Notes of ML Talk on 10/11/2017.

1. Notations: (Supervised Setting). (A Formal Learning Model.)

$\boxed{S} = (x_1, y_1), \cdots, (x_m, y_m)$ Training set.

$\boxed{\mathcal{X}}$: domain set.

$\boxed{\mathcal{Y}}$: label set.

$\boxed{\mathcal{H}}$: hypothesis class, a fixed subset of $\{$all functions $h : \mathcal{X} \to \mathcal{Y}\}$.

$\boxed{\mathcal{D}}$: a dist. on $\mathcal{X}$ (to be relieved later).

Assumption:

① $x_1, \cdots, x_m \in S$ is generated iid w/ $\mathcal{D}$.

② $y_1, \cdots, y_m$ are labelled by a certain $f : \mathcal{X} \to \mathcal{Y}$. ← labelling function

i.e. $y_i = f(x_i)$, $\forall i$. (to be relieved.)

$\boxed{\ell} : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$, a loss function, measuring how "preferred" a hypothesis is.

For $h \in \mathcal{H}$,

$$\boxed{L_S(h)} := \frac{1}{m} \sum_{i=1}^{m} \ell(h, (x_i, y_i)) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, (x_i, f(x_i))),$$ called training error, empirical error, or empirical risk.

$$\boxed{L_{\mathcal{D}}(h)} := \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\ell(h, (x, f(x)))],$$ called generalization error, risk, or true error.

2.

- An algorithm A is a process of generating an output hypothesis $A(S) \in \mathcal{H}$ given the input $S = (x_1, y_1), \cdots, (x_m, y_m)$.

- ERM (= Empirical Risk Minimization) is the algorithm s.t.

$$A(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h).$$

Def Given $\mathcal{D}$ over $\mathcal{X}$. $\mathcal{D}$ is called realizable by $\mathcal{H}$ if $\exists\, h^* \in \mathcal{H}$ s.t. $L_{\mathcal{D}}(h^*) = 0$.

Def (PAC = probably approximately correct).

Given $\mathcal{H}$ and $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$.

The hypothesis class $\mathcal{H}$ is called PAC learnable if

$\exists\, m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and algorithm A s.t.

$\forall$ dist. $\mathcal{D}$ on $\mathcal{X}$, $\varepsilon, \delta \in (0,1)$ and labelling function $f : \mathcal{X} \to \mathcal{Y}$,

if the realizable assumption holds, then

running A on $S \sim \mathcal{D}^m$ w/ $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, we have

$L_{\mathcal{D}}(A(S)) \leq \varepsilon$ w/ prob. $\geq 1-\delta$. (over $S \sim \mathcal{D}^m$).

Rmk.:

① ε : accuracy parameter, δ : confidence parameter
(approximately correct) (probably).

②
Minimum among such $m_{\mathcal{H}}$ is called the sample complexity.

Issues:

① Realizability assumption is NOT practical (i.e. $f \notin \mathcal{H}$ often happens).

② Labelling function is not realistic. (i.e. it's possible to have $x_i = x_j$ but $y_i \neq y_i$).

Sol'n to ②:

Instead of a dist. $\mathcal{D}$ on $\mathcal{X}$ and labelled by f, we can directly consider a dist. $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. Denote $\boxed{\mathcal{Z}} := \mathcal{X} \times \mathcal{Y}$ for simplicity.

To also solve issue ①, we modify PAC learnability as follows:

Def Given $\mathcal{H}$ and $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$.

$\mathcal{H}$ is called agnostic PAC learnable if $\exists\, m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and algorithm A

s.t. $\forall$ dist. $\mathcal{D}$ on $\mathcal{Z}$, $\varepsilon, \delta \in (0,1)$,

if we run A on $S \sim \mathcal{D}^m$ w/ $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, we have

$$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon, \text{ w/ prob.} \geq 1-\delta \text{ (over } S \sim \mathcal{D}^m),$$

where $\boxed{L_{\mathcal{D}}(h)} := \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(h, z)]$, $\forall\, h \in \mathcal{H}$.

3. VC dimension and Fundamental Thm of Statistical Learning.

Q: Which classes $\mathcal{H}$ are (agnostic) PAC learnable?

Def

$\mathcal{H}$: a hypothesis class of fun. from $\mathcal{X}$ to $\{0,1\}$.

$C = \{c_1, \cdots, c_m\} \subseteq \mathcal{X}$.

① $\boxed{\mathcal{H}_C} := \{ f|_C : C \to \{0,1\} \mid f \in \mathcal{H} \}$, called the restriction of $\mathcal{H}$ to C.

② $\mathcal{H}$ is said to shatter C if $\mathcal{H}_C = \{\text{all fun. from } C \text{ to } \{0,1\}\}$.

i.e. $|\mathcal{H}_C| = 2^{|C|}$.

③ The VC dimension of C, denoted $\boxed{\mathrm{VCdim}(C)}$, is the maximal size of $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

Rmk:

VC = Vapnik-Chervonenkis.

Thm (F.T. of S.L. - Qualitative).

$\mathcal{H}$: a hypothesis class of fun. from $\mathcal{X}$ to $\{0,1\}$, w/ 0-1 loss.

Then T.F.A.E.

(1) $\mathcal{H}$: agnostic PAC learnable.

(2) $\mathcal{H}$: PAC learnable.

(3) ERM is a successful agnostic PAC learner for $\mathcal{H}$.

(4) 〃 〃 〃 〃 PAC learner for $\mathcal{H}$.

(5) $\mathrm{VCdim}(\mathcal{H}) < \infty$.

Thm (F.T. of S.L. - Quantitative)

$\mathcal{H}$: hypo. class of fun. from $\mathcal{X}$ to $\{0,1\}$ w/ 0-1 loss, $\mathrm{VCdim}(\mathcal{H}) = d < \infty$.

Then $\exists$ abs. const. C_1 and C_2 s.t.

(1) $\mathcal{H}$: agnostic PAC learnable w/ $C_1 \cdot \dfrac{d + \log(1/\delta)}{\varepsilon^2} \le m_{\mathcal{H}}(\varepsilon, \delta) \le C_2 \cdot \dfrac{d + \log(1/\delta)}{\varepsilon^2}$.

(2) $\mathcal{H}$: PAC learnable w/ $C_1 \cdot \dfrac{d + \log(1/\delta)}{\varepsilon} \le m_{\mathcal{H}}(\varepsilon, \delta) \le C_2 \cdot \dfrac{d + \log(1/\delta)}{\varepsilon}$.