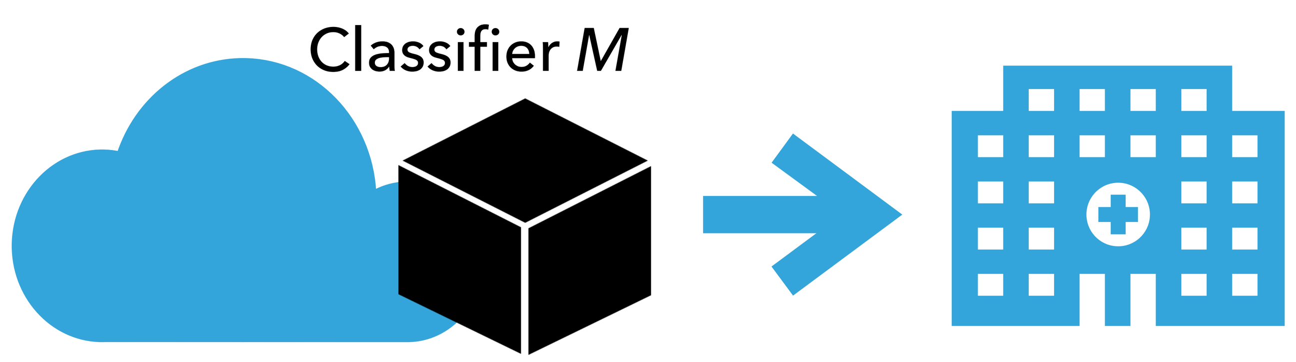
Active Bayesian Assessment for Black-Box Classifiers

Disi Ji[1], Robert L. Logan IV[1], Padhraic Smyth[1], Mark Steyvers[2]

[1]Department of Computer Science, [2]Department of Cognitive Sciences

University of California, Irvine

Objectives



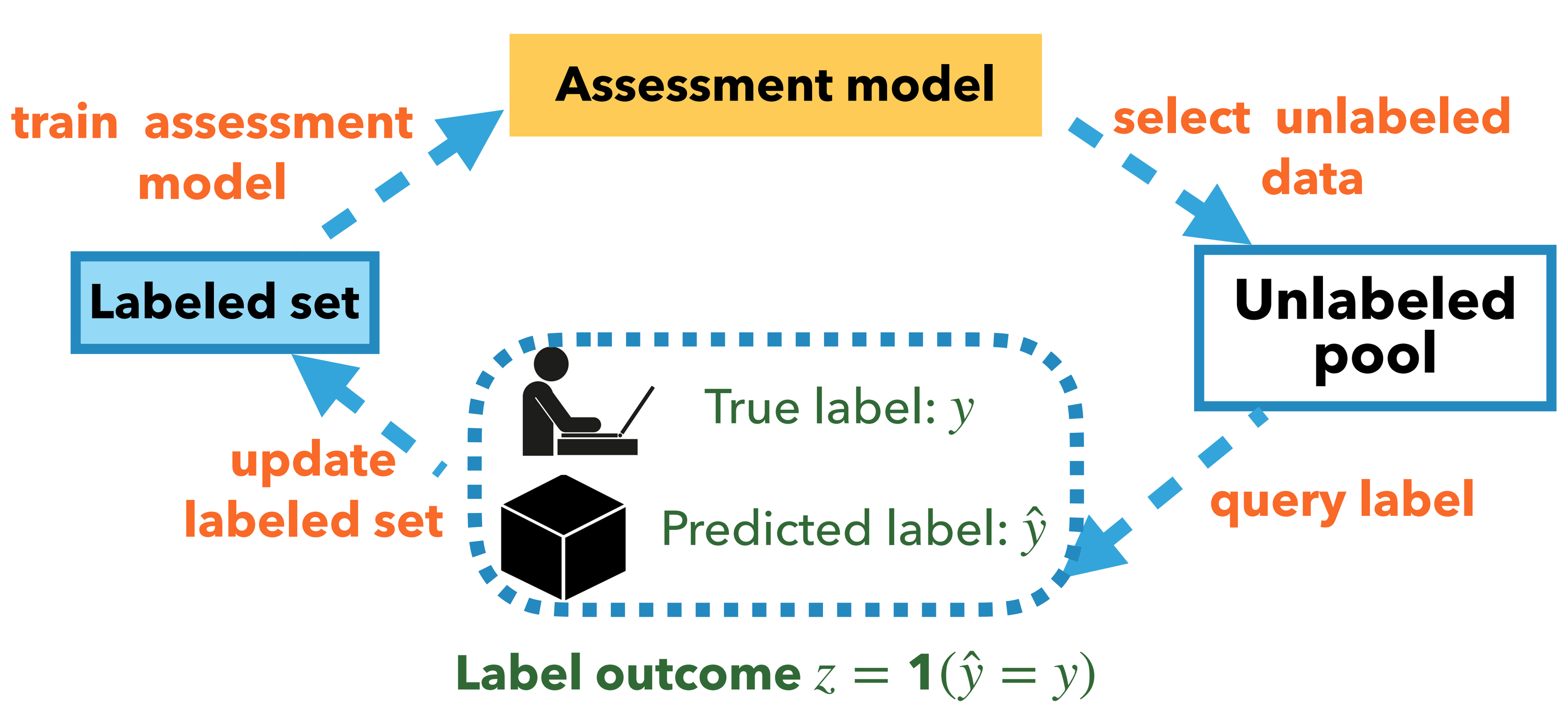
- **Estimation**: How accurate?
- **Identification**: Where is the model least accurate?
- **Comparison**: Is the model fair, e.g. equally accurate across different groups?

(Can replace accuracy with other performance metrics, e.g., calibration metrics)

Requires labeled data!

- How much **confidence** should we have in this assessment?
- How best to **increase our confidence** given a limited budget for labeled data?

Overview: Active Bayesian Assessment



- **Key assumption:** availability of a pool of unlabeled data
- **Main idea**: we propose to actively labeling data points by iterating between labeling and assessing
- **Assessment model:** Bayesian assessment for confidence quantification
- **Select unlabeled data:** Thompson Sampling

Bayesian Assessment with Uncertainty

Performance metric of interest θ

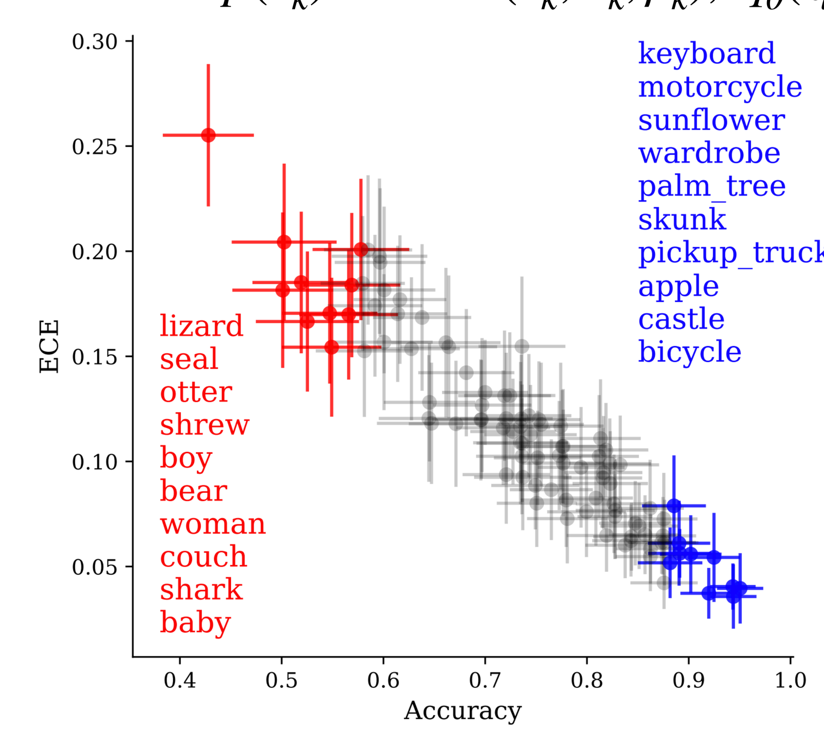
Labeled data: $D = \{(x_i, y_i) \mid i = 1,2,\cdots,N\}$, label outcome: $z_i = 1(y_i = \hat{y}_i)$

posterior / prior / label outcome likelihood

$$p(\theta|\mathcal{D}) = \frac{p(\theta)\cdot\prod_{i=1}^{N} q_\theta(z_i)}{\int_\theta p(\theta)\cdot\prod_{i=1}^{N} q_\theta(z_i)\, d\theta}$$

e.g. accuracy of the k-th predicted class:

$$p(\theta_k) = \mathrm{Beta}(\theta_k; \alpha_k, \beta_k),\;\; q_\theta(z_i) = \mathrm{Bern}(z_i; \theta_k)$$



CIFAR100

- 100 balanced classes
- 50,000 images for training
- 10,000 images for testing
- prediction model: ResNet model with110 layers
- overall accuracy on all test data: ~80%

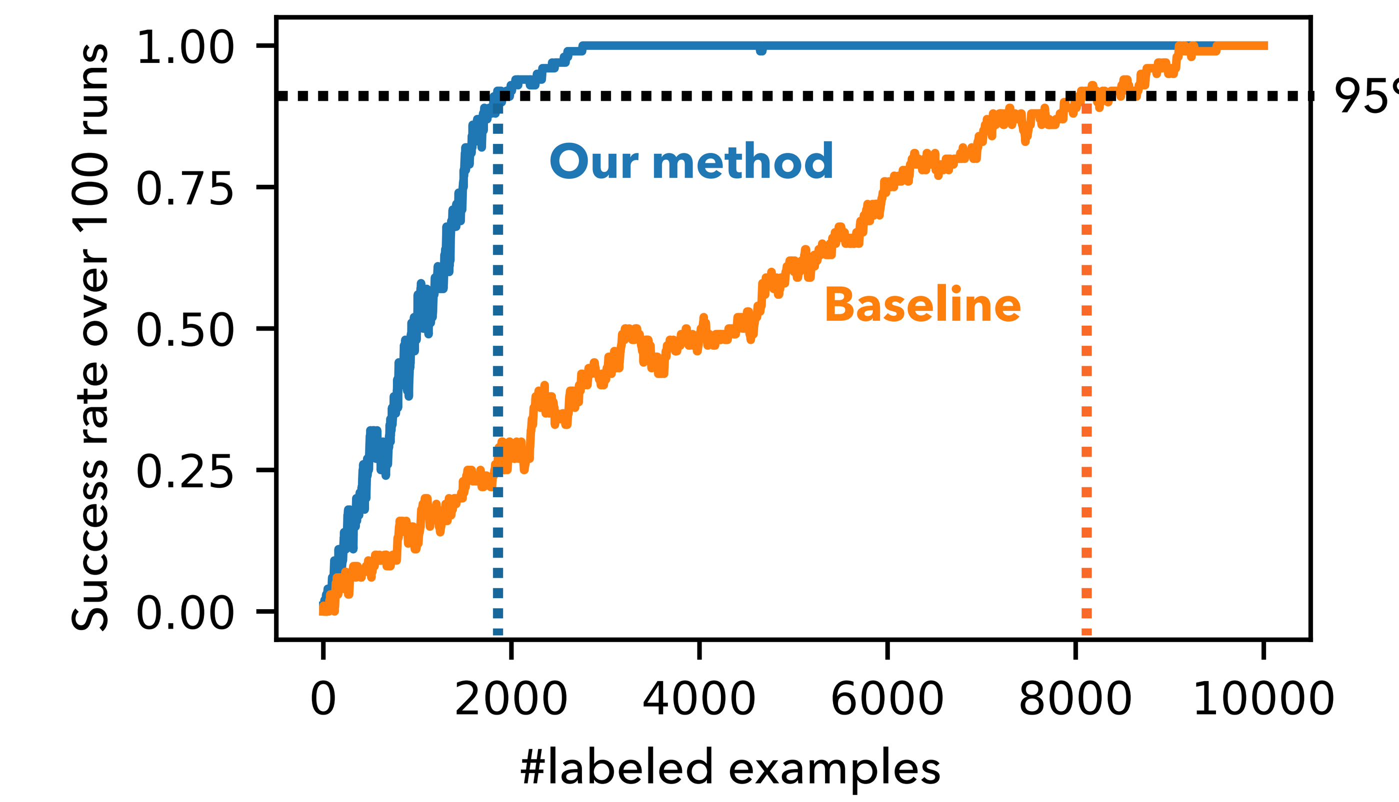
Active Assessment with Thompson Sampling

Algorithm 1 Thompson Sampling(p, q, r, M)

Initialize the priors on metrics $\{p_0(\theta_1), \ldots, p_0(\theta_g)\}$
for $i = 1, 2, \cdots$ **do**
  # Sample parameters for the metrics θ
  $\tilde{\theta}_g \sim p_{i-1}(\theta_g), g = 1, \ldots, G$
  # Select a group g (or arm) by maximizing expected reward
  $\hat{g} \leftarrow \arg\max_g \mathbb{E}_{q_{\tilde{\theta}}}[r(z|g)]$
  # Randomly select an input data point from group $\hat{g}$ and compute its predicted label
  $\mathbf{x}_i \sim \mathcal{R}_{\hat{g}}$
  $\hat{y}_i(\mathbf{x}_i) = \arg\max_k p_M(y = k|\mathbf{x}_i)$
  # Query to get a true label (pull arm $\hat{g}$)
  $z_i \leftarrow f(y_i, \hat{y}_i(\mathbf{x}_i))$
  # Update parameters of the $\hat{g}$th metric
  $p_i(\theta_{\hat{g}}) \propto p_{i-1}(\theta_{\hat{g}}) q(z_i|\theta_{\hat{g}})$
end for

- (p, q, r) are task specific. $p(\theta)$ is the prior distribution of metric θ, $q_\theta(z \mid g)$ is likelihood of the label outcome z for the g-th group, and $r(z \mid g)$ is the corresponding reward function.

Illustrative Results: Actively Identify the Least Accurate Class of CIFAR100



- Percentage of labeled samples needed to identify the least accurate classes **dropped by 71%**
- We obtained similar performance gain for other assessment tasks (full results in paper)

Our Contribution:

- Developed a general **Bayesian framework** to assess classification performance metrics, including
  - (1) accuracy, reliability diagram, ECE;
  - (2) performance difference;
  - (3) confusion matrix, misclassification cost, etc
- Developed an **active assessment framework** for
  - (1) estimation of model performance;
  - (2) identification of model deficiencies;
  - (3) performance comparison between groups
- Demonstrated that our proposed approaches need significantly fewer labels than baselines

Acknowledgements

This material is based upon work supported in part by the National Science Foundation under grants number 1900644 and 1927245, by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C002, and by a Qualcomm Faculty Award (PS). This work was also partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at the University of California, Irvine