Multi-Task Active Learning

Yi Zhang

Outline

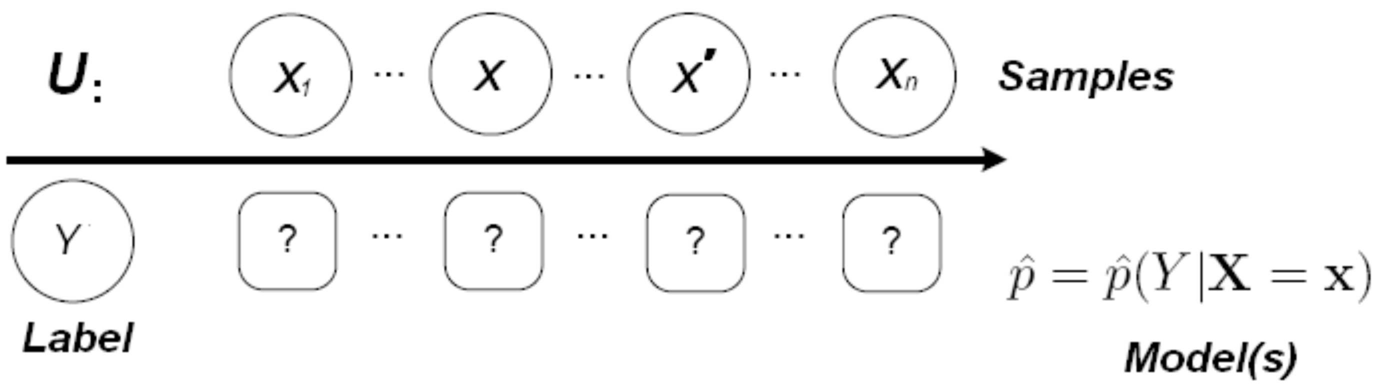
- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL’ 08)
  - Image Classification (CVPR’ 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories

Outline

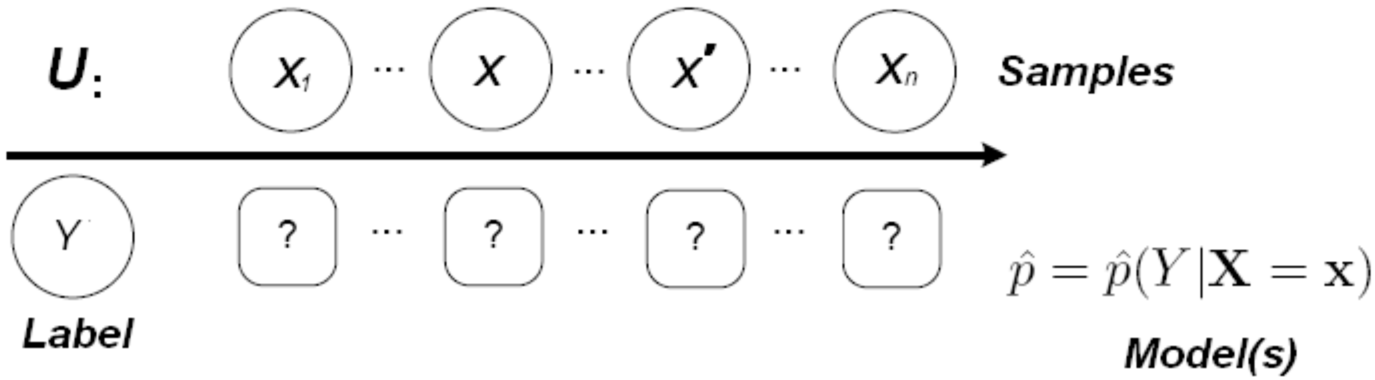
- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL' 08)
  - Image Classification (CVPR' 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories

Active Learning

- Select samples for labeling
  - Optimize model performance given the new label



Active Learning

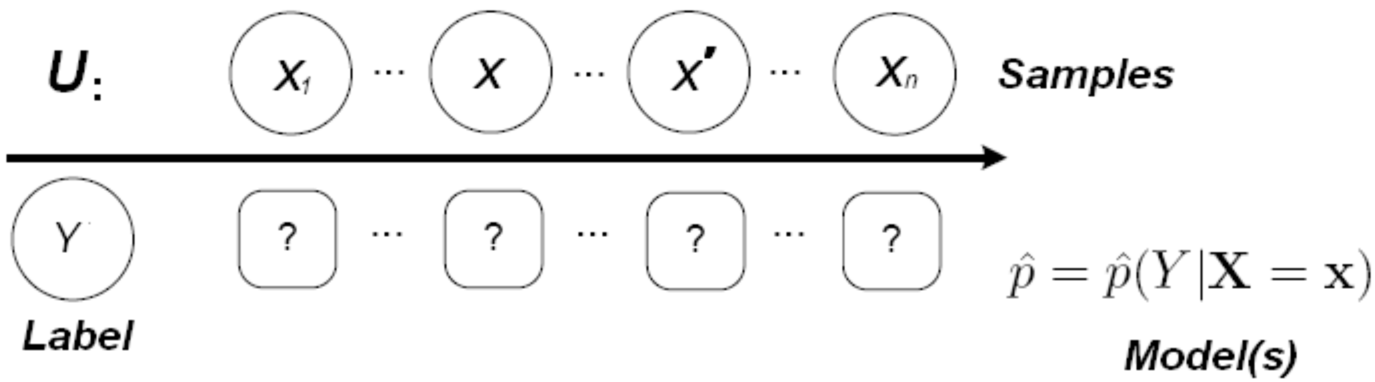


- Uncertainty sampling

$$\underset{\mathbf{x} \in \mathbf{U}}{\operatorname{argmax}} \left[- \sum_{y} \hat{p}(Y = y|\mathbf{x}) \log_2 \hat{p}(Y = y|\mathbf{x}) \right]$$

  - Maximize: *the reduction* of model entropy on ***x***

Active Learning

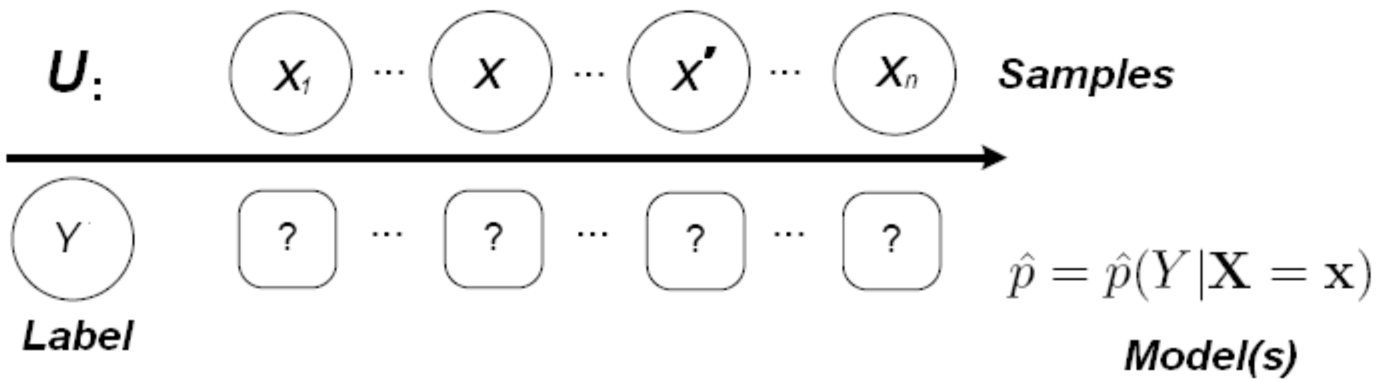


- Query by committee (e.g., vote entropy)

$$\underset{\mathbf{x} \in \mathbf{U}}{\operatorname{argmax}} \left[-\sum_{y} \hat{p}_C(Y = y|\mathbf{x}) \log_2 \hat{p}_C(Y = y|\mathbf{x}) \right]$$

  - Maximize: the reduction of version space

Active Learning

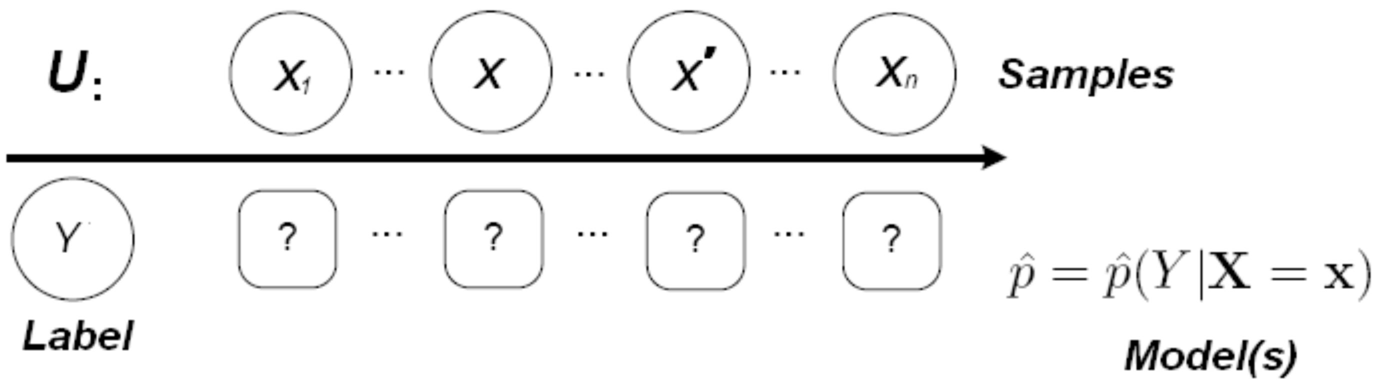


- Density-weighted entropy

$$\underset{\mathbf{x}\in\mathbf{U}}{\operatorname{argmax}} \ [\hat{P}_{\mathbf{U}}(\mathbf{x}) \cdot -\sum_{y} \hat{p}(Y = y|\mathbf{x}) \log_2 \hat{p}(Y = y|\mathbf{x})]$$

  - Maximize: approx. entropy *reduction* over ***U***

Active Learning



- Estimated error (uncertainty) reduction

$$\underset{\mathbf{x}\in\mathbf{U}}{\operatorname{argmin}} \ [\sum_{y} \hat{p}(Y = y|\mathbf{x}) \sum_{\mathbf{x}'\in\mathbf{U}} Uncertain^{+(\mathbf{x},y)}(\mathbf{x}')]$$

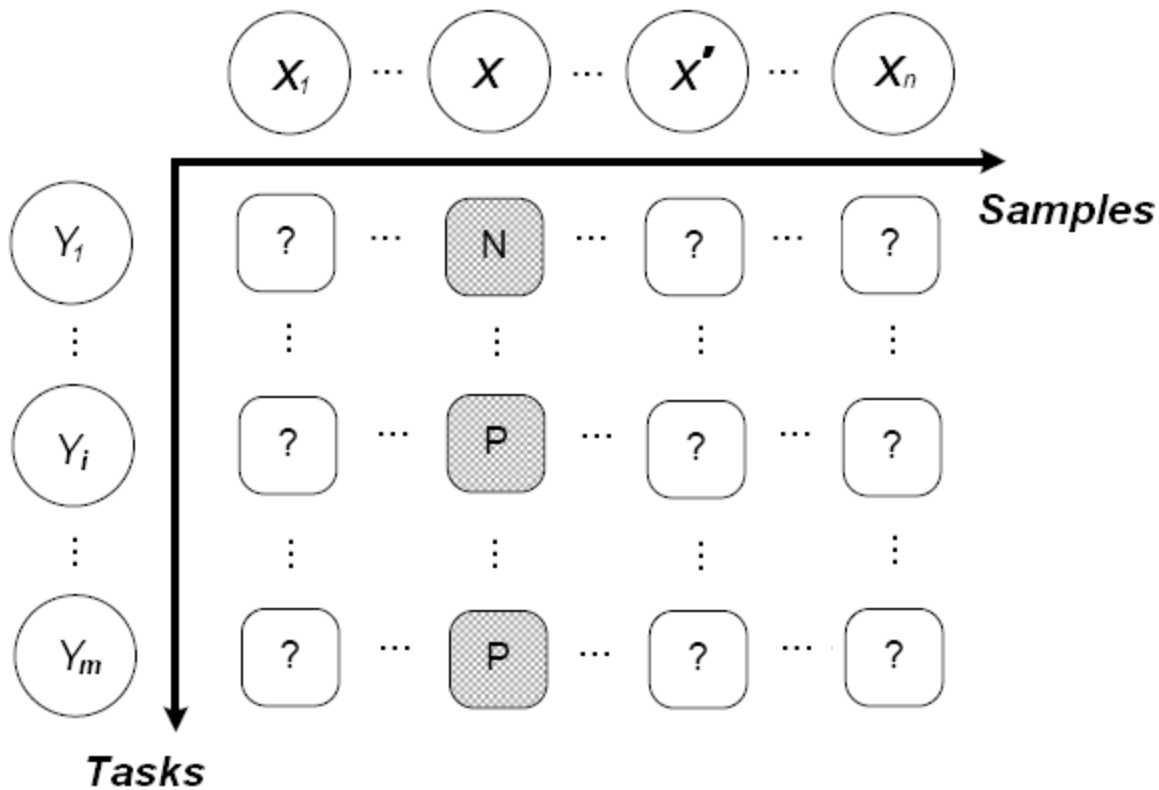
  - Maximize: reduction of uncertainty over ***U***

Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL' 08)
  - Image Classification (CVPR' 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories

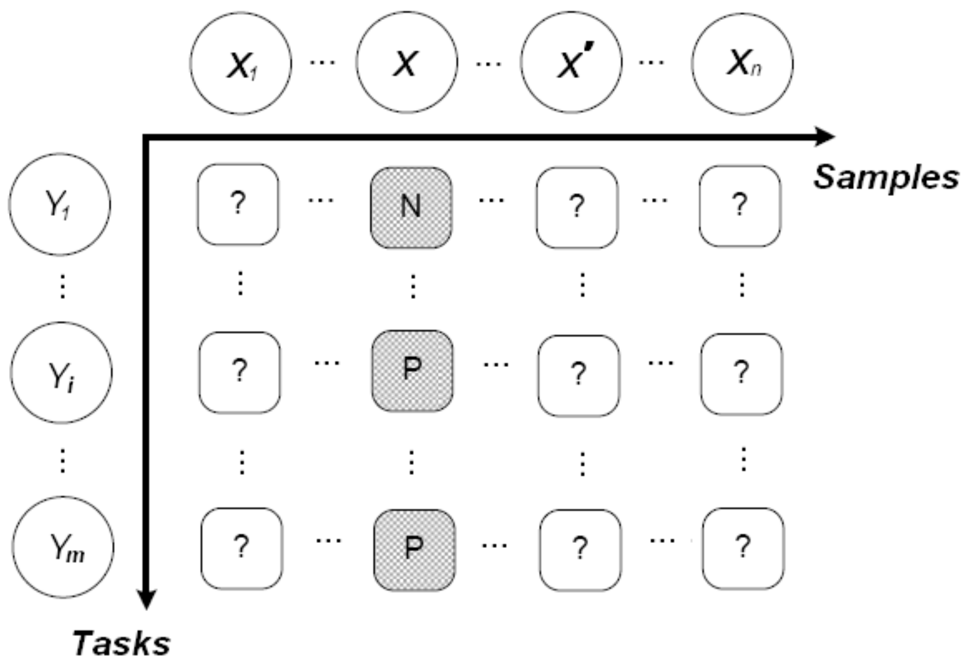
The Problem

- Select a sample → labeling all tasks



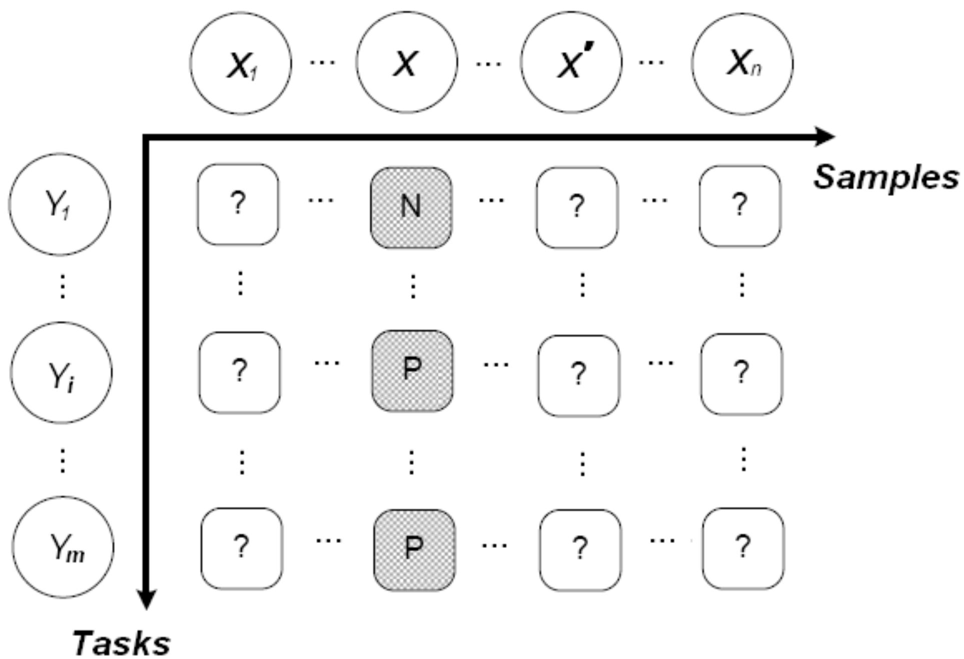
Methods

- Alternating selection
  - Iterate over tasks, sample a few from each task



Methods

- Rank combination
  - Combine rankings/scores from all single-task ALs



Experiments

- Learning two (dissimilar) tasks
  - Named entity recognition: CRFs
  - Parsing: Collins' parsing model
- Competitive AL methods
  - Random selection
  - One-side active learning: choose samples from one task, and require labels for all tasks
    - Separate AL in each task is not studied (!)
  - Alternating selection
  - Ranking combination

Unanswered Questions

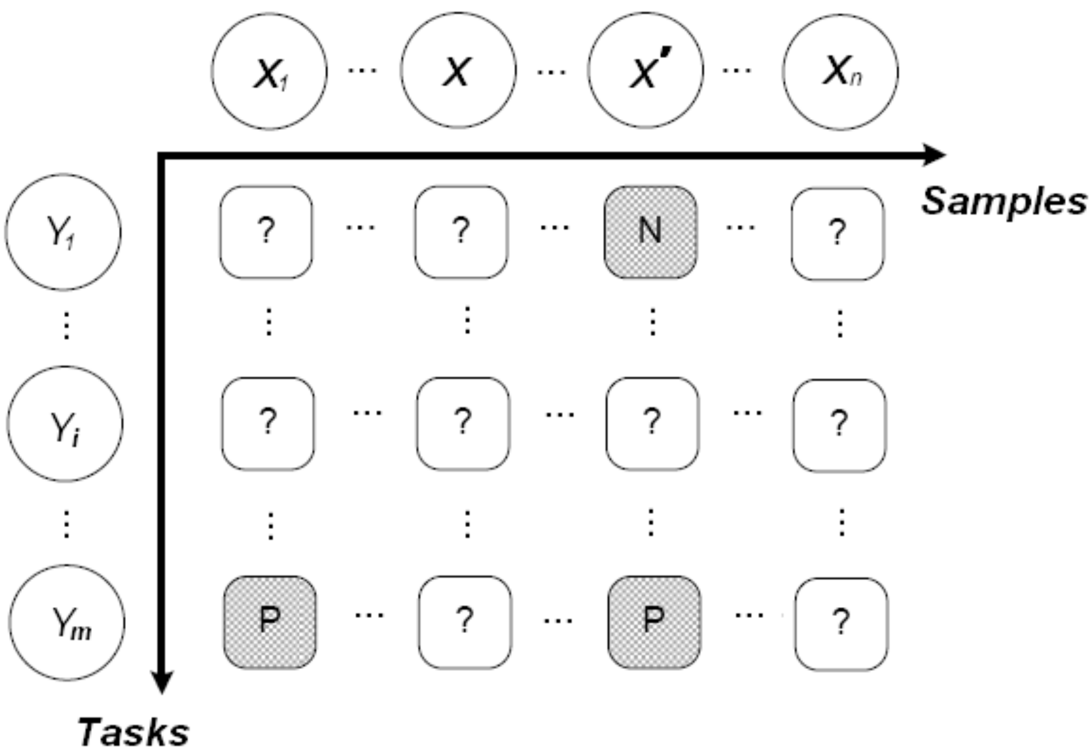
- Why “choose-one, labeling-all”?
  - Authors: annotators may prefer to annotate the same sample for all tasks
- Why learning two dissimilar tasks together?
  - Outputs of one task may be useful for the other
  - Not studied in the paper

Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL' 08)
  - Image Classification (CVPR' 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories

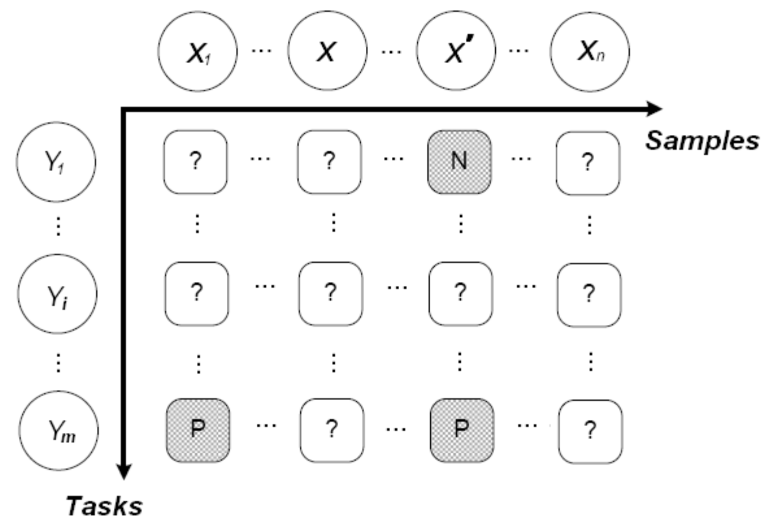
The Problem: Multi-Label Image Classification

- Select any sample-label pair for labeling



Proposed Method

$$\operatorname*{argmax}_{\mathbf{x}\in\mathbf{D},y_s\in\mathbf{U}(\mathbf{x})} \left[\sum_{i=1}^{m} MI(y_i, y_s | y_{\mathbf{L}(\mathbf{x})}, \mathbf{x})\right]$$

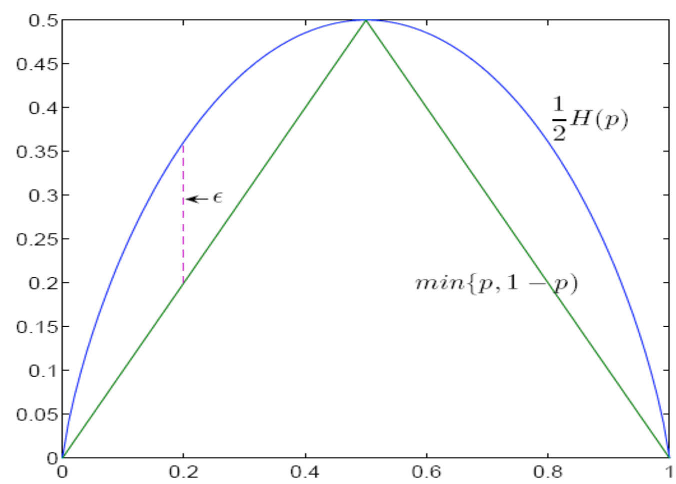


- ***D***: the set of samples
- ***x***: a sample in ***D***
- U(***x***): unknown labels of ***x***
- L(***x***): known labels of ***x***
- *m*: number of tasks
- y_s: a selected label from U(***x***)
- y_i: the label of the i^{th} task (for a sample ***x***)

Proposed Method

- Why maximizing Mutual Information?
  - Connecting Bayes (binary) classification error to entropy and MI (Hellman and Raviv, 1970)

$$\mathcal{E}\left(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}\right) \leq \frac{1}{2}H\left(y_i|y_s;y_{L(\boldsymbol{x})},\boldsymbol{x}\right)$$



Proposed Method

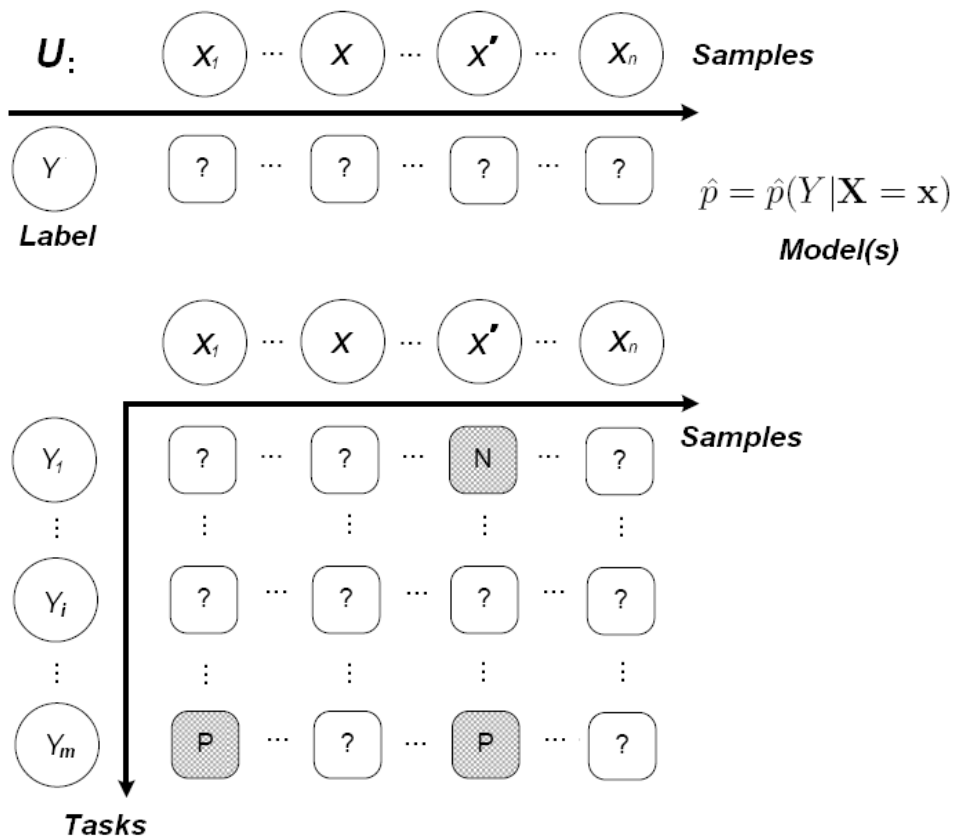
- Why maximizing Mutual Information?
  - Connecting Bayes (binary) classification error to entropy and MI (Hellman and Raviv, 1970)

$$\mathcal{E}\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \leq \frac{1}{2} H\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right)$$

$$\begin{aligned}
&\mathcal{E}\left(\boldsymbol{y} | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\
&\overset{(1)}{=} \frac{1}{m} \sum_{i=1}^{m} \mathcal{E}\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\
&\overset{(2)}{\leq} \frac{1}{2m} \sum_{i=1}^{m} H\left(y_i | y_s; y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \\
&\overset{(3)}{=} \frac{1}{2m} \sum_{i=1}^{m} \left\{ H\left(y_i | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) - MI\left(y_i; y_s | y_{L(\boldsymbol{x})}, \boldsymbol{x}\right) \right\}
\end{aligned}$$

Proposed Method

- Compare: maximize the reduction of entropy



Modeling Joint Label Probability

$$\underset{\mathbf{x}\in\mathbf{D}, y_s\in\mathbf{U}(\mathbf{x})}{\operatorname{argmax}} \left[\sum_{i=1}^{m} MI(y_i, y_s | y_{\mathbf{L}(\mathbf{x})}, \mathbf{x})\right]$$

- But how to compute:

$$MI(y_i, y_s | y_{\mathbf{L}(\mathbf{x})}, \mathbf{x})$$

- Need the *joint* conditional probability of labels

$$\hat{p}(\mathbf{y}|\mathbf{x}) = \hat{p}(y_1, y_2, \ldots, y_m|\mathbf{x})$$

Modeling Joint Label Probability

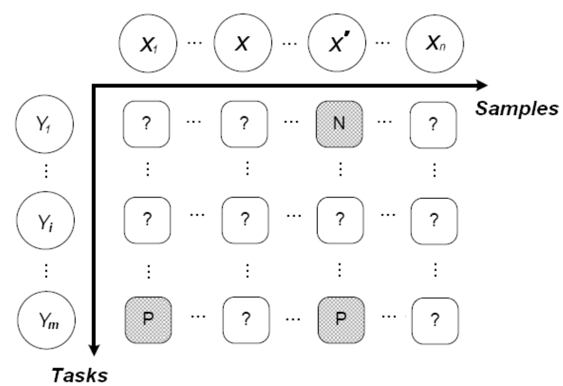
- Linear maximum entropy model

$$\hat{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{y}^T(\boldsymbol{b} + R\boldsymbol{y} + W\boldsymbol{x})\right)$$

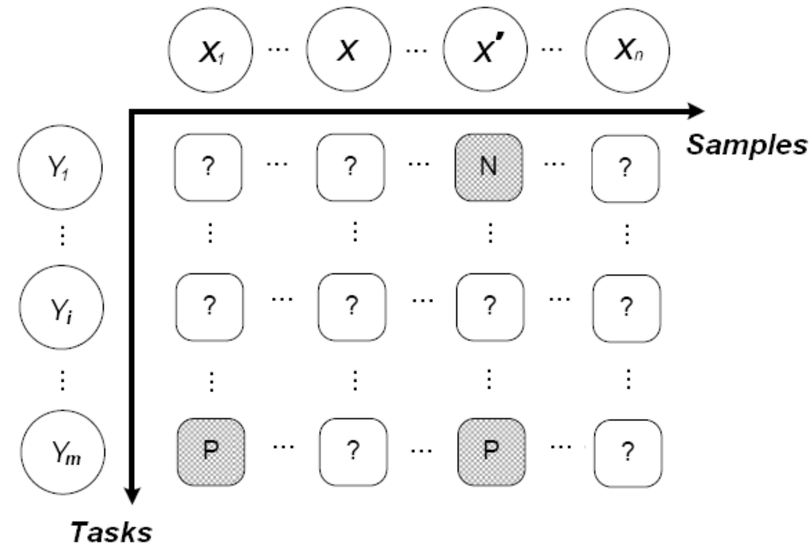
- Kernelized version

$$\hat{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{y}^T(\boldsymbol{b} + R\boldsymbol{y}) + \boldsymbol{y}^T K(W, \boldsymbol{x})\right)$$

- EM for incomplete labels



Experiments



- Data
  - Image scene classification
  - Gene function classification
- Two competitive AL methods
  - Random selection of sample-label pairs
  - Choose one sample, labeling all tasks for it
    - Separate AL in each task is not studied (!)

Discussion

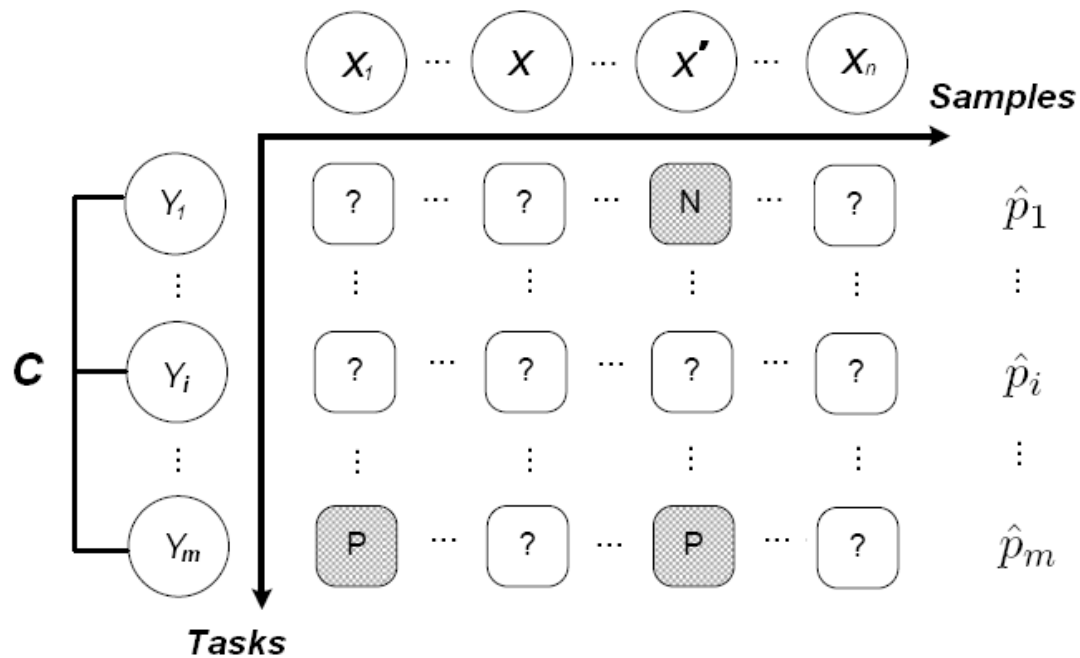
- Maximizing the joint mutual information is reasonable
- Directly estimate the *joint* label probability
  - Recognize the correlation between labels
  - Need more labeled examples
  - What if # tasks is large?
  - Cannot use specialized models for each task
  - Can we use external knowledge to couple tasks?

Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL' 08)
  - Image Classification (CVPR' 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories

Constraint-Driven Multi-Task Active Learning

- Multiple tasks Y_1, Y_2, …, Y_m
- Learners for each task $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m$
- A set of constraints *C* among tasks
- May have new tasks to launch



Value of Information (VOI) for Active Learning

- Single-task AL
  - Value of information (VOI) for labeling a sample ***x***

$$VOI(Y,\mathbf{x}) = VOI(\mathbf{x}) = \sum_{y \in Dom(Y)} P(Y = y|\mathbf{x}) R(Y = y, \mathbf{x})$$

Value of Information (VOI) for Active Learning

- Single-task AL
  - Value of information (VOI) for labeling a sample ***x***

$$VOI(Y, \mathbf{x}) = VOI(\mathbf{x}) = \sum_{y \in Dom(Y)} P(Y = y|\mathbf{x}) R(Y = y, \mathbf{x})$$

  - Reward *R*(*Y*=*y*, ***x***), e.g., how surprising it is?

$$R(Y = y, \mathbf{x}) = -\log_2 \hat{p}(Y = y|\mathbf{x})$$

Value of Information (VOI) for Active Learning

- Single-task AL
  - Value of information (VOI) for labeling a sample ***x***

$$VOI(Y,\mathbf{x}) = VOI(\mathbf{x}) = \sum_{y \in Dom(Y)} P(Y=y|\mathbf{x}) R(Y=y,\mathbf{x})$$

  - Reward $R(Y=y, \boldsymbol{x})$, e.g., how surprising it is?

$$R(Y=y,\mathbf{x}) = -\log_2 \hat{p}(Y=y|\mathbf{x})$$

  - Finally, replace $P(Y=y|\boldsymbol{x})$ with $\hat{p}$

$$VOI(\mathbf{x}) = \sum_{y \in Dom(Y)} -\hat{p}(Y=y|\mathbf{x}) \log_2 \hat{p}(Y=y|\mathbf{x})$$

Constraint-Driven Active Learning

- Multiple tasks with constraints

$$VOI(Y_i, \mathbf{x}) = \sum_{y_i \in Dom(Y_i)} P(Y_i = y_i|\mathbf{x}) R(Y_i = y_i, \mathbf{x})$$

$$\mathbf{x} \in \mathbf{D}, Y_i \in U(\mathbf{x})$$

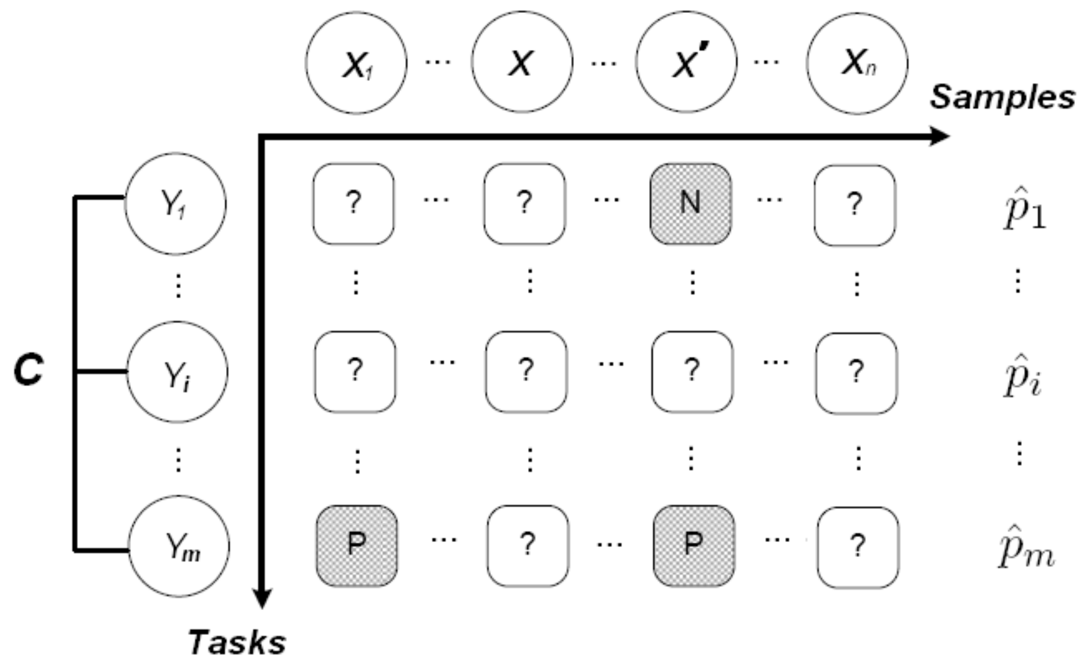
- Probability estimate of outcomes

$$P(Y_i = y_i|\mathbf{x}) = \begin{cases} \hat{p}_i(Y_i = y_i|\mathbf{x}) & \text{if } Y_i \text{ is a learned task} \\ \frac{1}{|Dom(Y_i)|} & \text{if } Y_i \text{ is a new task} \end{cases}$$

Constraint-Driven Active Learning

- Reward function $R(y,\mathbf{x})$ in:

$$VOI(Y_i, \mathbf{x}) = \sum_{y_i \in Dom(Y_i)} P(Y_i = y_i | \mathbf{x}) R(Y_i = y_i, \mathbf{x})$$

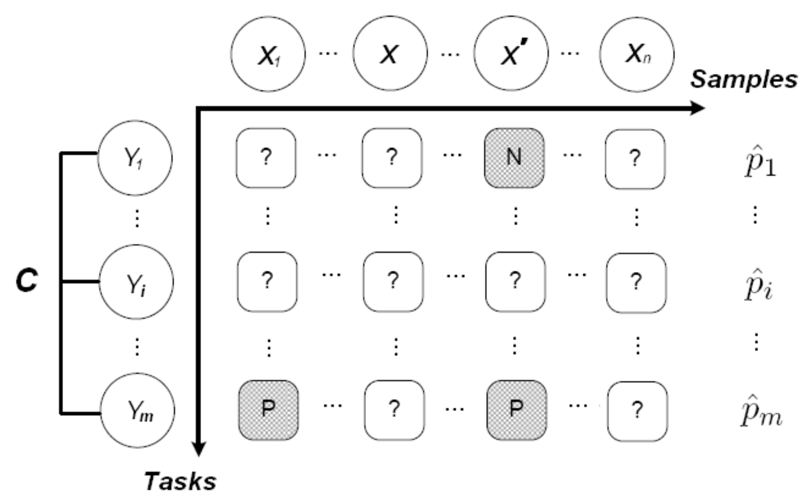


Constraint-Driven Active Learning

- Propagate rewards via constraints

$$R(Y_i = y_i, \mathbf{x}) = \sum_{(Y_j = y_j) \in Prop(Y_i = y_i, \mathbf{x}, \mathbf{C})} -\log_2 \hat{p}_j(Y_j = y_j | \mathbf{x})$$

$$Prop(Y_i = y_i, \mathbf{x}, \mathbf{C}) = \{(Y_j = y_j) | (Y_i = y_i) \dashrightarrow_{\mathbf{C}} (Y_j = y_j), Y_j \in \mathbf{U}(\mathbf{x})\}$$



Constraint-Driven Active Learning

- Multi-task AL with constraints

$$VOI(Y_i, \mathbf{x}) = \sum_{y_i \in Dom(Y_i)} \hat{p}_i(Y_i = y_i|\mathbf{x}) \sum_{(Y_j = y_j) \in Prop(Y_i = y_i, \mathbf{C})} -\log_2 \hat{p}_j(Y_j = y_j|\mathbf{x})$$

$$\mathbf{x} \in \mathbf{D}, Y_i \in U(\mathbf{x})$$

  - Recognize *inconsistency* of among tasks
  - Launch new tasks
  - Favor poorly performed tasks, and “pivot” tasks
  - Density-weighted measure?
  - Use state-of-the-art learners for single tasks

Experiments

- Four named entity recognition tasks
  - “Animal”
  - “Mammal”
  - “Food”
  - “Celebrity”
- Constraints
  - 1 inheritance, 5 mutual exclusion
  - Lead to 12 propagation rules (plus 1 identity rule)

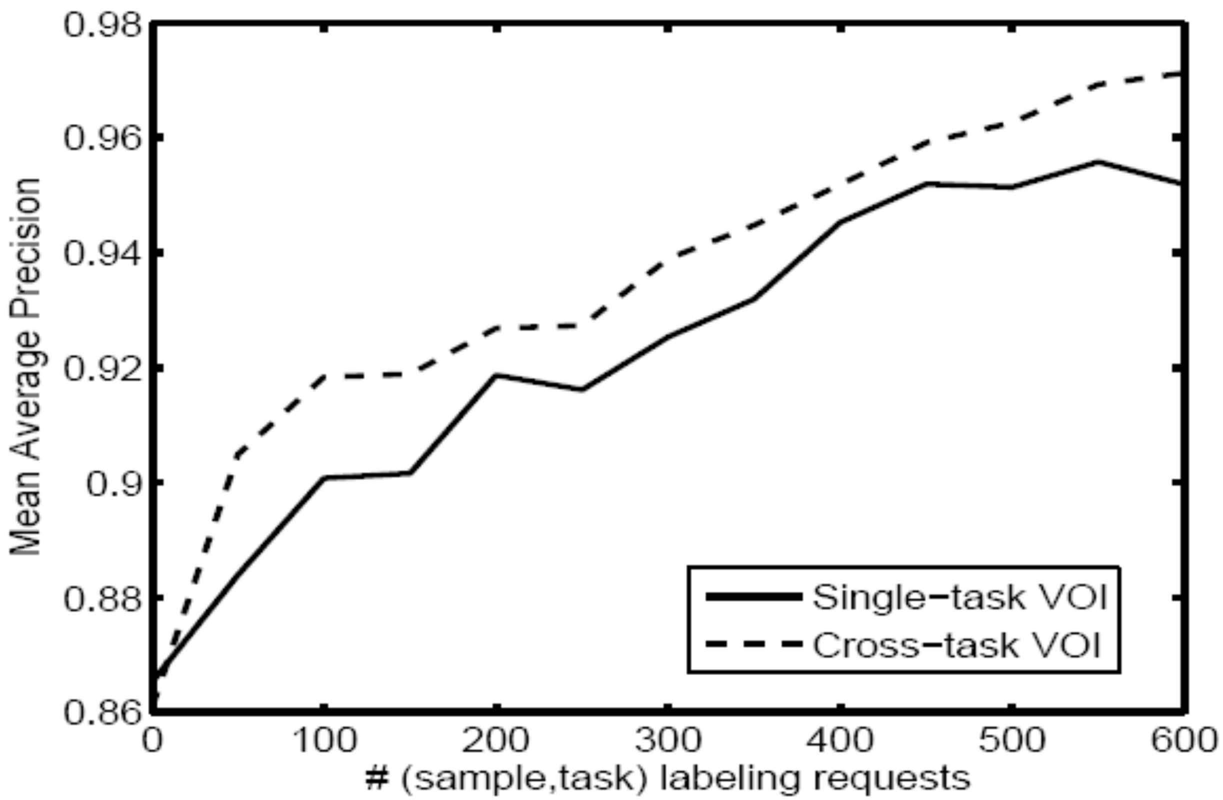
Experiments

- Competitive methods for AL
  - VOI of sample-task pairs with constraints
  - VOI of sample-task pairs without constraints
    - Single-task AL

$$VOI(Y_i, \mathrm{x}) = \sum_{y_i \in Dom(Y_i)} \hat{p}_i(Y_i = y_i|\mathrm{x}) \sum_{(Y_j = y_j) \in Prop(Y_i = y_i, \mathbf{C})} -\log_2 \hat{p}_j(Y_j = y_j|\mathrm{x})$$

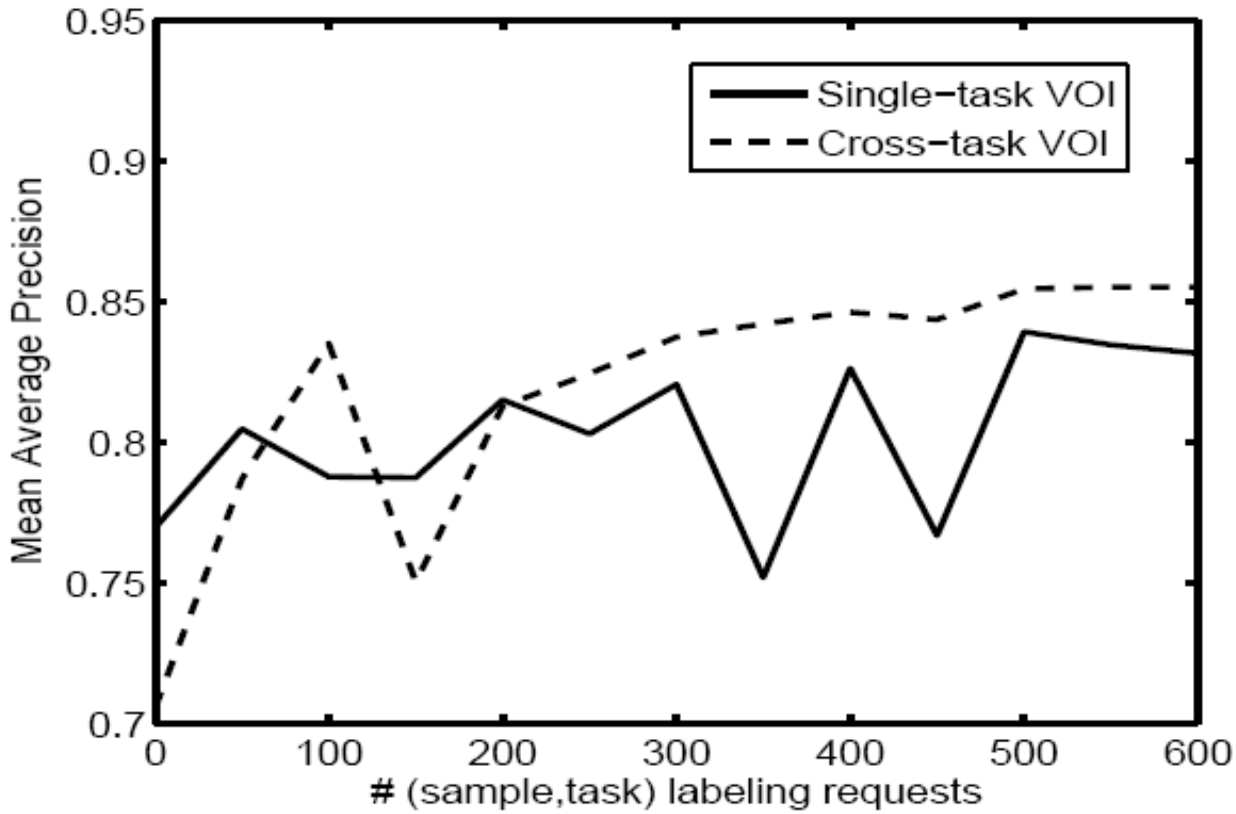
Experiments

- Results: MAP on animal, food and celebrity



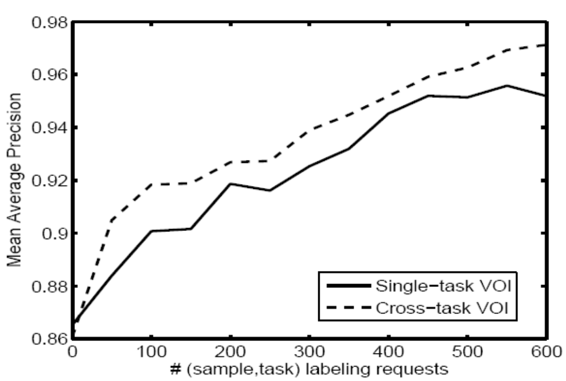
Experiments

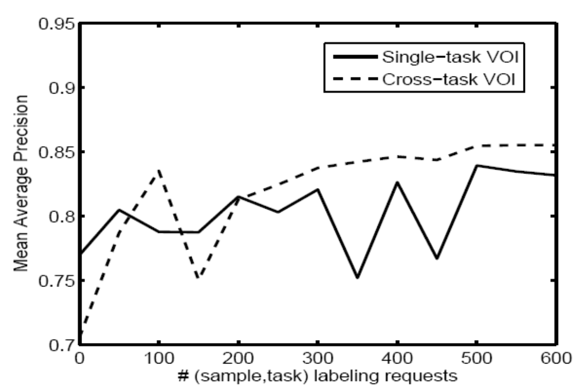
- Results: MAP on all four tasks



Experiments

- Analysis
  - True labels from the NNLL system
  - 90% precision for “mammal”
    - 10% label noise on the task “mammal”
  - Tasks are generally “easy”
    - Positive examples are highly homogenous



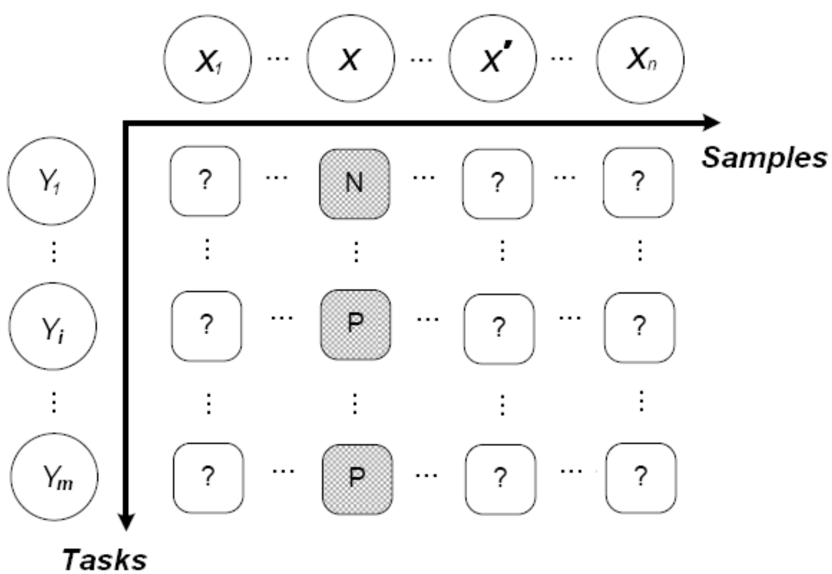


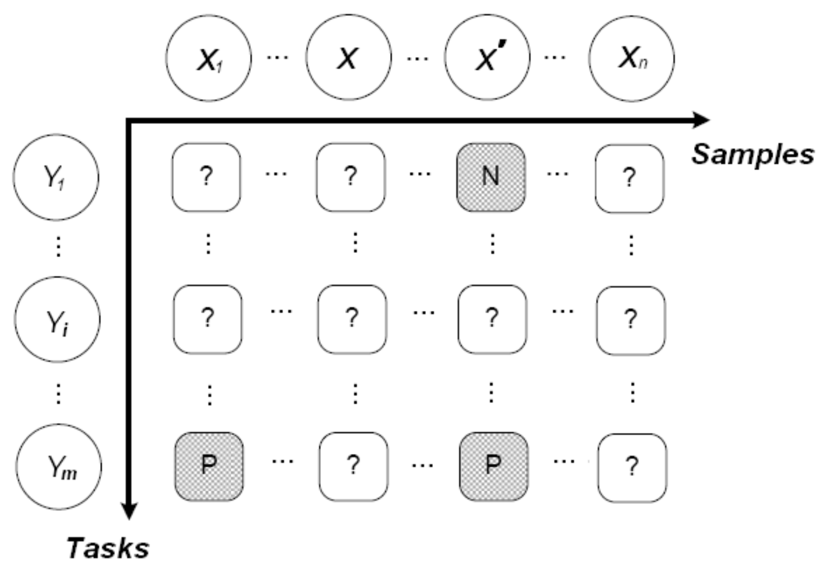
Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL' 08)
  - Image Classification (CVPR' 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories

Cost-Sensitive Active Learning Across Tasks

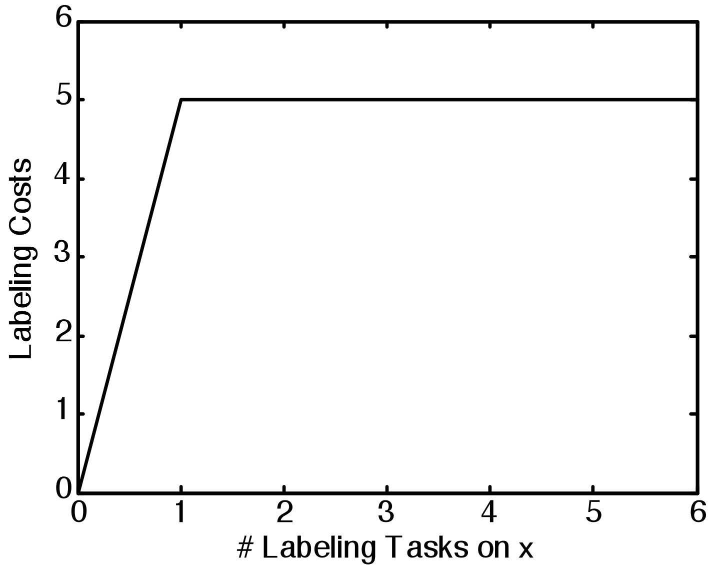
- Which scenario is reasonable?
  - Choose one sample, label all tasks
  - Arbitrary sample-label pairs

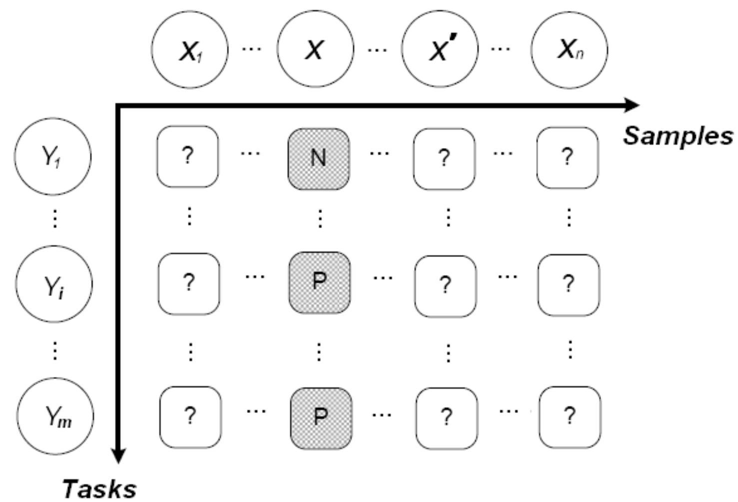




Cost-Sensitive Active Learning Across Tasks

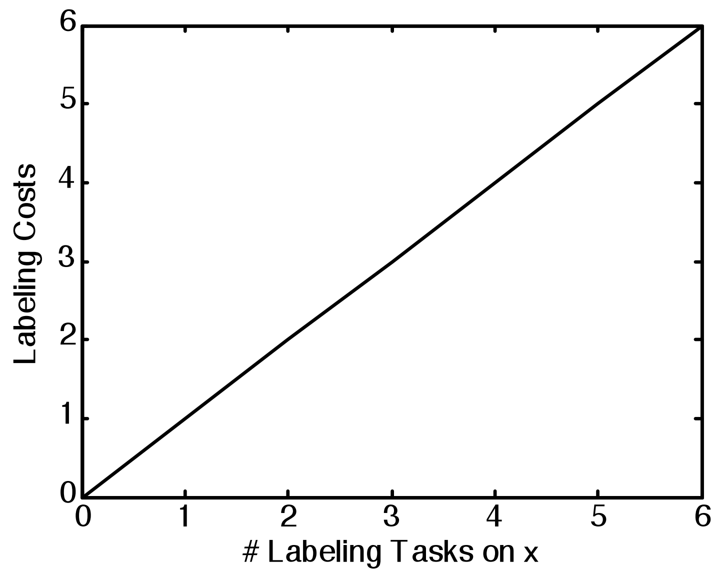
- Costs for labeling multi tasks on a sample ***x***
  - ***x*** is a long document

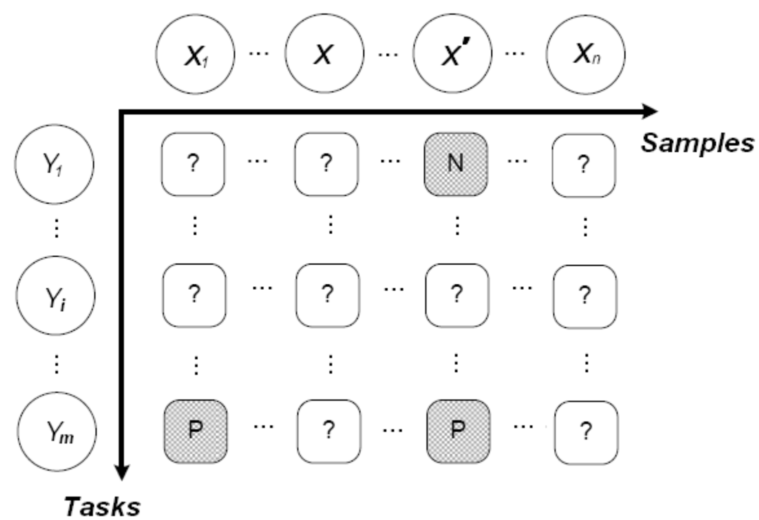




Cost-Sensitive Active Learning Across Tasks

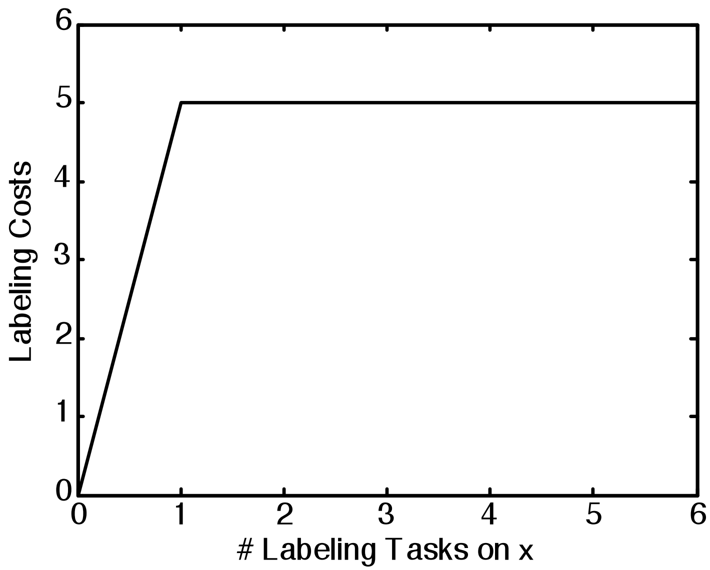
- Costs for labeling multi tasks on a sample ***x***
  - ***x*** is a word or an image

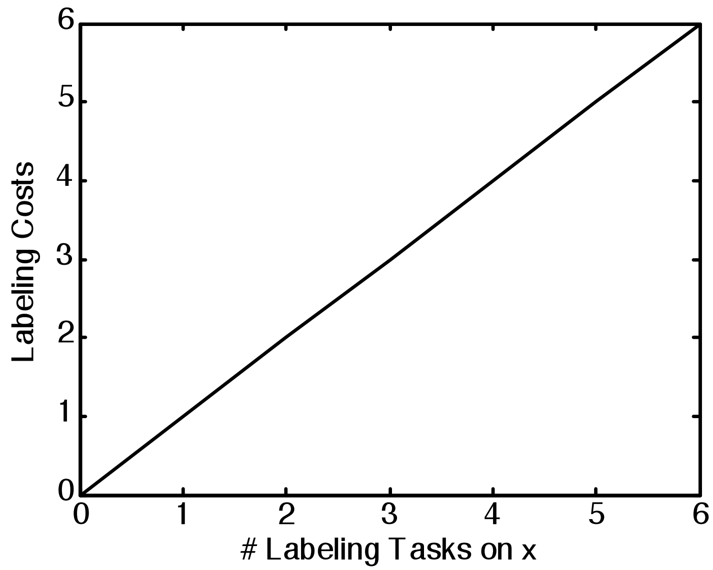




Cost-Sensitive Active Learning Across Tasks

- Learn a more realistic cost function?
- Active learning aware of labeling costs?





Outline

- Active Learning
- Multi-Task Active Learning
  - Linguistic Annotations (ACL' 08)
  - Image Classification (CVPR' 08)
- Current Work and Discussions
  - Constraint-Driven Active Learning Across Tasks
  - Cost-Sensitive Active Learning Across Tasks
  - Active Learning of Constraints and Categories

Active Constraint Learning

- New constraints/rules are highly valuable
- Find significant rules and avoid false discovery
  - Oversearching (Quinlan, et al. IJCAI’ 95)
  - Multiple comparisons (Jensen, et al. MLJ’ 00)
  - Statistical tests (Webb, MLJ’ 06)
- Combining first-order logic with graphical models
  - Bayesian logic programs (logic + BN)
  - Markov logic networks (logic + MRF)
  - Structure sparsity on graphs?

Active Category Detection

- Automatically detect new categories
- Clustering
  - High-dimensional space
  - Co-clustering/bi-clustering
  - Local search vs. global partition
- Subgraph/community detection
  - A huge bipartite graph
  - Optimize modularity of the graph
  - Overlapping communities?

Thanks!

- Questions?