VC Dimension and classification

John Duchi
Outline

I Setting: classification problems
II Finite hypothesis classes
   1 Union bounds
   2 Zero error case
III Shatter coefficients and Rademacher complexity
IV VC Dimension
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

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

small

Loss is always

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

Theorem

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

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

Corollary

Let $\mathcal{H}$ be a finite class, $\hat{h}_n \in \arg\min_h \hat{L}_n(h)$. Then (for numerical constant $C < \infty$)

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

w.p. $\geq 1 - \delta$
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* $t \geq 0$

$$\mathbb{P} \left( L(\hat{h}_n) \geq L(h^*) + \frac{\log |\mathcal{H}| + t}{n} \right) \leq e^{-t}. $$
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 $|f| \leq 1$ 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

$$\mathbb{P} \left( \max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)] \right| \geq 2R_n(\mathcal{F}) + t \right) \leq 2 \exp(-cnt^2)$$
Finite function classes: sub-Gaussianity

- Let $P_n$ be empirical distribution

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

- What about sum

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(x_i)$$
Proposition (Massart’s finite class bound)

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

$$\hat{R}_n(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log(2 \text{card}(\mathcal{F}))}{n}}.$$
Infinite classes with finite labels

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

Example (Sketchy)

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

$$\mathcal{F} := \{ f(x) = 1 \{ h_t(x) \leq 0 \} \}?$$
Complexity of function classes

Define

\[ \mathcal{F}(x_1:n) := \{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{F}\}. \]

Then

\[ \hat{R}_n(\mathcal{F}) = \hat{R}_n(\mathcal{F}') \]

whenever \( \mathcal{F}(x_1:n) = \mathcal{F}'(x_1:n) \)

Proposition

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

\[ R_n(\mathcal{F}) \leq c \cdot M \sup_{x_1,\ldots,x_n \in \mathcal{X}} \sqrt{\frac{\log \text{card}(\mathcal{F}(x_1:n))}{n}}. \]
Proof of complexity
Shatter coefficients

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

\[
s_n(\mathcal{F}) := \sup_{x_1, \ldots, x_n \in X} \text{card} (\mathcal{F}(x_1:n))
\]

\[
= \sup_{x_1:n \in X^n} \text{card} ((f(x_1), \ldots, f(x_n)) | f \in \mathcal{F})
\]

Example
Thresholds in $\mathbb{R}$
Proposition

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

$$R_n(\mathcal{F}) \leq cM \sqrt{\frac{\log s_n(\mathcal{F})}{n}}.$$
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

$$\text{VC}(\mathcal{H}) := \sup \{ n \in \mathbb{N} : s_n(\mathcal{H}) = 2^n \}.$$
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 \( \text{dim}(\mathcal{F}) = d \)

- Definition of dimension:

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

Example (Nonlinear functionals)
If \( \mathcal{F} = \{ f(x) = w^\top \phi(x), w \in \mathbb{R}^d \} \) then \( \text{dim}(\mathcal{F}) = d \)
VC dimension of finite dimensional classes

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

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

Proposition (Dimension bounds VC dimension)

$$\text{VC}(\mathcal{H}) \leq \dim(\mathcal{F})$$
Finite dimensional hypothesis classes: proof
Theorem

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

$$s_n(\mathcal{H}) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \begin{cases} 2^n & \text{if } n \leq d \\ \left(\frac{ne}{d}\right)^d & \text{if } n > d \end{cases}$$
Proposition

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

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

Proof is immediate (but a tighter result is possible):
Generalization bounds for VC classes

Proposition

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

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


5. Scribe notes for Statistics 300b: http://web.stanford.edu/class/stats300b/