Statistical learning framework.

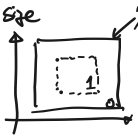
Learner:

Domain $\mathcal{X}$: set of objects you wish to label.

Label set $\mathcal{Y}$: the set of possible labels.

A prediction rule, $h: \mathcal{X} \to \mathcal{Y}$: used to label future examples.
also called predictor, hypothesis
classifier.

ex> Size



$\mathcal{X}$

$\mathcal{Y} = \{1: \text{taste good}, 0: \text{bad}\}$

Color Mango

Batch learning

Learner's input: Training data, $S = \underbrace{\{(x_i, y_i)\}_{i=1}^{m}}_{\text{set indexed by } i \in \{1,\dots,m\}} \in \underbrace{(\mathcal{X} \times \mathcal{Y})^m}_{\text{set of } m\text{-tuple of (object,label)}}$

Learner's output:

- A prediction rule, $h: \mathcal{X} \to \mathcal{Y}$

Goal: to get correct labels on future predictions

- suppose f is a correct classifier, then we want $f \approx h$
- One approach: define error

$$L_{D,f}(h) \underset{\text{definition}}{\triangleq} \mathbb{P}_{x \sim D}[h(x) \neq f(x)]$$

D is some unknown distribution.

↳ $\mathbb{P}_{x \sim D}(A)$ is probability that x falls in set $A \subseteq \mathcal{X}$

- Goal: find h s.t. $L_{D,f}(h)$ is small

We need some assumption on relation between D and f.

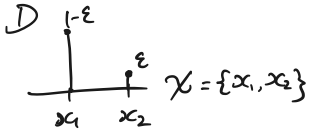
simple data generation

- Independent and identically distributed (i.i.d):
  each x_i is sampled independently from D
- Realizable:
  for every $i \in [m]$, $y_i = f(x_i)$
  ($[m] = \{1, \dots, m\}$)

Claim: this can only be **approximately** correct

i.e, we cannot find h s.t $L_{D,f}(h) = 0$ for all D, f.

Proof:



- $\mathbb{P}(\text{no } x_2 \text{ in } S_m) = (1-\varepsilon)^m \approx e^{-\varepsilon m} \approx 1.$ (↓ 0)
- letting $\varepsilon \ll \frac{1}{m}$, $m \uparrow$

⇒ you don't see any x_2, and cannot learn $f(x_2)$.

⇒ Relax goal to be approximate : $L_{D,f}(h) \leq \varepsilon$.

Claim: this can only be **Probably** correct.

i.e., no algorithm can guarantee $L_{D,f}(h) \leq \varepsilon$ $\underset{\text{almost surely}}{\text{w.p.1}}$

Recall that input is random

↳ there is a (small) chance that we see the same example again

⇒ Relax: we let algorithm fail with probability δ,
for some $\delta \in (0,1)$ chosen by the user

where the randomness is over the sampling of S.

Probably Approximately Correct (PAC) learning

- learner does not know D or f
- learner chooses accuracy parameter ε and confidence parameter δ
- learner gets training data S of size $m(\varepsilon, \delta)$
- learner outputs a hypothesis h s.t.

$$\mathbb{P}\left(L_{D,f}(h) \le \varepsilon\right) \ge 1-\delta$$

- then the learner is Probably (w.p. $1-\delta$) Approximately (up to ε) Correct.

Empirical Risk Minimization (ERM)

- Training error, empirical error, empirical risk

$$L_S(h) \triangleq \frac{|\{i \in [m] : h(x_i) \neq y_i|}{m}$$

$[m] = \{1, \dots, m\}$

$$\min_h L_S(h)$$

- what can go wrong? overfitting

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

↳ continuous D, $L_{D,f}(h_S) = \frac{1}{2}$. balanced y_i.

- Hypothesis class $\mathcal{H}$ → common measure of complexity $|\mathcal{H}|$

$h_S = ERM_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, L_S(h)$

Assumption 1. (Realizability of $\mathcal{H}$) $\exists h^* \in \mathcal{H}$ s.t. $L_{D,f}(h^*) = 0$

$$\Rightarrow L_S(h^*) = 0 \quad \text{w.p. } 1$$

Assumption 2. (iid samples) $\begin{cases} x_i \sim_{iid} D \\ y_i = f(x_i) \end{cases} \Rightarrow S \sim D^m$

Claim. For a finite hypothesis class $\mathcal{H}$, and some $\delta \in (0,1)$, and $\varepsilon > 0$

if $m \ge \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$, then for any f, and D, that are realizable,

$$L_{D,f}(h_S) \le \varepsilon \quad \text{w.p} \ge 1-\delta.$$

$$\to \mathbb{P}(L_{D,f}(h_S) > \varepsilon) \le \delta$$

- $ERM_{\mathcal{H}}$ is Probably Approximately Correct learning.

Define. "bad" hypotheses in $\mathcal{H}$.

$$\mathcal{H}_B \triangleq \{h \in \mathcal{H} : L_{D,f}(h) > \varepsilon\}$$

$$\mathbb{P}_S(L_{D,f}(h_S) > \varepsilon) \le \mathbb{P}_S\left(\bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}\right) \quad (*)$$

↑ realizable

$$\le \sum_{h \in \mathcal{H}_B} \mathbb{P}_S(L_S(h) = 0)$$

↑ Union bound

$$\le |\mathcal{H}_B| \max_{h \in \mathcal{H}_B} \prod_{i=1}^{m} \underbrace{\mathbb{P}(h(x_i) = y_i)}_{\le 1-\varepsilon}$$

$1-L_{D,f}(h)$

$$\leq |\mathcal{H}_B| \cdot (1-\varepsilon)^m$$

$\uparrow$ def $\mathcal{H}_B$

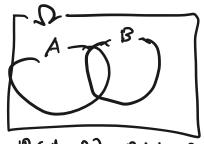
$$\leq |\mathcal{H}| \cdot e^{-\varepsilon m} \leq \delta$$

$\uparrow$ $\mathcal{H}_B \subseteq \mathcal{H}$, $\uparrow$ $m \geq \frac{1}{\varepsilon} \log \frac{|\mathcal{H}|}{\delta}$

* Union Bound

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

$$\mathbb{P}\left(\bigcup_{i \in [n]} A_i\right) \leq \sum_{i \in [n]} \mathbb{P}(A_i)$$



$\mathbb{P}(A \cup B) + \underbrace{\mathbb{P}(A \cap B)}_{\geq 0} = \mathbb{P}(A) + \mathbb{P}(B)$

(*) $$\mathbb{P}_S\left(L_{D,f}(h_S) > \varepsilon\right) \leq \mathbb{P}_S\left(\bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}\right)$$

$\uparrow$ ERM

$\exists h$ s.t.

$$\{S : L_{D,f}(h_S) > \varepsilon\} \subseteq \{S : L_{D,f}(h) > \varepsilon, L_S(h) = 0\}$$

$$= \bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}.$$

Define. Sample complexity $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$, $(\varepsilon, \delta) \to m_{\mathcal{H}}(\varepsilon, \delta)$

minimum number of samples needed to guarantee PAC learning.

Corollary. any finite hypothesis class $\mathcal{H}$ is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil$$

Complexity of $\mathcal{H}$ is measured by $|\mathcal{H}|$. → VC-dimension

End lecture 2.

Relaxing the realizability assumption

Data generating distribution $\mathcal{D}$ is over X, Y

True error: $L_{\mathcal{D}}(h) \triangleq \mathbb{P}_{\mathcal{D}}(h(x) \neq y)$

Empirical error: $L_S(h) \triangleq \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$

Goal: to find h that minimizes $L_{\mathcal{D}}(h)$

Bayes Optimal Predictor has lower error than any other predictor [HW1 Prob 2]

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}_{\mathcal{D}}(y=1 \mid x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

but requires knowledge of $\mathcal{D}$.

Agnostic PAC Learnability.

A class of hypotheses, $\mathcal{H}$, is Agnostic PAC Learnable if $\exists m_{\mathcal{H}}$ and a learning algorithm s.t. for $\varepsilon, \delta, \mathcal{D}$

running Algo on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ samples returns h s.t. w.p. $\geq 1-\delta$

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$

Relaxing Binary Classification

- multi-class classification $Y = [K]$, $l_h(x_i, y_i) = \mathbb{1}(h(x_i) \neq y_i)$
- Regression, $Y \subseteq \mathbb{R}$, loss $l_h(x_i, y_i) = (h(x_i) - y_i)^2$

Risk function $L_{\mathcal{D}}(h) \triangleq \mathbb{E}_{(x,y) \sim \mathcal{D}}[l(h, (x,y))]$

Empirical Risk function $L_S(h) \triangleq \frac{1}{m} \sum_{i=1}^{m} l(h, (x_i, y_i))$