Active Learning
Drifting Distributions, and
Convex Surrogate Losses

Liu Yang

Carnegie Mellon University
Outline

• Active Learning with a Drifting Distribution ([Yang11 NIPS])
Active Learning with a Drifting Distrib: Model

• Scenario:
  - Unobservable seq. of distrib.s \( D_1, D_2, \ldots \) with each \( D_t \in \mathcal{D} \)
  - Unobservable time-indep. regular cond. distrib. represent by fn \( \eta : X \rightarrow [0, 1] \)

• Active learning protocol
  - At each time \( t \), alg is presented with \( X_t \), and is required to predict a label \( \hat{Y}_t \) \in \{-1, +1\}, then it may optionally request to see true label value \( Y_t \)

• Interested in cumulative #mistakes up to time \( T \) and total #labels requested up to time \( T \)
Definition and Notations

- Instance space $X = \mathbb{R}^n$
- Distribution space $\mathcal{D}$ of distributions on $X$
- Concept space $C$ of classifiers $h: X \rightarrow \{-1,1\}$
  - Assume $C$ has VC dimension $vc < \infty$
- $D_t$: Data distrib. on $X$ at $t$
- Unknown target fn $h^*$: true labeling fn
- $\text{Err}_t(h) = P_{x \sim D_t} [h(x) \neq h^*(x)]$
- In realizable case, $h^* \in C$ and $\text{err}_t(h^*) = 0$
- For $V \subseteq C$, $\text{diam}_t(V) = \sup_{h,g \in V} D_t(\{x : h(x) \neq g(x)\})$
**Def: disagreement coefficient, tvd**

- The disagreement coefficient of $h^*$ under a distri. $P$ on $X$, is define as, $(r > 0)$

  $$\theta_P(\epsilon) = \sup_{r>\epsilon} \frac{P(DIS(B_P(h^*, r)))}{r}.$$  

  $DIS(V) = \{x \in X : \exists h, g \in V \text{ s.t. } h(x) \neq g(x)\}$

  $B_P(h, r) = \{g \in C : P(x : h(x) \neq g(x)) \leq r\}$

- Total variation distance of probability measures $P$ and $Q$ on a sigma-algebra $\mathcal{G}$ of subsets of the sample space is defined via

  $$\|P - Q\| = \sup_{A \in \mathcal{G}} |P(A) - Q(A)|$$
Assumptions

• Independence of the $X_t$ variables
• $V_c$-dim < $\infty$
• Assumption 1 (totally bounded): $\mathcal{D}$ is totally bounded (i.e. satisfies $\forall \epsilon > 0, |\mathcal{D}_\epsilon| < \infty$)
  - For each $\epsilon > 0$, $\mathcal{D}_\epsilon$ denote a minimal subset of $\mathcal{D}$ s.t.
  $\forall D \in \mathcal{D}, \exists D' \in \mathcal{D}_\epsilon$ s.t. $\|D - D'\| < \epsilon$ (i.e. a minimal $\epsilon$-cover of $\mathcal{D}$)
• Assumption 2 (poly-covers)

$$\forall \epsilon > 0, |\mathcal{D}_\epsilon| < c \cdot \epsilon^{-m}$$

where $c, m \geq 0$ are constants.
Realizable-case Active Learning

CAL

1. \( t \leftarrow 0, \mathcal{Q}_0 \leftarrow \emptyset \), and let \( \hat{h}_0 = A(\emptyset) \)
2. Do
3. \( t \leftarrow t + 1 \)
4. Predict \( \hat{Y}_t = \hat{h}_{t-1}(X_t) \)
5. If \( \max_{y \in \{-1,+1\}} \min_{h \in C} \hat{\epsilon}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\}) = 0 \)
6. Request \( Y_t \), let \( \mathcal{Q}_t = \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\} \)
7. Else let \( Y'_t = \arg\min_{y \in \{-1,+1\}} \min_{h \in C} \hat{\epsilon}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\}) \), and let \( \mathcal{Q}_t = \mathcal{Q}_{t-1} \cup \{(X_t, Y'_t)\} \)
8. Let \( \hat{h}_t = \arg\min_{h \in C} \hat{\epsilon}(h; \mathcal{Q}_t) \)

© Liu Yang 2012
Theorem. If $\mathcal{D}$ is totally bounded, then CAL achieves an expected mistake bound $\bar{M}_T = o(T)$

And if $\theta_D(\epsilon) = o(1/\epsilon)$, then CAL makes an $E[\#queries]$ $\bar{Q}_T = o(T)$

[Proof Sketch]:
Partition $\mathcal{D}$ into buckets of diam < eps.
Pick a time $T_{\epsilon}$ past all indices from finite buckets and all the infinite bucket has at least

$$L(\epsilon) = \lceil \frac{8}{\sqrt{\epsilon}} \left( d \ln \frac{24}{\sqrt{\epsilon}} + \ln \frac{4}{\sqrt{\epsilon}} \right) \rceil$$
Number of Mistakes

- **Alternative scenario:**
  - Let $P_i$ be in bucket $i$
  - Swap the $L(\varepsilon)$ samples for bucket $i$ with $L(\varepsilon)$ samples from $P_i$
  - $L(\varepsilon)$ large enough so $E[diam(V)]_{\text{alternative}} < \sqrt{\varepsilon}$.

- Note: $E[diam(V)] \leq E[diam(V)]_{\text{alternative}} + \sum_{\text{values}} L(\varepsilon) \cdot ||P_i - D_t|| < \sqrt{\varepsilon} + L(\varepsilon) \cdot \varepsilon$.

  So $E[diam] \to 0$ as $T \to \infty$

- $E[\#\text{mistake}] \leq \sum_{t=1}^{T} E[diam(V_{t-1})]$

- Since $E[diam(V_{t-1})] \to 0$, $\sum_{t=1}^{T} E[diam(V_{t-1})] = o(T)$
Number of Queries

- \( E[\#\text{queries}] = \sum_{t=1}^{T} P(\text{make query}) \)
- \( P(\text{make query}) = E[P(\text{DIS}(V_{t-1}))] \)
- \( E[\theta(r) \max\{diam, r\}] \leq \theta(r) E[diam] + \theta(r) \cdot r \)

- Let \( r_T = \frac{1}{T} \sum_{t=1}^{T} E[diam_t(V_{t-1})] \)
  
  \( r_t \to 0 \) and

\[ E[\#\text{queries}] \leq \theta(r_T) \sum_{t=1}^{T} E[diam_t(V_{t-1})] + \theta(r_T)r_T = \theta(r_T)r_T(T + 1) \]

- \( \theta(\epsilon) = o(1/\epsilon) \implies \theta(r_T)r_T \to 0 \implies \theta(r_T)r_T(T + 1) = o(T) \)
Explicit Bound: Realizable Case

Theorem. If poly-covers assumption is satisfied \(|D_\varepsilon| < (1/\varepsilon)^m\) then CAL achieves an expected mistake bound \(\hat{M}_T\) and \(E[\#queries] \hat{Q}_T\) such that

\[
\hat{M}_T = O\left(T \frac{m}{m+1} d \frac{1}{m+1} \log^2 T\right)
\]
\[
\hat{Q}_T = O\left(\theta_D(\varepsilon_T) T \frac{m}{m+1} d \frac{1}{m+1} \log^2 T\right)
\]

where \(\varepsilon_T = (d/T)^{\frac{1}{m+1}}\)

[Proof Sketch]
Fix any \(\varepsilon > 0\), and enumerate \(D_\varepsilon = \{P_1, P_2, \ldots, P_{|D_\varepsilon|}\}\)
For \(t\) in \(\mathbb{N}\), let \(K(t)\) be the index \(k\) of the closest \(P_k \in D_\varepsilon\) to \(D_t\).

Alternative data sequence:
Let \(\{X'_t\}_{t=1}^{\infty}\) be indep., with \(X_t \sim P_{K(t)}\)
This way all samples corresp. to distrib.\'s in a given bucket all came from same distri.
Let \(V'_t\) be the corresponding version spaces.
$$\mathbb{E}[\#\text{mistakes}] \leq \mathbb{E}\left[\sum_{t=1}^{T} \text{diam}_{P_K(t)}(V'_{t-1})\right] + \sum_{t=1}^{T} \|D_t - P_K(t)\|$$
$$\leq \sum_{t=1}^{T} \mathbb{E}[\text{diam}_{P_K(t)}(V'_{t-1})] + \epsilon T$$

Classic PAC bound $\Rightarrow$ $\mathbb{E}[\text{diam}_{P_K(t)}(V'_{t-1})] \leq O\left(\frac{d \log t}{|\{i \leq t: K(i) = K(t)\}|}\right)$

So $\sum_{t=1}^{T} \mathbb{E}[\text{diam}_{P_K(t)}(V'_{t-1})] \leq O(d \log T) \sum_{t=1}^{T} \frac{1}{|\{i \leq t: K(i) = K(t)\}|}$
$$\leq O\left(d \log T|\mathcal{D}_\epsilon| \sum_{u=1}^{T} \frac{1}{u} \leq O(d|\mathcal{D}_\epsilon| \log^2(T))\right)$$

(each bucket has at most $T$ samples)

So $\mathbb{E}[\#\text{mistakes}] \leq O\left(d\left(\frac{1}{\epsilon}\right)^m \log^2(T) + \epsilon T\right)$

Take $\epsilon = (T/d)^{-\frac{1}{m+1}}$ to get the stated theorem.

To bound $\mathbb{E}[\#\text{queries}]$, again it is
$$\leq \mathbb{E}\left[\sum_{t=1}^{T} D_t(DIS(V_{t-1}))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \theta(\epsilon) \max\{\text{diam}_t(V_{t-1}, \epsilon)\}\right]$$
$$\leq \theta(\epsilon) \mathbb{E}\left[\sum_{t=1}^{T} \text{diam}_t(V_{t-1})\right] + \theta(\epsilon)\epsilon T$$

just showed this is $\leq O\left(d\left(\frac{1}{\epsilon}\right)^m \log^2(T) + \theta(\epsilon)\epsilon T\right)$

So $O(\theta(\epsilon)d \left(\frac{1}{\epsilon}\right)^m \log^2(T) + \theta(\epsilon)\epsilon T)$

Again, taking $\epsilon = (T/d)^{-\frac{1}{m+1}}$ gives the stated result.
• **Strictly benign noise condition:**
  \[ h^* = \text{sign}(\eta - 1/2) \in C \quad \text{and} \quad \forall x, \eta(x) \neq 1/2 \]

• **Special case: Tsybakov's noise conditions**

• \( \eta \) satisfies strictly benign noise condition and for some \( c > 0 \) and \( \alpha \geq 0 \), \( \forall t > 0 \), \( P(|\eta(x) - 1/2| < t) < c \cdot t^\alpha \)

\[ P(h(x) \neq h^*(x)) \leq c' \left( er(h) - er(h^*) \right) \frac{\alpha}{\alpha + 1} \]

• **Unif Tsybakov assumption:** Tsybakov Assumption is satisfied for all \( D \in \mathcal{D} \) with the same \( c \) and \( \alpha \) values.
Agnostic CAL [DHM]

ACAL
1. $t \leftarrow 0$, $L_t \leftarrow \emptyset$, $Q_t \leftarrow \emptyset$, let $\hat{h}_t$ be any element of $\mathcal{C}$
2. Do
3. $t \leftarrow t + 1$
4. Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$
5. For each $y \in \{-1, +1\}$, let $h(y) = \text{LEARN}(L_{t-1} \cup \{(x_t, y)\}, Q_{t-1})$
6. If either $y$ has $h(-y) = \emptyset$ or
   \[
   \hat{e}_r(h(-y); L_{t-1} \cup Q_{t-1}) - \hat{e}_r(h(y); L_{t-1} \cup Q_{t-1}) > \hat{\epsilon}_{t-1}(L_{t-1}, Q_{t-1})
   \]
7. $L_t \leftarrow L_{t-1} \cup \{(X_t, y)\}$, $Q_t \leftarrow Q_{t-1}$
8. Else Request $Y_t$, and let $L_t \leftarrow L_{t-1}$, $Q_t \leftarrow Q_{t-1} \cup \{(X_t, Y_t)\}$
9. Let $\hat{h}_t = \text{LEARN}(L_t, Q_t)$
10. If $t$ is a power of 2
11. $L_t \leftarrow \emptyset$, $Q_t \leftarrow \emptyset$

Based on subroutine: $\text{LEARN}(L, Q) = \arg\min_{h \in \mathcal{C}} \hat{e}_r(h; Q)$ if $\min_{h \in \mathcal{C}} \hat{e}_r(h; L) = 0$, and otherwise $\text{LEARN}(L, Q) = \emptyset$. 

© Liu Yang 2012
Theorem. If $\mathcal{D}$ is totally bounded and $\eta$ satisfies strictly benign noise condition, then ACAL achieves an excess expected mistake bound

$$M_T - M_T^* = o(T)$$

and if additionally $\theta_\mathcal{D}(\epsilon) = o(1/\epsilon)$, then ACAL makes an expected number of queries $\bar{Q}_T = o(T)$

---

Theorem. If poly-covers Assumption and Unif Tsybakov assumption are satisfied, then ACAL achieves an expected excess number of mistakes ACAL achieves expected excess mistakes $\bar{M}$ and expected queries $\bar{Q}_T$ such that, for $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$

$$\bar{M}_T - M_T^* = \tilde{O} \left( T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}} \right)$$

$$\bar{Q}_T = \tilde{O} \left( \theta_\mathcal{D}(\epsilon_T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}} \right)$$
Outline

• Convex Losses in Active Learning
  (Joint work with Steve Hanneke)
Negative Results for AL with Convex Losses [AISTATS'10]

\[ F_{\text{offs}} : f : \mathcal{X} \to R \]

\[ \text{Loss fn } l : R \to [0, \infty) \]

Interested in convex nonincreasing loss

Data distri. \( D \) still on \( \mathcal{X} \times -1, +1 \)

Risk \( R_l(f) = E[l(f(x)Y)] \) for \( (X, Y) \sim D_{XY} \)

Question: How many labels needed to find \( \hat{f} \in \mathcal{F} \) with \( R_l(\hat{f}) - \inf_{f \in \mathcal{F}} R_l(f) \leq \epsilon \)?

We'll study "Bounded Noise" Scenario where \( \exists f \in \mathcal{F} \) s.t. \( P(Y = \text{sign}(f(x)) | x) > c \) for some \( c > 1/2 \)

These are easy for active learning under 0-1 loss. Now let us see about under convex losses.
\[ \mathcal{F} = \{ f_t(x) = x - t : t \in \mathbb{R} \} \quad D_X = \text{Uniform}(0, 4z) \]

Slope-one Linear fns. Corresponds to “threshold” classifiers when we take signs. e.g. would be intervals if has used quadratic fns instead.

Minimizer of \( R_l \)

Increasing fn of \( \nu \)

Calculus + convexity \[ R_l(f_t) - R_l(f_{t^*}) \geq c(t - t^*)^2 \]

Let \( \nu_t \) be the \( \nu \) that would make \( t^* = t \)

More calculus \[ (t - t^*)^2 \geq (\nu_t - \nu_{t^*})^2 \]

So \[ R_l(f_t) - R_l(f_{t^*}) \leq \epsilon \Rightarrow (\nu_t - \nu_{t^*})^2 < c\epsilon \]

estimating a Bernouli mean requires \( \Omega(1/\epsilon) \) samples
Definition: Surrogate Losses

[BJM06]: For $\eta_0$ in [0,1], define
\[
l^*(\eta_0) = \inf_{z \in R} (\eta_0 l(z) + (1 - \eta_0) l(-z))
\]
\[
l_-(\eta_0) = \inf_{z \in R: z(2\eta_0 - 1) \leq 0} (\eta_0 l(z) + (1 - \eta_0) l(-z))
\]

- **Loss l** is **classification-calibrated** if, for every $\eta_0$ in [0,1]\{1/2\},
\[
l^*(\eta_0) > l^*(\eta_0)
\]
Calibration means: fn with minimal surrogate loss $\Rightarrow$ fn with minimal err
\[
l^*(\eta(X)) : \text{minimum value of conditional-risk at } X \text{ s.t.}
\]
\[
\text{sign}(h(X)) \neq \text{sign} (\eta(X) - 1/2)
\]
\[
l^*(\eta(X)) : \text{minimum conditional l-risk at } X, \text{ s.t. } E[l^*(\eta(X))] = \inf_h R_l(h)
\]

- **$\Psi$-transform of a loss fn:**
- BJM06 defined a loss-dependent function $\Psi$ to convert excess surrogate risk bounds into excess error rate bounds, specifically,
\[
(\varepsilon_r(h) - \varepsilon_r(h^*))^{\alpha/(1+\alpha)} \Psi((\varepsilon_r(h) - \varepsilon_r(h^*))^{1/(1+\alpha)}) \leq R_l(h) - R_l(h^*)
\]

- **Modulus of convexity:**
\[
\delta(\epsilon) = \max \{(f(x) + f(y))/2f((x + y)/2) : |x - y| > \epsilon \}
\]
suppose $\delta(\epsilon) \geq \epsilon^p$
Alg: A modification on ACAL stream-based

ACAL
1. $t \leftarrow 0, \mathcal{L}_t \leftarrow \emptyset, Q_t \leftarrow \emptyset$, let $\hat{h}_t$ be any element of $\mathbb{C}$
2. Do
3. $t \leftarrow t + 1$
4. Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$
5. For each $y \in \{-1, +1\}$, let $h(y) = \text{LEARN}(\mathcal{L}_{t-1} \cup \{(x_t, y)\}, Q_{t-1})$
6. If either $y$ has $h(-y) = \emptyset$ or
   \[
   \hat{e}_r(h(-y); \mathcal{L}_{t-1} \cup Q_{t-1}) - \hat{e}_r(h(y); \mathcal{L}_{t-1} \cup Q_{t-1}) > \hat{E}_{t-1}(\mathcal{L}_{t-1}, Q_{t-1})
   \]
7. $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1} \cup \{(X_t, y)\}, Q_t \leftarrow Q_{t-1}$
8. Else Request $Y_t$, and let $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1}, Q_t \leftarrow Q_{t-1} \cup \{(X_t, Y_t)\}$
9. Let $\hat{h}_t = \text{LEARN}(\mathcal{L}_t, Q_t)$
10. If $t$ is a power of 2
11. $\mathcal{L}_t \leftarrow \emptyset, Q_t \leftarrow \emptyset$

Based on subroutine:

\[
\text{LEARN}(\mathcal{L}, Q) = \arg\min_{f \in F; \text{er}_\mathcal{L}(f) = 0} R_L(f; Q)
\]

\[
\text{LEARN}(\mathcal{L}, Q) = \arg\min_{h \in \mathbb{C}: \hat{e}_r(h; \mathcal{L}) = 0} \hat{e}_r(h; Q) \text{ if } \min_{h \in \mathbb{C}} \hat{e}_r(h; \mathcal{L}) = 0, \text{ and otherwise LEARN}(\mathcal{L}, Q) = \emptyset.
\]
Can we do it efficiently?

General Results

• In general, we have results on how many labels are required to obtain a given excess error rate with this method, for general classification calibrated losses.

• Generally, if \( \varepsilon_t \) denotes the solution of
  \[
  t = \tilde{O} \left( \left( \frac{1}{\varepsilon \Psi(1-\varepsilon)} \right)^{2-2/p} \right)
  \]
  for \( \varepsilon \) in terms of \( t \), then
  \[
  \mathbb{E}[\text{excess mistakes}] = \tilde{O} \sum_{t=1}^{T} \varepsilon_t
  \]
  \[
  \mathbb{E}[\#\text{queries}] = \tilde{O} \left( \sum_{t=1}^{T} \theta(\varepsilon_t^\alpha) \varepsilon_t^\alpha \right)
  \]
  e.g., when \( l \) is squared loss = \((1-x)^2\), \( \Psi(x) = x^2 \), \( p = 2 \)
Can we do it efficiently?
(Streamed-based, just for one distri.)

- Theorem. If loss is square loss, under surrogate loss assumption, optimal fn is in fn class, fn class is VC subgraph, satisfying Tsybkov noise with exponent \(\alpha/(1-\alpha)\), alg \(A'\) has excess \#mistake

\[
\begin{align*}
E[\text{excess #mistake}] &= \tilde{O}(T^{1-\alpha}/2) \\
E[\#queries] &= \tilde{O}(\theta(T^{2-\alpha})T^{2-2\alpha}/2)
\end{align*}
\]

[Proof Sketch] By BJM06 analysis,

- If \(t = \left(\frac{1}{e^\alpha \psi(1-e^{1-\alpha})}\right)^{2-2/p}\text{polylog}(\log 1/\epsilon)\) then excess err rate < \(\epsilon\). This is sample complexity of passive learning with surrogate loss.
- E excess error under 0-1 loss
- Solve for \(t = \frac{1}{e^\alpha \psi(1-e^{1-\alpha})} = T\)
- Get current excess error rate (as fn of \(t\), bound on excess error rate, excess mistake = pr(make mistake but optimal fn doesn't)
give excess err

E[excess #mistake] sublinear - if \(\theta = o(1/\epsilon)\),
E[#queries] sublinear.
Proof Sketch (cont.)

- If the loss is squared loss, fill in all the value, we get
  \[ \sum_{t=1}^{T} \left( \frac{1}{t} \right)^{\frac{1}{2-\alpha}} = T^{\frac{1-\alpha}{2-\alpha}} \]

- How to convert excess error to \( \Pr(\text{make a query}) \)
- use Tsybakov noise condition

- Take \( \left( \frac{1}{t} \right)^{\frac{1}{2-\alpha}} \), raise to the power of \( \alpha \), get diameter
- relate that to \( \Pr(\text{in DIS}) \) by multiplying with \( \theta \) (the disagreement coefficient, taking an argument)
- do that get
  \[ \sum_{t=1}^{T} \theta(t^{\frac{-\alpha}{2-\alpha}}) \left( \frac{1}{t} \right)^{\frac{\alpha}{2-\alpha}} \leq \theta(T^{\frac{-\alpha}{2-\alpha}}) \sum_{t=1}^{T} t^{\frac{-\alpha}{2-\alpha}} = T^{\frac{2-2\alpha}{2-\alpha}} \]
- plug in the bound on the diameter
- If \( \theta \) is \( o(1/\varepsilon) \), this is sublinear
Thanks!