Consider PAC-learning (probably approximate correct) – small generalization error with high probability. We see data $D_n = \{(X_i, Y_i)\}_{i=1}^n$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$. Assume that there is some joint distribution $P(X,Y)$. We want to learn $f : \mathcal{X} \to \mathcal{Y}$ such that. In this course we start with "realizable" setting: $Y = f(x)$, deterministic. Then we will consider "agnostic", "model-free" setting.

14.1 Realizable set

Consider set $\mathcal{X}$ and $\mathcal{P}$ – a set of distributions on $\mathcal{X}$. We have data D_n drawn from some $P \in \mathcal{P}$

14.1.1 Concept learning

- We assume that there is some class $\mathcal{C}$ of subsets of $\mathcal{X}$. $\mathcal{C}$ – concept classes.
- There exist a target $C^* \in \mathcal{C}$.
- We observe $Y_i = I(X_i \in C^*)$ – whether X_i belongs to C^*.
- We want to determine C^*.
- We want some procedure (algorithm) to do this: $\mathcal{A} = \{\, A_n \,\}_{n=1}^{\infty}$. $\mathcal{A}$ can depend on $\mathcal{P}$ but not on P.

Similar to statistics, but different language. We will denote $Z_i = (X_i, Y_i) \in \mathcal{Z}$.

Given D_n we want to produce $\widehat{C}_n = A_n(D_n) = A_n(Z_1, \ldots, Z_n) \in \mathcal{C}$. We evaluate $\widehat{C}_n$ by examination, whether $X_{n+1} \in \widehat{C}_n$ for new X_{n+1}. Two types of errors:

1. $X_{n+1} \in C^*$ but not in $\widehat{C}_n$.
2. $X_{n+1} \in \widehat{C}_n$ but not in C^*.

In other words, $X_{n+1} \in C^* \Delta \widehat{C}_n$ – symmetric difference of sets. We wan to estimate $P[X_{n+1} \in C^* \Delta \widehat{C}_n]$ (essentially this is a conditional random variable, i.e. $P(\ldots | X_1, \ldots, X_n)$).

Definition 14.1.

- $L_P(C, C^*) = P[X_{n+1} \in C \Delta C^*]$. *If* $L_P(\widehat{C}_n, C^*) \to 0$ *then* $\mathcal{A}$ *is good.*
- $r_{\mathcal{A}}(n, \epsilon, P) = \sup\limits_{C \in \mathcal{C}} P(Z^n \in \mathcal{Z}^n \mid L_P(A_n(z^n), C) \geq \epsilon)$.
- $R_{\mathcal{A}}(n, \epsilon, \mathcal{P}) = \sup\limits_{P \in \mathbb{P}} r_{\mathcal{A}}(n, \epsilon, P)$.

If P is fixed then $r_{\mathcal{A}}$ is a worst-case size of bad samples. We consider supremum over all C since we don't know C^*.

Definition 14.2.

- *$\mathcal{A}$ is PAC if* $\lim\limits_{n\to\infty} R_{\mathcal{A}}(n,\epsilon,\mathbb{P}) = 0$.
- *$\mathcal{C}$ is PAC-learnable if $\exists\mathcal{A}$ – is PAC.*

Goals

- Determine conditions for C to be PAC-learnable.
- Find PAC $\mathcal{A}$, determine sample complexity as function of C.

Equivalent definition of PAC:

$$\forall\epsilon,\delta>0\ \exists n_0(\epsilon,\delta)\ \forall n>n_0(\epsilon,\delta), C\in\mathcal{C}, P\in\mathcal{P}\quad P(D_n|L_P(A_n(D_n)\in C)\geq\epsilon)\leq\delta.$$

$\mathcal{C} = \{\, X\in[0;1] : \sin(\Theta X)>0, \Theta\in\mathbb{R}\,\}$

14.2 Function learning

- The same setup. $\mathcal{F}$ – class of functions.
- $Y\in f^*(X)\in[0;1]$.
- $\widehat{f}_n = A_n(D_n)$.
- $L_P(f,f^*) = \mathbb{E}\big[|f(X)-f^*(X)|_2^2\big] = \int\limits_{\mathcal{X}} |f(x)-f^*(x)|^2 P(dx) = ||f-f^*||^2_{L_2(P)}$.

For example, $L_P(C,C^*) = ||I_C - I_{C^*}||^2_{L_2(P)}$.

Example 14.3. *Take $\mathcal{X}=[0;1]^2$, $\mathbb{P}$ – all distributions on $\mathcal{X}$, C – set of all possible rectangles.*

$$C = \{\,[a_1;b_1]\times[a_2;b_2]\ \mid\ 0\leq a_1\leq b_1\leq 1,\ 0\leq a_2\leq b_2\leq 1\,\}\,.$$

To prove learnability need $\mathcal{A}$.

- $Z_i = (X_i, I(X_i\in C^*))$.
- $\widehat{C}_n = A_n(D_n)$ *– smallest $C\in\mathcal{C}$.*

Theorem 14.4.

$$R_{\mathcal{A}}(n,\epsilon,\mathcal{P}) \leq 4\left(1-\frac{\epsilon}{4}\right)^n$$

Proof. Since we select the smallest rectangle $\widehat{C}^n\in C^* \implies \widehat{C}^n\Delta C^* = C^*\setminus\widehat{C}^n$. If $P(C^*)<\epsilon$ then $P[C^*\setminus\widehat{C}^n]\leq P[C^*]\leq\epsilon$.

Consider the case when $P(C^*)\geq\epsilon$. Then $C^*\setminus\widehat{C}^n = V_1\cup V_2\cup H_1\cup H_2$, where V_1 is a strip between the left boundaries of the rectangles, with height bounded by largest rectangle, namely $V_1=[a_1^*,\widehat{a}_1]\times[a_2^*,b_2^*]$. V_2, H_1 and H_2 are similar strips corresponding to other boundaries (all derivations also will be the same). If P of each of these strips is less then $\frac{\epsilon}{4}$ then by union bound $P[V_1\cup V_2\cup H_1\cup H_2]<\epsilon$.

Consider probability that $V_1\geq\frac{\epsilon}{4}$. The probability that one sample is not in V_1 is at most $1-\frac{\epsilon}{4}$. Therefore the probability that all n samples are not in V_1 is at most $\left(1-\frac{\epsilon}{4}\right)^n$. By union bound probability that there exist a strip with no sample with it and $P\geq\frac{\epsilon}{4}$ is at most $4\left(1-\frac{\epsilon}{4}\right)^n$. ∎

Corollary:

$$n_0(\epsilon, \delta) \geq \frac{4 \log \frac{4}{\delta}}{\epsilon}.$$