Lecture 13: SQ learning and PAC learning with noise

Recap: Statistical Computation Trade-off

i) Computational Complexity

- Concept class k-CNF and hence 3-CNF is efficiently PAC learnable.
- Under widely believed assumption RP≠NP, 3-DNF is not efficiently PAC learnable.

However, note that by the distributive law of boolean operations, every $\phi \in 3\text{-DNF}$ can be represented as some $\varphi \in$ 3-CNF.

$$\phi = T_1 \vee T_2 \vee T_3 = \bigwedge_{\ell_1 \in T_1, \ell_2 \in T_2, \ell_3 \in T_3} (\ell_1 \vee \ell_2 \vee \ell_3) = \varphi$$

Thus, we can learn to output a 3-CNF instead, which is computationally feasible. So, if we are allowed to output a CNF, then there is no problem in learning 3-DNF.

Recap: Statistical Computation Trade-off

ii) Statistical Complexity Both 3-CNF and 3-DNF have a finite VC dimension and are hence PAC learnable (inefficiently for 3-DNF).

- $\mathbf{3\text{-DNF}}$ Using upper bound on sample complexity for PAC learning, $\phi \in 3\text{-DNF}_d$ can be inefficiently PAC learned with statistical complexity $O(\frac{d}{\epsilon})$ calls to example oracle.

- **3-CNF** By lower bound on sample compexity for PAC learning, any $\varphi \in$ 3-CNF$_d$ can be efficiently PAC learned with $\Omega\left(\frac{|\text{3-CNF}_d|}{\epsilon}\right) = \Omega\left(\frac{|d^3|}{\epsilon}\right)$ calls to example oracle.

- While 3-DNF_d could not be learned efficiently **properly**, it can be learned **efficiently improperly** with more samples — d^3 vs d.

- Whether this statistical gap can be reduced while maintaining computational efficiency remains an open question.

PAC learning with noise

- So far, we have considered only noiseless setting due to the definition of $\mathrm{Ex}\,(c; \mathbb{P}_x)$.
- In the noiseless setting, an efficient consistent learner for a hypothesis class implies an efficient PAC learning algorithm.
- However, this is not true when the dataset is noisy,
  - In the rectangle learning algorithm, a negative point mislabelled as positive can lead to an arbitrarily large rectangle.
  - For Conjunctions, a negative example labelled as positive can lead to the elimination of a large number of good literals.
- For the noisy case, we need to think of a different framework. We will look at two of them today
  - PAC with Random Classification Noise (RCN)
  - Statistical Query Learning (SQ)

PAC with Random Classification Noise

- For any $c \in \mathcal{C}$, distribution $\mathbb{P}_x$ over $\mathcal{X}$, and noise parameter $\eta < \frac{1}{2}$, a Noisy Example Oracle: $\mathrm{Ex}_\eta(c; \mathbb{P}_x)$ samples $x \sim \mathbb{P}_x$ and returns $(x, c(x))$ with probability $1 - \eta$ and $(x, 1 - c(x))$ with probability η.

Definition (PAC learning with RCN)

A concept class $\mathcal{C}$ is PAC learnable with RCN using hypothesis class $\mathcal{H}$ if there exists a learning algorithm $\mathcal{A}$ such that for all $d > 0$, all distributions $\mathbb{P}_x$ over $\mathcal{X}_d$, concept $c \in \mathcal{C}_d$, and $0 < \epsilon, \delta, \eta < \frac{1}{2}$ if $\mathcal{A}$ is given access to $\mathrm{Ex}_\eta(c; \mathbb{P}_x)$ and knows $\epsilon, \delta, \mathrm{size}(c), d$, and η_0 where $\frac{1}{2} > \eta_0 \geq \eta$ $\mathcal{A}$ returns $h \in \mathcal{H}$ such that with probability at least $1 - \delta$, we have that $\mathbb{P}_x[h(x) \neq c(x)] \leq \epsilon$. Further, the number of calls made to $\mathrm{Ex}(c; \mathbb{P}_x)$ should be polynomial in $\mathrm{size}(c), d, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{1-2\eta_0}$.

Efficient PAC learnability: $\mathcal{A}$ should run in time polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}, \mathrm{size}(c), d, \frac{1}{1-2\eta_0}$.

Learning Conjunctions with noise

- For any literal ℓ in c, we have that $\mathbb{P}_x[\ell(x) = 0 \wedge c(x) = 1] = 0$. We need to put all such literals that have a significant probability mass of being false in the distribution.
- **Significant Literal** A literal ℓ is significant if $\mathbb{P}_x[\ell(x) = 0] \geq \frac{\epsilon}{8d}$
- **Harmful Literal** A literal ℓ is harmful if $\mathbb{P}_x[\ell(x) = 0 \wedge c(x) = 1] \geq \frac{\epsilon}{8d}$
- Let h be a hypothesis that is a conjunction of all significant literals that are not harmful.
- Let L denote the set of all $2d$ literals, S the set of significant literals, and T the set of harmful literals. Then, $T \subseteq S \subseteq L$.

$$\begin{aligned}
\mathbb{P}_x[h(x) \neq c(x)] &= \mathbb{P}_x[h(x) = 0 \wedge c(x) \wedge 1] + \mathbb{P}_x[h(x) = 1 \wedge c(x) = 0] \\
&\leq \sum_{\ell \in S \setminus T} \mathbb{P}_x[\ell(x) = 0 \wedge c(x) \wedge 1] + \sum_{\ell \in L \setminus S} \mathbb{P}_x[\ell(x) = 0] \\
&\leq |S \setminus T| \frac{\epsilon}{8d} + |L \setminus D| \frac{\epsilon}{8d} \leq \frac{\epsilon}{2}
\end{aligned}$$

Show that the probability estimates of whether a literal is significant and/or harmful can be obtained using concentration bounds and is polynomial in all required quantities.

Statistical Query Learning

- Note that the above algorithm relied on computing statistics. We will utilise this directly here.
- Instead of having access to an example oracle, here the learning algorithm can access a statistical query oracle $\mathrm{STAT}(c; \mathbb{P}_x)$
- A statistical query is a tuple (χ, τ), where $\chi : \mathcal{X} \times \{0,1\} \to \{0.1\}$ is a boolean function and τ is the tolerance parameter.
- The response of $\mathrm{STAT}(c; \mathbb{P}_x)$ to a query (χ, τ) is a value $\nu \in [0,1]$ s.t.

$$\left|\mathbb{E}_{\mathbb{P}_x}[\chi(x, c(x))] - \nu\right| \leq \tau$$

Learning Conjunctions using statistical query oracle.

Todo: Show that the *insignificant* and *harmful* literals above can be identified with the statistical query oracle.

Statistical Query Learnability

Let $\mathcal{C}$ be a concept class and $\mathcal{H}$ be a hypothesis class.

Definition (SQ learnability)

We say $\mathcal{C}$ **is efficiently learnable from statistical queries using** $\mathcal{H}$ if there exists a learning algorithm $\mathcal{A}$ and polynomials $p(\cdot,\cdot,\cdot), q(\cdot,\cdot,\cdot)$, and $r(\cdot,\cdot,\cdot)$ such that for all $d \geq 1$ for every target $c \in \mathcal{C}_d$, for every distribution $\mathbb{P}_x$ over $\mathcal{X}_d$, for any accuracy parameter $\epsilon > 0$, if $\mathcal{A}$ is given access to the statistical query oracle, $\mathrm{STAT}(c; \mathbb{P}_x)$, and inputs ϵ and size(c) satisfies the following, $\mathcal{A}$ halts in time bounded by $p(d, \text{size}(c), \frac{1}{\epsilon})$ and outputs $h \in \mathcal{H}$ such that $\mathbb{P}_x[h(x) \neq c(x)] \leq \epsilon$.

Further, for any query (χ, τ) made by $\mathcal{A}$ to $\mathrm{STAT}(c; \mathbb{P}_x)$, the predicate χ must be evaluable in time $q(d, \text{size}(c), \frac{1}{\epsilon})$ and $\frac{1}{\tau}$ is bounded by $r(d, \text{size}\, c, \frac{1}{\epsilon})$.

- Why is there no δ here ?
  - The $\mathrm{STAT}(c; \mathbb{P}_x)$ is required to output a value within the tolerance parameter τ with probability one. However, if the algorithm were randomised, then a δ parameter would be required.
  - Intuitively, This separates the randomisation in the sampling of the data and the randomisation in the algorithm.

SQ Learnability implies PAC learnability

Theorem

If $\mathcal{C}$ is efficiently SQ-learnable using $\mathcal{H}$ then $\mathcal{C}$ is efficiently PAC learnable using $\mathcal{H}$

Proof Strategy

- Let $\mathcal{A}$ be the algorithm that learns $\mathcal{C}$ using $\mathcal{H}$ in the SQ model using k queries to $\mathrm{STAT}\left(c; \mathbb{P}_x\right)$
- Simulate $\mathcal{A}$ in the PAC model, by replacing each call to SQ oracle with an empirical estimate of the SQ χ using $m = \Theta\left(\frac{1}{\tau^2} \log \frac{k}{\delta}\right)$ samples $(x_1, c(x_1), \ldots, (x_m, c(x_m)))$ drawn from $\mathrm{Ex}\left(c; \mathbb{P}_x\right)$.
- Using Hoeffding's bound, $|\frac{1}{m} \sum_{i=1}^{m} \chi(x_i, c(x_i)) - \mathbb{E}_{\mathbb{P}_x}[\chi(x, c(x))]| \leq \tau$ holds w.p. $1 - \delta$.
- Applying a simple union bound over the k queries yields the desired result.

SQ Learnability implies PAC learnability with RCN

Theorem

If $\mathcal{C}$ is efficiently SQ-learnable using $\mathcal{H}$ then $\mathcal{C}$ is efficiently PAC learnable with Random Classification Noise (RCN) using $\mathcal{H}$

Proof sketch

We need to show that we can *simulate the statistical query* $\mathrm{STAT}\left(c; \mathbb{P}_x\right)$ for any query χ using polynomial calls to $\mathrm{Ex}_\eta\left(c; \mathbb{P}_x\right)$.

Extra notations for proof For simplicity, we will require the following—

- Assume the boolean functions are in $\{-1, 1\}$ instead of $\{0, 1\}$.
- To go from boolean to $\{-1, 1\}$, map 0 to 1 and 1 to -1.
- Assume the queries are of the form $\chi : \mathcal{X} \times \{-1, 1\} \to \{-1, 1\}$ so that $\mathbb{P}_x[\chi(x, c(x)) = -1] = \frac{1}{2} - \frac{1}{2}\mathbb{E}_{\mathbb{P}_x}[\chi(x, c(x))]$

Proof of SQ Learnability implies PAC learnability with RCN

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_x}[\chi(x, c(x))] =& \mathbb{E}_{\mathbb{P}_x}[\chi(x,1).\mathbb{1}(c(x)=1)] + \mathbb{E}_{\mathbb{P}_x}[\chi(x,-1).\mathbb{1}(c(x)=-1)] \\ =& \mathbb{E}_{\mathbb{P}_x}\left[\chi(x,1).\left(\frac{1+c(x)}{2}\right)] + \mathbb{E}_{\mathbb{P}_x}[\chi(x,-1)\left(\frac{1-c(x)}{2}\right)\right] \\ =& \frac{1}{2}\left(\mathbb{E}_{\mathbb{P}_x}[\chi(x,1)] + \mathbb{E}_{\mathbb{P}_x}[\chi(x,-1)]\right) \\ &+ \frac{1}{2}\left(\mathbb{E}_{\mathbb{P}_x}[\chi(x,1)c(x)] + \mathbb{E}_{\mathbb{P}_x}[\chi(x,-1)c(x)]\right)\end{aligned}$$

Note that there are two kinds of queries here

- $\mathbb{E}_{\mathbb{P}_x}[\chi(x,1)], \mathbb{E}_{\mathbb{P}_x}[\chi(x,-1)]$: Target independent queries. For any χ, using Hoeffding's bound, can be easily simulated using $\mathrm{Ex}_\eta(c;\mathbb{P}_x)$ for any η as the query is independent of target.
- $\mathbb{E}_{\mathbb{P}_x}[\chi(x,1)c(x)], \mathbb{E}_{\mathbb{P}_x}[\chi(x,-1)c(x)]$: Correlational queries. This computes the correlation between a function of x and the target. We will look into this in detail.

Proof of SQ Learnability implies PAC learnability with RCN

A correlation query has the form (φ,τ) where $\varphi : \mathcal{X} \to \{-1,1\}$,$\tau \in \{0,1\}$. The response of $\mathrm{STAT}(c;\mathbb{P}_x)$ is ν_φ such that $\left|\mathbb{E}_{\mathbb{P}_x}[\varphi(x)c(x)] - \nu_\varphi\right| \leq \tau$

Simulating responses to correlational queries

- Let $\sigma \sim B(\eta)$ be a r.v. that is 1 w.p. $1-\eta$ and -1 w.p. η. $\mathbb{E}[\sigma] = 1-2\eta$.
- Let $(x, c(x))$ be a random example from $\mathrm{Ex}(c;\mathbb{P}_x)$; then $(x, c(x)\sigma)$ is a random example from $\mathrm{Ex}_\eta(c;\mathbb{P}_x)$.

$$\mathbb{E}_{\mathrm{Ex}_\eta(c;\mathbb{P}_x)}[\varphi(x)y] = \mathbb{E}_{\mathbb{P}_x}\left[\mathbb{E}_\sigma[\varphi(x)c(x)\sigma]\right] = (1-2\eta)\mathbb{E}_{\mathbb{P}_x}[\varphi(x)c(x)]$$

- Draw m examples from $\mathrm{Ex}_\eta(c;\mathbb{P}_x)$, $((x_1,y_1)\ldots(x_m,y_m))$ and define $\hat{\nu} = \frac{1}{m}\sum_{i=1}^m \varphi(x_i)y_i$. Choose m s.t. $\left|\hat{\nu} - \mathbb{E}_{\mathrm{Ex}_\eta(c;\mathbb{P}_x)}[\varphi(x)y]\right| \leq \tau_1(1-2\eta)$ with prob. $1-\delta$, where we choose τ_1 later.

Proof of SQ Learnability implies PAC learnability with RCN

- Assume, we do not know the true η but some $\hat{\eta} \leq \eta_0$ (η_0 is an upper bound) such that $|\hat{\eta} - \eta| \leq \Delta$. Then

$$\begin{aligned}\left|\frac{\hat{\nu}}{1-2\eta} - \mathbb{E}_{\mathbb{P}_x}[\varphi(x)c(x)]\right| &\leq \left|\frac{\hat{\nu}}{1-2\eta} - \frac{\hat{\nu}}{1-2\eta} + \frac{\hat{\nu}}{1-2\eta} - \mathbb{E}_{\mathrm{Ex}_\eta(c;\mathbb{P}_x)}[\varphi(x)y]\right| \\ &\leq |\hat{\nu}|\frac{2\Delta}{(1-2\eta_0)^2} + \frac{1}{1-2\eta_0}\left|\hat{\nu} - \mathbb{E}_{\mathrm{Ex}_\eta(c;\mathbb{P}_x)}[\varphi(x)y]\right| \\ &\leq \frac{2\Delta}{(1-2\eta_0)^2} + \frac{\tau_1}{1-2\eta_0}\end{aligned}$$

- Make both term less than $\frac{\tau}{2}$. Set $m = O\left(\log(\frac{1}{\delta})\frac{1}{\sqrt{\tau(1-2\eta_0)}}\right)$ for the second term.
- For the first term, choose $\Delta \leq \frac{\tau}{2(1-2\eta_0)^2}$ and run the algorithm for all values of $\hat{\eta} = i\Delta$ for $i = 1, \ldots, \lfloor\frac{\eta_0}{\Delta}\rfloor$, let the corresponding output hypothesis be $h_1, \ldots, h_{\lfloor\frac{\eta_0}{\Delta}\rfloor}$.
- Finally, we can show that by testing each of the h_i on an independent sample of $\mathrm{Ex}_\eta(c; \mathbb{P}_x)$ and outputting the best one, solves our problem.

Conclusion

- We have seen that **SQ learnability implies PAC learnability**.
- We have also seen a much stronger result that **SQ learnability implies PAC learnability with RCN**.
- But does PAC learnability also imply SQ learnability ? No, PARITIES
- Does PAC learnability with noise imply SQ learnability ? No, Blum et. al. (2003)
- Thus **SQ learnability is a strictly weaker condition that both PAC and PAC with RCN**.
- People have used this implication to provide algorithms for learning with noise by providing an SQ learner and then simulating it with $\mathrm{Ex}_\eta(c; \mathbb{P}_x)$.
- For full proofs of everything, we have seen today refer to Chapter 5 in KV.