Selective Sampling (Realizable)

Ji Xu

October 2nd, 2017

Basic Settings

Model:

- D: a distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the input space and $\mathcal{Y} = \{\pm 1\}$ are the possible labels.
- $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a pair of random variables with joint distribution D.
- $\mathcal{H}$ be a set of hypotheses mapping from $\mathcal{X}$ to $\mathcal{Y}$. The error of a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ is

$$err(h) := \Pr(h(X) \neq Y).$$

- Let $h^* := \operatorname{argmin}\{err(h) : h \in \mathcal{H}\}$ be a hypothesis with minimum error in $\mathcal{H}$.

Basic Settings

Goal: with high probability, we return $\hat{h} \in \mathcal{H}$ such that

$$err(\hat{h}) \leq err(h^*) + \epsilon.$$

In realizable case, we have $err(h^*) = 0$, hence, we want

$$err(\hat{h}) \leq \epsilon.$$

Basic Settings

Passive VS Active:

- Passive setting:
  - At time t, observe X_t and choose $h_t \in \mathcal{H}$.
  - Make prediction $h_t(X_t)$ and then observe feedback Y_t.
  - Minimize the total number of mistakes of $h_t(X_t) \neq Y_t$.

Basic Settings

Passive VS Active:

- Active setting:
  - At time t, observe X_t.
  - We choose whether we need the feedback Y_t.
  - Minimize the number of mistakes of $\hat{h}$ and the total number of queries of the correct label Y_t.

Basic Settings

Passive VS Active:

- Active setting:
  - At time t, observe X_t.
  - We choose whether we need the feedback Y_t.
  - Minimize the number of mistakes of $\hat{h}$ and the total number of queries of the correct label Y_t.

Hence, intuitively, (X_t, Y_t) does not provide any information if $h(X_t)$ are the same for all the potential hypotheses at time t, and thus we should not query for such X_t.

Concepts

Definition

For a set of hypotheses $\mathcal{V}$, the region of disagreement $R(\mathcal{V})$ is

$$R(\mathcal{V}) := \{x \in \mathcal{X} : \exists h, h' \in \mathcal{V} \text{ such that } h(x) \neq h'(x)\}.$$

Definition

For a given set of hypotheses $\mathcal{H}$ and sample set

$$Z_T = \{(X_t, Y_t), t = 1 \cdots T\},$$

the uncertainty region $U(\mathcal{H}, Z_T)$ is

$$\begin{aligned} U(\mathcal{H}, Z_T) \quad := \quad & \{x \in \mathcal{X} : \exists h, h' \in \mathcal{H} \text{ such that } h(x) \neq h'(x) \\ & \text{and } h(X_t) = h'(X_t) = Y_t, \forall t \in [T]\}. \end{aligned}$$

Remarks

- Let $C = \{h \in \mathcal{H} : h(X_t) = Y_t, \forall t \in [T]\}$. Then we have

$$U(\mathcal{H}, Z_T) \ = \ R(C).$$

- Ideally, the area of the uncertainty region will be monotonically non-increasing by more training samples.
- If we can control the sampling procedure over X_t, it is better to only sample on $U(\mathcal{H}, Z_t)$. (Selective Sampling or Approximate Selective Sampling)
- Correctness of all labels Y_t for X_t not in the query. Need to query X_{t+1} if $X_{t+1} \in U(\mathcal{H}, Z_t)$.
- The complexity of finding a good set $\hat{\mathcal{H}}$ such that $h^* \in \hat{\mathcal{H}} \subseteq \mathcal{H}$ can be intuitively measured by the ratio between $\sup_{h \in \hat{\mathcal{H}}} err(h)$ and $\Pr(R(\hat{\mathcal{H}}))$.

Concepts

Definition

We redefine the region of disagreement by $R(h, r)$ of radius r around a hypothesis $h \in \mathcal{H}$ in the disagreement metric space $(\mathcal{H}, \rho)$ is

$$R(h, r) := \{x \in \mathcal{X} : \exists h' \in B(h, r) \text{ such that } h(x) \neq h'(x)\}.$$

where the disagreement (pseudo) metric ρ on $\mathcal{H}$ is defined by

$$\rho(h, h') \ := \ \Pr(h(X) \neq h'(X)).$$

Hence, we have $err(h) = \rho(h, h^*)$.

Concepts

Definition

We redefine the region of disagreement by $R(h, r)$ of radius r around a hypothesis $h \in \mathcal{H}$ in the disagreement metric space $(\mathcal{H}, \rho)$ is

$$R(h, r) := \{x \in \mathcal{X} : \exists h' \in B(h, r) \text{ such that } h(x) \neq h'(x)\}.$$

where the disagreement (pseudo) metric ρ on $\mathcal{H}$ is defined by

$$\rho(h, h') \;:=\; \Pr(h(X) \neq h'(X)).$$

Hence, we have $err(h) = \rho(h, h^*)$.

Remarks: We have $R(h^*, r) \subseteq R(B(h^*, r))$, but the reverse may not be true.

Concepts

Definition

The disagreement coefficient $\theta(h, \mathcal{H}, D)$ with respect to a hypothesis $h \in \mathcal{H}$ in the disagreement metric space $(\mathcal{H}, \rho)$ is

$$\theta(h, \mathcal{H}, D) \;:=\; \sup_{r>0} \frac{\Pr\left(X \in R(h, r)\right)}{r}.$$

Concepts

Definition

The disagreement coefficient $\theta(h, \mathcal{H}, D)$ with respect to a hypothesis $h \in \mathcal{H}$ in the disagreement metric space $(\mathcal{H}, \rho)$ is

$$\theta(h, \mathcal{H}, D) \ := \ \sup_{r>0} \frac{\Pr\left(X \in R(h, r)\right)}{r}.$$

Examples:

- X is uniform on $[0, 1]$. $\mathcal{H} = \{h = I_{X \geq r}, \forall r > 0\}$. Then $\theta(h, \mathcal{H}, D) \ = \ 2, \forall h \in \mathcal{H}$.
- Replace $\mathcal{H}$ by $\mathcal{H} = \{h = I_{X \in [a,b]}, \forall 0 < a < b < 1\}$. Then

$$\theta(h, \mathcal{H}, D) \ = \ \max(4, 1/Pr(h(X) = 1)), \quad \forall h \in \mathcal{H}.$$

Examples

Proposition

Let P_X be the uniform distribution on the unit sphere $S^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\} \subset \mathbb{R}^d$, and let $\mathcal{H}$ be the class of homogeneous linear threshold functions in $\mathbb{R}^d$, i.e,

$$\mathcal{H} = \{h_w : h_w(x) = sign(\langle w, x \rangle), \forall w \in S^{d-1}\}.$$

There is an absolute constant $C > 0$ such that

$$\theta(h, \mathcal{H}, P_X) \leq C \cdot \sqrt{d}.$$

Algorithm (CAL)

- Initialize: $Z_0 := \emptyset, \mathcal{V}_0 := \mathcal{H}$.
- For $t = 1, 2, \cdots, n$:
  - Obtain unlabeled data point X_t.
  - If $X_t \in R(\mathcal{V}_{t-1})$:
    (a) Then: Query Y_t, and set $Z_t := Z_{t-1} \bigcup \{(X_t, Y_t)\}$.
    (b) Else: Set $\tilde{Y}_t := h(X_t)$ for any $h \in \mathcal{V}_{t-1}$, and set $Z_t := Z_{t-1} \bigcup \{(X_t, \tilde{Y}_t)\}$ OR
    Set $Z_t := Z_{t-1}$.
  - Set $\mathcal{V}_t := \{h \in \mathcal{H} : h(X_i) = Y_i, \forall (X_i, Y_i) \in Z_t\}$.
- Return: any $h \in \mathcal{V}_n$.

Algorithm (Reduction-based CAL)

- Initialize: $Z_0 := \emptyset$.
- For $t = 1, 2, \cdots, n$:
  - Obtain unlabeled data point X_t.
  - If there exists both:
    - $h^+ \in \mathcal{H}$ consistent with $Z_{t-1} \bigcup \{(X_t, +1)\}$
    - $h^- \in \mathcal{H}$ consistent with $Z_{t-1} \bigcup \{(X_t, -1)\}$

    (a) Then: Query Y_t, and set $Z_t := Z_{t-1} \bigcup \{(X_t, Y_t)\}$.

    (b) Else: only h^y exists for some $y \in \{\pm 1\}$: Set $\tilde{Y}_t := y$ and set $Z_t := Z_{t-1} \bigcup \{(X_t, \tilde{Y}_t)\}$
- Return: any $h \in \mathcal{H}$ consistent with Z_n.

Algorithm (Reduction-based CAL)

- Initialize: $Z_0 := \emptyset$.
- For $t = 1, 2, \cdots, n$:
  - Obtain unlabeled data point X_t.
  - If there exists both:
    - $h^+ \in \mathcal{H}$ consistent with $Z_{t-1} \bigcup \{(X_t, +1)\}$
    - $h^- \in \mathcal{H}$ consistent with $Z_{t-1} \bigcup \{(X_t, -1)\}$

    (a) Then: Query Y_t, and set $Z_t := Z_{t-1} \bigcup \{(X_t, Y_t)\}$.

    (b) Else: only h^y exists for some $y \in \{\pm 1\}$: Set $\tilde{Y}_t := y$ and set $Z_t := Z_{t-1} \bigcup \{(X_t, \tilde{Y}_t)\}$
- Return: any $h \in \mathcal{H}$ consistent with Z_n.

Remark: Reduction-based CAL is equivalent to CAL.

Label Complexity Analysis

Theorem

The expected number of labels queried by Reduction-based CAL after n iterations is at most

$$O\left(\theta(h^*, \mathcal{H}, D) d \log^2 n\right),$$

where d is the VC-dimension of class $\mathcal{H}$. For any $\epsilon > 0$ and $\delta > 0$, if we have

$$n = O\left(\frac{1}{\epsilon}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta})\right),$$

then with probability $1 - \delta$, the return of Reduction-based CAL $\hat{h}$ satisfies that

$$err(\hat{h}) \leq \epsilon.$$

Proof

Note that, with probability $1-\delta_t$, any $h \in \mathcal{H}$ consistent with Z_t has error $err(h)$ at most

$$O\left(\frac{1}{t}\left(d\log t + \log\frac{1}{\delta_t}\right)\right) := r_t,$$

where $\delta_t > 0$ will be chosen later. (case when $P_n f_n = 0, Pf = 0$). This also implies that $n = O\left(\frac{1}{\epsilon}(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta})\right)$

Proof

Note that, with probability $1 - \delta_t$, any $h \in \mathcal{H}$ consistent with Z_t has error $err(h)$ at most

$$O\left(\frac{1}{t}\left(d \log t + \log \frac{1}{\delta_t}\right)\right) := r_t,$$

where $\delta_t > 0$ will be chosen later. (case when $P_n f_n = 0, Pf = 0$). This also implies that $n = O\left(\frac{1}{\epsilon}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta})\right)$

Let G_t is the event that described above happens. Hence, condition on G_t, we have

$$\{h \in \mathcal{H} : h \text{ is consistent with } Z_t\} \;\subseteq\; B(h^*, r_t).$$

Proof

Note that, we query Y_{t+1} if and only if

$$\exists h \in \mathcal{H} \text{ consistent with } Z_t \bigcup \{(X_{t+1}, -h^*(X_{t+1}))\},$$

(i.e., there is h disagree with h^*)
Hence, condition on G_t, if we query Y_{t+1}, then $X_{t+1} \in R(h^*, r_t)$.
Therefore, we have

$$\Pr(Y_{t+1} \text{ is queried} \big| G_t) \;\leq\; \Pr(X_{t+1} \in R(h^*, r_t) | G_t).$$

Proof

Let $Q_t = I_{\{Y_t \text{ is queried}\}}$. The expected total number of queries is

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{n} Q_t] &\leq 1 + \sum_{t=1}^{n-1} \Pr(Q_{t+1} = 1) \\
&= 1 + \sum_{t=1}^{n-1} \Pr(Q_{t+1} = 1 \big| G_t)\Pr(G_t) \\
&\quad + \sum_{t=0}^{n-1} \Pr(Q_{t+1} = 1 \big| \text{ not } G_t)(1 - \Pr(G_t)) \\
&\leq 1 + \sum_{t=1}^{n-1} \Pr(Q_{t+1} = 1 \big| G_t)\Pr(G_t) + \delta_t \\
&\leq 1 + \sum_{t=1}^{n-1} \Pr(X_{t+1} \in R(h^*, r_t) | G_t)\Pr(G_t) + \delta_t.
\end{aligned}$$

Proof

By definition of the coefficient of disagreement, we have

$$\Pr(X_{t+1} \in R(h^*, r_t) | G_t) \Pr(G_t) \leq \Pr(X_{t+1} \in R(h^*, r_t)) \leq r_t \cdot \theta(h^*, \mathcal{H}, D).$$

Hence, we have

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{n} Q_t] &\leq 1 + \sum_{t=1}^{n-1} r_t \cdot \theta(h^*, \mathcal{H}, D) + \delta_t \\
&= \sum_{t=1}^{n-1} O\left(\frac{\theta(h^*, \mathcal{H}, D)}{t}\left(d \log t + \log \frac{1}{\delta_t}\right) + \delta_t\right).
\end{aligned}$$

Proof

By definition of the coefficient of disagreement, we have

$$\Pr(X_{t+1} \in R(h^*, r_t) | G_t)\Pr(G_t) \leq \Pr(X_{t+1} \in R(h^*, r_t)) \leq r_t \cdot \theta(h^*, \mathcal{H}, D).$$

Hence, we have

$$\begin{aligned} \mathbb{E}[\sum_{t=1}^{n} Q_t] &\leq 1 + \sum_{t=1}^{n-1} r_t \cdot \theta(h^*, \mathcal{H}, D) + \delta_t \\ &= \sum_{t=1}^{n-1} O\left(\frac{\theta(h^*, \mathcal{H}, D)}{t}\left(d \log t + \log \frac{1}{\delta_t}\right) + \delta_t\right). \end{aligned}$$

Choose $\delta_t = \frac{1}{t}$, we have

$$\mathbb{E}[\sum_{t=1}^{n} Q_t] \leq O\left(\theta(h^*, \mathcal{H}, D) d \log^2 n\right).$$