Uniform concentration inequalities, martingales, Rademacher complexity and symmetrization

John Duchi

Outline

Standard recipe

In machine learning problem (say of predicting $y \in \mathcal{Y}$ from $x \in \mathcal{X}$)

1. Choose data representation for x and parameter space Θ (said differently, hypothesis class $\mathcal{H}$)
2. Choose loss function ℓ
3. Given sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, minimize

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; (X_i, Y_i)).$$

What is actual goal? Minimize risk/expected loss

$$L(\theta) := \mathbb{E}[\ell(\theta; (X, Y))] \text{ or } L(h) := \mathbb{E}[\ell(h; (X, Y))]$$

Does ML work?

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \widehat{L}_n(\theta) \quad \text{where} \quad \widehat{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; (X_i, Y_i))$$

Fixed θ:

$$\widehat{L}_n(\theta) \to L(\theta)$$

But $\widehat{\theta}_n$ *depends on data*.

Example: Failure when $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, I_d)$, $Y_i \perp X_i$, $\Theta = \mathbb{R}^d$

Does ML work?

Definition (Uniform law of large numbers)

$$\sup_{\theta \in \Theta} \left| \widehat{L}_n(\theta) - L(\theta) \right| \xrightarrow{p} 0$$

More generally, a collection of functions $\mathcal{F}$, $f : \mathcal{X} \to \mathbb{R}$, satisfies ULLN if

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \xrightarrow{p} 0.$$

Consequence for risk minimization

One picture of ULLNs

Covering idea (come back later)

Bounded differences and martingales

Definition (Martingales)

Let $X_1, X_2, \ldots$ be random vectors and $Z_1, Z_2, \ldots$ be a sequence of random variables. Then $\{X_n\}$ is a *martingale sequence adapted to* $\{Z_n\}$ if

i X_i is a function of $Z_1, Z_2, \ldots, Z_i$

ii $\mathbb{E}[X_i \mid Z_1, \ldots, Z_{i-1}] = X_{i-1}$.

Example: independent sums

Concentration of martingales

Definition (Martingale differences)

Let $\{X_i\}$ be a martingale adapted to $\{Z_i\}$ and define $D_i = X_i - X_{i-1}$. Then $\{D_i\}$ is a *martingale difference sequence*

Example: independent sums

Definition (Sub-gaussian martingale)

D_i is a σ^2*-sub-Gaussian martingale difference* if

$$\mathbb{E}[e^{\lambda D_i} \mid Z_{1:i-1}] \leq e^{\frac{\lambda^2 \sigma^2}{2}} \quad \text{for all } \lambda \in \mathbb{R}$$

Azuma-Hoeffding inequality

If D_i is a σ^2-sub-Gaussian martingale difference sequence, for $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} D_i \geq t\right) \leq \exp\left(-\frac{t^2}{2n\sigma^2}\right)$$
$$\mathbb{P}\left(\sum_{i=1}^{n} D_i \leq -t\right) \leq \exp\left(-\frac{t^2}{2n\sigma^2}\right)$$

Doob martingales

Let $X_1, \ldots, X_n \in \mathcal{X}$ be a sequence of independent random variables and $f : \mathcal{X}^n \to \mathbb{R}$. The *Doob martingale difference* is

$$D_i := \mathbb{E}[f(X_{1:n}) \mid X_{1:i}] - \mathbb{E}[f(X_{1:n}) \mid X_{1:i-1}]$$

Remark: look at expectations and sums

Concentration of functions with bounded differences

A function $f : \mathcal{X}^n \to \mathbb{R}$ has bounded differences if

$$\left| f(x_{1:i-1}, x_i, x_{i+1:n}) - f(x_{1:i-1}, x_i', x_{i+1:n}) \right| \le c_i$$

for all i and x, x'

Theorem (McDiarmid's or bounded-differences inequality)

Let f satisfy bounded differences and X_i be independent RVs. Then

$$\mathbb{P}\left(|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \ge t\right) \le \exp\left(-\frac{2t^2}{\|c\|_2^2}\right)$$

Proof of McDiarmid's inequality

Bounded differences in risk minimization

Let $\ell : \Theta \times \mathcal{X} \to [a, b]$. Then

$$\sup_{\theta \in \Theta} \left| \widehat{L}_n(\theta) - L(\theta) \right| = \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i) - L(\theta) \right|$$

satisfies bounded differences

From probability to expectation

Corollary

Let $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ *take values in* $[a, b]$. *Then*

$$\mathbb{P}\left(\sup_{\theta \in \Theta}\left|\widehat{L}_n(\theta) - L(\theta)\right| \ge \mathbb{E}\left[\sup_{\theta \in \Theta}\left|\widehat{L}_n(\theta) - L(\theta)\right|\right] + t\right) \le \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Symmetrization

Proposition (Symmetrization inequality)

Let $\mathcal{F}$ be a collection of $f : \mathcal{X} \to \mathbb{R}$ and $X_1, \ldots, X_n$ be independent. Then

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)]) \right|\right] \le 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|\right]$$

where $\varepsilon_i \in \{-1, 1\}$ are i.i.d. random signs

Rademacher complexity

Let $x_1, \ldots, x_n \in \mathcal{X}$ be arbitrary and $\mathcal{F}$ a collection of $f : \mathcal{X} \to \mathbb{R}$. The *empirical Rademacher complexity* of $\mathcal{F}$ on $x_{1:n}$ is

$$\widehat{R}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right]$$

where $\varepsilon_i \in \{-1, 1\}$ are i.i.d. random signs. The *Rademacher complexity* of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}\left[\widehat{R}_n(\mathcal{F})\right]$$

where expectation is over $X_1, \ldots, X_n$

Rademacher complexity and losses

Theorem (Concentration and Rademacher complexity)

Let $\mathcal{F} := \{\ell(\theta, \cdot) \mid \theta \in \Theta\}$ *(viewed as functions on* $\mathcal{X}$*) and* $\ell(\theta, x) \in [a, b]$*. Then*

$$\mathbb{P}\left(\exists \theta \in \Theta \text{ s.t. } \left|\widehat{L}_n(\theta) - L(\theta)\right| \geq 2R_n(\mathcal{F}) + t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Rademacher complexity of norm balls (ℓ_2)

Rademacher complexity of norm balls (ℓ_1)

Properties of Rademacher complexity

(1) Containment: if $\mathcal{F} \subset \mathcal{H}$ then $\widehat{R}_n(\mathcal{F}) \le \widehat{R}_n(\mathcal{H})$

(2) Convex hulls: $\widehat{R}_n(\mathcal{F}) = \widehat{R}_n(\mathrm{Conv}(\mathcal{F})) = \widehat{R}_n(\mathrm{absConv}(\mathcal{F}))$

(3) Single functions: $\widehat{R}_n(\{f\}) \le \frac{1}{\sqrt{n}} \|f\|_{L^2(P_n)}$

(4) Sums of function classes:
$\widehat{R}_n(\mathcal{F}_1 + \mathcal{F}_2 + \cdots + \mathcal{F}_k) \le \sum_{i=1}^k \widehat{R}_n(\mathcal{F}_i)$

(5) Contraction [Ledoux & Talagrand, Thm. 4.12]: if $\phi : \mathbb{R} \to \mathbb{R}$ is L_ϕ-Lipschitz and $\phi(0) = 0$, then

$$\widehat{R}_n(\phi \circ \mathcal{F}) \le 2L_\phi \widehat{R}_n(\mathcal{F})$$

Example: margin-based classification

Setting: Data $(X, Y) \in \mathbb{R}^d \times \{-1, 1\}$ where

$$\ell(\theta; (x, y)) = \phi(y\theta^T x) \quad \text{for } \phi : \mathbb{R} \to \mathbb{R}, \text{ non-increasing}$$

Margin-based classification and generalization

Theorem

Assume that ϕ is c_ϕ-Lipschitz, that $\|x\|_2 \le b_{\mathcal{X}}$ and $\|\theta\|_2 \le b_\Theta$. Then with probability at least $1 - \delta$, for all $\theta \in \Theta$

$$L(\theta) \le \widehat{L}_n(\theta) + O(1) \cdot \left[\frac{L_\phi b_{\mathcal{X}} b_\Theta}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} + \frac{\phi(0)}{\sqrt{n}} \right].$$

Proof of margin-based classifiers

Multiclass classification

Consider k-class classification problem,

$$\theta = \begin{bmatrix} \theta^1 & \theta^2 & \cdots & \theta^k \end{bmatrix} \in \mathbb{R}^{d \times k}$$

Let margin $s = \theta^T x \in \mathbb{R}^k$, loss $\phi : \mathbb{R}^k \to \mathbb{R}$ of form

$$\ell(\theta; x, y) = \phi(\Pi_y s) = \phi(\Pi_y \theta^T x)$$

for some "labeling" matrix Π_y

Multiclass margin-based losses

1. Multiclass logistic

2. Multiclass hinge/SVM

Additional comments

Reading and bibliography

1. M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer, 1991

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. S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence.* Oxford University Press, 2013