Exponential Computational Improvement by Reduction

John Langford @ Microsoft Research 

Computational Challenges Workshop, May 2

The Empirical Age

Speech Recognition

The Empirical Age

Speech Recognition
ImageNet

The Empirical Age

Speech Recognition
ImageNet
Deep Learning

The Empirical Age

Speech Recognition
ImageNet
Deep Learning
Neural Machine Translation

The Empirical Age

Speech Recognition
ImageNet
Deep Learning
Neural Machine Translation

What is a theorist to do?

The Contextual Bandit Setting

For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$
2. The learner chooses an action $a \in A$
3. The world reacts with reward $r_a \in [0, 1]$

Goal: Learn a good policy for choosing actions given context.

Reduction Results

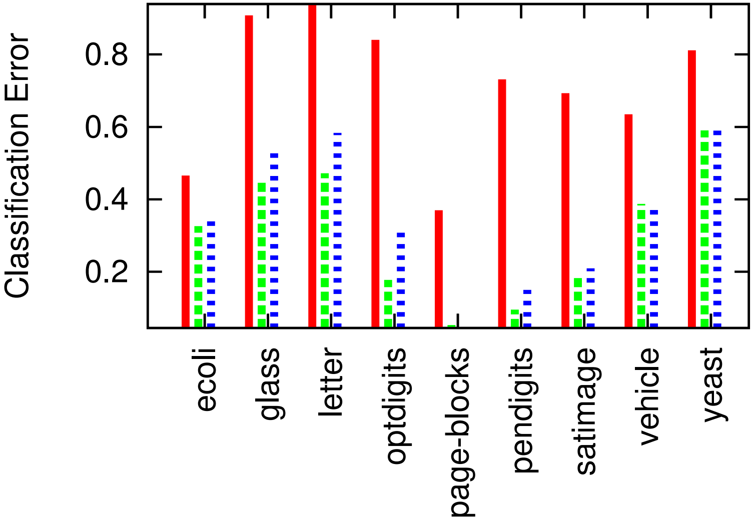
Algo	ϵ-greedy	Bagg	LinUCB	Online C.	Super.
Loss	0.095	0.059	0.128	0.053	0.051
Time	22	339	212×10^3	17	6.9

Progressive validation loss on RCV1.

The Offline Contextual Bandit Setting

Given exploration data $(x, a, r, p)^*$
Learn a good policy for choosing actions given context.

Offline Reduction Results



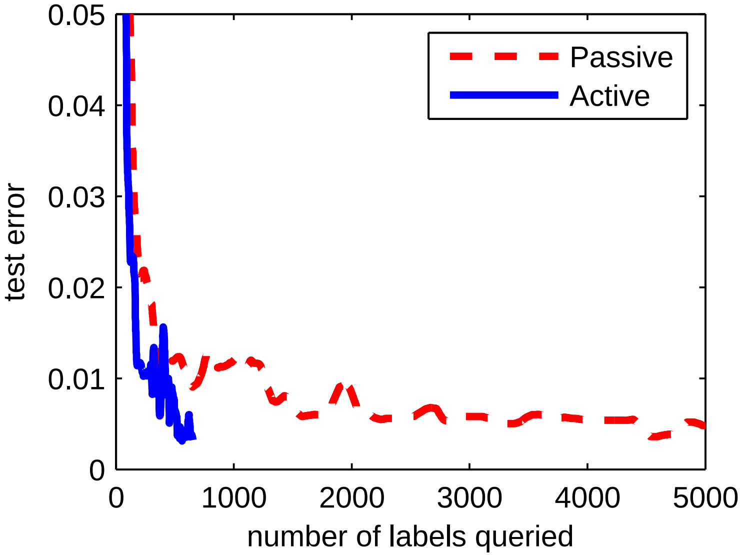
Agnostic Active Learning

For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$
2. The learner predicts a label $\hat{y} \in Y$
3. The learner chooses to request a label or not. If label requested:
    1. observe y
    2. update learning algorithm

Goal: Compete with supervised learning using all labels while requesting as few as possible.

AAL Reduction Results



Logarithmic Time Prediction

Repeatedly

1. See x
2. Predict $\hat{y} \in \{1, ..., K\}$
3. See y

Logarithmic Time Prediction

Repeatedly

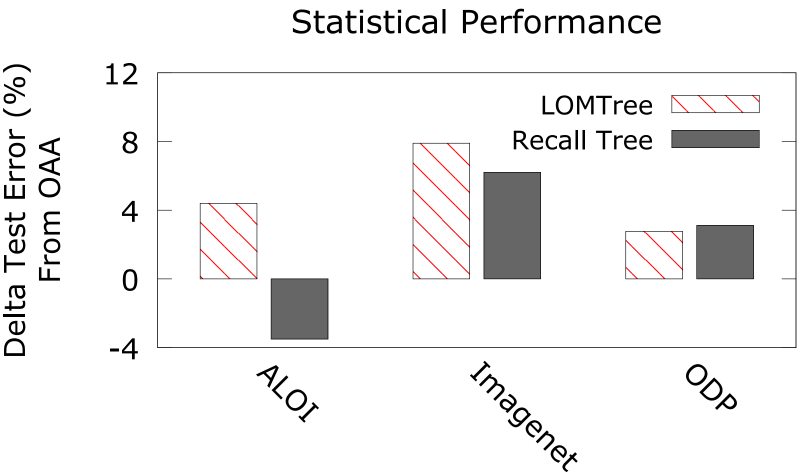
1. See x
2. Predict $\hat{y} \in \{1, \ldots, K\}$
3. See y

Goal: Find $h(x)$ minimizing error rate:

$$\Pr_{(x,y)\sim D}(h(x) \neq y)$$

with $h(x)$ in time $O(\log K)$.

Log-time prediction results



Summary

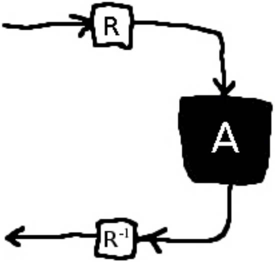
Problem	Learning Reductions	OCO	PAC
CB Explore	Yes	Sorta?	No
CB Learn	Yes	Sorta?	No
Agnostic Active	Yes	Sorta?	No
Log-time	Yes	No	No

Outline

1. Why Reductions
2. What is a learning reduction?
3. Exponential Improvements

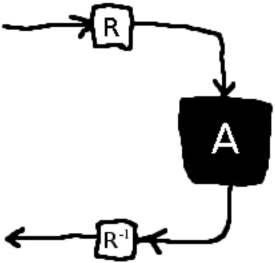
Learning Reduction Basics

Goal: minimize ℓ on D



Learning Reduction Basics

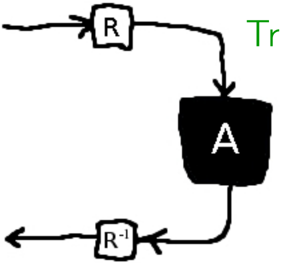
Goal: minimize ℓ on D



ℓ' optimizer

Learning Reduction Basics

Goal: minimize ℓ on D

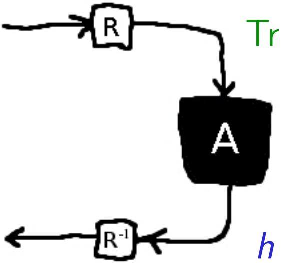


Transform D into D_R

ℓ' optimizer

Learning Reduction Basics

Goal: minimize ℓ on D



Transform D into D_R

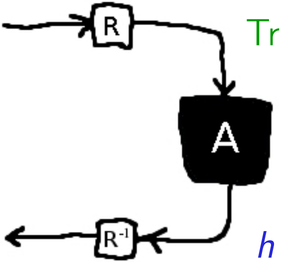
ℓ' optimizer

h

Transform h with small $\ell'(h, D_R)$ into R_h with small $\ell(R_h, D)$...

Learning Reduction Basics

Goal: minimize ℓ on D



Transform D into D_R

ℓ' optimizer

h

Transform h with small $\ell'(h, D_R)$ into R_h with small $\ell(R_h, D)$...

such that if h does well on (D_R, ℓ'), R_h is guaranteed to do well on (D, ℓ).

Error Reductions: the simplest possible

Prove: Small ℓ' error $\Rightarrow$ small ℓ error.

Error Reductions: the simplest possible

Prove: Small ℓ' error $\Rightarrow$ small ℓ error.

An issue: If R introduces noise, small ℓ' not possible.

Error Reductions: the simplest possible

Prove: Small ℓ' error $\Rightarrow$ small ℓ error.

An issue: If R introduces noise, small ℓ' not possible.

$\Rightarrow$ Must prove small ℓ' possible for nonvacuous statement.

Error Reductions: the simplest possible

Prove: Small ℓ' error $\Rightarrow$ small ℓ error.

An issue: If R introduces noise, small ℓ' not possible.

$\Rightarrow$ Must prove small ℓ' possible for nonvacuous statement.
$\Rightarrow$ Error reductions weak for noisy problems.

Regret Reductions: Dealing with noise

Let $\text{reg}_{\ell,D} = \ell(h, D) - \min_{h'} \ell(h', D)$

Regret Reductions: Dealing with noise

Let $\text{reg}_{\ell,D} = \ell(h, D) - \min_{h'} \ell(h', D)$

Prove: Small $\text{reg}_{\ell',D'} \Rightarrow$ small $\text{reg}_{\ell,D}$.

Regret Reductions: Dealing with noise

Let $\mathrm{reg}_{\ell,D} = \ell(h, D) - \min_{h'} \ell(h', D)$

Prove: Small $\mathrm{reg}_{\ell',D'} \Rightarrow$ small $\mathrm{reg}_{\ell,D}$.

Note: $\min_{h'}$ is over <u>all</u> functions.

Regret Reductions: Dealing with noise

Let $\text{reg}_{\ell,D} = \ell(h, D) - \min_{h'} \ell(h', D)$

Prove: Small $\text{reg}_{\ell',D'} \Rightarrow$ small $\text{reg}_{\ell,D}$.

Note: $\min_{h'}$ is over all functions.
$\Rightarrow$ User is responsible for choosing right hypothesis space.

Regret Reductions: Dealing with noise

Let $\mathrm{reg}_{\ell,D} = \ell(h,D) - \min_{h'} \ell(h',D)$

Prove: Small $\mathrm{reg}_{\ell',D'} \Rightarrow$ small $\mathrm{reg}_{\ell,D}$.

Note: $\min_{h'}$ is over all functions.
$\Rightarrow$ User is responsible for choosing right hypothesis space.
$\Rightarrow$ Unable to address information gathering

Oracle Reductions: Information gathering

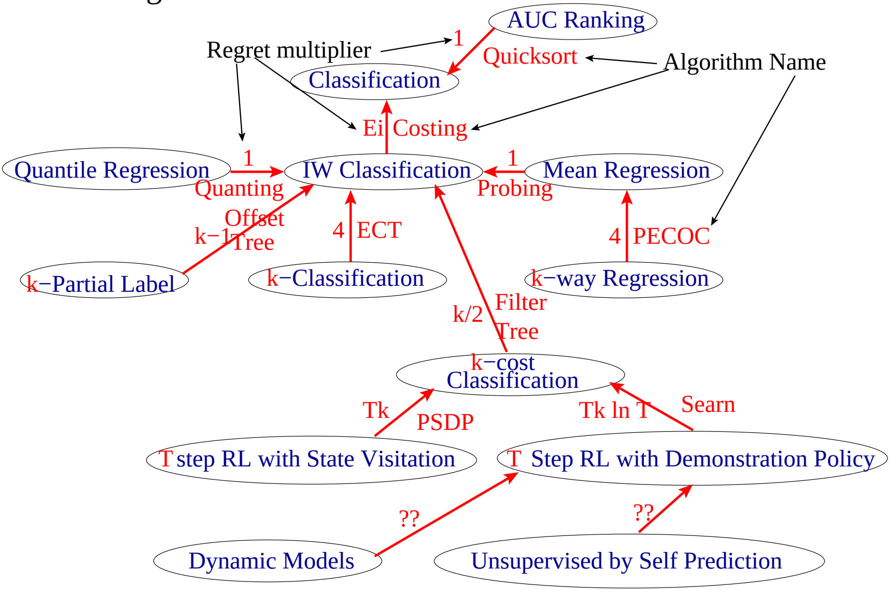
Assume an oracle which given samples S returns $\arg\min_{h \in H} \ell'(h, S)$

Oracle Reductions: Information gathering

Assume an oracle which given samples S returns $\arg\min_{h \in H} \ell'(h, S)$

Prove: Oracle (approximately) works $\Rightarrow$ Computationally efficient small online regret on original problem.

Regret Transform Reductions



Programming

Reductions $\Rightarrow$ modularity, code reuse $\Rightarrow$ good news for programming!

Vowpal Wabbit (http://hunch.net/ vw) uses this systematically.

An Open Problem: $1K reward!

Conditional Probability Estimation

Distribution D over $X \times Y$, where $Y = \{1, \ldots, k\}$.

Find a Probability estimator $h : X \times Y \to [0, 1]$ minimizing squared loss

$$\ell(h, D) = E_{(x,y) \sim D}[(h(y|x) - y)^2]$$

An Open Problem: \$1K reward!

Conditional Probability Estimation

Distribution D over $X \times Y$, where $Y = \{1, \ldots, k\}$.

Find a Probability estimator $h : X \times Y \to [0, 1]$ minimizing squared loss

$$\ell(h, D) = E_{(x,y)\sim D}[(h(y|x) - y)^2]$$

The problem: How can you do this in time $O(\log(k))$ with a constant regret ratio?

More Details

Beygelzimer, Langford, Daume, Mineiro, “Learning Reductions that Realy Work”, IEEE 104(1), 2016 https://arxiv.org/abs/1502.02704