Notes 17: Random Classification Noise and Statistical Query models

1. Random Classification Noise (RCN)

Variant of PAC model where labels may be corrupted with probability η $(0 \leqslant \eta < 1/2)$
Let $c \subseteq X$ be a concept and $\mathcal{D}$ be a distribution over instances space X
$\mathrm{EX}^\eta(c, \mathcal{D})$ = distribution of labeled samples $(x, y) \in X \times \{+1, -1\}$
x is drawn from $\mathcal{D}$
$y = c(x)$ with probability $1 - \eta$ (correct label)
$y = -c(x)$ with probability η (flipped label)
Definition: Algorithm A efficiently PAC-learns $\mathcal{C}$ with RCN if
for any target concept $c \in \mathcal{C}$, any distribution $\mathcal{D}$ over X
for any accuracy $\varepsilon > 0$, confidence $\delta > 0$, noise rate $0 \leqslant \eta < 1/2$
given samples from $\mathrm{EX}^\eta(c, \mathcal{D})$
with prob. $\geqslant 1 - \delta$, A outputs polynomially evaluatable hypothesis h with $\mathrm{err}_{\mathcal{D}}(h, c) \leqslant \varepsilon$
A runs in time $\mathrm{poly}(n, \mathrm{size}(c), 1/\varepsilon, 1/\delta, 1/(1 - 2\eta))$
If $\eta = 1/2$, label y is uniformly random and unrelated to x $\implies$ no learning is possible
$1 - 2\eta$ = distance to impossible learning
Strictly harder than (noiseless) PAC learning ($\eta = 0$ reduces to usual PAC)
error of h is still measured with respect to c, not y
η assumed to be known to A

---

2. Monotone Conjunctions

PAC-learning $\mathcal{C}$ = {monotone conjunctions} over $X = \{0, 1\}^n$ with RCN
Original algorithm (eliminate variables inconsistent with labeled samples) breaks down
Idea: individual examples cannot be trusted, but statistics of whole data set can
For each variable x_i, let $p_i = \mathbb{P}_{x\sim\mathcal{D}}[x_i = 0 \text{ and } c(x) = 1]$
If variable x_i belongs to $c(x)$, then $p_i = 0$
Each variable x_i not in $c(x)$ adds at most p_i to $\mathrm{err}_{\mathcal{D}}(h, c)$ if $h(x)$ contains x_i
Algorithm aims to (1) include all x_i in $c(x)$ (2) exclude all x_i with $p_i > \varepsilon/n$
Even if hypothesis $h(x)$ includes some x_i with $p_i \leqslant \varepsilon/n$, error is still $\leqslant \varepsilon$
Can estimate p_i by $\hat{p}_i$ using empirical samples $(x^1, y^1), \ldots, (x^m, y^m)$
If Algorithm instead gets noiseless samples from $\mathrm{EX}(c, \mathcal{D})$
Let $\hat{p}_i = \mathbb{E}_{j\in\{1,\ldots,m\}}[x_i^j = 0 \text{ and } y^j = 1]$
Hoeffding + Union bound: with prob. $\geqslant 1 - \delta$, every $\hat{p}_i$ is within $\pm\varepsilon/(2n)$ of p_i, provided
$m \geqslant \Omega\left(\frac{n^2}{\varepsilon^2} \ln \frac{n}{\delta}\right)$ (exercise)

Theorem 1 (Hoeffding). *Let $X_1, \ldots, X_n$ be independent random variables in $[0, 1]$. Let $\overline{X} = \frac{1}{n}\sum_{1\leqslant i\leqslant n} X_i$ be their empirical average. Then for any $t \geqslant 0$,*

$$\mathbb{P}[\overline{X} \geqslant \mathbb{E}[\overline{X}] + t] \leqslant \exp(-2nt^2) .$$

See Wikipedia page on Hoeffding's inequality for a proof if interested

---

But Algorithm only gets noisy samples from $\mathrm{EX}^\eta(c, \mathcal{D})$
$p_i = \mathbb{P}_{x\sim\mathcal{D}}[x_i = 0 \text{ and } c(x) = 1] = \mathbb{E}_{(x,y)\sim\mathrm{EX}(c,\mathcal{D})}[\varphi(x, y)]$ where $\varphi : X \times \{1, -1\} \to \{0, 1\}$ is

$$\varphi(x, y) = \mathbb{1}(x_i = 0)\mathbb{1}(y = 1) = \mathbb{1}(x_i = 0)\frac{1 + y}{2} = \underbrace{\frac{1}{2}\mathbb{1}(x_i = 0)}_{\text{independent of } y} + \underbrace{\frac{1}{2}\mathbb{1}(x_i = 0) \cdot y}_{\text{linear in } y}$$

1st term (independent of y) is the same under noisy and noiseless distributions
Algorithm can estimate expected value of 1st term within $\pm\varepsilon/(4n)$
Since $\mathbb{1}(x_i = 0)/2$ takes either 0 or 1/2 value

Estimate is accurate with prob $\geqslant 1 - \delta/2n$ using $O\left(\frac{n^2}{\varepsilon^2} \ln \frac{n}{\delta}\right)$ samples (Hoeffding)

2nd term (linear in y) has expectation

$$\mathop{\mathbb{E}}_{\mathrm{EX}^\eta(c,\mathcal{D})}\left[\frac{1}{2}\mathbb{1}(x_i = 0) \cdot y\right] = (1-\eta) \mathop{\mathbb{E}}_{\mathrm{EX}(c,\mathcal{D})}\left[\frac{1}{2}\mathbb{1}(x_i = 0) \cdot y\right] + \eta \mathop{\mathbb{E}}_{\mathrm{EX}(c,\mathcal{D})}\left[\frac{1}{2}\mathbb{1}(x_i = 0) \cdot -y\right]$$
$$= (1-2\eta) \mathop{\mathbb{E}}_{\mathrm{EX}(c,\mathcal{D})}\left[\frac{1}{2}\mathbb{1}(x_1 = 0) \cdot y\right]$$

i.e. expectation under noisy distribution = $(1-2\eta)$ expectation under noiseless distribution
Algorithm can estimate expectation of 2nd term (under noisy distribution) within $\pm\frac{\varepsilon}{4n}(1-2\eta)$
Then dividing this estimate by $1 - 2\eta$
$\Longleftrightarrow$ estimating expectation of 2nd term (under noiseless distribution) within $\pm\frac{\varepsilon}{4n}$
Since $\mathbb{1}(x_i = 0) \cdot y/2$ takes either 0 or $\pm 1/2$ value
Estimate is accurate with prob $\geqslant 1 - \delta/2n$ using $O\left(\frac{n^2}{\varepsilon^2(1-2\eta)^2} \ln \frac{n}{\delta}\right)$ samples (Hoeffding)
$m \geqslant \Omega\left(\frac{n^2}{\varepsilon^2(1-2\eta)^2} \ln \frac{n}{\delta}\right)$ suffices using Hoeffding + union bound

---

3. Statistical Query (SQ) model

Above algorithm for monotone conjunctions with RCN uses only statistics, hence robust to noise
We now define a model to capture this type of learning algorithms
In this model, algorithm does not get labeled samples $(x, c(x))$
Can only query statistics of **predicates** $\varphi : X \times \{+1, -1\} \to \{0, 1\}$ and get estimates for them
Denote $P_\varphi = \mathbb{P}_{x\sim\mathcal{D}}[\varphi(x, c(x)) = 1] = \mathbb{E}_{\mathrm{EX}(c,\mathcal{D})}[\varphi(x, c(x))]$
Algorithm in **Statistical Query** model can query an oracle (i.e. black-box function) $\mathrm{STAT}(c, \mathcal{D})$
about a predicate φ with **tolerance** $0 < \tau \leqslant 1$
$\mathrm{STAT}(c, \mathcal{D})$ returns an estimate $\hat{P}_\varphi$ such that $P_\varphi - \tau \leqslant \hat{P}_\varphi \leqslant P_\varphi + \tau$
A normal PAC learning algorithm can simulate $\mathrm{STAT}(c, \mathcal{D})$ using m samples from $\mathrm{EX}(c, \mathcal{D})$
succeeds with prob. $\geqslant 1 - \delta$ when $m \geqslant \Omega\left(\frac{1}{\tau^2} \ln \frac{1}{\delta}\right)$ (Hoeffding)
Definition: Algorithm A learns $\mathcal{C}$ from SQ's if
for any target concept $c \in \mathcal{C}$, any accuracy $\varepsilon > 0$, any distribution $\mathcal{D}$ over X
given access to $\mathrm{STAT}(c, \mathcal{D})$
A outputs hypothesis h with $\mathrm{err}_\mathcal{D}(h, c) \leqslant \varepsilon$
Definition: Algorithm A **efficiently** learns $\mathcal{C}$ from SQ's if in addition
For every query (φ, τ) of A to $\mathrm{STAT}(c, \mathcal{D})$
$\varphi(x, c(x))$ can be evaluated in time $\mathrm{poly}(n, \mathrm{size}(c), 1/\varepsilon)$ (assuming $X = \{0, 1\}^n$ or $\mathbb{R}^n$)
$\tau \geqslant 1/\mathrm{poly}(n, \mathrm{size}(c), 1/\varepsilon)$
A runs in time $\mathrm{poly}(n, \mathrm{size}(c), 1/\varepsilon)$
Each call to $\mathrm{STAT}(c, \mathcal{D})$ takes 1 unit time

---

Algorithm to learn monotone conjunctions from SQ's

For $i = 1, \ldots, n$
$\varphi_i = \mathbb{1}(x_i = 0)\mathbb{1}(y = 1)$
Query $\mathrm{STAT}(c, \mathcal{D})$ with $(\varphi_i, \varepsilon/2n)$ and get $\hat{P}_{\varphi_i}$
Output $h(x)$ = conjunction of all x_i such that $\hat{P}_{\varphi_i} \leqslant \varepsilon/2n$

Above algorithm runs in time $O(n)$
Exercise: Show that above algorithm learns $\mathcal{C}$ = {monotone conjunctions} from SQ's