Efficient Active Learning

Alina Beygelzimer, IBM Research
Daniel Hsu, Rutgers University
Nikos Karampatziakis, Cornell University
John Langford, Yahoo! Research
Tong Zhang, Rutgers University

Motivation

A lot of unlabeled data is plentiful and cheap, eg.

- documents off the web
- speech samples
- images and video

But labeling can be expensive.

Active Learning: Can interaction help us learn effectively?

The Active Learning Setting

Repeatedly:

1. Observe unlabeled example x and predict.
2. Decide whether to query label.
3. If yes, observe label y.

Goal: Get great classifier with few requested labels

Typical heuristics for active learning

Start with few labels
Repeat
  Fit a classifier to the labels seen so far
  Query the unlabeled point that is closest to the boundary
  (or most uncertain, or most likely to decrease overall uncertainty,...)

Fail to capture tradeoff between exploration and exploitation.

Manifestation in practice, eg. Schutze et al 03.

A Theoretically Sound Approach

- Maintain a version space.
- Each new label eliminates some hypotheses.
- Noise free case easy (CAL 1994)
- Noisy setting via confidence bounds (A^2 algorithm BBL 2006).

Importance Weighted Active Learning

$$\mathrm{err}(h, S_n) := \frac{1}{n} \sum_{(X_i, Y_i, 1/P_i) \in S_n} \frac{1}{P_i} \cdot \mathbb{I}[h(X_i) \neq Y_i] \quad (1)$$

Algorithm: Basic Agnostic Active Learner

Initialize: $S_0 := \emptyset$.

For $k = 1, 2, \ldots, n$:

1. Obtain unlabeled data point X_k.
2. Let
   $h_k := \arg\min\{\mathrm{err}(h, S_{k-1}) : h \in \mathcal{H}\}$, and
   $h'_k := \arg\min\{\mathrm{err}(h, S_{k-1}) : h \in \mathcal{H} \ \wedge \ h(X_k) \neq h_k(X_k)\}$.
   Let $G_k := \mathrm{err}(h'_k, S_{k-1}) - \mathrm{err}(h_k, S_{k-1})$, and

$$P_k := \begin{cases} 1 \text{ if } G_k \leq \sqrt{\frac{C_0 \log k}{k-1}} + \frac{C_0 \log k}{k-1} \\ s \text{ otherwise} \end{cases}$$

   where $s \in (0, 1)$ is the positive solution to the equation

$$G_k = \left(\frac{c_1}{\sqrt{s}} - c_1 + 1\right) \cdot \sqrt{\frac{C_0 \log k}{k-1}} + \left(\frac{c_2}{s} - c_2 + 1\right) \cdot \frac{C_0 \log k}{k-1}. \quad (2)$$

3. Toss a biased coin with $\Pr(\text{heads}) = P_k$.
   If heads, query Y_k, and let $S_k := S_{k-1} \cup \{(X_k, Y_k, 1/P_k)\}$.
   Else, let $S_k := S_{k-1}$.

Return: $h_{n+1} := \arg\min\{\mathrm{err}(h, S_n) : h \in \mathcal{H}\}$.

Theorems

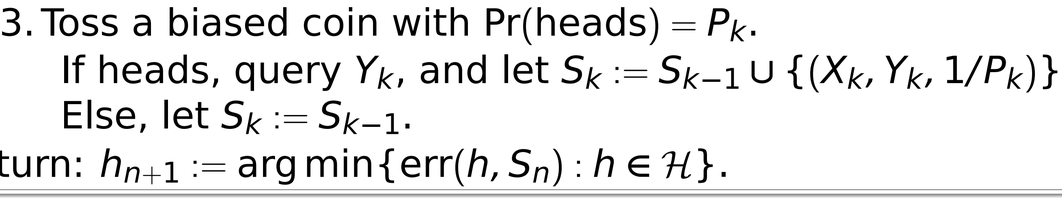
(Consistency, BDL 2009) For all methods choosing $p > 0$, the algorithm is consistent.

(Accuracy, BHLZ 2010) With high probability, the IWAL reduction has a similar error rate to supervised learning on k points.

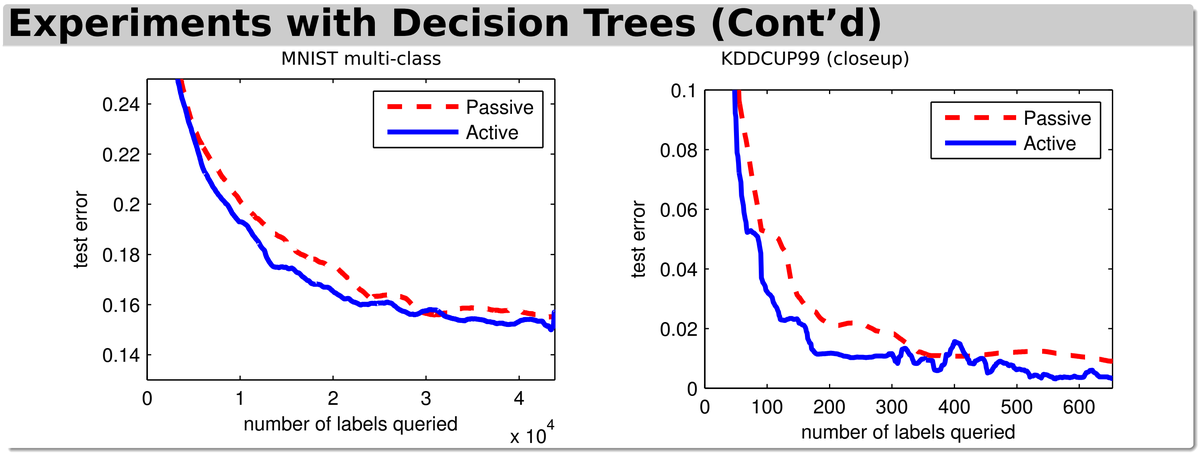
(Efficiency, BHLZ 2010) If there is a small disagreement coefficient θ, the queried labels are only $O\left(\theta\sqrt{k \log k}\right)$ + a minimum due to noise (K 2006).

Experiments with Decision Trees

- ERM hypothesis h = output of C4.5
- Alternative hypothesis = flip x_k's leaf in h.



Experiments with Decision Trees (Cont'd)



Active Learning with Online Gradient

- ERM hypothesis h = current iterate w_k
- Alternative hypothesis h' not necessary
- Only need: Δ difference in error rates
- Tricky: large importance weights (small probabilities)

Principle: Having an example with importance weight i should be equivalent to having the example i times in the dataset.

Importance Invariant Updates

- Losses for linear models $\ell(w^\top x, y)$. $\nabla_w \ell = \frac{\partial \ell(p,y)}{\partial p} x$
- Update must be given by $w_{t+1} = w_t - s(p, i)x$
- $s(p, i)$ must satisfy $\frac{\partial s}{\partial i} = \eta \left.\frac{\partial \ell(p,y)}{\partial p}\right|_{p=(w_t - s(p,i)x)^\top x}$ $\quad s(p, 0) = 0$
- Properties
  - Invariance: $s(p, a + b) = s(p, a) + s(p - s(p, a)||x||^2, b)$
  - Safety, No regret (fallback), Closed form for many losses

Implicit updates

- Solve $\mathrm{argmin} \frac{1}{2}||w - w_t||^2 + i\eta\ell(w^\top x_t, y_t)$
- Safe - No regret - Closed form only for squared, hinge, quantile loss - Invariant only for hinge, quantile loss

Estimating Difference in Error Rates

Suppose x_k didn't have the label preferred by h.

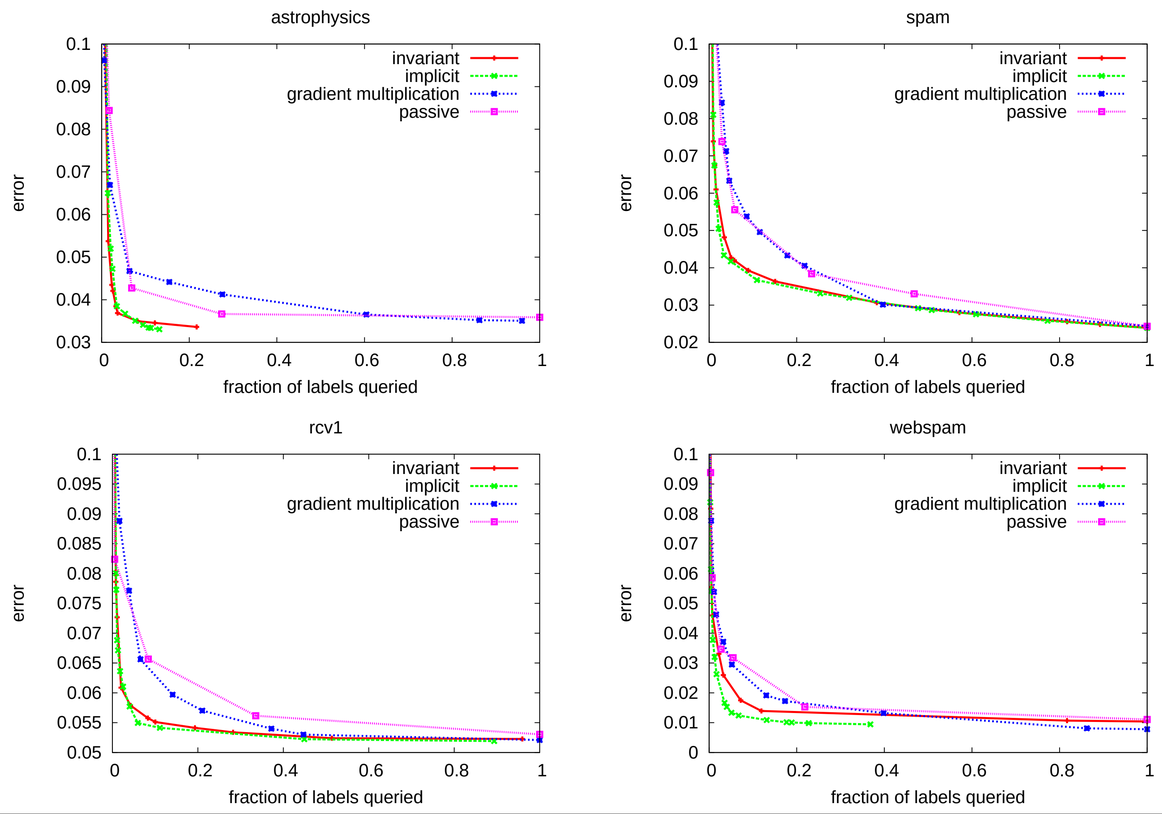
$y_a = -\mathrm{sign}(h(x_k))$.

Present (x_k, y_a) repeatedly i_k times until $h(x_k) = y_a$.

For the invariant updates $i_k = \frac{1}{\eta_k} \int_0^{\frac{w_k^\top x_k}{||x_k||^2}} \frac{dt}{\left.\frac{\partial \ell(p,y_a)}{\partial p}\right|_{p=w_k^\top x_k - t||x_k||^2}}$

- No unified formula for implicit. For logistic loss: $i_k = -\frac{2w_k^\top x_k}{\eta_k y_a ||x_k||^2}$
- Then $\Delta_k \approx i_k / k$

Online Gradient Descent Results



How fast is it?

- As fast as (passive) online gradient descent
  - Train on RCV1 (≈780K docs 77 features/doc): 2.6 sec.
  - Passive online gradient descent takes 2.5 sec
  - 91K queries (11%) 0.045 squared error (vs. 0.041)
- Faster!
  - On RCV1 with 3500 features/doc 2 min.
  - Passive gradient descent takes 3 min.
  - 89K queries 0.042 squared error (vs. 0.038)

References

Agnostic Active Learning without Constraints, BHLZ NIPS'10.
Online Importance Weight Aware Updates, KL UAI'11.
Implementation available in Vowpal Wabbit

VW: Fork it from `http://github.com/JohnLangford/vowpal_wabbit`