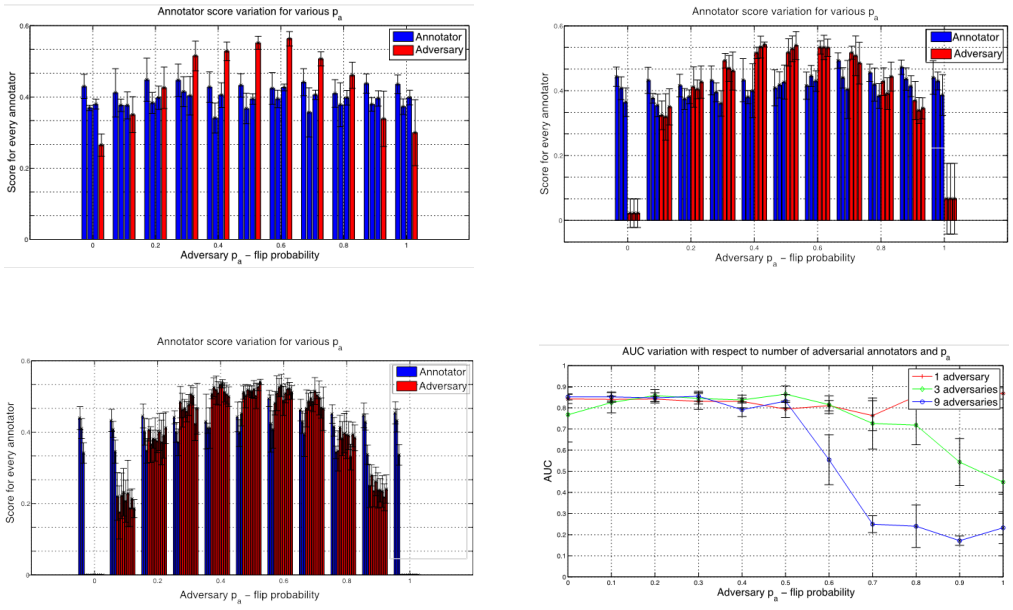


Figure: Breast cancer diagnosis from X-ray images [N=75, d=8, A=3]

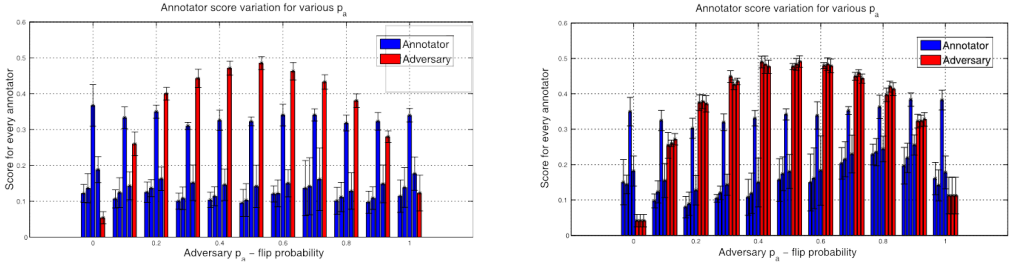


Figure: Atrial Fibrillation Dataset [N=1081, d=231, A=4]

Experiments [score/pa/A]. UCI Data

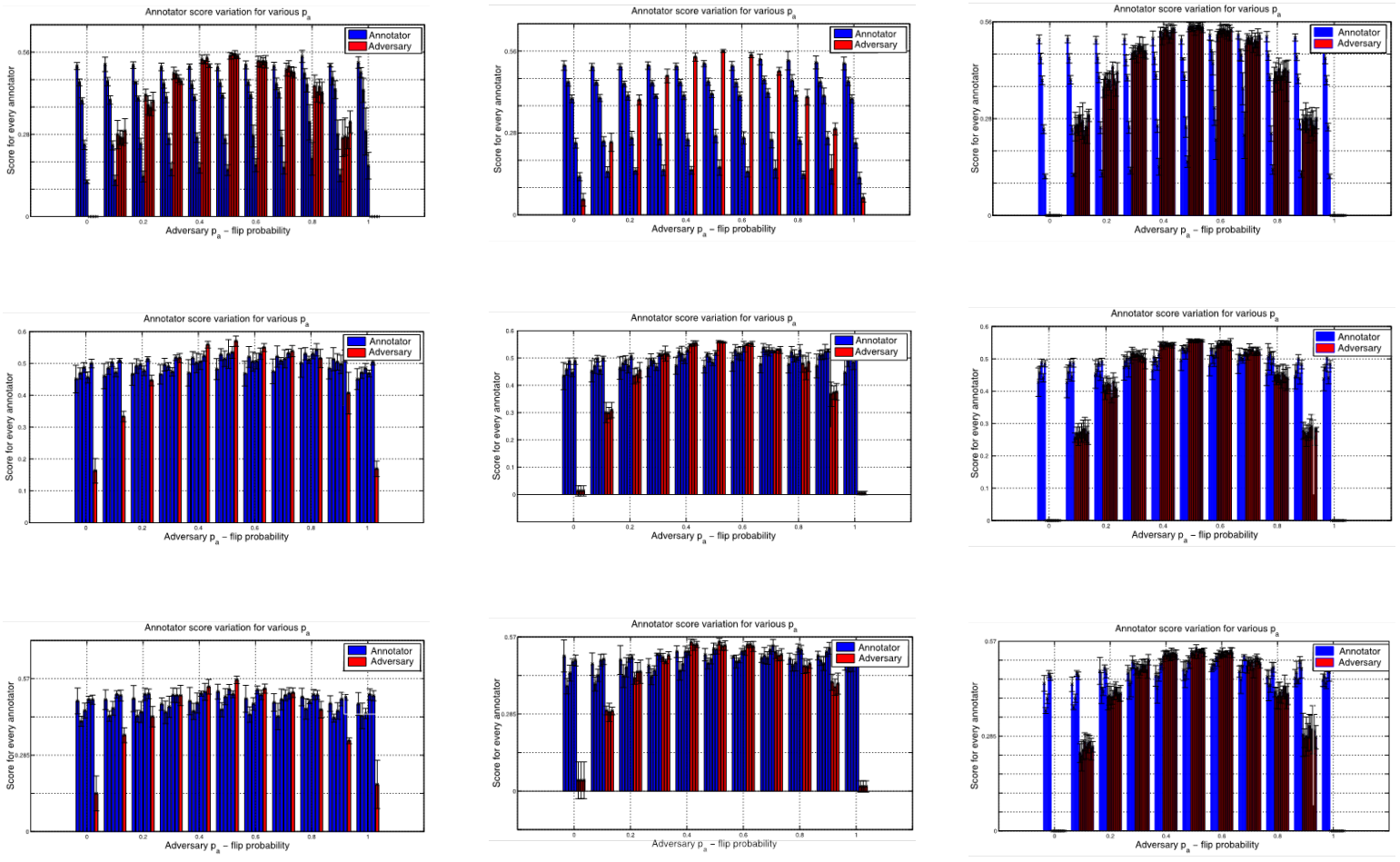
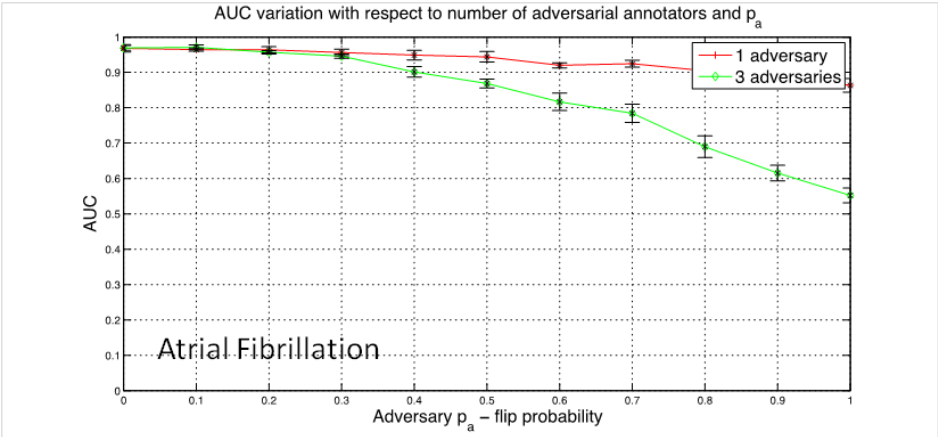
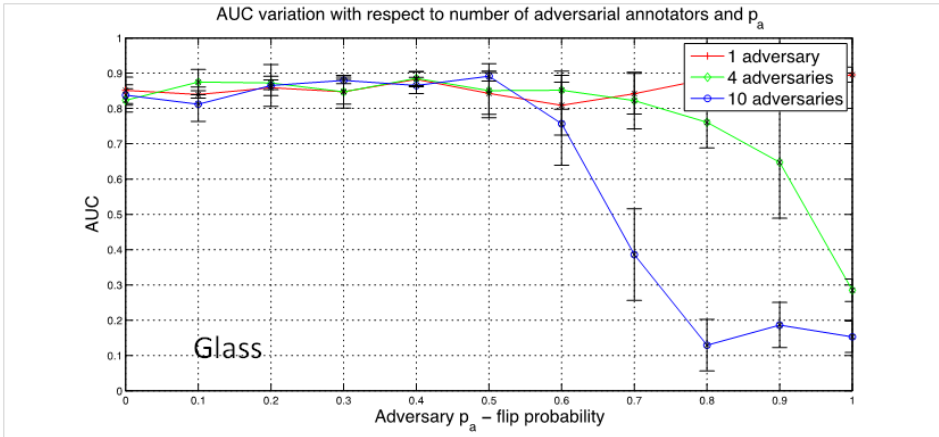
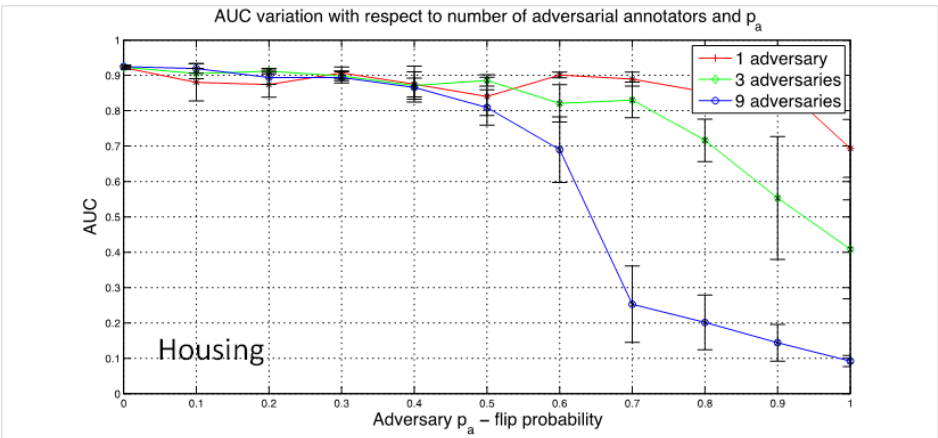


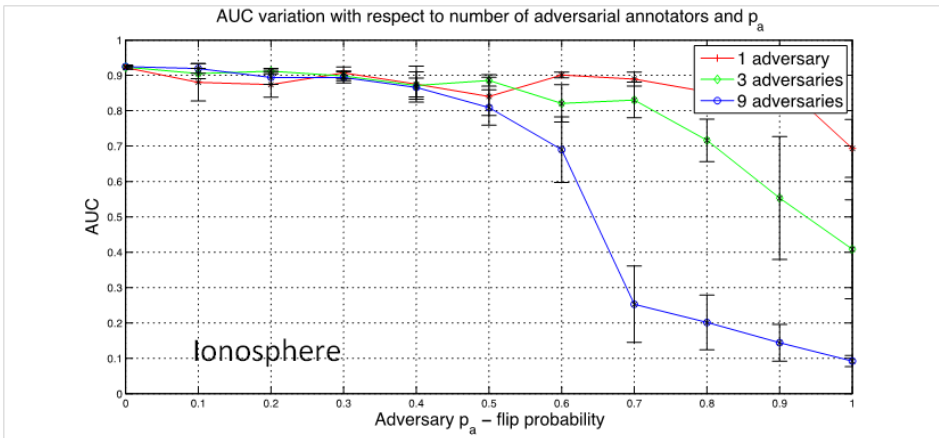
Figure: Row 1: UCI Glass Dataset [N=214, d=9, A=sim],
Row 2: UCI Housing Dataset [N=506,d=13, A=sim],
Row 3: UCI Ionosphere Dataset [N=351,d=34, A=sim]

Experiments [AUC/pa/A]. Multiple Datasets









Conclusion

- Proposed a novel approach to identify unhelpful or adversarial participants in a crowdsourcing setting where multiple labelers are available.
- Perfected a robust scoring scheme to evaluate the participants.
- Demonstrated experimentally the strength of our approach in response to number of adversaries, their degree of adversarial behavior, and their effect on the classification accuracy.



References

[1] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy., *"Learning from crowds,"* Journal of Machine Learning, 11:1297-1322, 2010.

[2] Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L., and Dy, J. *"Modeling annotator expertise: Learning when everybody knows a bit of something,"* in International Conf. on Artificial Intelligence and Statistics, 2010.

[3] Ofer Dekel, Ohad Shamir, *"Good Learners for Evil Teachers,"* in 26th International Conference on Machine Learning, 2009.

[4] Mark Harris, *"How a Lone Hacker Shredded the Myth of Crowdsourcing,"* in Backchannel, Feb 9, 2015.

[5] Harper, *"Crowdsourcing isn't broken,"* Backchannel, Feb. 2015.

[6] Tim Gowers, Michael Nielsen, *"Massively collaborative mathematics,"* in Nature, October 15, 2009.

[7] Short Survey on Multi-Teacher Learning shared by O.Dekel.


When radiologists take a selfie
memes.com