Passive Learner

Theorem 10. *Suppose that we have a finite set of hypotheses $\mathcal{H}$ (i.e., $|\mathcal{H}| < \infty$) and $\hat{h}_n = \arg\min_{h\in\mathcal{H}} \hat{R}_n(h)$. Also, assume that the data is separable (i.e., the perfect classifier h^* with no error exists). For any $\epsilon, \delta \in (0,1)$, we have $\Pr\left(R(\hat{h}_n) > \epsilon\right) \leq \delta$ whenever $n \geq \epsilon^{-1}\log(\frac{|\mathcal{H}|}{\delta})$. In other words, for any $\epsilon, \delta \in (0,1)$, with probability $1-\delta$, we have $R(\hat{h}_n) \leq \frac{\log(\frac{|\mathcal{H}|}{\delta})}{n}$*

$(x_i, y_i) \in \mathcal{X} \times \{0,1\}$ $\quad$ $(x_i, y_i) \overset{iid}{\sim} \mathcal{D}$ $\quad$ $i = 1, 2, \dots, n$

Assume $h^* = \arg\min_h R(h)$ and $R(h^*) = 0$.

$$R(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\mathbb{1}\{h(X) \neq Y\}\right]$$

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(x_i) \neq y_i\}$$

Active Learner for streaming/sampling-oracle setting.

Separable setting

Assume $\min_h R(h) = 0$.

Algorithm 1 CAL

```
Initialize: Z_0 = ∅, V_0 = H
for t = 1, 2...n do
    Nature reveals unlabeled data point x_t
    if x_t ∈ DIS(V_{t-1}) then
        Query y_t, and set Z_t = Z_{t-1} ∪ (x_t, y_t)
    else
        Z_t = Z_{t-1}
    end if
    V_t = {h ∈ H : h(x_i) = y_i ∀(x_i, y_i) ∈ Z_t}   ← "Version space"
end for
return any h ∈ V_n
```

Definition 6. *For some hypothesis class $\mathcal{H}$ and subset $V \subset \mathcal{H}$ where for each $h \in \mathcal{H}, h : \mathcal{X} \to \{0,1\}$, the* region of disagreement *is defined as*

$$DIS(V) = \{x \in \mathcal{X} : \exists\, h, h' \in \cancel{\mathcal{H}}\, V \text{ s.t. } h(x) \neq h'(x)\}$$

which is the set of unlabeled examples x for which there are hypotheses in V that disagree on how to label x.

$h^* \in V_t \quad \forall t$

Algorithm 1 CAL

Initialize: $Z_0 = \emptyset$, $V_0 = \mathcal{H}$
for $t = 1, 2 \ldots n$ **do**
  Nature reveals unlabeled data point x_t
  if $x_t \in DIS(V_{t-1})$ **then**
    Query y_t, and set $Z_t = Z_{t-1} \cup (x_t, y_t)$
  else
    $Z_t = Z_{t-1}$ // $y_t = h^*(x_t) = h(x_t) \quad \forall h \in V_{t-1}$
  end if
  $V_t = \{h \in \mathcal{H} : h(x_i) = y_i \ \ \forall (x_i, y_i) \in Z_t\}$
end for
return any $h \in V_n$

Algorithm 2 Efficient CAL

Initialize: $Z_0 = \emptyset$
for $t = 1, 2 \ldots n$ **do**
  Nature reveals unlabeled data point x_t
  if for $\widehat{y} \in \{0, 1\}$ $\exists h_{\widehat{y}} \in \mathcal{H} : h_{\widehat{y}}(x_s) = y_s$, $\forall (x_s, y_s) \in Z_{t-1} \cup (x_t, \widehat{y})$ **then**
    Query y_t, and set $Z_t = Z_{t-1} \cup (x_t, y_t)$
  else
    $Z_t = Z_{t-1}$
  end if
end for
return $\arg\min_{h \in \mathcal{H}} \sum_{(x,y) \in Z_t} \mathbf{1}\{h(x) \neq y\}$.

$B(h, r) = \{h' \in \mathcal{H} : \mathbb{E}_x[\mathbb{1}\{h(x) \neq h'(x)\}] \leq r\}$

Definition 7. *The* disagreement coefficient *of $h \in \mathcal{H}$ with respect to a hypothesis class $\mathcal{H}$ and distribution $\mathcal{D}_X$ is defined as*

$$\theta_{h^*} = \sup_r \frac{\mathbb{P}_{X \sim \mathcal{D}_X}(X \in DIS(B(h^*, r)))}{r}$$

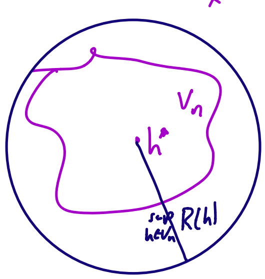
$\theta^*(\varepsilon) = \sup_{r > \varepsilon} \cdots \leq \frac{1}{\varepsilon}$

Theorem 11. *Let $h^* = \arg\min_{h \in \mathcal{H}} R(h)$ and assume $R(h^*) = 0$. Suppose n iid labeled examples $\{(x_i, y_i)\}_{i=1}^n$ are drawn from $\mathcal{D}$ and $V_n = \{h \in \mathcal{H} : h(x_i) = y_i \ \forall i \in [n]\}$. If we request λ additional labels only when the samples lie in the disagreement region $DIS(V_n)$, where $\lambda = 2\theta_{h^*} \log(|\mathcal{H}|/\delta)$, then, with probability greater than $1 - \delta$ we have $\sup_{h \in V_{n+\lambda}} R(h) \leq \sup_{h \in V_n} \frac{1}{2} R(h)$.*

$$\frac{\mathbb{P}_x(X \in DIS(V_n))}{\sup_{h \in V_n} R(h)} \leq \frac{\mathbb{P}(X \in DIS(B(h^*, \sup_{h \in V_n} R(h))))}{\sup_{h \in V_n} R(h)}$$

$$\leq \theta^*_{h^*} \implies P_x(X \in DIS(V_n)) \leq \theta^*_{h^*} \cdot \sup_h R(h)$$

$$\mathbb{E}_x[\mathbb{1}\{h(x) \neq h^*(x)\}] = \mathbb{E}_x[\mathbb{1}\{h(x) \neq Y\}] = R(h)$$



$$\sup_{h\in V_{n+\lambda}} R(h) = \sup_{h\in V_{n+\lambda}} P(Y\neq h(X))$$

$$= \sup_{h\in V_{n+\lambda}} P(Y\neq h(X) \mid X\in DIS(V_n))\,P(X\in DIS(V_n))$$

$$+ \overset{=0}{\cancel{P(\underset{\sim}{Y}\neq h(X) \mid X\notin DIS(V_n))}}\,P(X\notin DIS(V_n))$$

$Y = h^*(X)$, $h, h^* \in V_n$, $X \notin DIS(V_n)$ $\Rightarrow h(X) = h^*(X)$

$$= \underbrace{\sup_{h\in V_{n+\lambda}} P(Y\neq h(X) \mid X\in DIS(V_n))}_{\leq \frac{\log(|\mathcal{H}|/\delta)}{\lambda}}\,\underbrace{P(X\in DIS(V_n))}_{\leq \theta^*_{h^*}\cdot \sup_{h\in V_n} R(h)}$$

$$\leq \frac{1}{2}\sup_{h\in V_n} R(h).$$

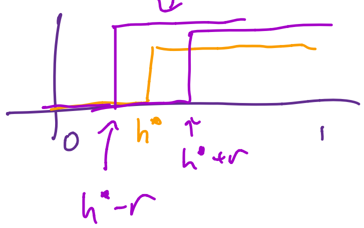
Conclude that every λ labels we halve the risk. If we take $n = k\lambda$ labels total then w.p. $1-k\delta$, $R(\hat{h}) \leq 2^{-k} = 2^{-n/\lambda}$

$$\leq \exp\left(-\frac{n}{2\theta^* \log(|\mathcal{H}|/\delta)}\right)$$

Equivalently, $R(\hat{h}) \leq \varepsilon$ after $\theta^* \log(|\mathcal{H}|/\delta)\log(1/\varepsilon)$ labels.

} cases where $\theta^* = \Theta(1)$.

Ex. $\mathcal{H}$ are thresholds on unit interval ($\mathcal{D}_X$ is <u>uniform</u>($[0,1]$))



$B(h^*, r)$

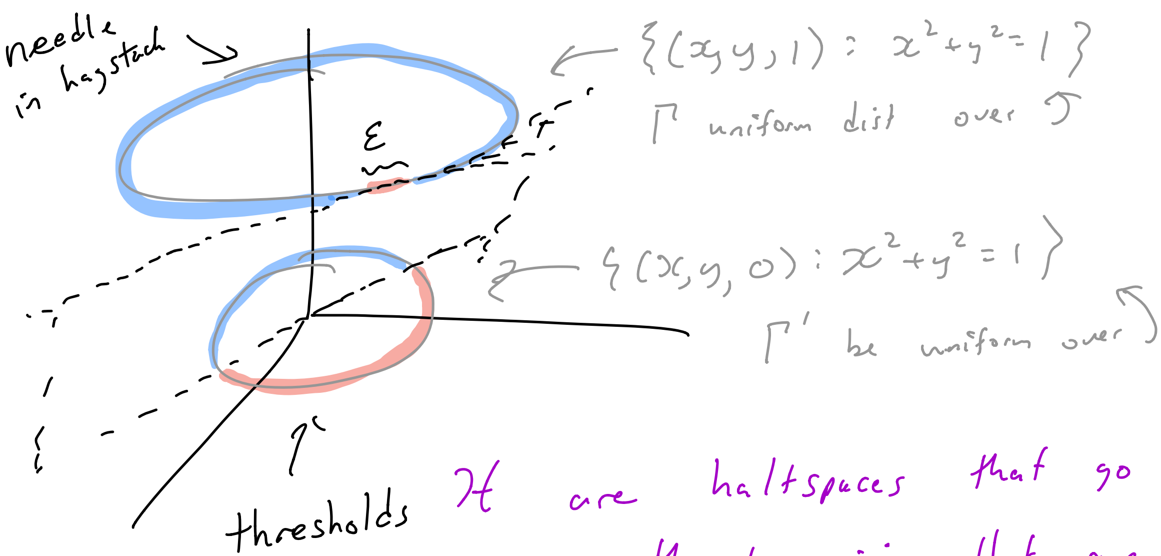
$\mathbb{P}(X \in DIS(B(h^*, r))) = 2r$

$\Rightarrow \theta^*(\varepsilon) \leq \frac{2r}{r} = 2.$

If $\mathcal{H}$ halfspaces going through origin

and $\mathcal{D}_X$ is uniform on sphere

then $\theta^*(\varepsilon) \lesssim \sqrt{d}$



$\mathcal{H}$ are halfspaces that go through origin that are tilted to chop off a tiny bit of Γ.

Assume $h^* \in \mathcal{H}$ w/ $Y = h^*(X)$

$$\mathcal{D}_X = (1-\zeta)\,\Gamma + \zeta\,\Gamma' \quad \text{for } \zeta \ll 1$$

If $\zeta = 1$ CAL achieves $\log(1/\varepsilon)$ labels

If $\zeta = 0$ any algorithm requires $1/\varepsilon$ labels.

If you run CAL and wait for samples only from Γ', requires $\log(1/\varepsilon)$ labels

and $\frac{1}{\zeta}\log(1/\varepsilon)$ unlabeled data

Consider a finite hypothesis space $\mathcal{H}$ and consider any $Q \subset \binom{\mathcal{H}}{2}$ where $(h, h') \in Q$ can be considered an edge connecting any two hypotheses. For any $\widehat{y} \in \{0, 1\}$ define $\mathcal{H}_{(x,\widehat{y})} = \{h \in \mathcal{H} : h(x) = \widehat{y}\}$. We say an example x *ρ-splits* Q if requesting its label reduces the number of edges by at least a fraction $\rho \in (0, 1)$:

$$\max\{|Q \cap \mathcal{H}_{(x,0)}|, |Q \cap \mathcal{H}_{(x,1)}|\} \leq (1 - \rho)|Q|.$$

We are now ready to introduce the splitting index.

Definition 8. *Fix any subset $S \subset \mathcal{H}$ and $Q \subset \binom{S}{2}$ such that $\mathbb{P}(h(X) \neq h'(X)) \geq \epsilon, \forall (h, h') \in Q$. Then we say S is (ρ, ϵ, τ)-*splittable *if $\mathbb{P}(X$ splits $Q) \geq \tau$.*

Basically, the definition is saying that to reduce the number of pairs of hypotheses that differ by at least ϵ by a fraction at least ρ, requires $1/\tau$ unlabeled data. If $\mathcal{H}$ is finite, and $\mathcal{H}$ is (ρ, ϵ, τ)-splittable, then it is almost immediate that there exists an algorithm that requires $1/(\tau\rho)$ unlabeled data and $1/\rho$ labels to identify an ϵ-good classifier ([Dasgupta, 2005b] suggests one, though it is computationally intractable). What is more important is the reproduced lower bound:

Theorem 12 ([Dasgupta, 2005b])**.** *Fix any hypothesis space $\mathcal{H}$ and distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$. Suppose that for some $\rho \in (0, 1)$, $\epsilon \in (0, 1)$ and some $\tau \in (0, 1/2)$, the set $S \subset \mathcal{H}$ is not (ρ, ϵ, τ)-splittable. Then any active learning strategy that achieves an accuracy of $\epsilon/2$ on all target hypotheses in S must, with probability at least $3/4$ (taken over the random sampling of data), either draw $\geq 1/\tau$ unlabeled samples, or must request $\geq 1/\rho$ labels.*