Binary Classification, Part 1

(X, Y) $\quad X \in \mathcal{X}$ (feature)
$\quad\quad\quad\quad Y \in \{0,1\}$ (label)

classifiers: $\quad \mathcal{X} \to \{0,1\}$

$\quad 1_{\{X \in C\}}$ – prediction

$$L(C) := \mathbb{P}[Y \neq 1_{\{X \in C\}}]$$

$P = \mathcal{L}((X,Y))$ known: Bayes optimal classifier

$$f^*(x) = 1_{\{\eta(x) \geq \frac{1}{2}\}} \quad \longleftrightarrow \quad C^* = \{x \in \mathcal{X} : \eta(x) \geq \tfrac{1}{2}\}$$

$$\eta(x) = \mathbb{P}[Y = 1 \mid X = x)$$
$$= \mathbb{E}[Y \mid X = x]$$

P unknown, iid samples given:

try to 'learn the "best" C in some $\mathcal{C}$

$$\text{data} \longrightarrow \boxed{LA} \longrightarrow \hat{C} \in \mathcal{C}$$

$$L(\hat{C}) \approx \inf_{C \in \mathcal{C}} L(C) \quad \text{whp}$$

Preview: $\mathcal{C}$ has to be a VC class

Ex.: $\quad \mathcal{X} = \mathbb{R}^d$

$$C = \{x \in \mathbb{R}^d : \langle w, x \rangle + b \geq 0\}$$
$$w \in \mathbb{R}^d \setminus \{0\}$$
$$b \in \mathbb{R}$$

– linear discriminant rules $\quad\quad$ VC-dim $= d+1$

ERM, binary classification version

$\mathcal{X}$: feature space

$\mathcal{C}$: collection of subsets of $\mathcal{X}$

$(X_1, Y_1), \dots, (X_n, Y_n) \overset{iid}{\sim} P$ on $\mathcal{X} \times \{0,1\}$

ERM: $\quad C \in \mathcal{C} \longrightarrow \ell_C(x,y) := 1_{\{y \neq 1_C(x)\}}$

$$\hat{C}_n = \underset{C \in \mathcal{C}}{\operatorname{argmin}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_C(X_i, Y_i)}_{\substack{\text{fraction of}\\ \text{mistakes by } C\\ \text{on the data}}}$$

Abstract ERM (reminder):

$Z_1, \dots, Z_n \overset{iid}{\sim} P$ on $\mathcal{Z}$

$\mathcal{F}$: a class of fcns $f : \mathcal{Z} \to [0,1]$

$$\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} f(Z_i)$$

Here: $\mathcal{Z} = \mathcal{X} \times \{0,1\} \qquad z = (x,y)$

$\mathcal{F} = \mathcal{F}_{\mathcal{C}} := \{(x,y) \mapsto \ell_C(x,y) : C \in \mathcal{C}\}$

Thm With prob. $\geq 1-\delta$,

$$L(\hat{C}_n) \leq \inf_{C \in \mathcal{C}} L(C) + 4\,\mathbb{E}R_n(\mathcal{F}_{\mathcal{C}}(Z^n)) + \sqrt{\frac{2\log(1/\delta)}{n}}.$$

Fix $z^n = (z_1, \dots, z_n)$ $\quad z_i = (X_i, Y_i)$

$$\mathcal{F}_{\mathcal{C}}(z^n) = \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}_{\mathcal{C}}\}$$
$$= \{(1_{\{Y_1 \neq 1_{\{X_1 \in C\}}\}}, \dots, 1_{\{Y_n \neq 1_{\{X_n \in C\}}\}}) : C \in \mathcal{C}\}$$
$$\subseteq \{0,1\}^n$$

FCL: $R_n(\mathcal{F}_{\mathcal{C}}(z^n)) \leq 2\sqrt{\dfrac{\log|\mathcal{F}_{\mathcal{C}}(z^n)|}{n}}$

– if $\mathcal{F}_C$ is a VC class $[V(\mathcal{F}_C) < \infty]$,
$\log|\mathcal{F}_{\mathcal{C}}(z^n)| \leq V(\mathcal{F}_C)\log(n+1)$ [Sauer–Shelah]

Key risk bound for ERM: w.p. $\geq 1-\delta$,

$$L(\hat{C}_n) \leq \inf_{C \in \mathcal{C}} L(C) + 8\sqrt{\frac{V(\mathcal{F}_C)\log(n+1)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

NB: using Dudley's chaining,

$$\sqrt{\frac{V(\mathcal{F}_C)\log(n+1)}{n}} \longrightarrow \text{const}\sqrt{\frac{V(\mathcal{F}_C)}{n}}.$$

$$\mathcal{F}_{\mathcal{C}} = \{(x,y) \mapsto 1_{\{y \neq 1_{\{x \in C\}}\}} : C \in \mathcal{C}\}$$

$\mathcal{C}$ – fundamental object

<u>Lemma</u> $\quad V(\mathcal{F}_C) = V(\mathcal{C})$

$\mathcal{F}_C$: class of subsets of $\mathcal{X} \times \{0,1\}$

$\mathcal{C}$: class of subsets of $\mathcal{X}$

E.g.: $\mathcal{C}$ are indicators of half-spaces in $\mathbb{R}^d$,

$V(\mathcal{F}_{\mathcal{C}}) = V(\mathcal{C}) = d+1$

— excess risk $O\left(\sqrt{\frac{d+\log(1/\delta)}{n}}\right)$.

Proof (of lemma)

0) Assume $V(\mathcal{C}), V(\mathcal{F}_C) < \infty$

1) A set $\{x_1, \dots, x_n\} \subset \mathcal{X}$ is shattered by $\mathcal{C}$ iff $\{(x_1, 0), \dots, (x_n, 0)\} \subset \mathcal{X} \times \{0,1\}$ is shattered by $\mathcal{F}_C$

$\forall C \in \mathcal{C}$: $\quad l_C(x,y) = 1_{\{y \neq 1_C(x)\}}$

$$l_C(x,0) = 1_{\{0 \neq 1_C(x)\}}$$

$$= 1_{\{x \in C\}}$$

$(1_{\{x_1 \in C\}}, \dots, 1_{\{x_n \in C\}}) = (b_1, \dots, b_n)$ iff

$(l_C(x_1, 0), \dots, l_C(x_n, 0)) = (b_1, \dots, b_n)$

2) $V(\mathcal{C}) \le V(\mathcal{F}_{\mathcal{C}})$

$n = V(\mathcal{C})$

$\exists \{x_1, \dots, x_n\} \subset \mathcal{X}$ shattered by $\mathcal{C}$

$\Rightarrow \{(x_1, 0), \dots, (x_n, 0)\} \subset \mathcal{X} \times \{0,1\}$ shattered by $\mathcal{F}(\mathcal{C})$

3) $V(\mathcal{F}_C) \le V(\mathcal{C})$

$n = V(\mathcal{F}_C)$

$\exists \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathcal{X} \times \{0,1\}$ shattered by $\mathcal{F}_C$

$\downarrow$

$\forall\, b = (b_1, \dots, b_n) \in \{0,1\}^n \quad \exists\, c \in \mathcal{C}$ s.t.

$l_C(x_i, y_i) = 1_{\{y_i \neq 1_C(x_i)\}} = b_i \qquad i \in [n]$

<u>Goal</u>: show that $\{x_1, \dots, x_n\} \subset \mathcal{X}$ shattered by $\mathcal{C}$

— need x_i's to be distinct — can assume this:

$\{(x, 0), (x, 1)\}$ can't be shattered by $\mathcal{F}_\mathcal{C}$

$l_C(x, 0) \neq l_C(x, 1) \quad \forall x \in \mathcal{X}$

$$
\begin{aligned}
l_C(x_i, y_i) &= 1_{\{y_i \neq 1_C(x_i)\}} \\
&= y_i \oplus 1_{\{x_i \in C\}} \\
&= y_i \oplus l_C(x_i, 0)
\end{aligned}
$$

$l_C(x, y) = \begin{cases} 0, & y = 1_C(x) \\ 1, & y \neq 1_C(x) \end{cases}$

Know that $\{(x_1, 0), \dots, (x_n, 0)\}$ shattered by $\mathcal{F}_\mathcal{C}$

$\Rightarrow \{x_1, \dots, x_n\}$ shattered by $\mathcal{C}$

Let $b = (b_1, \dots, b_n) \in \{0,1\}^n$ be given;

take $b_i' := y_i \oplus b_i \qquad \forall i \in [n]$

$\exists\, C \in \mathcal{C}$ s.t. $l_C(x_i, y_i) = b_i' \quad \forall i$

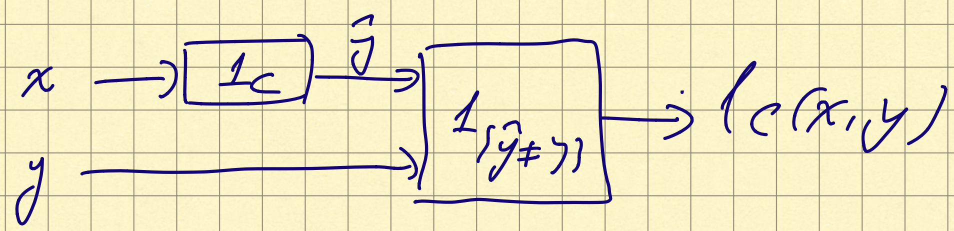
$$y_i \oplus 1_{\{x_i \in C\}} = y_i \oplus b_i$$

$$1_{\{x_i \in C\}} = b_i \,.$$

$\blacksquare$

Some intuition:

① $\ell_C(x,y)$ depends on x only through $1_{\{x \in C\}}$



② $(X, Y) = (X, 0)$ w.p. 1

$$L(C) = \mathbb{P}[1_C(X) = 1] = \mathbb{P}(X \in C]$$

---

1) If $\mathcal{C}$ is a VC-class, then the corresp. learning problem is PAC-learnable (by ERM).

2) The converse is also true: if binary classif. problem (for a given $\mathcal{C}$) is PAC-learnable, then $V(\mathcal{C}) < \infty$

$$\text{PAC:} \quad \sup_P \Big\{ L_P(\hat{C}_n) - \inf_{C \in \mathcal{C}} L_P(C) \Big\} \to 0$$

as $n \to \infty$

Examples

1) linear discriminant rules

$\mathcal{X} = \mathbb{R}^d$

$\mathcal{C}$ – indicators of half-spaces $\qquad V(\mathcal{C}) = d+1$

$$L(\hat{C}_n) - \inf_{C \in \mathcal{C}} L(\mathcal{C}) = O\left(\sqrt{d/n}\right) \qquad (\hat{C}_n : \text{ERM})$$

2) Dudley classifiers

$\mathcal{X}$ arbitrary

$\mathcal{G}$: linear space of fcns $\mathcal{X} \to \mathbb{R}$ spanned by lin. ind. $\{\psi_1, \dots, \psi_m\}$

$$g(x) = \sum_{j=1}^{m} c_j \psi_j(x)$$

$$\mathcal{X} \xrightarrow{(\psi_1, \dots, \psi_m)} \underset{\text{(feature space)}}{\mathbb{R}^m} \xrightarrow{(c_1, \dots, c_m)} g(x) = \sum_{j=1}^{m} c_j \psi_j(x)$$

$$C_g = \{ x \in \mathcal{X} : \ g(x) \geq 0 \} = \text{pos}(g)$$

$$V(\text{pos}(\mathcal{G})) = m, \quad \text{excess risk} = O\left(\sqrt{m/n}\right)$$

<u>Limitations</u>

1) Expressivity: how small can $\inf_{C \in \mathcal{C}} L(C)$ be?

2) Computational tractability:

finding ERM classifier may be expensive

e.g. : complexity of finding a half-space in $\mathbb{R}^d$ to minimize classification error on n points is $O(n^d)$ — prohibitive when $d \geq 5$.

<u>Next lecture:</u> surrogate losses via penalty fcns