CS6781 - Theoretical Foundations of Machine Learning

Lecture 24: Statistical Queries and Random Classification Noise

May 5th, 2020

Lecturer: Nika Haghtalab *Readings: N/A*
Scribe: Aaron Tucker

Last time

- Defined Random Classification Noise (RCN)
- Defined statistical query model (SQ)

STAT$(h^*, \mathcal{D})$ gets queries of the form (ψ, τ), where $\psi : \mathcal{X} \times \{0,1\} \to \{0,1\}$ and $0 < \tau < 1$. The answer to a query is v such that

$$\left| v - \mathbf{E}\left[\psi\left(x, h^*(x)\right)\right] \right| \leq \tau.$$

Today

Formally see the connection between the RCN and STAT model, and how learnability in these two settings are related.

1 Warmup: Simulating statistical queries

Theorem 1.1. *If $\mathcal{H}$ is efficiently learnable in the statistical query model, it is also efficiently learnable in the realizable PAC model. In particular, if we use M queries each with a tolerance of at least $\tau > 0$ to learn a hypothesis of $err_{\mathcal{D}}(h) \leq \epsilon$, then m samples are sufficent to get $err_{\mathcal{D}}(h) \leq \epsilon$ with probability $1 - \delta$ for*

$$m \in O\left(\frac{1}{\tau^2}\log\left(\frac{M}{\delta}\right)\right).$$

Proof. Consider using m samples $\left\{(x_i, y_i)\right\}_{i=1}^m$ to learn a single query (ψ, τ) using the empirical estimate

$$v = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left(\psi(x_i, y_i) = 1\right).$$

Note that $\frac{1}{m}\sum \mathbf{1}\left(\psi(x_i, y_i) = 1\right)$ is the empirical estimate of $\Pr\left(\psi(x, h^*(x)) = 1\right)$, which is the same as $\mathbf{E}\left(\psi(x, h^*(x))\right)$ because ψ has range $\{0,1\}$. We can use the Hoeffding bound to show that

$$\left| \frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{1}\left(\psi(x_i, y_i) = 1\right)\right) - \mathbf{E}\left[\psi\left(x, h^*(x)\right)\right] \right| \leq \exp\left(\frac{-m\tau^2}{2}\right) \leq \delta.$$

If we have M queries, then we need to use a union bound to ensure that all of them are true with probability δ, which means we need to solve

$$\exp\left(\frac{-m\tau^2}{2}\right) \leq \delta/M.$$

Solving, we see that if we can answer all M queries with τ accuracy with probability $1-\delta$ as long as

$$m \geq \frac{1}{\tau^2}\log\left(\frac{M}{\delta}\right),$$

as desired. □

2 SQ-Learnability and RCN-Learnability

STAT-learnability implies learnability in the PAC with random classification noise model.

Theorem 2.1. *If $\mathcal{H}$ is efficiently learnable in the statistical query model, it is also efficiently learnable in the PAC model with random classification noise. In particular, if we use M queries each with a tolerance of at least $\tau > 0$ to learn a hypothesis of $err_{\mathcal{D}}(h) \leq \epsilon$, then m samples are sufficent to get $err_{\mathcal{D}}(h) \leq \epsilon$ with probability $1-\delta$ for*

$$m \in O\left(\frac{1}{\tau^2(1-2\eta)^2}\ln\left(\frac{M}{\delta}\right)\right).$$

Remark 2.2. For ease of this next proof, instead of $\psi : \mathcal{X} \times \{0,1\} \to \{0,1\}$ we use

$$\psi : \mathcal{X} \times \{-1,1\} \to \{-1,1\}.$$

This doesn't change the expressivity of the model, but makes the math easier for us.

2.1 Simplifying $\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x, h^*(x)\right)\right]$

For any query ψ, we want to separate it into a part that cares only about the marginal distributions (but not the labels), and a part that also cares about the labels. This is similar to our previous proof for monotone ands – there were parts that cared about the relative frequency of 0 and 1 but not about the label or the noise, and other parts that cared about the labels given that certain features were set to 1.

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x,h^*(x)\right)\right] &= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x,1\right)\cdot\mathbf{1}\left(h^*(x)=1\right)\right] &&+ \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x,-1\right)\cdot\mathbf{1}\left(h^*(x)=-1\right)\right] \\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x,1\right)\cdot\frac{1+h^*(x)}{2}\right] &&+ \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x,-1\right)\cdot\frac{1-h^*(x)}{2}\right] \\
&= \frac{1}{2}\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x,1\right)+\psi\left(x,-1\right)\right] &&+ \frac{1}{2}\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\psi\left(x,1\right)h^*(x)-\psi\left(x,-1\right)h^*(x)\right] \\
&\quad\text{target independent} &&\quad\text{correlation of } h^*(x) \text{ with some other function}
\end{aligned}$$

This works because

$$\mathbf{1}\left(h^*(x)=1\right)=\frac{1+h^*(x)}{2} \text{ and } \mathbf{1}\left(h^*(x)=-1\right)=\frac{1-h^*(x)}{2}$$

This substitution lets us split any query ψ into a target independent term that cares only about the marginal distribution of x_i, and a correlational term that cares about the correlation between $h^*(x)$ and some other function.

2.2 Target Independent and Correlational Queries

We can use these two types of queries without loss of generality.

Definition 2.3 (Correlational queries)**.** A query of the form $\varphi : \mathcal{X} \to \{-1, 1\}$ and tolerance $\tau \in (0, 1)$ which returns a v as follows is called a correlational query.

$$\left|\mathbf{E}\left[\varphi(x)\cdot h^*(x)\right]-v\right|\leq\tau$$

This gives us information about h^*.

Definition 2.4 (Target independent queries)**.** A query of the form $\phi : \mathcal{X} \to \{-1, 1\}$ and tolerance $\tau \in (0, 1)$ which returns a v as follows is called a target independent query.

$$\left|\mathbf{E}\left[\phi(x)\right]-v\right|\leq\tau$$

This is sufficient to answer questions about the marginal distribution.

2.3 Simulating Queries in the RCN Model

Simple case: Target independent queries

Target independent $\phi : \mathcal{X} \to \{-1, 1\}$. The noise has no effect, so given $\{(x_1, y_1), ..., (x_m, y_m)\}$, just throw out the y_i and compute

$$v=\frac{1}{m}\sum_{i=1}^{m}\phi(x_i).$$

This is just the same as the warm up exercise, since throwing out y_i throws out the source of noise, and it is the same as in the case for deterministic labels.

$$\left|v-\mathbf{E}\left[\phi(x)\right]\right|\leq\tau.$$

Given $\frac{1}{\tau^2}\log\left(\frac{M}{\delta}\right)$ data points, we succeed with probability $1-\delta$.

Correlational queries

Recall that $\mathcal{D}(\eta)$ is the noisy version of $\mathcal{D}$, with random classification noise η. Answering the query φ means approximating $\mathbf{E}_{x\sim\mathcal{D}}\left[\varphi(x)\cdot h^*(x)\right]$. However we can only have direct access to $\mathbf{E}_{(x,y)\sim\mathcal{D}(\eta)}\left[\varphi(x)\cdot y\right]$. How can we relate these?

Claim 2.5. $\mathbf{E}_{x\sim\mathcal{D}}\left[\varphi(x)\cdot h^*(x)\right]$ *and* $\mathbf{E}_{(x,y)\sim\mathcal{D}(\eta)}\left[\varphi(x)\cdot y\right]$ *are related by:*

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}(\eta)}\left[\varphi(x)\cdot y\right] = (1-2\eta)\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\varphi(x)\cdot h^*(x)\right].$$

Proof. Because φ has range $\{-1,1\}$, we can say that

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}(\eta)}\left[\varphi(x)\cdot y\right] = \Pr_{\mathcal{D}(\eta)}\left[\varphi(x)=y\right] - \Pr_{\mathcal{D}(\eta)}\left[\varphi(x)\neq y\right]$$

The first term is the probability that $\varphi(x)$ agrees with y. That can either happen because $\varphi(x)$ agrees with $h^*(x)$ and y wasn't flipped, or because $\varphi(x)$ disagrees with $h^*(x)$, but y was flipped.

$$\begin{aligned}\Pr_{\mathcal{D}(\eta)}\left[\varphi(x)=y\right] &= \Pr\left(\varphi(x)=h^*(x)\right)\cdot\Pr\left(h^*(x)=y\right) + \Pr\left(\varphi(x)\neq h^*(x)\right)\cdot\Pr\left(h^*(x)\neq y\right)\\ &= \Pr\left(\varphi(x)=h^*(x)\right)(1-\eta) + \left(1-\Pr\left(\varphi(x)=h^*(x)\right)\right)\eta\\ &= (1-2\eta)\Pr\left(\varphi(x)=h^*(x)\right)+\eta.\end{aligned}$$

The second term is the probability that $\varphi(x)$ disagrees with y. That can either happen because $\varphi(x)$ disagrees with $h^*(x)$ and y wasn't flipped, or because $\varphi(x)$ agrees with $h^*(x)$ but y was flipped.

$$\begin{aligned}\Pr_{\mathcal{D}(\eta)}\left[\varphi(x)\neq y\right] &= \Pr\left(\varphi(x)\neq h^*(x)\right)\cdot\Pr\left(h^*(x)=y\right) + \Pr\left(\varphi(x)=h^*(x)\right)\cdot\Pr\left(h^*(x)\neq y\right)\\ &= \Pr\left(\varphi(x)\neq h^*(x)\right)(1-\eta) + \left(1-\Pr\left(\varphi(x)\neq h^*(x)\right)\right)\eta\\ &= (1-2\eta)\Pr\left(\varphi(x)=h^*(x)\right)+\eta.\end{aligned}$$

Notice that both terms are now expressed in terms of the probability of how $\varphi(x)$ relates to the true $h^*(x)$, rather than to the noisy y.

$$\begin{aligned}\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}(\eta)}\left[\varphi(x)\cdot y\right] &= \Pr_{\mathcal{D}(\eta)}\left[\varphi(x)=y\right] - \Pr_{\mathcal{D}(\eta)}\left[\varphi x\neq y\right]\\ &= (1-2\eta)\Pr\left(\varphi(x)=h^*(x)\right)+\not{\eta} - \left((1-2\eta)\Pr\left(\varphi(x)=h^*(x)\right)+\not{\eta}\right)\\ &= (1-2\eta)\left(\Pr\left(\varphi(x)=h^*(x)\right)-\Pr\left(\varphi(x)=h^*(x)\right)\right)\\ &= (1-2\eta)\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\varphi(x)\cdot h^*(x)\right]\end{aligned}$$

as desired. □

Number of samples needed to estimate v

If we want to estimate a correlational query $v = \mathop{\mathbf{E}}_{\mathcal{D}}\left[\varphi(x) \cdot h^*(x)\right]$ to accuracy τ, we should estimate $v' = \mathop{\mathbf{E}}_{\mathcal{D}(\eta)}\left[\varphi(x) \cdot h^*(x)\right]$ to accuracy τ', where τ' is a parameter to be set later. Consider m samples $\left\{(x_i, y_i)\right\}_{i=1}^m$ from the RCN oracle $\mathbb{E}X^\eta(h^*, \mathcal{D})$, and compute v' as

$$v' = \frac{1}{m}\sum_{i=1}^{m} \varphi(x_i) \cdot y_i.$$

setting m such that

$$\left| v' - \mathop{\mathbf{E}}_{\mathcal{D}(\eta)}\left[\varphi(x) \cdot y\right] \right| \leq \tau'.$$

Setting $v = \frac{v'}{1-2\eta}$ as our estimate for $\mathop{\mathbf{E}}_{\mathcal{D}}\left[\varphi(x) \cdot h^*(x)\right]$, we get

$$\begin{aligned}
\left| v - \mathop{\mathbf{E}}_{\mathcal{D}}\left[\varphi(x) \cdot h^*(x)\right] \right| &= \left| \frac{v'}{1-2\eta} - \mathop{\mathbf{E}}_{\mathcal{D}}\left[\varphi(x) \cdot h^*(x)\right] \right| && (1)\\
&= \left| \frac{v'}{1-2\eta} - \frac{1}{1-2\eta}\mathop{\mathbf{E}}_{\mathcal{D}(\eta)}\left[\varphi(x) \cdot y\right] + \frac{1}{1-2\eta}\mathop{\mathbf{E}}_{\mathcal{D}(\eta)}\left[\varphi(x) \cdot y\right] - \mathop{\mathbf{E}}_{\mathcal{D}}\left[\varphi(x) \cdot h^*(x)\right] \right| && (2)\\
&= \frac{1}{1-2\eta}\left| v' - \mathop{\mathbf{E}}_{\mathcal{D}(\eta)}\left[\varphi(x) \cdot y\right] \right| && (3)\\
&\leq \frac{1}{1-2\eta}\tau' && (4)\\
& && (5)
\end{aligned}$$

At step 2 we add and subtract $\frac{1}{1-2\eta}\mathop{\mathbf{E}}_{\mathcal{D}(\eta)}\left[\varphi(x) \cdot y\right]$. We can cancel out the last two terms of line (2) because of Claim 2.5. Setting $\tau' = (1-2\eta)\,\tau$, we get

$$\left| v - \mathop{\mathbf{E}}_{\mathcal{D}}\left[\varphi(x) \cdot h^*(x)\right] \right| \leq \tau$$

We know from Theorem 1.1 that the sample complexity of learning v' is $\frac{1}{\tau'^2}\log\left(\frac{M}{\delta}\right)$, so if we substitute in our expression for τ' we get a sample complexity of

$$\frac{1}{\tau^2(1-\eta)^2}\log\left(\frac{M}{\delta}\right).$$

Conclusion

This shows that anything you can learn efficiently in the statistical query model, you can learn in the PAC+RCN model, and if you needed M queries of tolerance τ, then all you need is $m \in O\left(\frac{1}{\tau^2(1-\eta)^2}\log\left(\frac{M}{\delta}\right)\right)$ data points to learn in the PAC+RCN model.

3 Discussion

What if η is unknown?

We can guess and try it. For example, if we have an effective upper bound $\eta' \geq \eta$, and $\eta' - \eta = \Delta$ for a small Δ, then as long as Δ is small compared to τ it won't really affect our accuracy. We just need a few more samples.

So, we can try our algorithm with $\eta' \in \{0, \Delta, 2\Delta, ..., 1/2\}$. We know that one of these η' is close enough to be okay, and we can look at the error that we observe.

Why are correlational queries interesting?

- They give us insight into how much of the difficulty of the learning ask is coming from its labels, as opposed to coming from its marginal distribution. If I knew $\mathcal{D}$ the marginal, the only type of queries that you need to ask are correlational. This gives rise to a statistical query dimension.

- SQ-Dim$_{\mathcal{D}}(\mathcal{H}) = d$ if I have functions $f_1, f_2, ..., f_d \in \mathcal{H}$, such that for all $i \neq j$ we have $\mathop{\mathbf{E}}_{\mathcal{D}}\left[f_i(x) \cdot f_j(x)\right] \leq \frac{1}{d} \approx 0$. This lets us say that f_i are effectively an orthogonal basis vectors for $\mathcal{H}$. If you have SQ-Dim$_{\mathcal{D}}(\mathcal{H}) = d$, then you need $\Omega(\tau^2 d)$ queries of τ accuracy. Since the f_i are not correlated with each other, any query φ will be measuring one of them, but not the rest of them. So there has to be many of these queries in order to learn. You will formalize this in HW5.