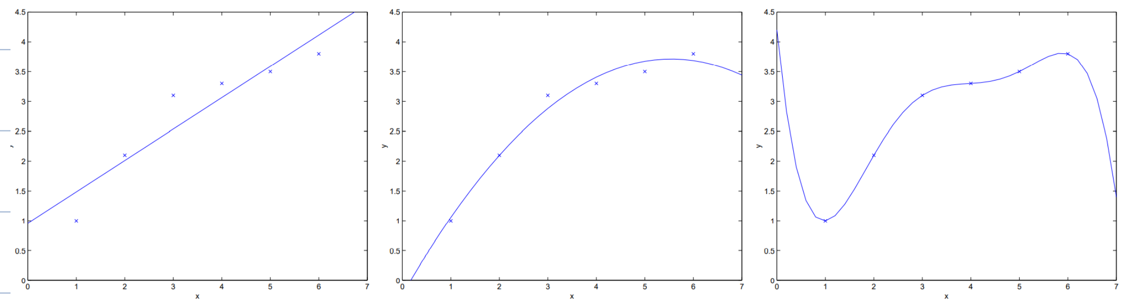
Lecture 4. Learning Theory.

- Bias / Variance trade-off:

  (1) generalization error.

  model → apply to other data.



underfitting overfitting.

  (2) bias: failed to capture structure. (simple)

  varience: don't reflect wider pattern. (complex)

- Preliminary:

  (1) Questions:

  ① formalize bias—variance trade-off.

  ② relationship between training error & generalization.

  ③ standard of learning alogrithm.

  (2) Two useful lemmas:

  ① Union bound.

  $P(A_1 \cup A_2 \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots P(A_n)$

  ② Hoeffding inequality:

$z_1, z_2 \cdots z_m \sim$ i.i.d. Bernoulli (Φ)

$\hat{\Phi} = \frac{1}{m} \sum_{i=1}^{m} z_i$ (mean)

$P(|\Phi - \hat{\Phi}| > \gamma) \leq 2\exp(-2\gamma^2 m)$

*: estimate of $\hat{\Phi}$: $m \to$ large converge exp.

*: another understanding: estimating biased coin.

(3) empirical error:

$\hat{\varepsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} 1(h(x_i) \neq y_i)$

generalization error:

$\varepsilon(h) = P_{(x,y)\sim D}(h(x) \neq y)$

*: probability of misclassifying new data

(4) PAC assumption:

① training / testing on same distribution

② independently drawn from training example.

hypothesis class H.

*: probably approximately correct.

∘ The case of finite H:

(1) $H = \{h_1, h_2 \cdots h_k\}$.

$X \to \{0, 1\}$.

(2) Strategy:

① $\hat{\varepsilon}(h)$ reliable for all h

② $\hat{\varepsilon}(h)$ upper bound.

(3) $Z_j = 1(h_i(x_j) \neq y_j)$

$\hat{\varepsilon}(h_i) = \frac{1}{m} \sum_{j=1}^{m} Z_j$

Z_j : drawn i.i.d.

for a fixed index i :

$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2\exp(-2\gamma^2 m)$

(4) for all $h \in H$:

A_i : $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$

$P(\exists h \in H : |\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma)$

$= P(A_1 \cup A_2 \cdots A_k)$

$\leq \sum_{i=1}^{k} P(A_i) \leq 2k\exp(-2\gamma^2 m)$

$\therefore$ $P(\forall h \in H ; |\varepsilon(h) - \hat{\varepsilon}(h)| < \gamma)$

$= 1 - 2k\exp(-2\gamma^2 m)$

*: uniform convergence.

(5) application: $m \cdot \gamma \cdot \delta$ (error)

bound one interms of two.

$1-\delta \rightarrow \delta = 2k\exp(-2\gamma^2 m)$

$\therefore$ $m = \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$ $O(\log k)$

*: sample complexity

(6) generalization:

$h^* = \arg\min_{h \in H} \varepsilon(h)$ (min generalization error)

$\varepsilon(\hat{h}) \leq \hat{\varepsilon}(\hat{h}) + \gamma$

$\leq \hat{\varepsilon}(h^*) + \gamma \leq \varepsilon(h^*) + 2\gamma$

We can get the theorem:

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

*: bias-variance trade-off.

larger hypothesis space:

$\varepsilon(h^*) \downarrow$ bias $\downarrow$

$2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}} \uparrow$ variance $\uparrow$

- The case of infinite H:

(1) parameterized by real numbers (d)

IEEE double-precision → 64 bits.

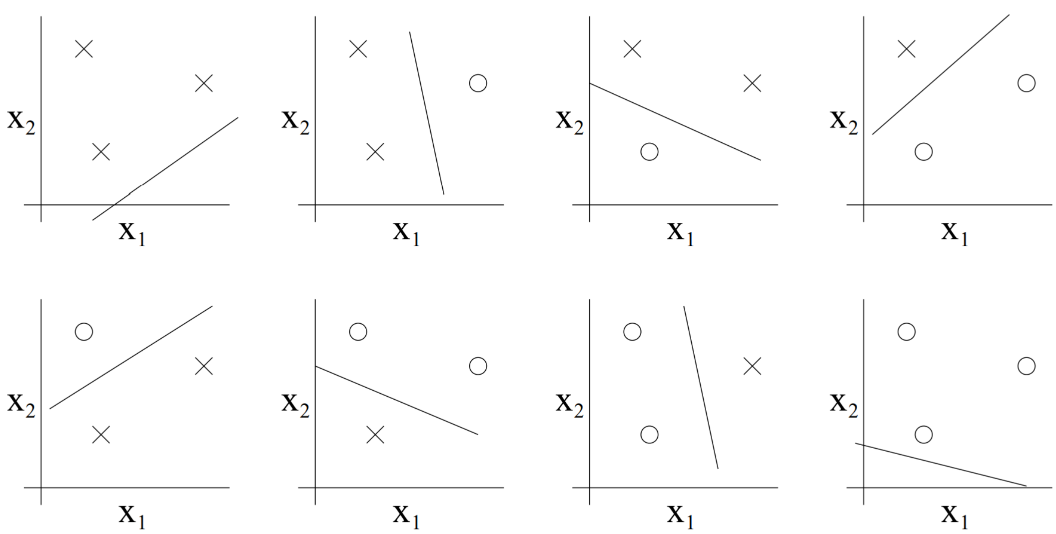
$k = 2^{64d} \Rightarrow O(\log k) = O(d) \Rightarrow$ linear

(2) finite: relies on parameterization of H.

(3) Definition: VC-dimension zero training error.

H shatters S. (realizing any label)

VC(H): largest size of S.



*: prove VC(H) = d:

① $\exists\, |S_0| = d$. H can shatter

② $\forall\, |S| > d$. H can't shatter.

(4) Theorem: $d = VC(H)$ probability $1-\delta$

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \le O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right.$$