Good Learners for Evil Teachers

Ofer Dekel, Microsoft Research
Ohad Shamir, Hebrew University

ICML 2009

Classification with Multiple Teachers

Problem setting:

- $\mathcal{X}$ is an instance space
- $\mathcal{D}$ is a distribution over $\mathcal{X} \times \{-1, 1\}$
- labels provided by k teachers, some malicious
- data generation:
  (1) sample $S = \{\mathbf{x}_i\}_{i=1}^{m}$ i.i.d. from $\mathcal{D}|_{\mathbf{x}}$
  (2) S is randomly split into $S_1, \ldots, S_k$
  (3) teacher t labels S_t.

Examples

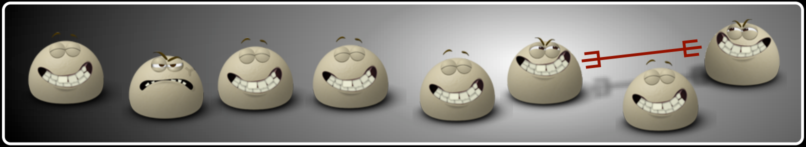
- collecting labels over the Internet (e.g. *Mechanical Turk*): scripts and *bots* masquerade as real people
- learning from search engine logs: scripts, SEOs (search engine optimizers)

Label Collection Common Practices

- repeated labeling - multiple teachers label each example, not always possible, wasteful
- honeypots - test each teacher, not always possible, requires "truth set"
- challenge-response tests - e.g. captchas, not always possible, often more difficult than the labeling task itself
- outlier detection - verify labeling speed, IP address, label distribution, easy to pass this test

Are these techniques necessary?

Theoretical Model: Good vs. Evil



- a teacher is either good ($t \in G$) or evil ($t \in E$)
- good teachers label according to $\mathcal{D}|_{(y|\mathbf{x})}$
- evil teachers are malicious, allowed to collude
- evil teachers don't see the examples labeled by good teachers

The SVM Algorithm

define $F(\mathbf{w}|S,\lambda) \;=\; \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\left[1-y_i\langle\mathbf{x}_i,\mathbf{w}\rangle\right]_+$

where

- $[\alpha]_+ = \max\{\alpha, 0\}$ is the hinge-loss function
- λ is a positive parameter
- $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ is the training set

the SVM algorithm: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \; F(\mathbf{w}|S,\lambda)$

The "SVM+Oracle" Algorithm

- define: the set of good examples $S_G = \cup_{t \in G} S_t$, the set of bad examples $S_E = \cup_{t \in E} S_t$
- ideally, an oracle reveals G and E, and we train our favorite binary classifier (e.g. *SVM*) on S_G

$$\text{SVM:} \quad \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \; F(\mathbf{w} | S, \lambda)$$

$$\text{SVM+Oracle:} \quad \mathbf{w}^{\star} = \arg\min_{\mathbf{w}} \; F(\mathbf{w} | S_G, \tfrac{m}{|S_G|}\lambda)$$

OUR GOAL: to approximate SVM+Oracle (without knowing G)

Main Idea

- How many support vectors does each teacher contribute?
- if all teachers are good, expect equal contribution
- our algorithm: enforce "equal contribution" as a constraint

The SVM Dual

$$\text{primal:} \quad \min_{\mathbf{w}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\left[1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\right]_+$$

$$\text{dual:} \quad \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \quad \sum_{i=1}^{m}\alpha_i - \frac{1}{2\lambda}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\langle \mathbf{x}_i, \mathbf{x}_j\rangle$$

$$\text{s.t.} \quad \forall i \in [m] \;\; 0 \le \alpha_i \le \frac{1}{m}$$

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i$$

We say that $(\mathbf{x}_i, y_i)$ is a support vector if $\alpha_i > 0$.

Our Algorithm: A Modified SVM

$$\text{primal:} \quad \min_{\mathbf{w}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\left[1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\right]_+$$

$$\text{dual:} \quad \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \quad \sum_{i=1}^{m}\alpha_i - \frac{1}{2\lambda}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\langle \mathbf{x}_i, \mathbf{x}_j\rangle$$

$$\text{s.t.} \quad \forall i \in [m] \;\; 0 \le \alpha_i \le \frac{1}{m}$$

$$\forall t \in [k] \quad \underbrace{\frac{1}{|S_t|}\sum_{i\in S_t}\alpha_i}_{t\text{'s contibution}} \le \underbrace{\frac{1}{m}\sum_{i=1}^{m}\alpha_i}_{\text{average contribution}} + \underbrace{\frac{\epsilon}{m\sqrt{|S_t|}}}_{\text{small}}$$

Theorem 1

If $\epsilon > \frac{|S_E|\sqrt{|S_t|}}{m}$ for all $t \in G$, then with probability at least

$$1 - \underbrace{\sum_{t \in G} \exp\left(-2|S_t|\left(\frac{\epsilon}{\sqrt{|S_t|}} - \frac{|S_E|}{m}\right)^2\right)}_{\text{small}}$$

over the assignment of examples to teachers, the "equal contribution" constraint is non-binding for all $t \in G$.

Theorem 2

- $F(\mathbf{w}|S,\lambda) \;=\; \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\left[1-y_i\langle\mathbf{x}_i,\mathbf{w}\rangle\right]_+$
- SVM: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\; F(\mathbf{w}|S,\lambda)$
- SVM+Oracle: $\mathbf{w}^\star = \arg\min_{\mathbf{w}}\; F(\mathbf{w}|S_G,\frac{m}{|S_G|}\lambda)$
- our algorithm (with S,λ) : $\mathbf{w}'$

$$\underbrace{F\left(\hat{\mathbf{w}}\middle|S_G,\tfrac{m}{|S_G|}\lambda\right)}_{\text{SVM's objective on } S_G} \;-\; \underbrace{F\left(\mathbf{w}^\star\middle|S_G,\tfrac{m}{|S_G|}\lambda\right)}_{\text{best possible objective on } S_G} \;\leq\; \frac{|S_E|}{|S_G|}C$$

Theorem 3

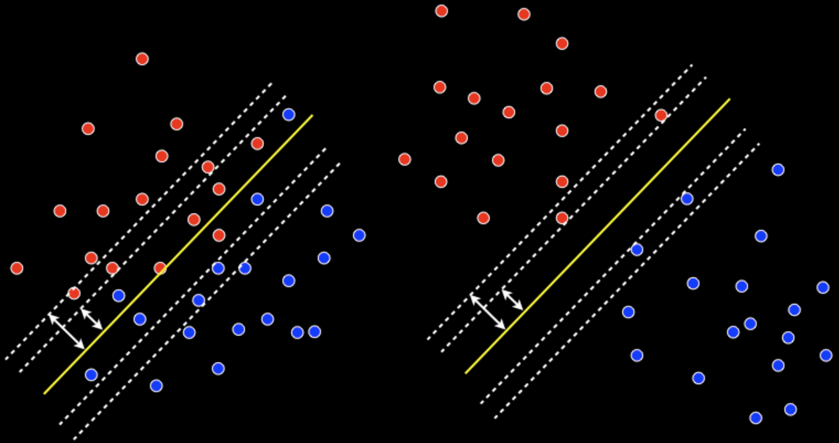
- $F(\mathbf{w}|S,\lambda) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\left[1 - y_i\langle \mathbf{x}_i, \mathbf{w}\rangle\right]_+$
- SVM: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \; F(\mathbf{w}|S,\lambda)$
- SVM+Oracle: $\mathbf{w}^\star = \arg\min_{\mathbf{w}} \; F(\mathbf{w}|S_G, \frac{m}{|S_G|}\lambda)$
- our algorithm (with S, λ) : $\mathbf{w}'$

$$F\left(\hat{\mathbf{w}}\middle|S_G, \tfrac{m}{|S_G|}\lambda\right) - F\left(\mathbf{w}^\star\middle|S_G, \tfrac{m}{|S_G|}\lambda\right) \le \frac{|S_E|}{|S_G|}C$$

$$\underbrace{F\left(\mathbf{w}'\middle|S_G, \tfrac{m}{|S_G|}\lambda\right)}_{\text{our alg's objecitve on } S_G} - \underbrace{F\left(\mathbf{w}^\star\middle|S_G, \tfrac{m}{|S_G|}\lambda\right)}_{\text{best possible objective on } S_G} \le \frac{|S_E|}{|S_G|}CV$$

Where $V \approx \frac{1}{|S_G|}\left|\left\{(\mathbf{x}, y) \in S_G \;:\; y\langle \mathbf{w}^\star, \mathbf{x}\rangle \le 1+\gamma\right\}\right|$

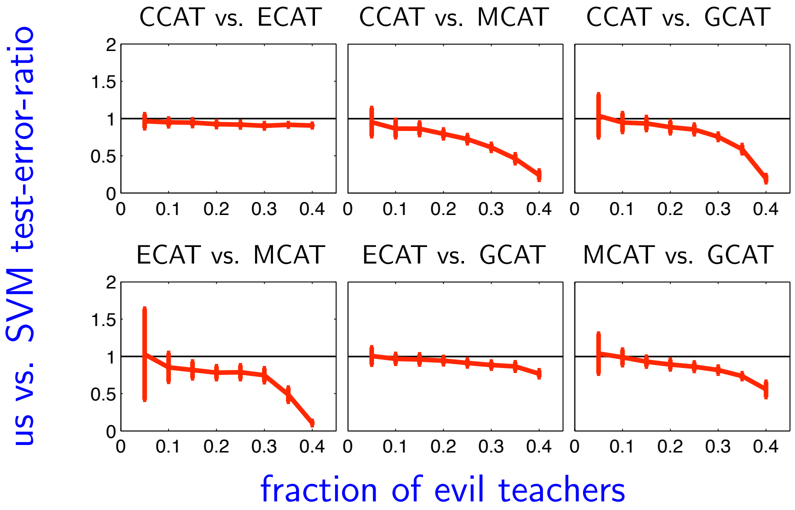
Theorems 2/3 - Cartoon Version



our algorithm *won't* improve over SVM

our algorithm *will* improve over SVM

Experiments with RCV1



Final Remarks

- take-home message: all we need is the teacher identity - no repeated labels, prior knowledge, pre-labeled "truth sets", etc.
- more in the paper - a second algorithm, experiments where S is partitioned by subtopic
- related work - our COLT09 paper *"Vox Populi: Collecting High Quality Labels from a Crowd"*, talk on Sunday afternoon