Identifiability of Priors from Bounded Sample Sizes with Applications to Transfer Learning

Liu Yang

Joint work with Steve Hanneke & Jaime Carbonell

Carnegie Mellon University

Outline

- Self-Verifying Bayesian Active Learning

  (AISTATS 2010)

- Transfer Learning

  (COLT 2011 and Machine Learning Journal)

Notation

- Instance space $X = R^n$
- Concept space C of classifiers $h: X \to \{0,1\}$
  - Assume C has VC dimension $vc < \infty$
- Data Distribution D on X
- Unknown target function h^*: the true labeling function (Realizable case: h^* in C)
- Assume $\rho(h, g) = P_{x \sim D}[h(x) \neq g(x)]$ for any classifiers h, g, is a **metric** on C
- $Err(h) = P_{x \sim D}[h(x) \neq h^*(x)]$

"Active" means Label Request

- Label request:
  have a pool of unlabeled exs, pick any x and receive $h^*(x)$, repeat
- Motivation:
  labeled data is expensive to get
- Using label request, can do **Active Learning**:
  find h has small err(h)

Self-Verifying Bayesian Active Learning

Self-verifying

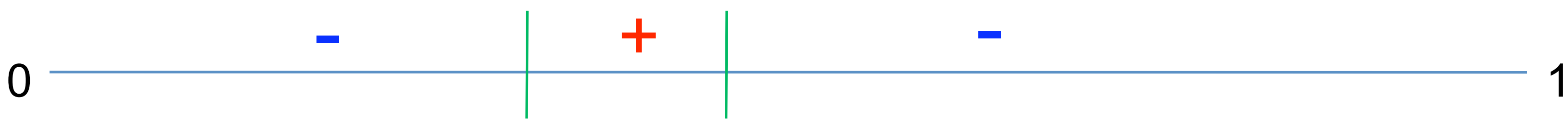
(a special type of stopping criterion)

- Given ε, adaptively decides # of query, then halts
- has the property that $E[err] < \varepsilon$ when halts

Question: Can you do with $E[\#query] = o(1/\varepsilon)$? (passive learning need $1/\varepsilon$ labels)

Example: Intervals

Suppose D is uniform on [0,1]



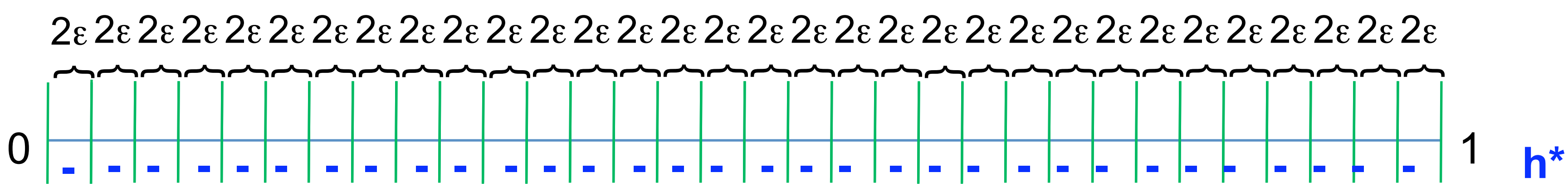
Example: Intervals

Verification Lower Bound

In non-Bayesian setting, supposing h^* is empty interval.

Given any classifier h,
just to verify $\mathrm{err}(h) < \varepsilon$,
Need to verify h^* is not an interval of width 2ε.

Need an example in $\Omega(1/\varepsilon)$ regions to verify this fact.



Suppose h^* is empty interval, D is uniform on [0,1]

Learning with a prior

- Suppose we know a distribution the target is sampled from, call it prior

Interval Example with prior

- - - - - |+++++++|- - - - - -

- **Algorithm:** Query random pts till find first +, do binary search to find end-pts. Halt when reach a pre-specified prior-based query budget. Output posterior's Bayes classifier.

- Let budget N be high enough so $E[err] < \varepsilon$
  - $N = o(1/\varepsilon)$ sufficient for $E[err|w^*>0] < \varepsilon$: if $w^* > 0$, even prior-independent analysis needs only $E[\#queries|w^*] = O(1/w^* + \log(1/\varepsilon)) = o(1/\varepsilon)$.
  - $N = o(1/\varepsilon)$ sufficient for $E[err|w^*=0] < \varepsilon$: if $P(w^*=0)>0$, then after some $L = O(\log(1/\varepsilon))$ queries, w.p.$>1-\varepsilon$, most prob. mass on empty interval, so posterior's Bayes classifier has 0 error rate

Can do o(1/eps) for any VC-class

Theorem: With the prior, can get $o(1/\varepsilon)$ QC

- There are methods that find a good classifier in o(1/eps) queries (though they aren't self-verifying) [BHW08]
- Need set a stopping criterion for those alg
- The stop criterion we use : budget
- Set the budget to be just large enough so $E[err] < \varepsilon$.

Outline

- Self-Verifying Bayesian Active Learning (AISTATS2010)
- Transfer Learning (COLT 2011 and Machine Learning Journal)

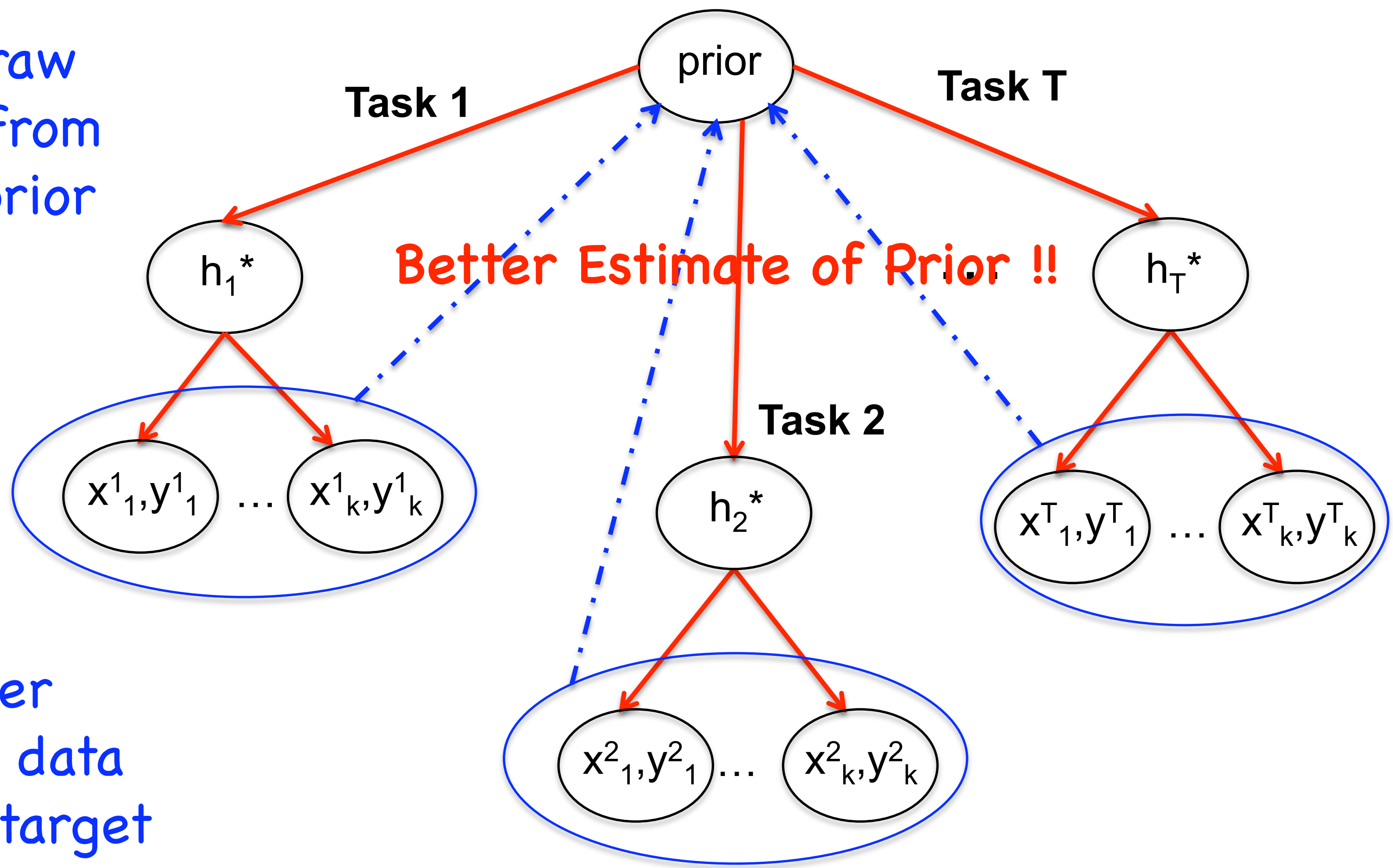
Transfer Learning

- Principle: solving a new learning problem is easier given that we've solved several already !
- How does it help?
  - New task directly ``related" to previous task
    [e.g., Ben-David & Schuller 03; Evgeniou, Micchelli, & Pontil 2005]
  - Previous tasks give us useful sub-concepts [e.g., Thrun 96]
  - Can gather statistical info on the variety of concepts
    [e.g., Baxter 97; Ando & Zhang 04]
- Example: Speech Recognition
  - After training a few times, figured out the dialects.
  - Next time, just identify the dialect.
  - Much easier than training a recognizer from scratch

Model of Transfer Learning

Motivation: Learners often Not Too Altruistic

Layer 1: draw task i.i.d. from unknown prior



Layer 2: per task, draw data i.i.d. from target

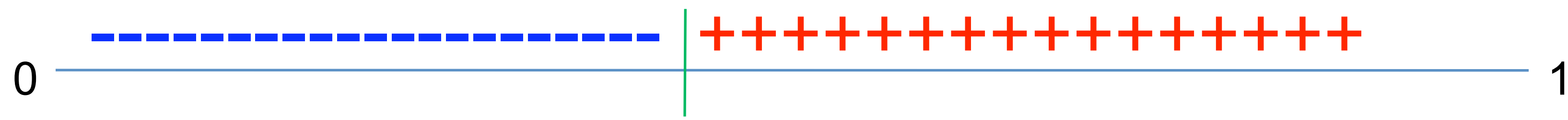
Identifiability of priors from joint distribs

- Let prior π be any distribution on C
  - example: (w, b) ~ multivariate normal
- Target $h^*_\pi \sim \pi$
- Data $X = (X_1, X_2, \ldots)$ i.i.d. D indep h^*_π
- $Z(\pi) = ((X_1, h^*_\pi (X_1)), (X_2, h^*_\pi (X_2)), \ldots)$.
- Let $[m] = \{1, \ldots, m\}$.
- Denote $X_I = \{X_i\}_{i \in I}$ (I : subset of natural numbers)
- $Z_I (\pi) = \{(X_i, h^*_\pi (X_i))\}_{i \in I}$

Theorem: $Z_{[VC]} (\pi_1) =_d Z_{[VC]} (\pi_2)$ **iff** $\pi_1 = \pi_2$.

Identifiability of priors by VC-dim joint distri.

- Threshold:



- for two points x_1, x_2, if $x_1 < x_2$, then
Pr(+,+)=Pr(+.), Pr(-,-)=Pr(.-), Pr(+,-)=0,
So Pr(-,+)=Pr(.+)-Pr(++) = Pr(.+)-Pr(+.)
- for any k > 1 points, can directly to reduce number of labels in the joint prob from k to 1

P(-----------(-+)++++++++++++++++)
= P((-+))
= P((.+)) - P((++))
= P((.+)) - P((+.))
+ P((+-)) (unrealized labeling !!)
= P((.+)) - P((+.))

- **Theorem: $Z_{[VC]}(\pi_1) =_d Z_{[VC]}(\pi_2)$ iff $\pi_1 = \pi_2$.**

Proof Sketch

- Let $\rho_m(h,g) = 1/m \; \Sigma_{i=1}^{m} \; II(h(X_m) \neq g(X_m))$
  Then vc < ∞ implies w.p.1 forall h, g € C with $h \neq g$
  $\lim_{m \to \infty} \rho_m(h,g) = \rho(h,g) > 0$
- ρ is a metric on C by assumption,
  so w.p.1 each h in C labels ∞-seq $(X_1, X_2 \ldots)$ distinctly $(h(X_1), h(X_2), \ldots)$
- => w.p.1 conditional distribution of the label seq $Z(\pi)|X$ identifies π
  => distrib of $Z(\pi)$ identifies π
  i.e. $Z_\infty(\pi_1) =_d Z_\infty(\pi_2)$ implies $\pi_1 = \pi_2$

Identifiability of Priors from Joint Distributions

$$\textbf{Theorem: } Z_{[\text{vc}]}(\pi_1) \overset{d}{=} Z_{[\text{vc}]}(\pi_2) \Leftrightarrow \pi_1 = \pi_2.$$

Proof Sketch:

Fix any $m > \text{vc}$, $x_1, \ldots, x_m \in \mathcal{X}$, $y_1, \ldots, y_m \in \{0,1\}$.

Note $\mathbb{C}$ cannot shatter $(x_1, \ldots, x_m)$.

Let $\tilde{y}_1, \ldots, \tilde{y}_m \in \{0,1\}$ be s.t. $\nexists h \in \mathbb{C}$ with $\forall i, h(x_i) = \tilde{y}_i$.

Clearly $\mathbb{P}\left(Z_{[m]}(\pi) = \{(x_i, \tilde{y}_i)\}_{i \in [m]} \middle| \mathbb{X}_{[m]} = \{x_i\}_{i \in [m]}\right) = 0.$

If $\exists k$ s.t. $y_k \neq \tilde{y}_k$, then letting $y'_i = y_i$ for $i \neq k$, and $y'_k = \tilde{y}_k$,

$$\mathbb{P}\left(Z_{[m]}(\pi) = \{(x_i, y_i)\}_{i \in [m]} \middle| \mathbb{X}_{[m]} = \{x_i\}_{i \in [m]}\right)$$

lower–dim cond distrib

$$= \boxed{\mathbb{P}\left(Z_{[m]\setminus\{k\}}(\pi) = \{(x_i, y_i)\}_{i \in [m]\setminus\{k\}} \middle| \mathbb{X}_{[m]\setminus\{k\}} = \{x_i\}_{i \in [m]\setminus\{k\}}\right)}$$

$$- \boxed{\mathbb{P}\left(Z_{[m]}(\pi) = \{(x_i, y'_i)\}_{i \in [m]} \middle| \mathbb{X}_{[m]} = \{x_i\}_{i \in [m]}\right)}.$$

y' closer to ỹ

Induction: $\mathbb{P}\left(Z_{[m]}(\pi) = \cdot \middle| \mathbb{X}_{[m]}\right)$ function of $\mathbb{P}\left(Z_{[\text{vc}]}(\pi) = \cdot \middle| \mathbb{X}_{[\text{vc}]}\right)$.

Identifiability of Priors from Joint Distributions

Theorem: $Z_{[\mathrm{vc}]}(\pi_1) \stackrel{d}{=} Z_{[\mathrm{vc}]}(\pi_2) \Leftrightarrow \pi_1 = \pi_2$.

Proof Sketch:

By the above,
$Z_{[\mathrm{vc}]}(\pi_1) \stackrel{d}{=} Z_{[\mathrm{vc}]}(\pi_2) \Rightarrow \forall m \in \mathbb{N}, Z_{[m]}(\pi_1) \stackrel{d}{=} Z_{[m]}(\pi_2)$.

Classic result:
set of distribs of $Z_{[m]}(\pi) : m \in \mathbb{N}$ identify distrib of $Z(\pi)$, so
$Z_{[m]}(\pi_1) \stackrel{d}{=} Z_{[m]}(\pi_2), \forall m \in \mathbb{N} \Rightarrow Z(\pi_1) \stackrel{d}{=} Z(\pi_2)$.

Showed above that
$Z(\pi_1) \stackrel{d}{=} Z(\pi_2) \Rightarrow \pi_1 = \pi_2$. $\square$

Identifiability of Priors from Joint Distributions

Theorem: $Z_{[\mathrm{vc}]}(\pi_1) \overset{d}{=} Z_{[\mathrm{vc}]}(\pi_2) \Leftrightarrow \pi_1 = \pi_2$.

Theorem: $\exists \mathcal{D}$, $\pi_1 \neq \pi_2$ s.t. $\forall m < \mathrm{vc}$, $Z_{[m]}(\pi_1) \overset{d}{=} Z_{[m]}(\pi_2)$.

Proof Sketch:

Let $(x_1, \ldots, x_{\mathrm{vc}})$ be shattered by $\mathcal{H} = \{h_1, \ldots, h_{2^{\mathrm{vc}}}\} \subseteq \mathbb{C}$.
Let $\mathcal{D}$ be uniform on $\{x_1, \ldots, x_{\mathrm{vc}}\}$,
let π_1 be uniform on $\mathcal{H}$.
Let $\mathcal{H}' = \{h'_1, \ldots, h'_{2^{\mathrm{vc}-1}}\} \subset \mathcal{H}$ shatter $(x_1, \ldots, x_{\mathrm{vc}-1})$
s.t. $h'_i(x_{\mathrm{vc}}) = \mathrm{Parity}(\{h'_i(x_1), \ldots, h'_i(x_{\mathrm{vc}-1}))$.
Let π_2 be uniform on $\mathcal{H}'$.
Clearly $\pi_1 \neq \pi_2$.

But for $m < \mathrm{vc}$, $Z_{[m]}(\pi_1) \overset{d}{=} Z_{[m]}(\pi_2)$:
unif cond on labels given distinct $X_1, \ldots, X_m$. $\square$

Transfer Learning Setting

- Collection Π of distribs on C. (known)
- Target distrib $\pi^* \in \Pi$. (unknown)
- Indep target fns $h_1^*, \ldots, h_T^* \sim \pi^*$ (unknown)
- Indep i.i.d. D data sets $X^{(t)} = (X_1^{(t)}, X_2^{(t)}, \ldots)$, $t \in [T]$.
- Define $Z^{(t)} = ((X_1^{(t)}, h_t^*(X_1^{(t)})), (X_2^{(t)}, h_t^*(X_2^{(t)})), \ldots)$.
- Learning alg. "gets" $Z^{(1)}$, then produces $\hat{h}_1$, then "gets" $Z^{(2)}$, then produces $\hat{h}_2$, etc. in sequence.
- Interested in: values of $\rho(\hat{h}_t, h^*(t))$, and the number of $h^*_t(X_j^{(t)})$ value alg. needs to access.

Estimating the prior

- Principle: learning would be easier if know π^*
- Fact: π^* is identifiable by distrib of $Z_{[VC]}^{(t)}$
- Strategy: Take samples $Z_{[VC]}^{(i)}$ from past tasks 1, …, t-1, use them to estimate distrib of $Z_{[VC]}^{(i)}$, convert that into an estimate π'_{t-1} of π^*,
- Use π'_{t-1} in a prior-dependent learning alg for new task h_t^*
- Assume Π is totally bounded in total variation
- Can estimate π^* at a bounded rate:

  $\|\pi^* - \pi'_t\| < \delta_t$ converges to 0 (holds whp)

Transfer Learning

- Given a prior-dependent learning $A(\varepsilon, \pi)$, with $E[\#\text{ labels accessed}] = \Lambda(\varepsilon, \pi)$ and producing $\hat{h}$ with $E[\rho(\hat{h}, h^*)] \leq \varepsilon$

For $t = 1, \ldots, T$

If $\delta_{t-1} > \varepsilon/4$,

run prior-indep learning on $Z_{[VC/\varepsilon]}^{(t)}$ to get $\hat{h}_t$

Else let $\pi''_t = \mathrm{argmin}_{\pi \in B(\pi'_{t-1}, \delta_{t-1})} \Lambda(\varepsilon/2, \pi)$

and run $A(\varepsilon/2, \pi''_t)$ on $Z^{(t)}$ to get $\hat{h}_t$

Theorem: Forall t, $E[\rho(\hat{h}_t, h_t^*)] \leq \varepsilon$, and

$\limsup_{T \to \infty} E[\#\text{labels accessed}]/T \leq \Lambda(\varepsilon/2, \pi^*) + vc$.

Is this Better than without Transfer ?

- The question becomes:
  - How much does knowledge of target distrib π^* help?
- There are some (constant factor) gains for passive learning [e.g. HKS1992]
- It really helps in Active learning:
  - Earlier, we showed can get $o(1/\varepsilon)$ for all π
- For many C (e.g. linear separators), no prior-indep alg has this guarantee.
- Plugging in that method, transfer method accesses $o(1/\varepsilon)$ labels on avg.

Remarks

- Not too many extra labels per task (vc extra)
- Subroutine A can be fairly arbitrary (supervised, semi-supervised, active, …)
- π estimation may be useful for other things too
- Open problem: calculate the rate of convergence

Thanks !