FML LN2

Assumptions as before:

1) x_i drawn iid from unknown $\mathcal{D}$ on X.

2) $y_i = f(x_i)$, $i = 1, \dots, m$ for unknown f.

Goal: find f, at least approximately.

(Later, we'll assume $(x_i, y_i) \in X \times Y$ drawn iid from unknown $\mathcal{D}$ on $X \times Y$.)

Learner selects $h \in \mathcal{H} \subset \{g : X \to Y\}$.

I.e, seek h : $R(h)$ minimized. Problem: can't compute R since we don't know $\mathcal{D}$.

Def: Empirical risk. Let $h : X \to Y = \{0, 1\}$, the (true) labeling function $f : X \to Y$, sample $S = (x_1, \dots, x_m)$ (with $x_i \in X$), empirical risk is:

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{h(x_i) \neq f(x_i)\}} = \frac{1}{m} \# \{ i \in [m] : h(x_i) \neq f(x_i) \}.$$

$[m] := \{1, \dots, m\}$; for finite set A, $\#A = |A|$.

Drawing $x_{i, \dots, m}$ @ $\mathcal{D}$ $\rightsquigarrow$ draw $S = (x_1, \dots, x_m) \sim \mathcal{D}^m$, we obtain a rv denoted $\hat{R}(h)$.

Lemma 3. If $x_1, \dots, x_m$ are drawn iid accord. to $\mathcal{D}$, then for any (measurable) $h : X \to \{0, 1\}$,

$$\mathbb{E}[\hat{R}(h)] = R(h)$$

linearity

Proof: $\mathbb{E}\,\hat{R}(h) = \mathbb{E}\, \frac{1}{m} \sum_{i=1}^{m} \underbrace{\mathbb{1}_{h(x_i) \neq f(x_i)}}_{= \begin{cases} 1 & \text{if } h(x_i) \neq f(x_i) \\ 0 & \text{o/w} \end{cases}} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\, \mathbb{1}_{h(x_i) \neq f(x_i)}$

iid each $x_i \sim \mathcal{D}$

$$= \frac{1}{m} \sum_{i=1}^{m} \underbrace{\mathbb{E}\, \mathbb{1}_{h(x) \neq f(x)}}_{= R(h)} = \frac{1}{m} \sum_{i=1}^{m} R(h) = \frac{1}{m} m R(h) = R(h)$$

Def. 4: **Empirical Risk Minimization.** Let $f: X \to Y = \{0,1\}$ be the true labeling function and let $\mathcal{H} \subset \{h: X \to Y\}$ a hypothesis set. Given a sample $S = (x_1, \dots, x_m) \in X^m$ with corr. labels $y_i = f(x_i)$, for $i \in \{1, \dots, m\}$, empirical risk minimization consists of selecting a minimizer $h \in \mathcal{H}$ of $\hat{R}$, ie. selecting an h realizing

$$\min_{h \in \mathcal{H}} \hat{R}(h) = \min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{h(x_i) \neq f(x_i)}$$

some LLN argument made to justify?

We must be careful: overfitting.

Choice of suitable $\mathcal{H}$ is important.

Example. $X \subset \mathbb{R}^2$ an axis-aligned rect, eg

$$X = [0,1]^2.$$

B another axis aligned rect, eg $A = X \setminus B$.

$\mathbb{P}$: cts prob distr. on X $(\mathbb{P}(\{x\}) = 0 \ \forall x \in X)$, and s.t. $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$

e.g.: $X = [0,1]^2$, $B = [0, 1/2] \times [0,1]$,

$$\mathbb{P}(M) = \text{Area}(M).$$

Let $f: X \to \{0,1\}$ be given by

$$f(x) = \begin{cases} 0 & \text{if } x \in A \\ 1 & \text{if } x \in B \end{cases}$$

Given $S = (x_1, \dots, x_m) \in X^m$, labels $y_i = f(x_i)$, choose

$$h_S(x) = \begin{cases} y_i & \text{if } x = x_i \\ 0 & \text{otherwise.} \end{cases}$$

Here,

$$\hat{R}(h_S) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{h_S(x_i) \neq f(x_i)} = \frac{1}{m} \cdot 0 = 0.$$

But the true risk is

$$\mathbb{P}(h_S(x) \neq f(x)) = \mathbb{P}(x \neq x_1, \dots, x_m \text{ and } f(x) = 1)$$
$$= \mathbb{P}(B \setminus \{x_1, \dots, x_m\}) = \mathbb{P}(B) = \tfrac{1}{2}.$$

$\mathbb{P}(\{x_i\}) = 0$

Not better than random guessing.

$\rightsquigarrow$ $\hat{R}(h_S)$ minimal but $R(h)$ bad!

Def 5: PAC-learning, consistent case.

A hyp. class $\mathcal{H} \subset \{h : X \to Y = \{0,1\}\}$ is PAC-learnable if $\exists$ a fxn $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A w/ the following property:

For every $\varepsilon, \delta \in (0,1)$, $\forall$ prob. distrs. $\mathcal{D}$ over X, for every labeling function $f \in \mathcal{H}$, the following holds:

If $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ and $S = (x_1, \dots, x_m)$ is an iid sample, $S \sim \mathcal{D}^m$, then given the data $(x_i, y_i) = (x_i, f(x_i))$, $i = 1, \dots, m$, the algo A returns a hypothesis

$$h_S \in \mathcal{H}$$

s.t. with probability at least $1-\delta$ (over $S \sim \mathcal{D}^m$),

$$R(h_S) \leq \varepsilon.$$

Remarks.

- Analogous defn's of emp risk, ER mn, PAC-lrnbl for any Y consisting of 2 elts.
- Sample complexity $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ dets. # of of required training data to learn $\mathcal{H}$.

- Def'n not requires algo $\rightsquigarrow$ fast.
- ERM possible algo, may not be optimal.

See slides for more details, Also books.

Thm 6 (Finite $\mathcal{H}$, consistent case). Let $\mathcal{H}$ = finite set of functions $h: X \to Y = \{0,1\}$. Assume that the labeling function f belongs to $\mathcal{H}$ and let A be an algorithm that for each iid sample $S = (x_1, \dots, x_m)$ and labeled training data $(x_i, y_i) = (x_i, f(x_i)),\ i = 1, \dots, m$ returns a consistent hyp.

$$h_S \in \mathcal{H},$$

ie $\hat{R}(h_S) = 0$. Then for $\varepsilon, \delta \in (0,1)$, if

$$m \geq \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)\right)$$

the inequality $\mathbb{P}[R(h_S) \leq \varepsilon] \geq 1 - \delta$ holds.

Proof. $R(h_S)$ depends on S and is difficult to eval. directly, instead bound it as follows.

Let $0 < \varepsilon < 1$, $\mathcal{H} = \{h_1, \dots, h_n\}$, $n = \#\mathcal{H}$.

$$\mathbb{P}(R(h_S) > \varepsilon) = \mathbb{P}(\hat{R}(h_S) = 0 \text{ and } R(h_S) > \varepsilon)$$

transfer rv from $R(h_S) \rightsquigarrow \hat{R}_{(S)}(h)$. ie, transfer S-dep. $\rightsquigarrow$

$$\overset{*\text{key}}{\leq} \mathbb{P}(\hat{R}(h) = 0 \text{ and } R(h) > \varepsilon \text{ for some } h \in \mathcal{H})$$

$$= \mathbb{P}(\{\hat{R}(h_1) = 0 \ \& \ R(h_1) > \varepsilon\} \cup \{\hat{R}(h_2) = 0 \cdots\} \cup \cdots)$$

$$\{S \in X^m : \hat{R}_S(h_1) = 0 \ \& \ R(h_1) > \varepsilon\}$$

$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}(\{\hat{R}(h) = 0\} \cap \underbrace{\{R(h) > \varepsilon\}}_{= \left\{ X^m \text{ or } \emptyset \right.})$$

$$= \sum_{h : R(h) > \varepsilon} \mathbb{P}(\hat{R}(h) = 0)$$

$$\rightsquigarrow \mathbb{P}(\hat{R}(h) = 0) = \mathbb{P}(h(x_j) = f(x_j)\ \forall j = 1, \dots, m)$$

$$\overset{\text{iid}}{=} \prod_{j=1}^{m} \mathbb{P}(h(x_j) = f(x_j)) = \prod_{j=1}^{m} (1 - R(h))$$

$$= (1 - R(h))^m \leq (1-\varepsilon)^m$$

for $R(h) > \varepsilon$

$$\rightsquigarrow \mathbb{P}(R(h_S) > \varepsilon) \leq \sum_{h: R(h) > \varepsilon} (1-\varepsilon)^n \leq |\mathcal{H}|(1-\varepsilon)^m \leq |\mathcal{H}| e^{-\varepsilon m}$$

$\forall x \in \mathbb{R}$, $1 + x \leq e^x$ (exercise)

$$\mathbb{P}(R(h_S) \leq \varepsilon) \geq 1 - |\mathcal{H}| e^{-\varepsilon m}$$

with $\delta := |\mathcal{H}| e^{-\varepsilon m}$,

$$\mathbb{P}(R(h_S) \leq \varepsilon) \geq 1 - \delta$$

$\rightsquigarrow$ $\delta \in (0,1)$ and $m \geq \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)\right)$,

$$R(h_S \leq \varepsilon) \geq 1 - \delta$$

solve $\delta = |\mathcal{H}| e^{-\varepsilon m}$ for m $\rightsquigarrow$ $e^{\varepsilon m} = |\mathcal{H}|/\delta$

$$\varepsilon m = \log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)$$

$$m = \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\varepsilon}\right)\right)$$

Conclusion: for finite hypothesis set $\mathcal{H}$, a consistent learning algorithm $\mathcal{A}$ is a PAC learning algorithm with sample complexity polynomial (even linear) in $1/\varepsilon$, and logarithmic in $1/\delta$ and $|\mathcal{H}|$.

$\log|\mathcal{H}|$ may be interpreted as number of bits to represent $\mathcal{H}$ (up to a constant factor).

Note that $f \in \mathcal{H}$ guarantees that ERM always returns an h_S with $\hat{R}(h_S)$.

("consistent": $\hat{R}(h) = 0$. "Consistent case": $f \in \mathcal{H}$.)