Realizable Active Learning Notes

Kevin Jamieson

November 11, 2010

Definition 1. *For a hypothesis class $\mathcal{H}$ and dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in X$ and $y_i \in \{-1, 1\}$ we call $\hat{h}$ the empirical risk minimizer if $\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{R}(h)$ where*

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

Definition 2. *If a finite dataset is identically and independently distributed from some distribution, that is, $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}_{X,Y}$ then the true risk of a hypothesis $h \in \mathcal{H}$ is denoted*

$$R(h) = P(h(X) \neq Y) = \int_X \mathbf{1}\{h(X) \neq Y\} dP(X)$$

Definition 3. *After observing n labels we call the set of all hypotheses still "in the running" for being the empirical risk minimizer over all observed data the version space and denote it $V_n \subset \mathcal{H}$. If $f = \arg\min_{h\in\mathcal{H}} R(h)$ and $R(f) = 0$ then*

$$V_n = \{h \in \mathcal{H} : \hat{R}(h) = 0\}$$

and $f \in V_n$ for all $n \in \mathbb{N}$.

Theorem 1. *Vapnik (1982) For a hypothesis class $\mathcal{H}$ with finite VC dimension d and an iid sample $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}_{X,Y}$ with $f = \arg\min_{h\in\mathcal{H}} R(h)$ and $R(f) = 0$ then with probability greater than $1 - \delta$*

$$\sup_{h\in V_n} R(h) \leq \frac{4d \log \frac{2en}{d} + \log \frac{2}{\delta}}{n}.$$

Definition 4. *For some hypothesis class $\mathcal{H}$ and set $\mathcal{X}$ where for $h \in \mathcal{H}$, $h : \mathcal{X} \to \{-1, 1\}$, the region of disagreement is defined as*

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

Observation 1. *Assume $f = \arg\min_{h\in\mathcal{H}} R(h)$ and $R(f) = 0$. If after n labeled examples we observe an $X \notin DIS(V_n)$ then we need not ask for its label because no two hypotheses in V_n disagree on its label and because $f \in V_n$ its label is deterministic.*

Definition 5. *For some hypothesis class $\mathcal{H}$ and the marginal $\mathcal{D}_X$ of $\mathcal{D}_{X,Y}$ the closed ball centered at $h \in \mathcal{H}$ with radius r is defined as*

$$B(h, r) = \{h' \in \mathcal{H} : P_{X\sim\mathcal{D}}(h(X) \neq h'(X)) \leq r\}.$$

Definition 6. *The disagreement coefficient of $h \in \mathcal{H}$ with respect to $\mathcal{H}$ and $\mathcal{D}_{X,Y}$ is*

$$\theta_h = \sup_r \frac{P(DIS(B(h, r)))}{r}.$$

Lemma 1. *If $f = \arg\min_{h\in\mathcal{H}} R(h)$ and $R(f) = 0$ then after n labeled examples there is a nonempty subset $V_n \subset \mathcal{H}$ s.t. $V_n = \{h \in \mathcal{H} : \hat{R}(h) = 0\}$ and if we simulate samples $x \sim \mathcal{D}_X$ requesting labels only when $x \in DIS(V_n)$ then with probability greater than $1 - \delta/n$ after at most $\lambda_n = 4\theta_f(4d \log(15e\theta_f) + \log(2n/\delta))$ label requests*

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

for some $m \geq n + \lambda_n$.

Proof. The disagreement coefficient allows for a bound that relates the region of disagreement to the true risk of any $h \in V_n$:

$$\begin{aligned}\frac{P(DIS(V_n))}{\sup_{h\in V_n} R(h)} &\leq \frac{P(DIS(B(f, \sup_{h\in V_n} R(h))))}{\sup_{h\in V_n} R(h)} \\ &\leq \sup_r \frac{P(DIS(B(f,r)))}{r} \\ &= \theta_f\end{aligned}$$

which implies

$$\frac{P(DIS(V_n))}{\theta_f} \leq \sup_{h\in V_n} R(h) \tag{1}$$

and there exists some $m > n$ such that

$$\begin{aligned}\sup_{h\in V_m} R(h) &\leq \sup_{h\in V_m} R(h(X)|X \in DIS(V_n))P(DIS(V_n)) &(2)\\ &\leq \frac{4d\log\frac{2e\lambda_n}{d} + \log\frac{2n}{\delta}}{\lambda_n} \cdot P(DIS(V_n)) &(3)\\ &\leq \frac{P(DIS(V_n))}{2\theta_f} &(4)\\ &\leq \sup_{h\in V_n} \frac{R(h)}{2} &(5)\end{aligned}$$

where (2) follows from Observation 1, (3) and (4) follow from Theorem 1 and the definition of λ_n and finally (5) follows from (1). □

Theorem 2. *Under the same setting and assumptions of Lemma 1, for any $t \in \mathbb{N}$ with probability greater than $1-\delta$ the true risk of $\hat{h}$ after t label requests satisfies*

$$R(\hat{h}) \leq 2 \cdot exp\Big\{ - \frac{t}{6\theta_f(4d\log(44\theta_f) + \log(2t/\delta))} \Big\}.$$

Proof. Lemma 1 says that after just λ_t labeled examples we have $\sup_{h\in V_{n+\lambda_t}} R(h) \leq \sup_{h\in V_n} R(h)/2$ with probability greater than $1-\delta/t$. So after $t \geq \lambda_t\lceil\log_2(1/\epsilon)\rceil$ labeled examples $\sup_{h\in V_n} R(h) \leq \epsilon$ with probability greater than $1-\delta$ by taking a union bound over $n = \lambda_t, 2\lambda_t, \ldots, \lceil t/\lambda_t\rceil$. Solving for ϵ in terms of t gives the result. □

Some disagreement coefficients

With the exception of very nice situations (uniform distribution, symmetric geometry, etc) the disagreement coefficient is often impossible to calculate. However, Theorem 2 shows that θ just needs to be finite to learn with $O(\log(1/\epsilon))$ labels and this is often done in the literature.

- Thresholds on $\mathbb{R}$: $\theta = 2$
- Homogeneous hyperplanes in $\mathbb{R}^d$ with data uniformly distributed on a sphere: $\theta \leq \sqrt{d}$
- General hyperplanes in $\mathbb{R}^d$ with the data density bounded below: $\theta = O(d)$
- Intervals $[a, b]$ on $\mathbb{R}$: $\theta = \infty$