Review: Concept / Function Learning

(X, Y) $\quad X$ - features

$Y = f(X)$ - deterministic label / or response

$f : \mathcal{X} \to \{0,1\}$ - concept learning

$f : \mathcal{X} \to [0,1]$ - function learning

Training data:

$(X_1, f^*(X_1)), (X_2, f^*(X_2)), \ldots, (X_n, f^*(X_n))$

where • $X_1, \ldots, X_n \overset{iid}{\sim} P \in \mathcal{P}$

• f^* is an element of a given class $\mathcal{F}$

Learning algo: data $\longrightarrow \hat{f}_n \in \mathcal{F}$

$L_P(f, f^*) := \mathbb{E}_P[|f(X) - f^*(X)|^2]$

$L_P(\hat{f}_n, f^*)$ - random variable, depends on data

Goal: make $L_P(\hat{f}_n) \sim 0$ (with high prob.)

Learning algo is Probably Approximately Correct (PAC) if:

$$\sup_{P \in \mathcal{P}} \sup_{f \in \mathcal{F}} P_f^n\left(L_P(\hat{f}_n, f) > \varepsilon\right) \longrightarrow 0 \quad \text{as } n \to \infty$$

for any $\varepsilon > 0$.

Concept learning: $\mathcal{C}$ – family of sets $C \subseteq \mathcal{X}$

$$L_p(C, C^*) = P(C \Delta C^*)$$

$$\mathcal{C} \longleftrightarrow \mathcal{F} = \{1_C : C \in \mathcal{C}\}$$

$$1_C(x) = \begin{cases} 1, & \text{if } x \in C \\ 0, & \text{if } x \notin C \end{cases}$$

$$P(C \Delta C^*) = \mathbb{E}_P\left[|1_C(X) - 1_{C^*}(X)|^2\right]$$

Limitations:

1) deterministic relationship b/w Y and X
   $Y = f(X)$ – what about $Y = f(X) + \text{noise}$
2) the class $\mathcal{F}$ is known – what about mismatch between our assumption on $\mathcal{F}$ and ground truth?

→ realizable setting

Agnostic (model-free) formulation [D. Haussler]:

- $\mathcal{X}, \mathcal{Y}$ – feature and label space
  $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$
- $\mathcal{U}$ – action (decision) space
- $\mathcal{F}$ – class of functions $f : \mathcal{X} \to \mathcal{U}$
  [hypothesis space]
- $\ell : \mathcal{Y} \times \mathcal{U} \to [0, 1]$ – loss fctn
- $\mathcal{P}$: class of prob. dist. on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

Ideal case: $Z = (X, Y) \sim P$, known

$$L_P(f) := \mathbb{E}_P[\ell(Y, f(X))]$$

$$L_P^*(\mathcal{F}) := \inf_{f \in \mathcal{F}} L_P(f)$$

Learning: "Nature" selects $P \in \mathcal{P}$

$$\underbrace{(X_1, Y_1)}_{Z_1}, \dots, \underbrace{(X_n, Y_n)}_{Z_n} \overset{iid}{\sim} P$$

learning algo $A_n : Z^n \longrightarrow \hat{f}_n \in \mathcal{F}$

$$L_P(\hat{f}_n) = \int_{X \times Y} P(dx, dy)\, \ell(y, \hat{f}_n(x))$$

$$0 \le L_P^*(\mathcal{F}) \le L_P(\hat{f}_n) \le 1$$

Goal: construct learning algo s.t.

$$L_P(\hat{f}_n) - L_P^*(\mathcal{F}) \longrightarrow 0 \quad \text{as } n \to \infty$$

regardless of P.

PAC: for a fixed $\varepsilon > 0$,

$$\lim_{n \to \infty} \underbrace{\sup_{P \in \mathcal{P}} P^n\left(L_P(\hat{f}_n) \ge L_P^*(\mathcal{F}) + \varepsilon \right)}_{:= r_A(n, \varepsilon)} = 0$$

A - learning algo.

Ex (function learning)

$\mathcal{X}$ - feature space

$\mathcal{Y} = \mathcal{U} = [0,1]$

$\mathcal{F}$ - target functions $\quad [Y = f(X) \text{ for some } f \in \mathcal{F}]$

$l(y,u) = |y-u|^2$

$\mathcal{P}$ - all distributions of (X,Y) where $\mathcal{L}(X)$ comes from some P_0 (of dist. on $\mathcal{X}$) and $Y = f(X)$ for some $f \in \mathcal{F}$

$$P_f((X,Y) \in A) = \int_A P(dx)\, 1_{\{y = f(x)\}}$$

$$\Rightarrow \mathcal{P} = \{P_f : \; P \in P_0, \; f \in \mathcal{F}\}$$

$$L_{P_f}(g) = E_{P_f} |Y - g(X)|^2$$

$$= \int_{\mathcal{X}\times\mathcal{Y}} P(dx)\, 1_{\{y=f(x)\}} |y - g(x)|^2$$

$$= \int_{\mathcal{X}} P(dx)\, |f(x) - g(x)|^2 = L_P(f,g)$$

$$L_{P_f}(f) = 0 \quad \Rightarrow \quad L^*_{P_f}(\mathcal{F}) = 0$$

— PAC definitions coincide.

Agnostic case can cover more general situations:

$Y = f(X) + W \qquad EW = 0$, indep. of X

f may not be in $\mathcal{F}$

When is PAC learning possible in model-free case?

$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P$

$\ell_f : \mathcal{Z} \to [0,1]$: $\ell_f(z) := \ell_f((x,y)) = \ell(y, f(x))$

$P(\ell_f) := \int P(dx, dy)\, \ell_f((x,y))$ - expected loss of f

$P_n(\ell_f) := \frac{1}{n} \sum_{i=1}^{n} \ell_f((X_i, Y_i)) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$ - empirical loss of f on data

- can be computed for every $f \in \mathcal{F}$

$$\mathbb{P}\left\{ |P_n(\ell_f) - P(\ell_f)| \geq \varepsilon \right\} \leq 2e^{-2n\varepsilon^2}$$

(Hoeffding)

- in particular, if $f^* \in \mathcal{F}$ achieves $L_P^*(\mathcal{F})$, then

$$\mathbb{P}\left\{ |P_n(\ell_{f^*}) - L_P^*(\mathcal{F})| \geq \varepsilon \right\} \leq 2e^{-2n\varepsilon^2}.$$

Candidate learning algo: Empirical Risk Minimization (ERM)

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} P_n(\ell_f) = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

Claim: (to be proved in next lecture)

if ℓ (loss), $\mathcal{P}$, and $\mathcal{F}$ (hypothesis class) are such that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} P^n\left(\sup_{f\in\mathcal{F}} \left| P_n(\ell_f) - P(\ell_f) \right| \geq \varepsilon \right) = 0$$

for any $\varepsilon > 0$, then ERM is a PAC learning algo.

Exercise: prove that this holds if $|\mathcal{F}| < \infty$.