VC Dimension and classification

John Duchi

Outline

Setting for the lecture

Binary classification problems: data $X \in \mathcal{X}$ and labels $Y \in \{-1, 1\}$. Hypothesis class $\mathcal{H} \subset \{h : \mathcal{X} \to \mathbb{R}\}$.
Goal: Find $h \in \mathcal{H}$ with

$$L(h) := \mathbb{E}[\mathbf{1}\{h(X)Y \le 0\}]$$

small
Loss is always

$$\ell(h; (x, y)) = \mathbf{1}\{h(x)y \le 0\} = \begin{cases} 1 & \text{if } \operatorname{sign}(h(x)) \neq y \\ 0 & \text{if } \operatorname{sign}(h(x)) = y \end{cases}$$

Finite hypothesis classes

Theorem

Let $\mathcal{H}$ be a finite class. Then

$$\mathbb{P}\left(\exists h \in \mathcal{H} \text{ s.t. } |L(h) - \widehat{L}_n(h)| \geq \sqrt{\frac{\log|\mathcal{H}| + t}{2n}}\right) \leq 2e^{-t}.$$

Finite hypothesis classes: generalization

Corollary

Let $\mathcal{H}$ be a finite class, $\widehat{h}_n \in \operatorname{argmin}_h \widehat{L}_n(h)$. Then (for numerical constant $C < \infty$)

$$L(\widehat{h}_n) \le \min_{h \in \mathcal{H}} L(h) + C\sqrt{\frac{\log \frac{|\mathcal{H}|}{\delta}}{n}}$$

w.p. $\ge 1 - \delta$

Finite hypothesis classes: perfect classifiers

Possible to give better guarantees if there are good classifiers! We won't bother looking at bad ones.

Theorem

Let $\mathcal{H}$ *be a finite hypothesis class and assume* $\min_h L(h) = 0$*. Then for* $t \geq 0$

$$\mathbb{P}\left(L(\widehat{h}_n) \geq L(h^\star) + \frac{\log|\mathcal{H}| + t}{n}\right) \leq e^{-t}.$$

Do not pick the bad ones

Finite function classes: Rademacher complexity

Idea: Use Rademacher complexity to understand generalization even for these?

Let $\mathcal{F}$ be finite with $|f| \le 1$ for $f \in \mathcal{F}$. Then

$$R_n(\mathcal{F}) := \mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i) \right| \right]$$

satisfies

$$\mathbb{P}\left(\max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] \right| \ge 2 R_n(\mathcal{F}) + t \right) \le 2 \exp(-cnt^2)$$

Finite function classes: sub-Gaussianity

- Let P_n be empirical distribution

- Define $\|f\|_{L^2(P_n)}^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2$

- What about sum

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(x_i)$$

Finite function classes: Rademacher complexity

Proposition (Massart's finite class bound)

Let $\mathcal{F}$ be finite with $M := \max_{f \in \mathcal{F}} \|f\|_{L^2(P_n)}$. Then

$$\widehat{R}_n(\mathcal{F}) \le \sqrt{\frac{2M^2 \log(2 \operatorname{card}(\mathcal{F}))}{n}}.$$

Infinite classes with finite labels

What if we had a classifier $h : \mathcal{X} \to \{-1, 1\}$ that could only give a certain number of different labelings to a data set?

Example (Sketchy)

Say $\mathcal{X} = \mathbb{R}$ and $h_t(x) = \mathrm{sign}(x - t)$. Complexity of

$$\mathcal{F} := \{f(x) = \mathbf{1}\{h_t(x) \leq 0\}\}?$$

Complexity of function classes

Define

$$\mathcal{F}(x_{1:n}) := \{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{F}\}.$$

Then

$$\widehat{R}_n(\mathcal{F}) = \widehat{R}_n(\mathcal{F}')$$

whenever $\mathcal{F}(x_{1:n}) = \mathcal{F}'(x_{1:n})$

Proposition

Rademacher complexity depends on values of $\mathcal{F}$: if $|f(x)| \le M$ for all x then

$$R_n(\mathcal{F}) \le c \cdot M \sup_{x_1, \ldots, x_n \in \mathcal{X}} \sqrt{\frac{\log \operatorname{card}(\mathcal{F}(x_{1:n}))}{n}}.$$

Proof of complexity

Shatter coefficients

Given function class $\mathcal{F}$, shattering coefficient (growth function) is

$$\begin{aligned} \mathsf{s}_n(\mathcal{F}) &:= \sup_{x_1,\ldots,x_n \in \mathcal{X}} \operatorname{card}\left(\mathcal{F}(x_{1:n})\right) \\ &= \sup_{x_{1:n} \in \mathcal{X}^n} \operatorname{card}\left(\left(f(x_1), \ldots, f(x_n)\right) \mid f \in \mathcal{F}\right) \end{aligned}$$

Example

Thresholds in $\mathbb{R}$

Shatter coefficients and Rademacher complexity

Proposition

For any function class $\mathcal{F}$ with $|f(x)| \le M$ we have

$$R_n(\mathcal{F}) \le cM\sqrt{\frac{\log \mathsf{s}_n(\mathcal{F})}{n}}.$$

VC Dimension

How do we use shatter coefficients to give complexity guarantees?

Definition (VC Dimension)

Let $\mathcal{H}$ be a collection of boolean functions. The *Vapnik Chervonenkis (VC) Dimension* of $\mathcal{H}$ is

$$\mathsf{VC}(\mathcal{H}) := \sup \left\{ n \in \mathbb{N} : \mathsf{s}_n(\mathcal{H}) = 2^n \right\}.$$

VC Dimension: examples

Example (Thresholds in $\mathbb{R}$)

Example (Intervals in $\mathbb{R}$)

VC Dimension: examples

Example (Half-spaces in $\mathbb{R}^2$)

Finite dimensional hypothesis classes

Let $\mathcal{F}$ be functions $f : \mathcal{X} \to \mathbb{R}$ and suppose $\dim(\mathcal{F}) = d$

- Definition of dimension:

Example (Linear functionals)

If $\mathcal{F} = \{f(x) = w^\top x, w \in \mathbb{R}^d\}$ then $\dim(\mathcal{F}) = d$

Example (Nonlinear functionals)

If $\mathcal{F} = \{f(x) = w^\top \phi(x), w \in \mathbb{R}^d\}$ then $\dim(\mathcal{F}) = d$

VC dimension of finite dimensional classes

Let $\mathcal{F}$ have $\mathsf{dim}(\mathcal{F}) = d$ and let

$$\mathcal{H} := \{h : \mathcal{X} \to \{-1, 1\} \text{ s.t. } h(x) = \mathrm{sign}(f(x)), f \in \mathcal{F}\}.$$

Proposition (Dimension bounds VC dimension)

$$\mathsf{VC}(\mathcal{H}) \le \mathsf{dim}(\mathcal{F})$$

Finite dimensional hypothesis classes: proof

Sauer-Shelah Lemma

Theorem

Let $\mathcal{H}$ be boolean functions with $\mathsf{VC}(\mathcal{H}) = d$. *Then*

$$\mathsf{s}_n(\mathcal{H}) \le \sum_{i=0}^{d} \binom{n}{i} \le \begin{cases} 2^n & \textit{if } n \le d \\ \left(\frac{ne}{d}\right)^d & \textit{if } n > d \end{cases}$$

Rademacher complexity of VC classes

Proposition

Let $\mathcal{H}$ be collection of boolean functions with $\mathsf{VC}(\mathcal{H}) = d$*. Then*

$$R_n(\mathcal{H}) \le c\sqrt{\frac{d \log \frac{n}{d}}{n}}.$$

Proof is immediate (but a tighter result is possible):

Generalization bounds for VC classes

Proposition

Let $\mathcal{H}$ have VC-dimension d and $\ell(h;(x,y)) = \mathbf{1}\{h(x) \neq y\}$. Then

$$\mathbb{P}\left(\exists\ h \in \mathcal{H} \text{ s.t. } |\widehat{L}_n(h) - L(h)| \geq c\sqrt{\frac{d \log \frac{d}{n}}{n}} + t\right) \leq 2e^{-nt^2}$$

Things we have not addressed

- Multiclass problems (Natarajan dimension, due to Bala Natarajan; see also Multiclass Learnability and the ERM Principle by Daniely et al.)
- Extending "zero error" results to infinite classes
- Non-boolean classes

Reading and bibliography

1. M. Anthony and P. Bartlet. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 1999
2. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002
3. S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005
4. A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York, 1996 (Ch. 2.6)
5. Scribe notes for Statistics 300b: `http://web.stanford.edu/class/stats300b/`