Generalization theory

1 Big picture

(a) True error vs training error

$\mathcal{H}$ = set of classifiers, e.g. {linear separators in $\mathbb{R}^d$}

Training set of n points: $(x_1, y_1), \ldots, (x_n, y_n)$

Training error $\hat{err}(h) = \frac{1}{n} \sum_{i=1}^{n} 1(h(x_i) \neq y_i)$

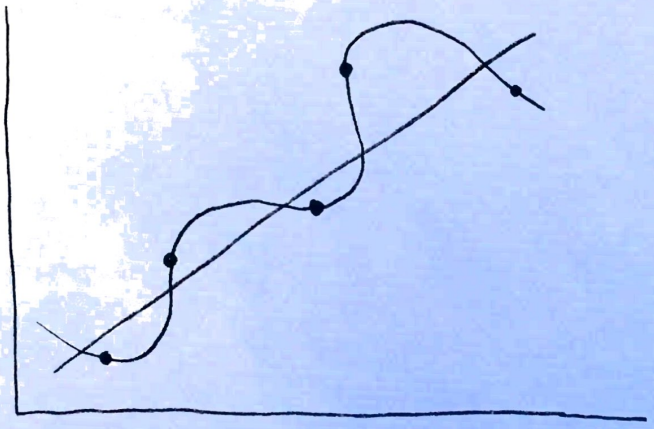
How does this relate to true error, $err(h)$?

Very roughly:

$$err(h) \leq \hat{err}(h) + \frac{c(\mathcal{H})}{n} + \sqrt{\hat{err}(h) \frac{c(\mathcal{H})}{n}}$$

where $c(\mathcal{H})$ is a measure of the size, or complexity, of $\mathcal{H}$.

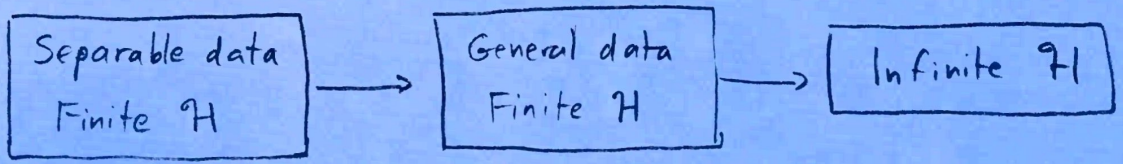
(b) Occam's razor: simpler hypotheses are preferable.



ML version of this: "smaller" hypothesis classes generalize better.

We'll look at some sophisticated notions of size. [A far cry from parameter-counting.]

(c) Cases of interest



2. Separable data, finite $\mathcal{H}$

(a) $\mathcal{X}$: input space

$\mathcal{Y}$: label space

$\mathcal{H}$: hypothesis class

Separable : $y = h^*(x)$ for some $h^* \in \mathcal{H}$

(b) Distribution P over $\mathcal{X} \times \mathcal{Y}$.

True error of $h \in \mathcal{H}$:

$$err(h) = Pr_{(x,y)\sim P}\left(h(x) \neq y\right)$$

(c) Thm Pick any $0 < \delta < 1$. Suppose $(x_1, y_1), \ldots, (x_n, y_n) \sim P$ i.i.d., ~~and we pick $\hat{h} \in \mathcal{H}$ consistent with them~~. With probability at least $1-\delta$ over the choice of training set, for any h consistent with the data,

$$err(h) \leq \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta}.$$

Pf. Define $\varepsilon = \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta}$.

(i) Fix any $h \in \mathcal{H}$ with $err(h) > \varepsilon$.

Pick $(x_1, y_1), \ldots, (x_n, y_n) \sim P$.

$$Pr\left(h \text{ consistent with } (x_1, y_1), \ldots, (x_n, y_n)\right) < (1-\varepsilon)^n$$

$$< e^{-\varepsilon n} = \frac{\delta}{|\mathcal{H}|}.$$

(ii) $\therefore Pr(\exists\, h \in \mathcal{H} : err(h) > \varepsilon \text{ and } h \text{ consistent with training set})$

$$\leq \sum_{\substack{h \in \mathcal{H} \\ err(h) > \varepsilon}} Pr(h \text{ consistent with training set})$$

$$\leq |\mathcal{H}| \cdot \frac{\delta}{|\mathcal{H}|} = \delta. \quad \blacksquare$$

(d) Example: $\mathcal{H}$ = k-sparse disjunctions over d variables.

[3] Non-separable case, finite $\mathcal{H}$

(a) Concentration of empirical averages

Fix any hypothesis $h \in \mathcal{H}$.

How close is (training error on n pts) to $\text{err}(h)$?

(training error on n pts): empirical average

(b) Central limit theorem

Informal version Suppose $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$.

Then the distribution of $\dfrac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$ approaches $N(0,1)$

as $n \to \infty$.

Example A coin of bias p is flipped n times. Let $\hat{p}$ be the fraction of heads observed. What is the distribution of $\hat{p}$ for large n?

- Define $X_i = \begin{cases} 1 & \text{if } i^{th} \text{ coin heads} \\ 0 & \text{otherwise.} \end{cases}$

Then $\mu = EX_i = p$ and $\sigma^2 = \text{var}(X_i) = p(1-p)$.

- By CLT:

$$\frac{n\hat{p} - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

$$\Rightarrow \hat{p} - p \sim \sqrt{\frac{p(1-p)}{n}} \cdot N(0,1)$$

$$\Rightarrow \hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

$\therefore \hat{p}$ has roughly a Gaussian distribution with mean p and standard deviation $\sqrt{\dfrac{p(1-p)}{n}} \leq \dfrac{1}{2\sqrt{n}}$.

$\therefore$ W.p. $\geq 95\%$, $\hat{p} = p \pm \dfrac{1}{\sqrt{n}}$. (for n large enough)

Hoeffding's inequality

Thm Let $Z_1, \ldots, Z_n$ be i.i.d. random variables with $a \le Z_i \le b$ and $\mu = \mathbb{E} Z_i$.

Then:

$$\Pr\left(\left| \frac{Z_1 + \ldots + Z_n}{n} - \mu \right| \ge \varepsilon \right) \le 2e^{-2\varepsilon^2 n/(b-a)^2}$$

Another way to write it: With probability $\ge 1-\delta$,

$$\left| \frac{Z_1 + \ldots + Z_n}{n} - \mu \right| \le (b-a)\sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$$

④ Application: relate true error to training error for finite $\mathcal{H}$.

Thm Suppose a training set $\{(x_i, y_i)\}$ is drawn i.i.d. from P. Define:

- $\hat{err}(h)$ = training error of h
  $= \frac{1}{n} \sum_{i=1}^{n} 1(h(x_i) \ne y_i)$
- $err(h)$ = true error of h = $\Pr_{(x,y)\sim P}(h(x) \ne y)$

Then with prob $\ge 1-\delta$,

$$\max_h \left| \hat{err}(h) - err(h) \right| \le \sqrt{\frac{1}{2n} \ln \frac{2|\mathcal{H}|}{\delta}}.$$

Pf. Fix any $h \in \mathcal{H}$. Apply Hoeffding's inequality with

$$Z_i = 1(h(x_i) \ne y_i).$$

Then take a union bound over $\mathcal{H}$. ⊠

In particular, let h^* be the best classifier:

h^* minimizes $err(h)$

and let $\hat{h}$ be the learned classifier.

$\hat{h}$ minimizes $\hat{err}(h)$.

Then:

$$err(\hat{h}) \le err(h^*) + \sqrt{\frac{2}{n} \ln \frac{2|\mathcal{H}|}{\delta}}.$$

(spell out the argument)

We have seen generalization bounds of the form

$$\max_h \left| \hat{err}(h) - err(h) \right| \leq \sqrt{\frac{c(\mathcal{H})}{n}}$$

where $c(\mathcal{H}) = \log |\mathcal{H}|$.

What if $\mathcal{H}$ is infinite?

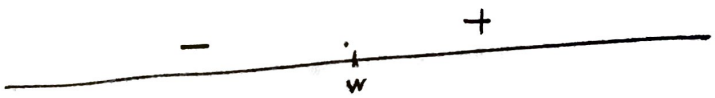
4 VC dimension

(a) Example

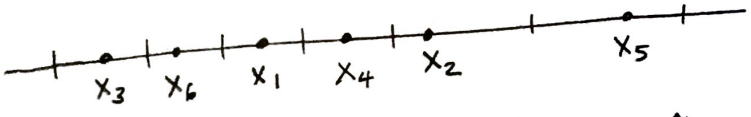
(i) $\mathcal{X} = \mathbb{R}$

$\mathcal{H} = \{\text{threshold classifiers } h_w : w \in \mathbb{R}\}$

$$h_w(x) = \begin{cases} + & \text{if } x \geq w \\ - & \text{if } x < w \end{cases}$$



(ii) Say we have n pts $x_1 \ldots x_n$.



Effectively there are just $n+1$ classifiers to consider, even though $\mathcal{H}$ is infinite.

∴ Think of $|\mathcal{H}|$ as $n+1$.

(b) Effective hypothesis class

Pretend you get your training data in two phases:

(i) First, $x_1, x_2, \ldots, x_n$

(ii) Then, the labels are revealed

After seeing (i), effective hypothesis class is

We have seen generalization bounds of the form

$$\max_h \left| \hat{err}(h) - err(h) \right| \leq \sqrt{\frac{c(\mathcal{H})}{n}}$$

where $c(\mathcal{H}) = \log |\mathcal{H}|$.

What if $\mathcal{H}$ is infinite?

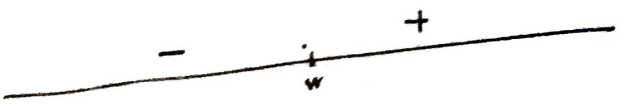
VC dimension

(a) Example

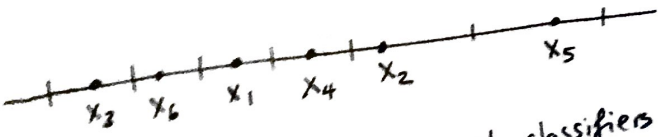
$\mathcal{X} = \mathbb{R}$

$\mathcal{H} = \{$threshold classifiers $h_w : w \in \mathbb{R}\}$

$$h_w(x) = \begin{cases} + & \text{if } x \geq w \\ - & \text{if } x < w \end{cases}$$



(i) Say we have n pts $x_1 \ldots x_n$.



Effectively there are just $n+1$ classifiers to consider, even though $\mathcal{H}$ is infinite.

∴ Think of $|\mathcal{H}|$ as $n+1$.

(b) Effective hypothesis class

Pretend you get your training data in two phases:

(i) First, $x_1, x_2, \ldots, x_n$

(ii) Then, the labels are revealed

After seeing (i), effective hypothesis class is

$$\tilde{\mathcal{H}} = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$$

eg. If $\mathcal{H}$ = thresholds, $|\tilde{\mathcal{H}}| = n+1$.

We can use $c(\mathcal{H}) = \log |\tilde{\mathcal{H}}|$ in our generalization bounds. $\left[\sqrt{\frac{c(\mathcal{H})}{n}}\right]$

(c) How big is $\tilde{\mathcal{H}}$, in general?

- Clearly, $|\tilde{\mathcal{H}}| \leq 2^n$, but this gives a vacuous bound
- Any hypothesis class $\mathcal{H}$ has an associated parameter called the Vapnik-Chervonenkis (VC) dimension, $VC(\mathcal{H})$.

  And: $|\tilde{\mathcal{H}}| \leq O(n^{VC(\mathcal{H})})$
- Can use $c(\mathcal{H}) = VC(\mathcal{H})$, as a measure of the complexity of $\mathcal{H}$.

(d) VC dimension

(i) Def $\mathcal{H}$ shatters a set of points $x_1, \ldots, x_m \in \mathcal{X}$ if it can realize all 2^m possible labelings of these points.

Ex. $\mathcal{H} = \{\text{linear seporators in } \mathbb{R}^2\}$

$\mathcal{H}$ shatters any 3 (non-collinear) points in $\mathbb{R}^2$.

(ii) Def The VC dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Ex. $\mathcal{H} = \{\text{linear separators in } \mathbb{R}^2\}$

$\mathcal{H}$ cannot shatter any 4 pts

$\therefore VC(\mathcal{H}) = 3$

(d) Ex. $\mathcal{H} = \{\text{thresholds in } \mathbb{R}\}$.

$VC(\mathcal{H}) = 1$.

(iv) Can show: $VC(\text{linear separators in } \mathbb{R}^d) = d+1$.

(e) Sauer's lemma If $VC(\mathcal{H}) = V$ then for any $x_1, \ldots, x_n \in \mathcal{X}$,

$$\tilde{\mathcal{H}} = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$$

has size at most $O(n^V)$.

(f) Thm Suppose $\mathcal{H}$ has VC dimension $V < \infty$. Suppose a training set of n pts is drawn i.i.d. from the underlying distribution P on $\mathcal{X} \times \mathcal{Y}$ and that $\hat{err}$ denotes training error and err true error. Then, with probability $\geq 1-\delta$,

For all $h \in \mathcal{H}$: $$|\hat{err}(h) - err(h)| \leq C_0 \sqrt{\frac{V + \log 1/\delta}{n}}$$

where C_0 is some constant.

5 More refined notions of hypothesis class size

(a) $\mathcal{H} = \{\text{linear separators in } \mathbb{R}^d\} = \{h_w : w \in \mathbb{R}^{d+1}\}$ $h_w(x) = \text{sign}(w \cdot x)$

(i) We have seen $VC(\mathcal{H}) = d+1$, so we get generalization bounds of the form $\sqrt{d/n}$. But what if we look at separators with large margin?

(ii) Fix any $x_1, \ldots, x_n \in \mathbb{R}^d$ and define

$$\tilde{\mathcal{H}}_\gamma = \{(h_w(x_1), \ldots, h_w(x_n)) : w \in \mathbb{R}^{d+1}, \|w\| = 1, |w \cdot x_i| \geq \gamma \text{ for all } i\}$$

What is $|\tilde{\mathcal{H}}_\gamma|$?

For any labeling in $\tilde{\mathcal{H}}_\gamma$, the Perceptron alg will converge within $1/\gamma^2$ updates. $\therefore$ There are $\leq n^{1/\gamma^2}$ choices for the Perceptron's final classifier $\Rightarrow$ $|\tilde{\mathcal{H}}_\gamma| \leq n^{1/\gamma^2}$