Modeling Annotator Expertise

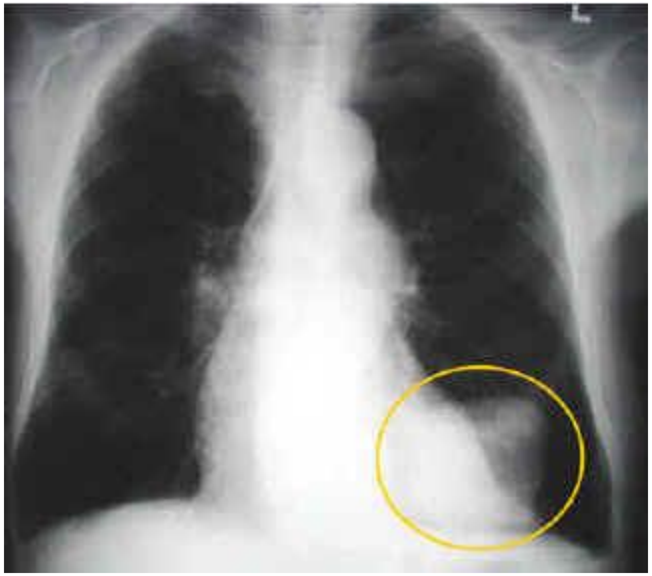
-Learning when everybody knows a bit of something

Y. Yan[1] R. Rosales [2] G. Fung[2] M. Schmidt[3]
G. Hermosillo [2] L. Bogoni [2] L. Moy[4] J. Dy [1]

[1]Northeastern Univ., Boston, MA USA [2]Siemens Healthcare, Malvern, PA USA
[3]Univ. of British Columbia, Vancouver, BC Canada
[4]Univ. of Penn., Philadelphia, PA USA

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?*

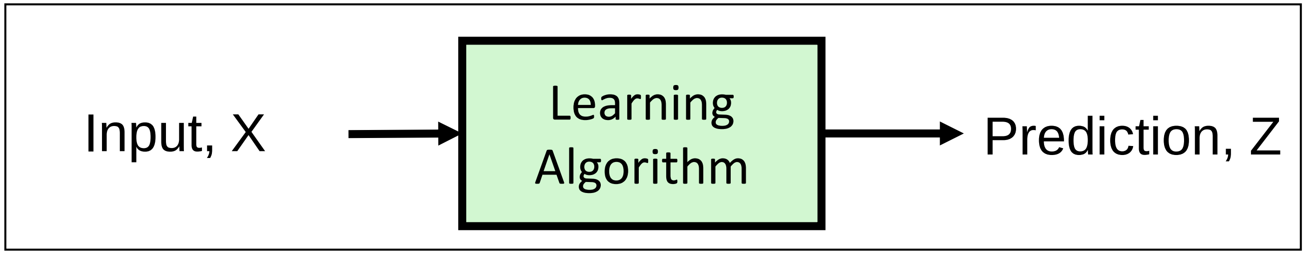
Model Assumptions

- 1. Multiple yet unreliable annotators.

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

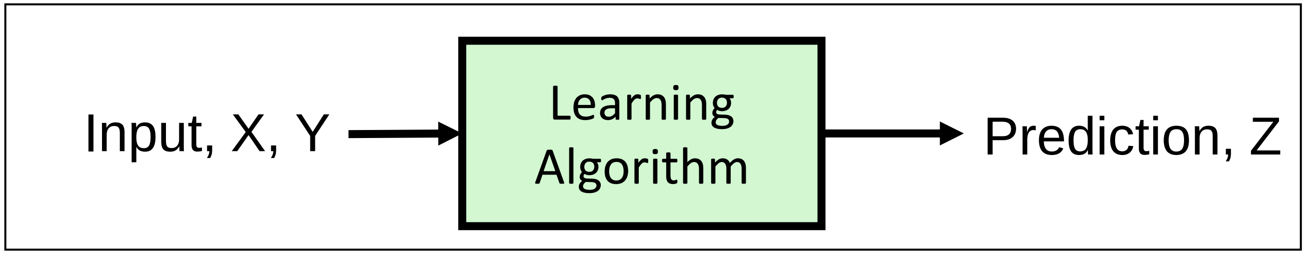
Typical Classification Problem

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

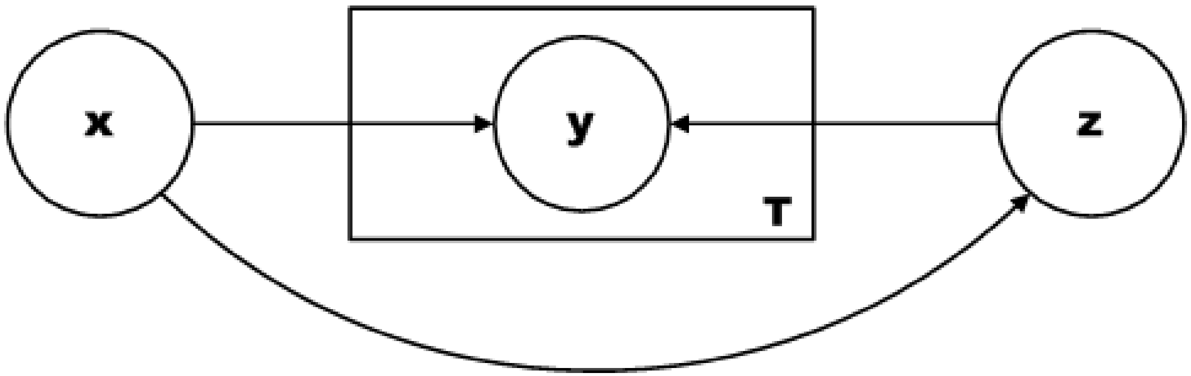


The Multiple Expert Problem

	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



Probabilistic Model



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

-$\mathbf{x}_i \in \mathbf{R}^D, i = 1,2, \ldots, N$ observations;
-$y_i^{(t)} \in \mathbf{R}, i = 1,2, \ldots, N; t = 1,2, \ldots, T$ annotation by t for sample i;
-$z_i \in \mathbf{R}, i = 1,2, \ldots, N$ true (hidden) label for sample i.

Probabilistic Model

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

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

Classifier: Logistic regression model

Probabilistic Model

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

Bernoulli Model:

$$\boxed{p\left(y_i^{(t)}\middle|z_i\right) = (1 - \eta_t)^{\left|y_i^{(t)} - z_i\right|} {\eta_t}^{1 - \left|y_i^{(t)} - z_i\right|}}$$

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

Gaussian Model:

$$\boxed{p\left(y_i^{(t)}\middle|z_i\right) = \boldsymbol{N}\left(y_i^{(t)}; z_i, \sigma_t\right)}$$

σ_t: **How labeler t deviates from the true label z**

Probabilistic Model

when annotator's performance vary with data

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

Bernoulli Model:

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

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

Gaussian Model:

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

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

Implementation

Maximum Likelihood Estimation:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \prod_t \prod_i p\left(y_i^{(t)} \middle| \mathbf{x}_i; \boldsymbol{\theta}\right)$$
$$= \arg\max_{\theta} \sum_t \sum_i \log \sum_{z_i} p\left(y_i^{(t)}, z_i \middle| \mathbf{x}_i; \boldsymbol{\theta}\right)$$

Since z_i are *hidden*, **EM** algorithm is utilized:

E-step Compute:

$$\begin{aligned} \tilde{p}(z_i) &\triangleq p(z_i|\mathbf{x}_i, y_i) \\ &\propto p(z_i, y_i|\mathbf{x}_i) \\ &= \prod_t p\left(y_i^{(t)} \middle| \mathbf{x}_i, z_i\right) p(z_i|\mathbf{x}_i) \end{aligned}$$

M-step Maximize:

$$\sum_t \sum_i E_{\tilde{p}(z_i)}[\log p(y_i^{(t)}, z_i|\mathbf{x}_i)]$$

to update $\widetilde{\boldsymbol{\theta}} = \{\boldsymbol{\alpha}, \beta, \{\mathbf{w}_t\}, \{\gamma_t\}\}$

Insights on Classification Model

$$LLR(\{y^{(t)}\}, \mathbf{x}) = \log \frac{p(z = 1 | \{y^{(t)}\}, \mathbf{x})}{p(z = 0 | \{y^{(t)}\}, \mathbf{x})}$$

- ***Bernoulli Case***

$$LLR = \boxed{\alpha^T \mathbf{x} + \beta} + \boxed{\sum_t (-1)^{(1-y^{(t)})} (\mathbf{w}_t^T \mathbf{x} + \gamma_t)}$$

by general learnt classifier *by each annotator*

- ***Gaussian Case***

$$LLR = \boxed{\alpha^T \mathbf{x} + \beta} + \boxed{T^+ - T^- + \sum_t (-1)^{(1-y^{(t)})} \exp(-\mathbf{w}_t^T \mathbf{x} - \gamma_t)}$$

by general learnt classifier *by each annotator*

Missing Annotators

- *When not all annotators provided a label for a particular sample, the true label is predicted based on:*

- 1. $p\left(z \middle| \left\{y^{t\backslash k}\right\}, \mathbf{x}\right) = \dfrac{\prod_{t\backslash k} p\left(y^{(t)} \middle| z, \mathbf{x}\right) p(z|\mathbf{x})}{\sum_z \prod_{t\backslash k} p\left(y^{(t)} \middle| z, \mathbf{x}\right) p(z|\mathbf{x})}$

- 2. $p(z = 1|\mathbf{x}) = (1 + \exp(-\boldsymbol{\alpha}^T \mathbf{x} - \beta))^{-1}$

Predicting Ground Truth without Observation

- *Estimate hidden label purely on annotations when observation is not available.*

- $p\left(z \middle| \{y^{(t)}\}\right) = \int \prod_t p\left(y^{(t)} \middle| z, \mathbf{x}\right) p(z|\mathbf{x}) dp(\mathbf{x})$

$$\approx \frac{1}{S} \sum_{s=1}^{S} p(z|\mathbf{x}_s) \prod_t p(y^{(t)}|z, \mathbf{x}_s)$$

Approximation is reached by sampling.

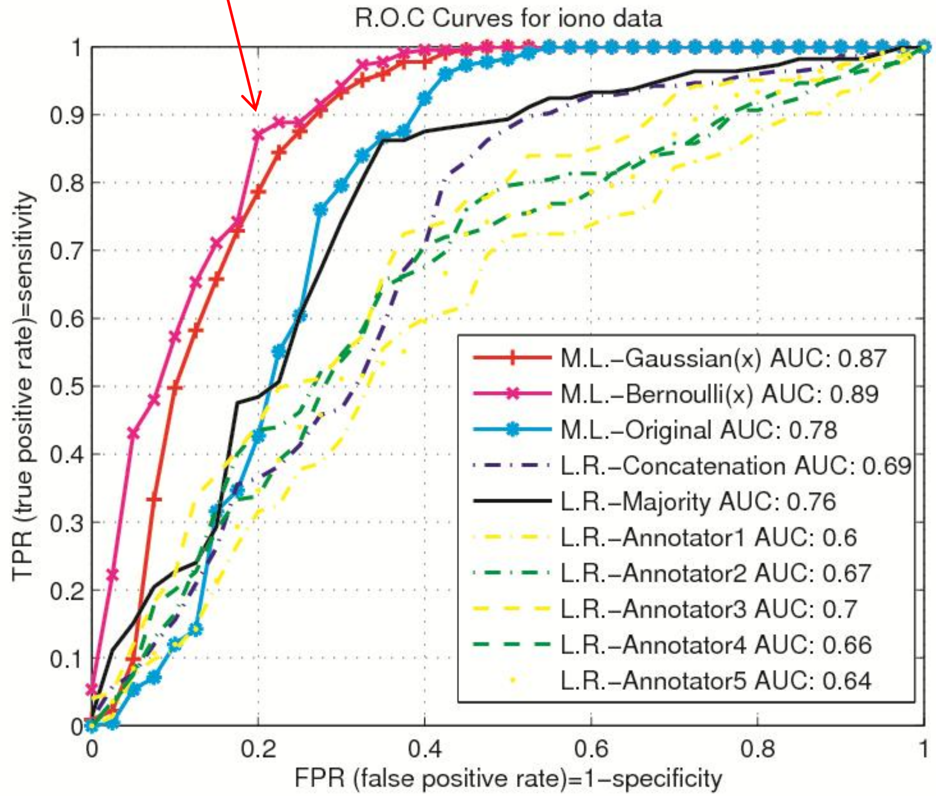
Evaluating Annotators

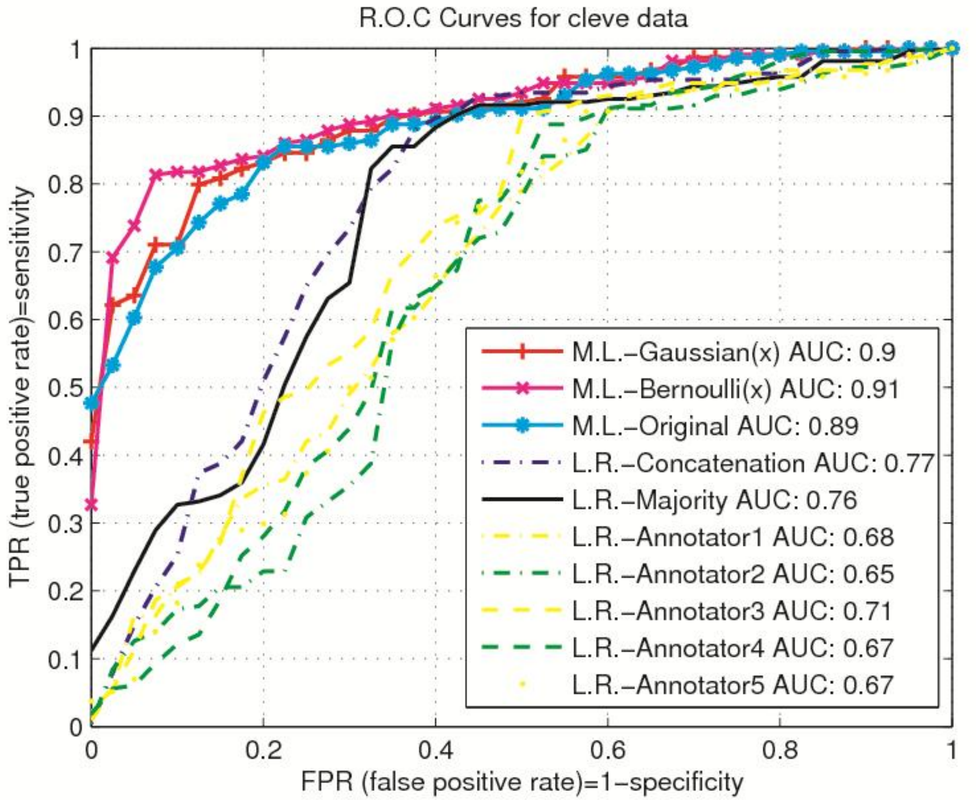
- *Is it possible to evaluate annotators without ground truth?*

- $$p\left(y^{(k)}\middle|\left\{y^{(t\backslash k)}\right\},\mathbf{x}\right)=\frac{p(\left\{y^{(t)}\right\}|\mathbf{x})}{p(\left\{y^{(t\backslash k)}\right\}|\mathbf{x})}$$
$$=\frac{\sum_{z}p\left(\left\{y^{(t)}\right\}\middle|z,\mathbf{x}\right)p(z|\mathbf{x})}{\sum_{z}p\left(\left\{y^{(t\backslash k)}\right\}\middle|z,\mathbf{x}\right)p(z|\mathbf{x})}$$

UCI Data Classification

Our method (x)

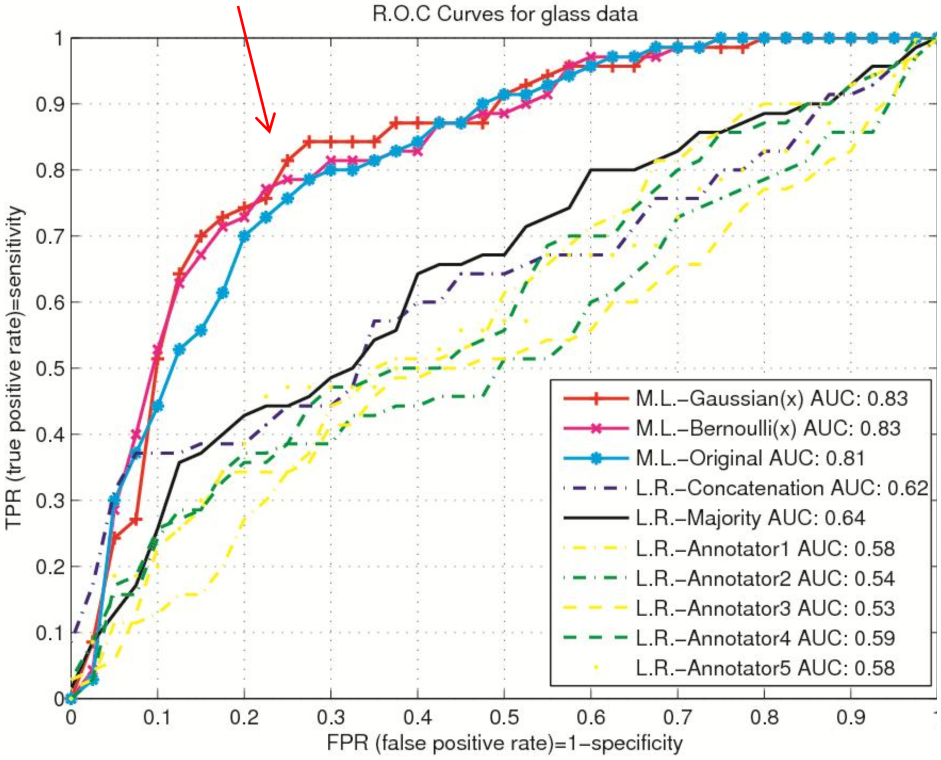


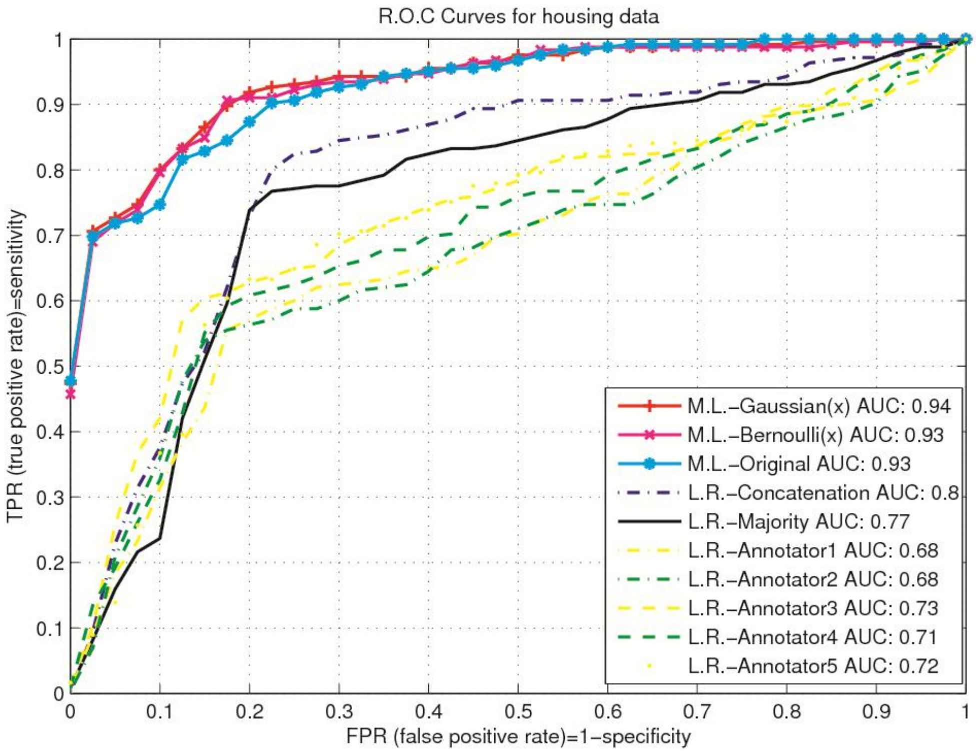


Data tested: Ionosphere, Cleveland Heart.

UCI Data Classification

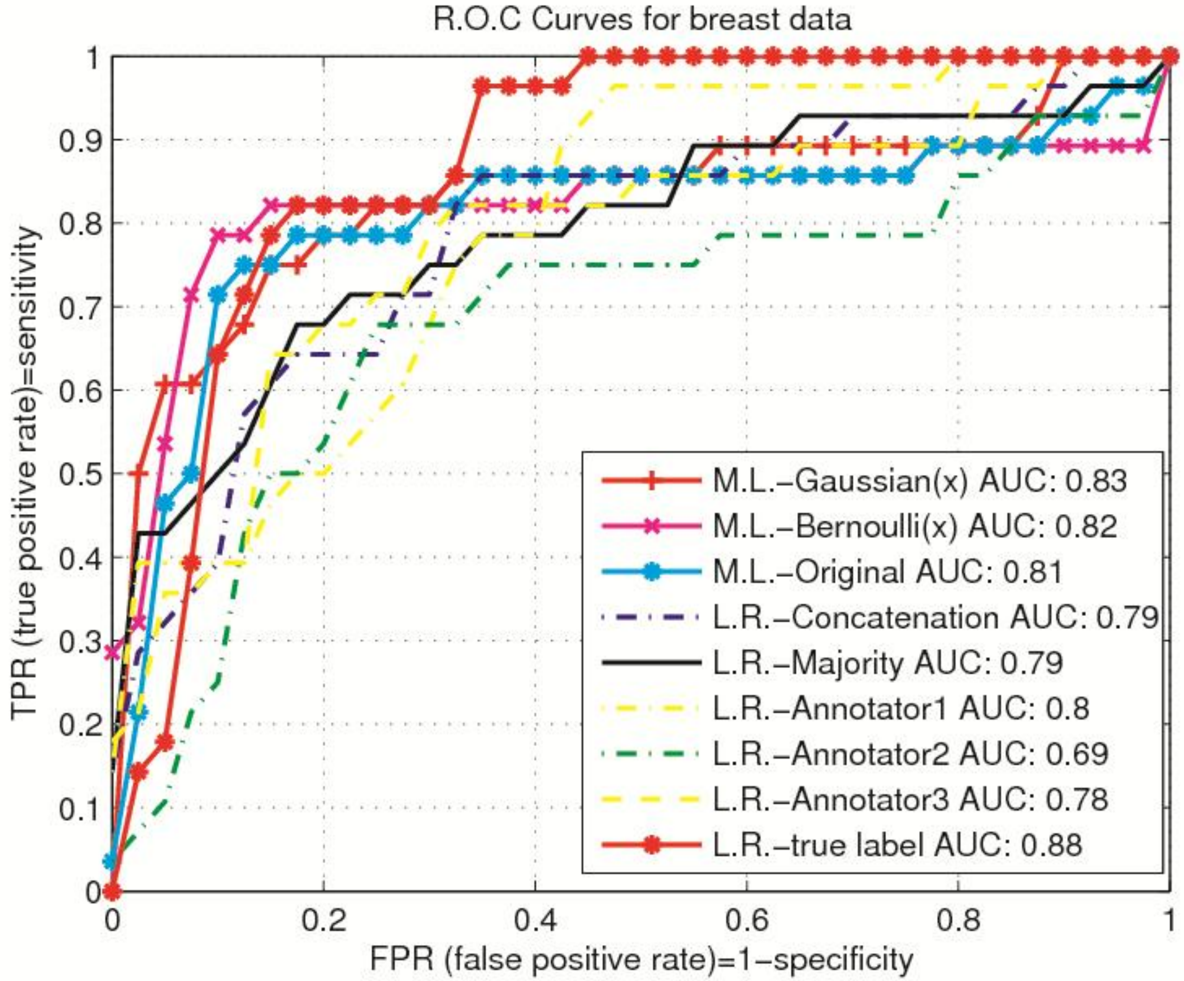
Our method (x)





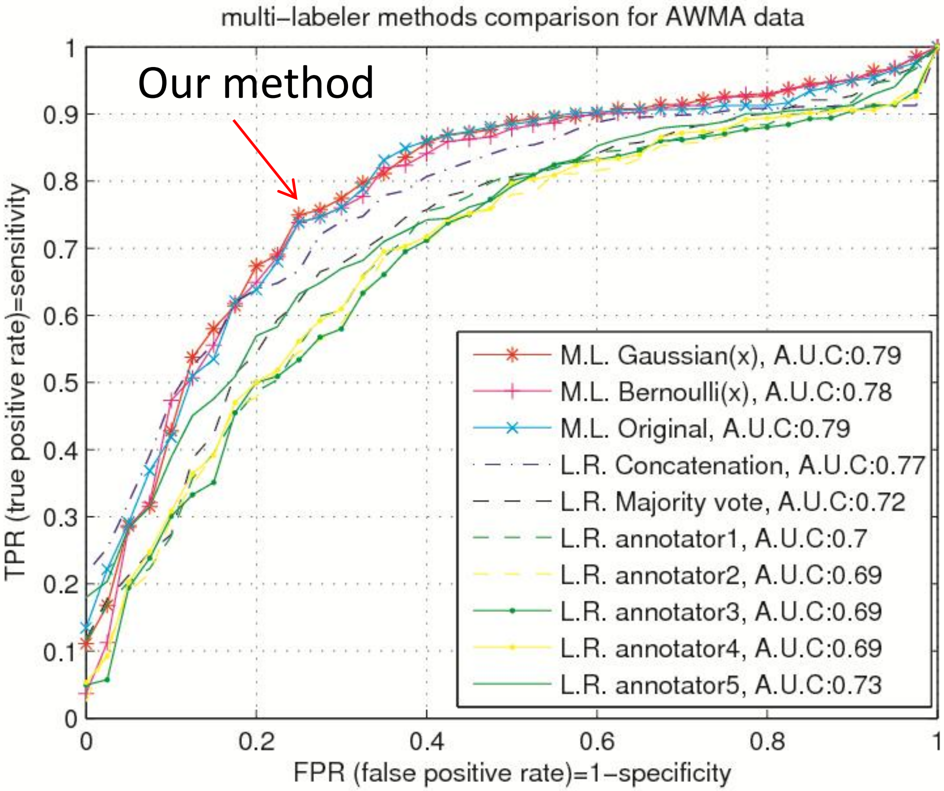
Data tested: Glass, and Housing.

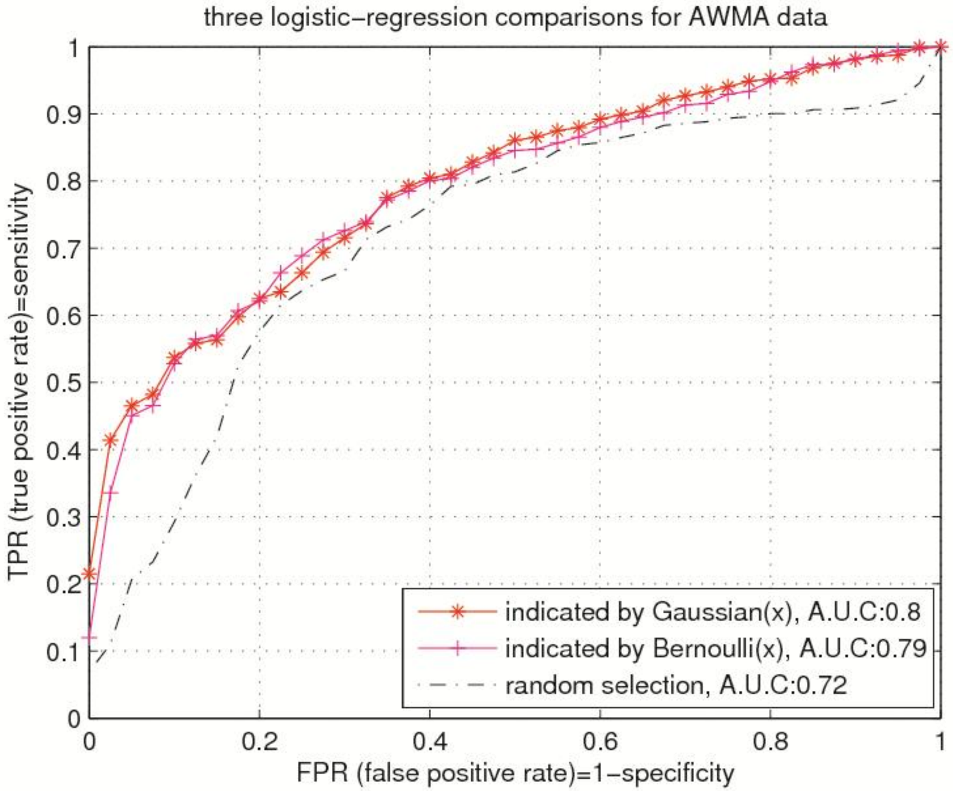
Breast Cancer Detection



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

Cardiac Wall Motion Anomaly Detection





220 cases, 16 LV heart wall segments, 5 annotators (doctors), binary labels (-/+1)

Conclusions

- We provided a probabilistic model that allows learning from multiple annotators whose annotations may be noisy;
- Our model takes into account that the quality of annotation may vary with data;
- We show that 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.

Thanks for Listening

Questions?

References

- K. Crammer, et. al. (2008). Learning from multiple sources. *J. of Machine Learning Research*, 9: 1757-1774
- O. Dekel, and O. Shamir (2009). Good learners for evil teachers. In *Int. Conf. on Machine Learning*.
- J. Howe (2008). *Crowdsourcing: why the power of the crowd is driving the future of business*. Crown Business
- R. Jin, and Z. Ghahramani (2003). Learning with multiple labels. In *Adv. Neural Inf. Processing Systems*.
- V. Raykar, and et. al. (2009). Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Int. Conf. on Machine Learning.*
- V. S. Sheng, and et. al. (2008). Get another label? Improve data quality and data mining using multiple, noisy labelers. In *Knowledge Discovery and Data Mining (KDD)*.
- R. Snow, and et. al. (2008). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Conf. Empirical Methods on Natural Language Processing (EMNLP)*.