Machine Learning Algorithms for Biomedical Data

Learning from the Crowd

Jennifer G. Dy

Dept. of Electrical and Computer Engineering

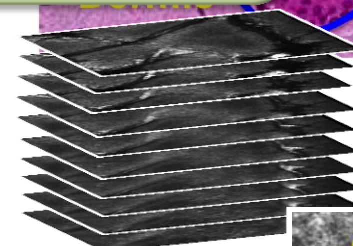
Northeastern University

Machine Learning & Data Mining

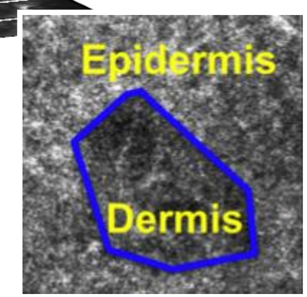
- Clustering, Unsupervised Learning
- Dimensionality Reduction, Feature Selection, Sparse Models/Methods

Crowdsourcing,
Learning from Multiple Annotators

Current Projects

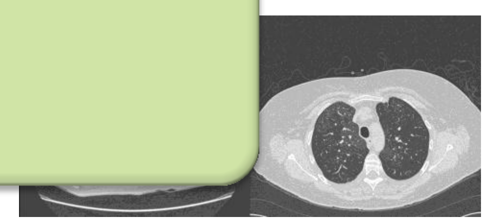


- 3D Confocal Skin Image Segmentation
  (with Prof. Brooks and Memorial Sloan Kettering Cancer Center)
- Subtyping COPD (with Brigham and Women's Hospital)
- Emotion Detection (with Draper Labs)
- Road Defect Detection (with VOTERS)



of. Aslam)

Computer Aided Diagnosis,
Automated Labeling of Medical Text

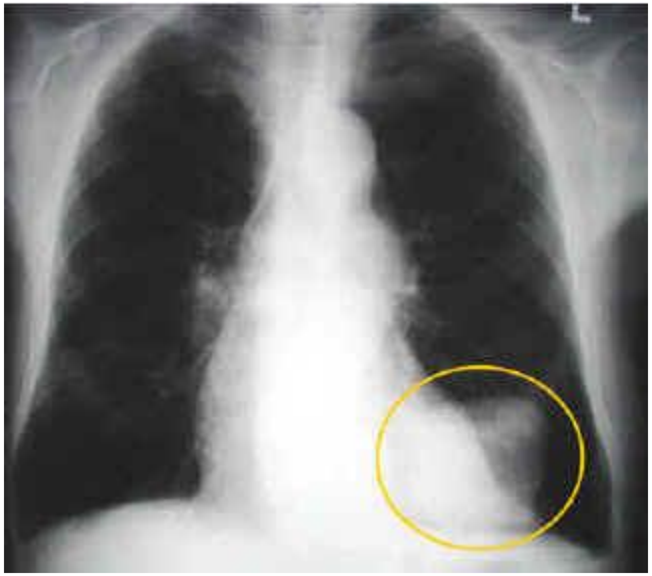


Conclusions

- We provided a probabilistic model that allows learning from multiple annotators/crowd whose annotations may be noisy;
- Our model takes into account that the quality of annotation may vary with data;
- This model can deal with missing annotators/data;
- Our model can also be utilized to evaluate annotators even when ground truth is not available; and
- We can also utilize our model to select the most trustworthy/accurate annotator for each new instance labeling
- We've developed an approach that can intelligently select samples to label and the associated annotators to query (active learning from multiple annotators).

Motivation

- *Multiple Expert Diagnoses*
- *Amazon Mechanical Turk*



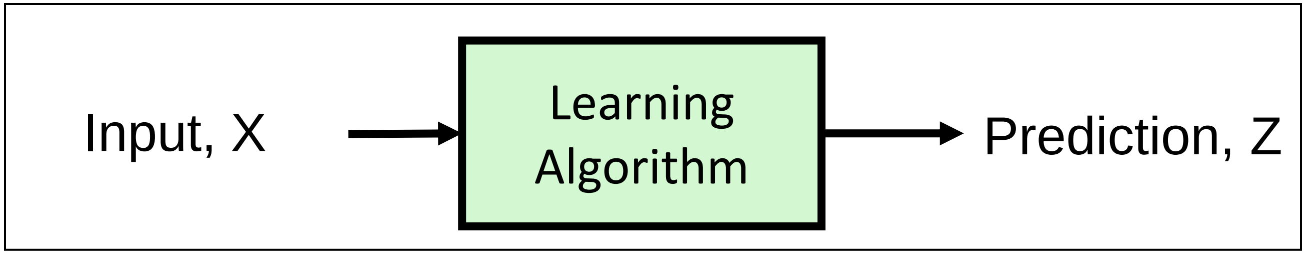
1. *How should the patients be diagnosed when doctors disagree?*
2. *How do we evaluate the doctors' diagnoses?*

Challenges

- 1. Multiple yet unreliable annotators/sources.

- 2. Varying performance on types of data.
  - *Due to different expertise.*
  - *Due to quality of data.*

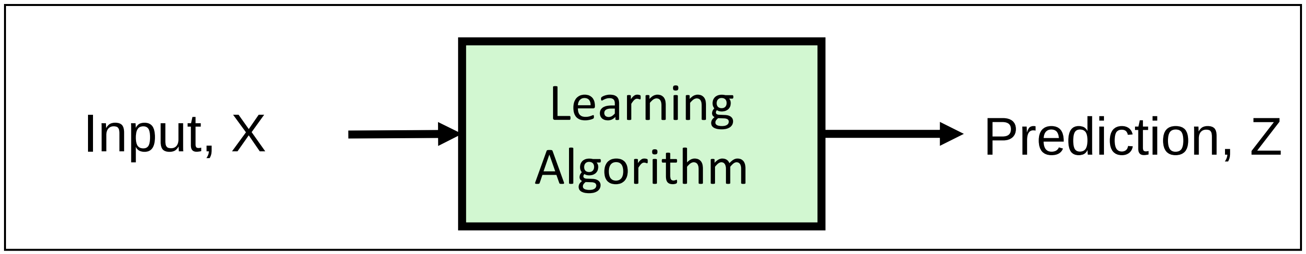
Standard Supervised Learning Problem



Ground Truth

	Age	Temp.	Symptoms…	Z
Patient 1	1	96	…	sick
Patient 2	50	102	…	not sick
…			…	
Patient N	65	95	…	not sick

Multiple Annotator Learning Problem



	Age	Temp.	Symptoms…	Ann. Y_1	Ann. Y_2	Ann. …	Ann. Y_T
Patient 1	1	96	…	not sick	sick	…	sick
Patient 2	50	102	…	sick	sick	…	sick
…			…				
Patient N	65	95	…	not sick	not sick	…	sick

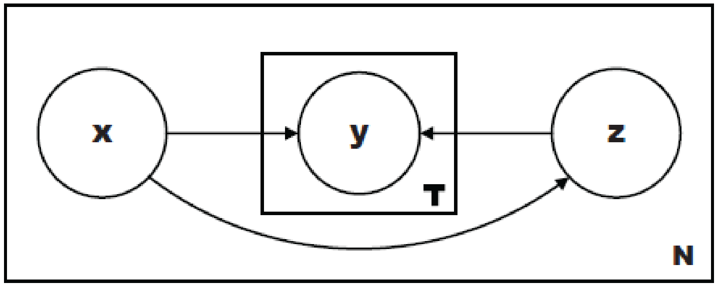
- No objective ground truth.
- Multiple inexpensive annotators may be available, but will depend on annotator idiosyncrasies.

Challenges

- 1. Multiple yet unreliable annotators/sources.
- 2. Varying performance on types of data.
  - *Due to different expertise.*
  - *Due to quality of data.*

Probabilistic Model for Multiple Annotators

(Yan et al., AISTATS 2010)



x: samples **z**: true labels *y*: labels from annotators

Joint Conditional Distribution:

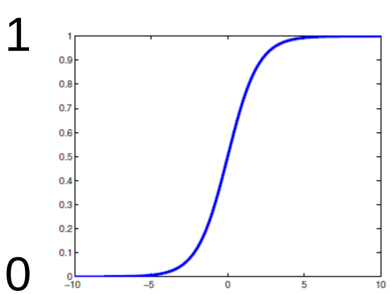
$$p(Y,Z|X) = \prod_i \boxed{p(z_i|\mathbf{x}_i)} \prod_t \boxed{p(y_i^{(t)}|\mathbf{x}_i, z_i)}$$

Classifier Model:

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

Logistic regression model

$$\frac{1}{1+\exp(\rho)}$$



$$\rho = (-\alpha^T x_i - \beta)$$

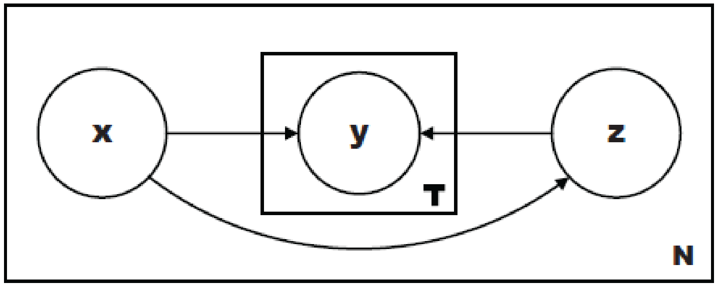
Annotator Model: **Bernoulli model**

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

η_t : **Probability of labeler t to be correct**

Probabilistic Model for Multiple Annotators

(Yan et al., AISTATS 2010)



x: samples **z**: true labels y: labels from annotators

Joint Conditional Distribution:

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

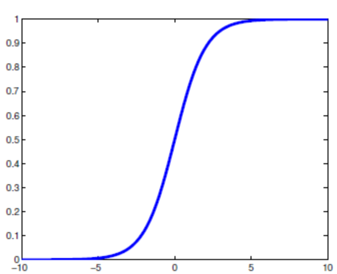
Classifier Model:

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

Logistic regression model

$$\frac{1}{1+\exp(\rho)}$$

1



0

$\rho = (-\alpha^T x_i - \beta)$

Annotator Model:

Bernoulli model

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

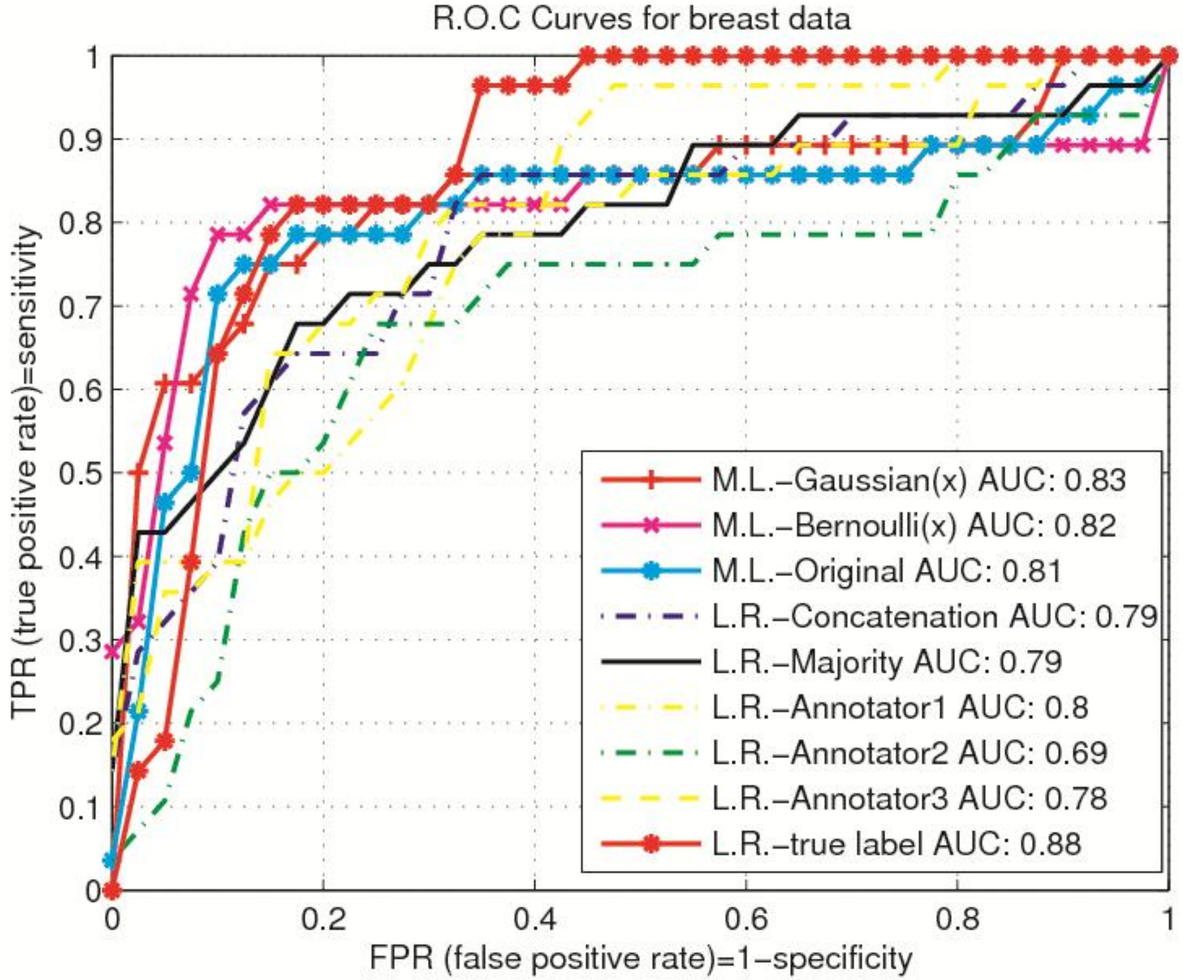
η_t : **Probability of labeler t to be correct**

Confidence Model:

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

When annotator's performance vary with data

Breast Cancer Detection

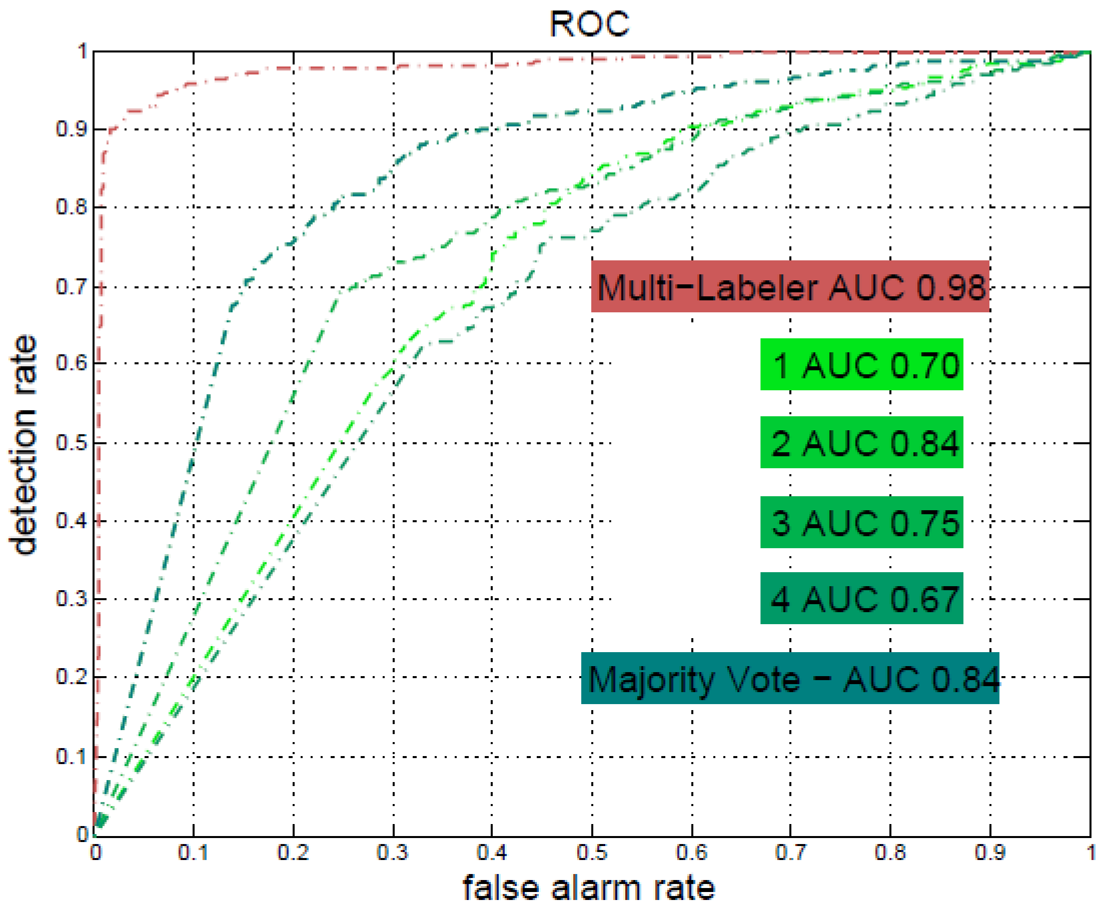


75 cases, 8 texture features, 3 annotators (radiologists)

The Atrial Fibrillation Data

- Atrial Fibrillation (cardiac arrhythmia of abnormal heart rhythm) from unstructured medical text.
- We are using actual electronic medical records (EMR) from various medium/large-size hospitals. Our dataset consists of a set of 1058 passages from a medical database containing a variety of different medical records: discharge notes, visit notes, bills, etc.
- The passages have been annotated by an expert labeler (nurse abstractor) and four non-expert labelers.
- Each passage is labeled into one of two categories: whether the passage is relevant in determining (or providing clear evidence) that the patient has a history of atrial fibrillation or not.
- After preprocessing, cleaning and normalization of the resulting representative vectors, we ended up with 998 samples and 323 features.

Atrial Fibrillation Detection from EMR



998 passages, 323 (metadata and text) features, 4 (non-expert) annotators , ground truth based on expert (nurse abstractor)

Active Learning

Even though we may have access to many annotators,

- it is still expensive to label
- not all annotators have the same level of expertise or confidence

Instead of having annotators label all the training data, we would like to intelligently choose instances to be labeled -- called ***active learning.***

New Paradigm:
Active Learning from Multiple Annotators

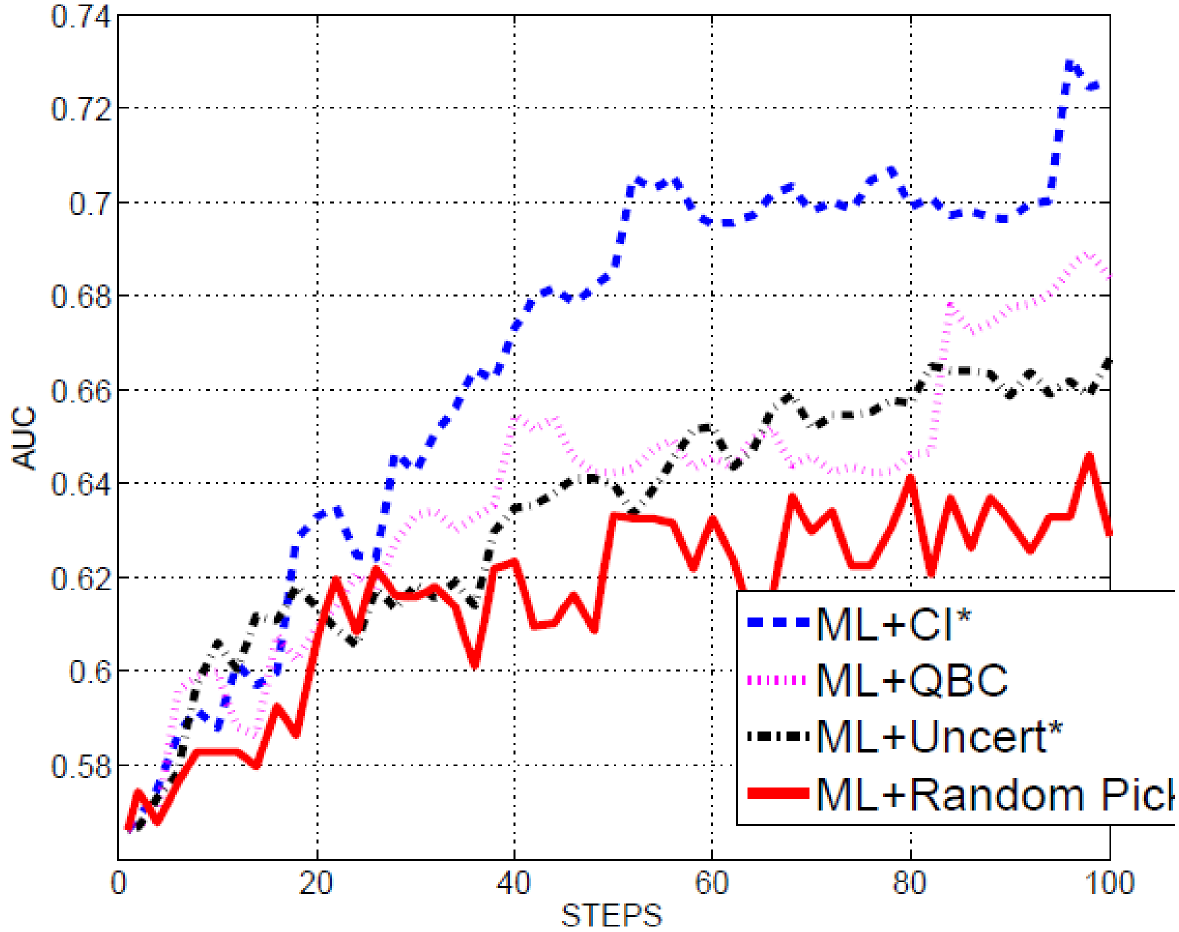
New Challenges:

- Intelligently choose instances to be labeled.
- Intelligently decide which annotator(s) to query from.

Two Strategies:

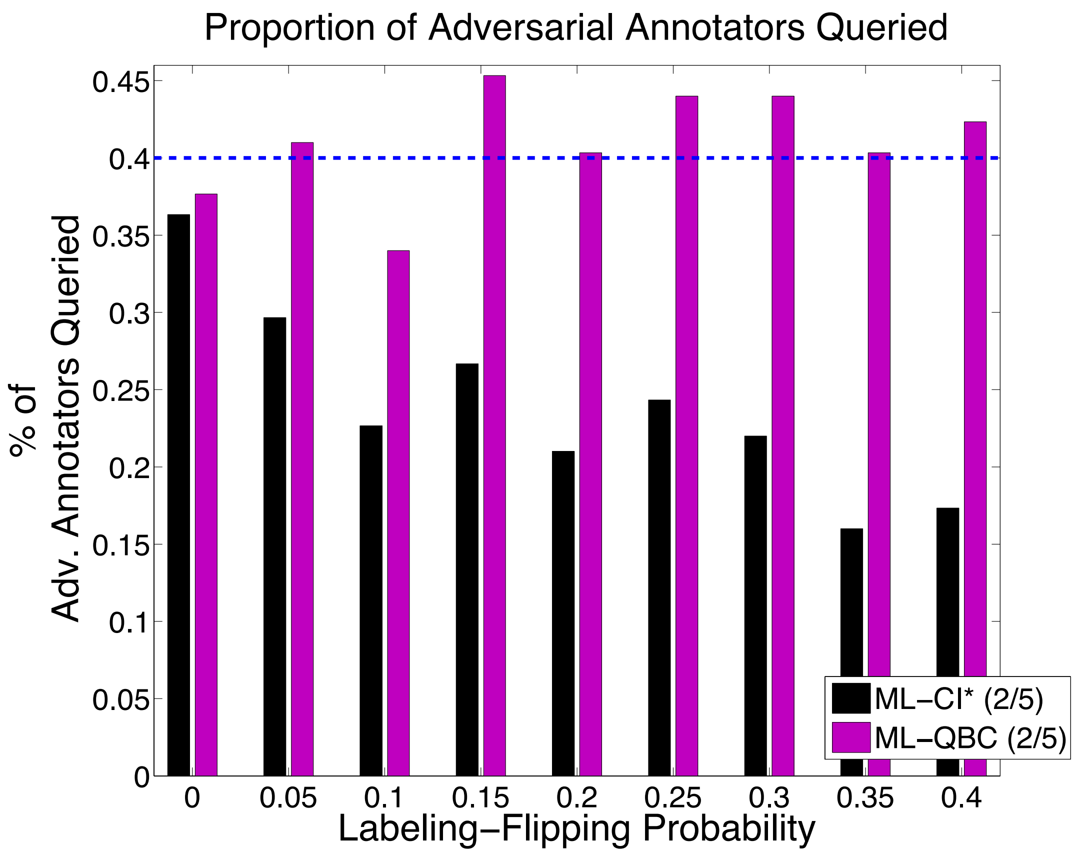
1. Uncertainty Sampling (ICML 2011)
2. Most Informative Sample and Annotator (AISTATS 2012)

Atrial Fibrillation Detection from EMR



998 passages, 323 (metadata and text) features, 4 (non-expert) annotators , ground truth based on expert (nurse abstractor), 30 random initial training, 300 active pool, the rest as test

Experiments (Query Efficiency)



Adversaries chosen at random (rate = 0.4: 2/5)
Label flip with probability p_f in {0.1, 0.2, 0.3, 0.4}

Conclusions

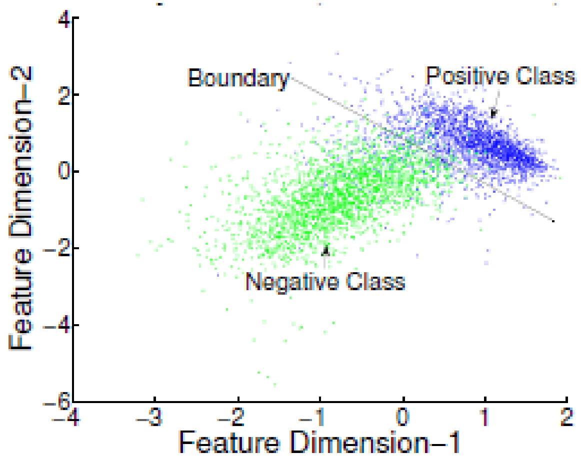
- We provided a probabilistic model that allows learning from multiple annotators/crowd whose annotations may be noisy;
- Our model takes into account that the quality of annotation may vary with data;
- This model can deal with missing annotators/data;
- Our model can also be utilized to evaluate annotators even when ground truth is not available; and
- We can also utilize our model to select the most trustworthy/accurate annotator for each new instance labeling
- We've developed an approach that can intelligently select samples to label and the associated annotators to query (active learning from multiple annotators).

An Illustrative Example

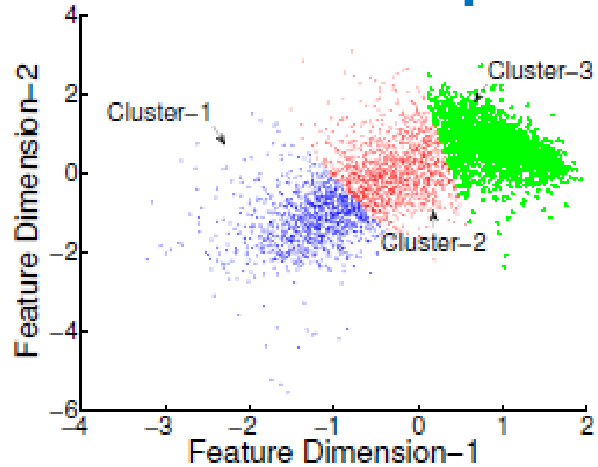
Galaxy Dim Data

- Can our model find the correct annotator to query?

True Labels

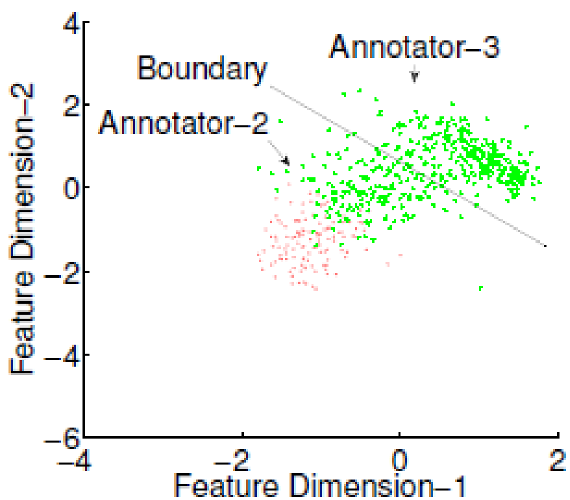


Annotator Expertise

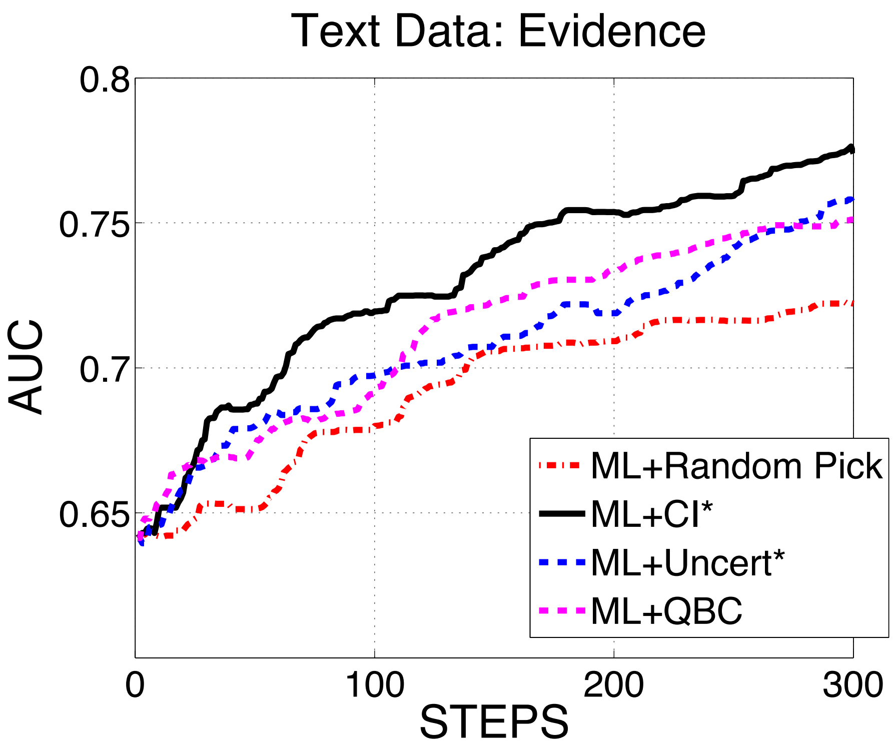


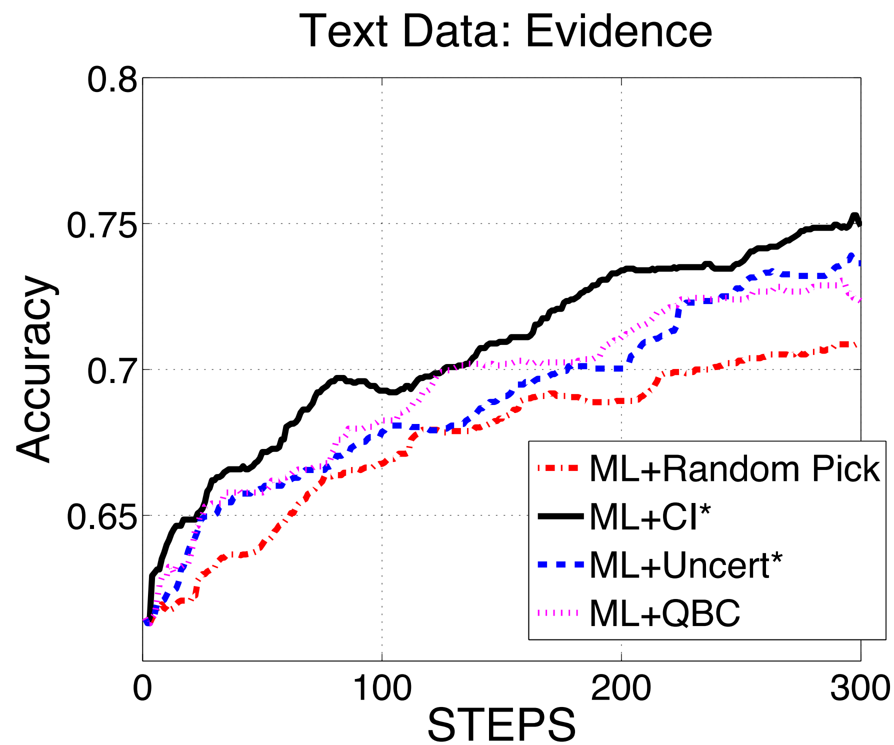
Annotators Queried for each Sample Selected by our Method

Active Learning



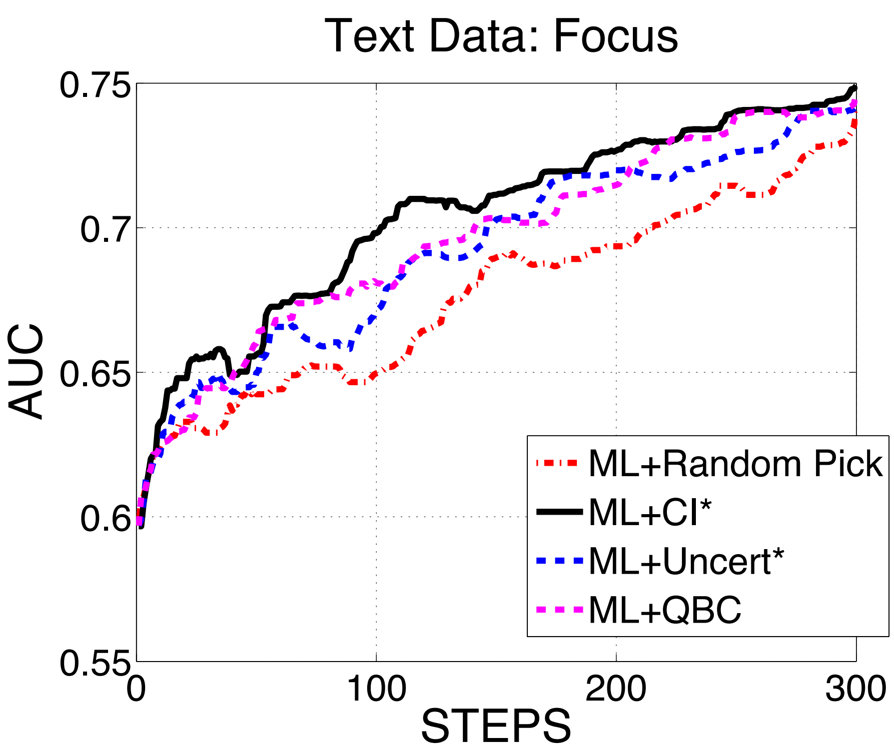
Experiments (Learning Rate)

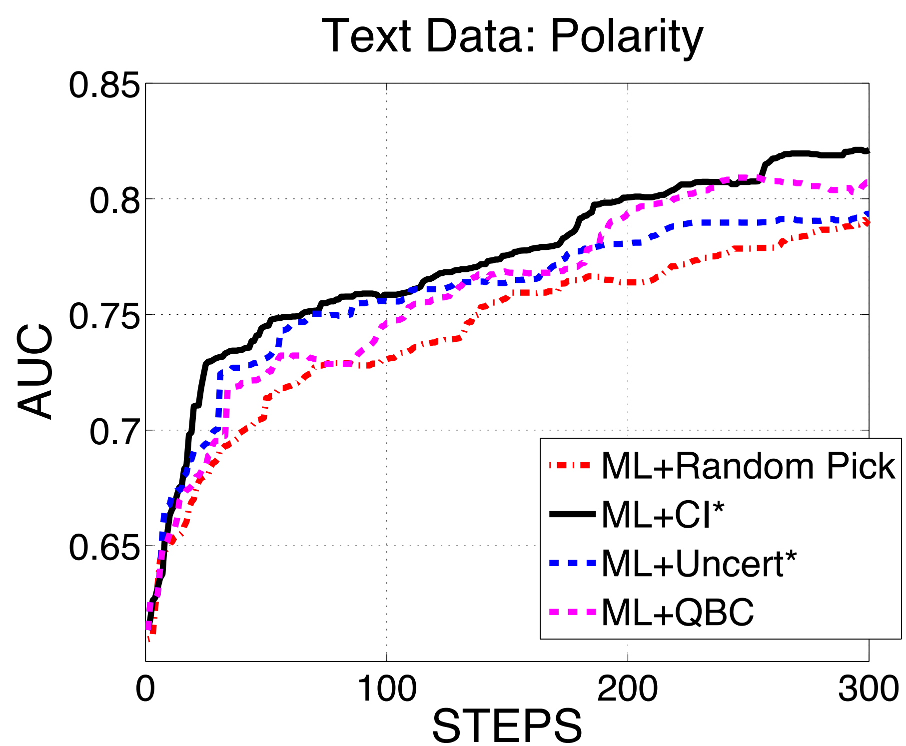




Randomized runs (10 fold) : 300 initial training, 300 active pool, 400 testing
Input space: 292 features (bow)

Experiments (Learning Rate)





(ICML 2011)

Active Learning from Multiple Annotators

- We'd like to select samples in which our classifier model is most uncertain.

Classifier Model:

$$p(z_i = 1 | x_i) = (1 + \exp\boxed{(-\alpha^T x_i - \beta)})^{-1}$$

Most uncertain when, $p(z_i = 1 | x_i) = 0.5$

-> the smaller $(-\alpha^T x_i - \beta)$ is, the more uncertain the classifier is.

- We don't have an oracle, we would like to pick the sample that our annotators are most confident in labeling.

Confidence Model:

$$\eta_t(x) = (1 + \exp\boxed{(-w_t^T x_i - \gamma_t)})^{-1}$$

The larger $(-w_t^T x_i - \gamma_t)$ is for each annotator, the more confident the annotator is.

(ICML 2011)

Active Learning from Multiple Annotators

Objective Function

$$\min_{\mathbf{x},\mathbf{p}} \overbrace{C(\alpha'\mathbf{x}+\beta)^2}^{\text{uncertainty}} + \overbrace{\mathbf{p}'[\mathbf{w}_1,\ \mathbf{w}_2,\ldots,\ \mathbf{w}_T]'\mathbf{x}+\mathbf{p}'\gamma}^{\text{annotator confidence}}$$

$$\texttt{constrained to}: \quad C \geq 0, \quad \mathbf{p} \geq \mathbf{0}, \quad \sum_t \mathbf{p} = 1,$$

$$\texttt{where}: \quad \mathbf{p} \triangleq [p_1,\ p_2,\ldots,\ p_T]', \quad \gamma \triangleq [\gamma_1\ \gamma_2,\ldots,\ \gamma_T]',$$

(AISTATS 2012)

Active Learning from Multiple Annotators

- We'd like to select samples and the corresponding annotator that

maximize the information about the true label value.

Criterion:

$$[k^*, s^*] = \arg\max I(z_k; [y_k^{(s)}, x_k] \mid X, Y_O)$$

$$= \arg\max \sum_{z_k, y_k^s} p(z_k \mid [y_k^s, x_k]; \theta) \log p(z_k \mid [y_k^s, x_k]; \theta)$$

$$- \sum_{z_k} p(z_k \mid \theta) \log p(z_k \mid \theta)$$