CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Annealed Entropy and Growth Function

Lecturer: Michael I. Jordan *Scribe: Gregory Lawrence*

Theorem

$$\Pr\left(\sup_{\alpha\in\Lambda}\left|R(\alpha)-R_{emp}(\alpha_l)\right|>\epsilon\right)\le 4\exp\left\{\left(\frac{H^{\Lambda}_{\text{ann}}(2l)}{l}-(\epsilon-\frac{1}{l})^2\right)l\right\}$$

There is a trade off between ϵ and our sample size l. If we assume that the $\frac{H^{\Lambda}_{\text{ann}}(2l)}{l}$ term goes to zero for a large enough training set, keeping $\epsilon\sim\frac{1}{\sqrt{l}}$ will keep η constant where $\eta=4\exp\left\{\left(\frac{H^{\Lambda}_{\text{ann}}(2l)}{l}-(\epsilon-\frac{1}{l})^2\right)l\right\}$

Proof Sketch

$$\epsilon_* \quad \triangleq \quad \epsilon-\frac{1}{l}$$

Split $(z_1,\dots,z_l,z_{l+1},\dots,z_{2l})$ into half samples

Using Chernoff's bounds (for a fixed α) we get:

$$\Pr\left\{\left|\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)-\frac{1}{l}\sum_{i=l+1}^{2l}Q(z_i,\alpha)\right|>\epsilon_*\right\}\le 2e^{-\epsilon_*^2 l}$$

For any fixed sample z^{2l}, the number of distinguishable functions is finite.

Denote representatives of these equivalent classes by $\alpha^*\in\Lambda^*$.

This set has cardinality $N^{\Lambda^*}(z_1,\dots,z_{2l})\triangleq N^{\Lambda}(z_1,\dots,z_{2l})$

$$\begin{aligned}
&\Pr\left\{\sup_{\alpha\in\Lambda}\left|\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)-\frac{1}{l}\sum_{i=l+1}^{2l}Q(z_i,\alpha)\right|>\epsilon_*\middle|z_1,\dots,z_{2l}\right\}\\
&=\ \Pr\left\{\sup_{\alpha^*\in\Lambda^*}\left|\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha^*)-\frac{1}{l}\sum_{i=l+1}^{2l}Q(z_i,\alpha^*)\right|>\epsilon_*\middle|z_1,\dots,z_{2l}\right\}\\
&\le\ \sum_{\alpha^*\in\Lambda^*}\Pr\left\{\left|\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha^*)-\frac{1}{l}\sum_{i=l+1}^{2l}Q(z_i,\alpha^*)\right|>\epsilon_*\middle|z_1,\dots,z_{2l}\right\}\qquad\text{Union bound}\\
&\le\ 2N^{\Lambda^*}(z_1,\dots,z_{2l})\exp(-\epsilon_*^2 l)
\end{aligned}$$

Now average wrt $dF(z_1,\dots,z_l)$

$$\Pr\left\{\sup_{\alpha\in\Lambda}\left|\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)-\frac{1}{l}\sum_{i=l+1}^{2l}Q(z_i,\alpha)\right|>\epsilon_*\right\}\quad\le\quad 2\mathrm{E}N^{\Lambda}(z_1,\dots,z_{2l})\exp(-\epsilon_*^2 l)$$

$$
\begin{aligned}
&= 2\exp\left(\log \mathrm{E} N^{\Lambda}(z_1,\ldots,z_{2l}) - \epsilon_*^2 l\right) \\
&= 2\exp\left(\left(\frac{H^{\Lambda}_{\mathrm{ann}}(2l)}{l} - \epsilon_*^2\right) l\right)
\end{aligned}
$$

This puts a bound on the difference between two empirical distributions. We can use the **Basic lemma** introduced last time to incorporate the true risk function, finishing our proof.

$$
\Pr\left(\left|R(\alpha) - R_{\mathrm{emp}}(\alpha)\right| > \epsilon\right) \quad \leq \quad 2\Pr\left(\left|R_{\mathrm{emp},1} - R_{\mathrm{emp},2}\right| > \epsilon_*\right) \qquad \text{Basic Lemma}
$$

Therefore

$$
\Pr\left(\sup_{\alpha\in\Lambda}\left|R(\alpha) - R_{emp}(\alpha_l)\right| > \epsilon\right) \leq 4\exp\left\{\left(\frac{H^{\Lambda}_{\mathrm{ann}}(2l)}{l} - (\epsilon - \frac{1}{l})^2\right) l\right\}
$$

(For a more rigorous proof, read the chapter).

Distribution-independent bounds

Replace $H^{\Lambda}_{\mathrm{ann}}(l)$ with the growth function.

$$
G^{\Lambda}(l) \quad = \quad \sup_{z_1,\ldots,z_l} \log N^{\Lambda}(z_1,\ldots,z_l) \qquad \text{“Worst case sequence”}
$$

$$
H^{\Lambda}(l) \overset{\text{Jensen's}}{\leq} H^{\Lambda}_{\mathrm{ann}}(l) \overset{\text{trivial}}{\leq} G^{\Lambda}(l)
$$

Learnability: We say that a class of functions Λ is learnable if

$$
\lim_{l\to\infty} \frac{G^{\Lambda}(l)}{l} \quad = \quad 0
$$

VC dimension

The VC dimension h of a family Λ is the largest integer s.t.

$$
N^{\Lambda}(z_1,\ldots,z_l) = 2^h \text{ for some } (z_1,\ldots,z_l)
$$

VC lemma (Sauer's lemma)

$$
G^{\Lambda}(l) = \begin{cases} \log 2^l & l \leq h \\ \leq \log \sum_{i=0}^{h} \binom{l}{i} & l > h \end{cases}
$$

The above formula says that the growth function grows linearly when the training set size is $\leq$ the VC dimension. Afterwards the growth function grows logarithmically. Later we will prove

$$
\sum_{i=0}^{h} \binom{l}{i} \leq \left(\frac{el}{h}\right)^h
$$

Rectangle example

Suppose we want to classify a set of points in the plane by fitting a rectangle around the members of one class. The VC dimension of this classifier is the highest number of points that may be “shattered”. A family Λ can shatter n points if it is possible to label these n points in all possible ways. Since we can only shatter up to 4 points, the VC dimension of the rectangle family is 4. We'll continue this discussion next time.

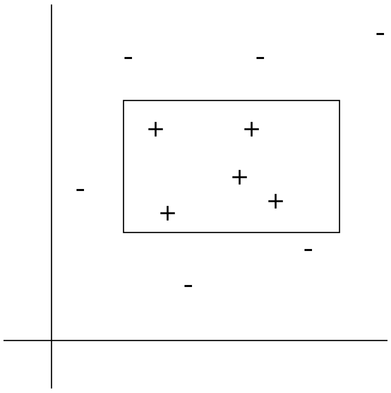

Figure 1: Rectangle family of classifiers