1 Lecture 5: 2017.02.20

Outline for today:

1. VC theory rcap
2. Matching lower bound
3. VC++:
   - unrealizable/agnostic case
   - ERM/SRM
   - Rademacher complexity, sparsity, etc.
   - general loss functions
4. Learning with noise

1.1 VC theory review

Recall from last time that we defined $\Pi_{\mathcal{C}}(m)$ as the max. # of labelings induced by $\mathcal{C}$ on m points, by definition we have the naive bound $\Pi_{\mathcal{C}}(m) \leq 2^m$. One of the highlights is that we were able to prove that

$$\Pi_{\mathcal{C}}(m) \leq \phi_d(m) \leq O(m^d). \tag{1.1}$$

This bound is saying that past the VCD dimension, the number of possible labelings we can achieve is only a increasingly small fraction of the total number (2^m). This was the key step to prove the following theorem

Theorem 1.1. *Given $c \in \mathcal{C}$ and a distribution, D, over X. If we take $m \geq c_0 \frac{d}{\varepsilon} \ln \frac{1}{\delta}$ examples, where $VCD(\mathcal{C}) = d$, then, with probability $\geq 1 - \delta$, any $h \in \mathcal{C}$ that is consistent with the sample has $Err(h) \leq \varepsilon$.*

Remark 1.2. *This is a purely "information theory" statement, we are ignoring how hard is to actually find an appropriate h.*

1.2 Matching lower bound

The converse of the theorem above is not true, in other words, we can find distributions such that is not necessary to take that many examples. For example, consider the case of the triangles in dimension 2 and a distribution that has support on a single point. This motivates the next result.

Theorem 1.3. *Given $\mathcal{C}$ with $VCD(\mathcal{C}) = d$, there exists D a distribution over X such that $\Omega(d/\varepsilon)$ examples are required to learn with error $\leq \varepsilon$.*

Proof. Consider $\{x_1, \dots, x_d\}$ a shattered set and let D be the uniform distribution over this set. Now, since the set is shattered, we can choose a function $c_i \in \mathcal{C}$ with $i = 1, \dots, 2^d$ for each of the possible labeling of these d points. Now we randomly choose the target concept, c, among these c_i. This is equivalent to flipping a fair coin d times to determine the labeling induced by c on S.

Now we let L be any PAC learning algorithm with the data above. Given error parameter $\varepsilon \leq \frac{1}{8}$ suppose we only draw $m < d$ examples. Say that the number of distinct points that we have seen is $m' \leq m$, for the remaining of the points the problem is equivalent to predicting a fair coin toss. Therefore, the expected error on the whole sample is $\frac{\frac{d-m'}{2}}{d} = \frac{d-m'}{2d}$. By Chebyshev's inequality,

$$\mathbb{P}\left(Error \geq \frac{d-m'}{4d}\right) \leq \frac{1}{2}. \tag{1.2}$$

In particular, if $m = \frac{d}{2}$, the error is at least $\frac{1}{8}$ with probability at least $\frac{1}{2}$ and the algorithm fails the conclusion of the theorem when c is chosen randomly. This implies that there exist some target c on which L fails, so we need at least $\Omega(d)$ sample complexity lower bound.

Finally, to incorporate the ε onto the lower bound, we need only scale the distribution above. Modify D so that the point x_1 has probability $1-8\varepsilon$ and let the rest have probability $\frac{8\varepsilon}{d-1}$. This results in needing to draw more points to be in the same set up as before. Details are left as exercise or can be checked on the textbook in page 63. □

1.3 VC++

1.3.1 Unrealizable/Agnostic setting

So far we have assumed that the the target $c \in \mathcal{C}$. That is, that the truth is perfectly represented by our chosen class. Now we drop this assumption. We have a *hypothesis* class $\mathcal{H}$ but $c \notin \mathcal{H}$, i.e. the target might not be representable by $\mathcal{H}$. The first thing we need to lower our expectations for learning.

Definition 1.4. $h^* := \underset{h \in \mathcal{H}}{argmin} \{Err(h)\}$, where $Err(h) = \mathbb{P}_{X \sim D}[h(x) \neq c(x)]$. We also define $\varepsilon^* = Err(h^8) =$ best possible error in $\mathcal{H}$.

Let's set up some additional notation. Given $h \in \mathcal{H}$ and a labeled sample

$$S = \{\langle x_1, c(x_1)\rangle, \dots, \langle x_m, c(x_m)\rangle\}, \tag{1.3}$$

we have some different types of errors:

- $Err(h)$, the *true error* as aboe
- $\widehat{Err}(h) = \frac{1}{m}\left|\{\langle x_i, y_i\rangle \in S \ : \ h(x_i) \neq y_i\}\right|$, the *empirical error.*

Theorem 1.5. *Given any target c (not necessarily in $\mathcal{H}$) and a distribution D over X, if we take $m \geq c_0 \frac{d}{\varepsilon^2} \ln \frac{1}{\delta}$ where $VCD(\mathcal{H}) = d$, then with probability $\geq 1-\delta$, for all $h \in \mathcal{H}$ we have* uniform *convergence*

$$\left|\widehat{Err}(h) - Err(h)\right| \leq \varepsilon. \tag{1.4}$$

Proof. According to Prof. Kearns is a modification of the analogous result where $c \in \mathcal{H}$. He mentioned that some extra materials might be uploaded to the course website. □

Remark 1.6. *To apply this theorem, in applications, we find/compute* $\hat{h}^* = \underset{h\in\mathcal{H}}{argmin}\left\{\widehat{Err}(h)\right\}$. *By the theorem, we have* $\left|\widehat{Err}(\hat{h}^*) - Err(\hat{h}^*)\right| \leq \varepsilon$ *and* $\left|\widehat{Err}(h^*) - Err(h^*)\right| \leq \varepsilon$. *This implies that,*

$$\left|Err(\hat{h}^*) - Err(h^*)\right| \leq 3\varepsilon. \tag{1.5}$$

1.3.2 Structural Risk Minimization

Fix some target c, a distribution D and a sample of size m S.Suppose we have $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_d \subset \ldots$. For example, each hypothesis could be a neural network with an increasing number of hidden layers. Note that the VCD dimension can only increase along the nested sequence. For simplicity, let's assume that $VCD(\mathcal{H}_d) = d$. Define

$$\begin{aligned} Err(d) &\coloneqq \min_{h\in\mathcal{H}_d}\{Err(h)\} && (1.6)\\ &= \text{ best true error in } \mathcal{H}_d. && (1.7)\\ \widehat{Err}(d) &\coloneqq \min_{h\in\mathcal{H}_d}\{\widehat{Err}(h)\} && (1.8)\\ &= \text{ best empirica error on } S \text{ in } \mathcal{H}_d. && (1.9) \end{aligned}$$

Remark 1.7. *From the theorem above,* $|Err(d) - \widehat{Err}(d)| \lesssim \sqrt{\frac{d}{m}}$

If we use minimization of the empirical error as the criterion to choose our model, we might run into trouble in the form of overfitting. Basically, the empirical error can only decrease as the complexity of the model increases but it can deviate from the true error. Luckily for us, the theorem above gives a way to correct for this by instead finding

$$\min_d \left\{\widehat{Err}(d) + \sqrt{\frac{d}{m}}\right\}. \tag{1.10}$$

1.3.3 Refinements

In the theorem from last time, $\Pi_{\mathcal{C}}(S) \to \Pi_{\mathcal{C}}(m) \leq O(m^d)$. But if one is more careful in the proof, it is possible to arrive to a bound on $\mathbb{E}_{S\sim D,c}[\ln\left|\Pi_{\mathcal{C}}(S)\right|]$ where $|S| = m$, the *VC entropy*. (cf. Book)

1.4 General loss functions

The assumptions are as follows:

- observations $z \sim D$ iid. E.g. $z = \langle x, y\rangle$ with $y \in \{0, 1\}$; $z = \langle x, y\rangle$ with $y \in \mathbb{R}$; $z = x$.
- models $h \in \mathcal{H}$.

- $\mathbb{R}$-valued loss function, $L(h, z)$. E.g.

  1. Classification: $L(h, \langle x, y\rangle) = 1$ if $h(x) \neq y$, or 0 otherwise.
  2. Linear regression: $L(h, \langle x, y\rangle) = (h(x) - y)^2$
  3. $L(h, x) = \ln \frac{1}{h(x)}$

  We'd like to minimize $\mathbb{E}_{z\sim D}[L(h, z)]$

Remark 1.8. *Although we are not going to touch on that, there is a theorem like the one we proved for this setting as well.*

1.4.1 Analog to VCD

Take d observations $z_1, \ldots, z_d$ and consider the following set $T = \{\langle L(h, z_1), \ldots, L(h, z_d)\rangle : h \in \mathcal{H}\}$. The analog to the VC dimension would be to look at this cloud of points a see if it is "space-filling" or that not all points lie on some lower dimensional subspace. The criterion is that you want all the d-dimensional orthants to be intersected by T.