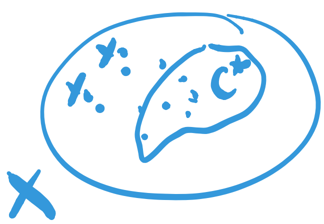
Chapter 5 continued – formulation of learning problem, ERM, and uniform Convergence of empirical distributions (UCERM)

We saw $\underline{(X, P_X, \mathcal{C})}$ - realizable concept learning



Observe $Z^n = ((x_1, 1_{(x_1 \in C^*)}), \ldots, (x_n, 1_{(x_n \in C^*)}))$

Learning algorithm $A = (A_n)_{n \geq 1}$

$A_n(Z^n) = \hat{C}$

$P(\hat{C} \triangle C^*)$ = error probability of using $\hat{C}$ on a fresh sample.

A is PAC if for any $\varepsilon, \delta > 0$ there exist $n(\varepsilon, \delta)$ so for any $P \in P_X$, and $C^* \in \mathcal{C}$, if $n \geq n(\varepsilon, \delta)$ then with prob. at least $1-\delta$, $P(\hat{C}_n \triangle C^*) \leq \varepsilon$, if $n \geq n(\varepsilon, \delta)$.

Sect. 5.1.2 $(X, P_X, \mathcal{F})$ with mean square error.

"realizable function learning"

$\mathcal{F}$ is a set of functions $f: X \to [0,1]$.

$$Z^n = ((X_1, f^*(X_1)), (X_2, f^*(X_2)), \dots, (X_n, f^*(X_n)))$$

where $f^* \in \mathcal{F}$, f^* is the true function.

$$A_n(Z^n) = \hat{f} \in \mathcal{F} \qquad \ell(f(x), f^*(x)) = (f(x) - f^*(x))^2.$$

$X_1, \dots, X_n$ iid P_X

$$L_P(\underbrace{\hat{f}_n}_{= A(Z^n)}, f^*) = E_P[|\hat{f}_n(X) - f^*(X)|^2]$$

A is PAC if for any $\varepsilon, \delta > 0$ there exists $n(\varepsilon, \delta)$ so for any $P \in P_X$, any $f^* \in \mathcal{F}$ if $\hat{f}_n = A(Z^n)$ then with prob. at least $1 - \delta$

$$L_P(\hat{f}_n, f^*) \le \varepsilon \quad \text{if } n \ge n(\varepsilon, \delta)$$

5.3 Function learning agnostic (model free) case

$(X, Y, U, \mathcal{P}, \mathcal{F}, l)$

- X = space of feature vectors
- Y : space of labels
- U = space of output labels (usually $U = Y$)
- $\mathcal{P}$ = set of probability dist^{n} on $Z = X \times Y$.
- $\mathcal{F}$ = set of functions $f : X \to U$
- l : loss function $l(y, u)$ = loss for output u when label is y.

Training samples $Z^n = (Z_1, \dots, Z_n)$

$Z_i = (X_i, Y_i)$ independent, dist^{n} $P \in \mathcal{P}$

$$A_n(Z^n) = \hat{f}_n \in \mathcal{F}.$$

Performance of $\hat{f}_n$ on a fresh sample is

$$L_P(\hat{f}_n) = E_P[l(Y, \hat{f}_n(X))]$$

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

$\mathcal{A}$ is PAC if for any $\varepsilon, \delta > 0$ there exists $n(\varepsilon, \delta)$ so that for any $P \in \mathcal{P}$ if $\hat{f}_n \in \mathcal{A}(Z^n)$ and $n \geq n(\varepsilon, \delta)$ then with probability at least $1-\delta$:

$$L_P(\hat{f}_n) \leq \underbrace{L_P^*(\mathcal{F})}_{= \min_{f \in \mathcal{F}} L_P(f)} + \varepsilon$$

Section 5.3.2 Learning to classify with <u>noisy labels</u>.

Start with realizable model $(X, \mathcal{P}_X, \mathcal{C})$

Let $\eta \in (0, \frac{1}{2})$

$\mathcal{P}$: distributions on $X \times Y$ of the following form:

- select $P \in \mathcal{P}_X$
- select $C^* \in \mathcal{C}$

W is $\mathrm{Be}(\eta)$
$P(W=1) = \eta$
$P(W=0) = 1-\eta$

X has distⁿ P, $Y = 1_{\{X \in C^*\}} \oplus W$

$A_n(Z^n) = \hat{C}_n \in \mathcal{C}$.

How good is a given $C \in \mathcal{C}$

(for $P \in \mathcal{P}_X$, $C^* \in \mathcal{C}$, η)

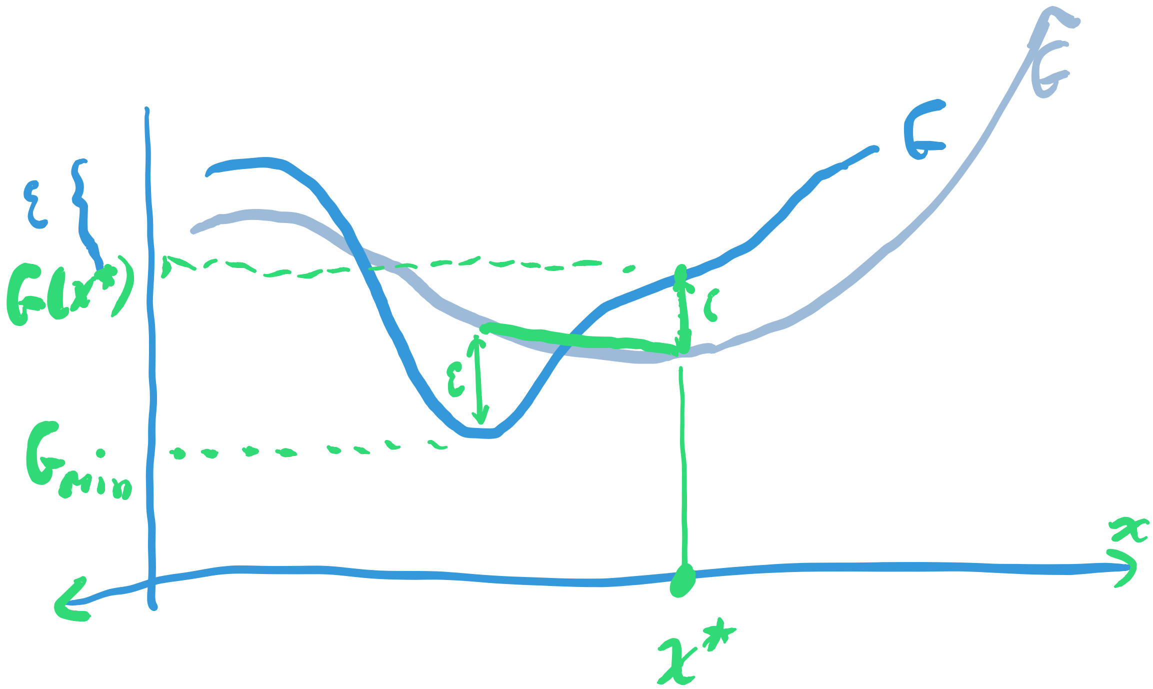
$$L_{P_X, \mathcal{C}^*}(C) = P_X\left(\underline{1_{\{X \in C\}} \neq 1_{\{X \in C^*\}}} \oplus W\right)$$

$$= P_X\left(W \neq 1_{\{X \in C^* \triangle C\}}\right)$$

$$= (1-\eta) P_X(C^* \triangle C)$$

$$\eta\left(1 - P_X(C^* \triangle C)\right)$$

$$= \eta + \underbrace{(1-2\eta)}_{>0} P_X(C^* \triangle C)$$



ε
G(x*)
G_min
ε
ε
G
G̃
x
x*