30 August 2021

CS4932(F) Comp. Learning Theory – The PAC Learning Framework (Ch2)

1. Overview / Setting

(1) In this chapter, we consider the problem of learning a classifier from examples, and study theoretical tools for analysing such a problem.

(2) A classification problem is described as follows:

① $\mathcal{X}$ ... space for the input. Usually, $\mathcal{X} = \mathbb{R}^I$ for some $I \in \mathbb{N}$.

② $\mathcal{Y}$ ... space for the outputs. We assume $\mathcal{Y} = \{0,1\}$.

③ unknown concept $c : \mathcal{X} \to \mathcal{Y}$ from a concept class $\mathcal{C}$.

④ unknown distribution D on $\mathcal{X}$.

⑤ Sample $S = (x_1, x_2, \cdots, x_m)$ and $(y_1, \cdots, y_m)$ s.t.
(or training set) $x_1, \cdots, x_m$ are independently drawn from D and

$$y_i = c(x_i) \text{ for all } i \in [m] = \{1, 2, 3, \cdots, m\}.$$

⑥ Hypothesis set $\mathcal{H} \subseteq [\mathcal{X} \to \mathcal{Y}]$

⑦ Given ①, ②, ⑤, ⑥, find a good $h \in \mathcal{H}$ that behaves like c for the inputs drawn from D.

(3) To analyse such a problem and its candidate solution rigorously, we need mathematical formalisation of "good" or "behave like", which expresses a desired property of $h \in \mathcal{H}$. We want this formalisation also talks about computational resources need to find such a good h. By computational resources, we mean time and space (the usual suspect in CS), as well as the number of examples in S needed. The latter is called sample complexity.

and is new in these learning problems.

(4) The PAC learning framework is such a formalism. It also naturally leads to the idea of generalization bound, which appears repeatedly in the course.

(5) The purpose of this chapter is to learn key concepts and results in the PAC framework, and to prove generalization bounds for some particular classification problems.

2. The PAC learning model.

(1) Reminder: $\mathcal{X}, \mathcal{Y}, \mathcal{C}, \mathcal{H}, D, S, c \in \mathcal{C}$

$\mathcal{X}$: input space; $\mathcal{Y}$: "labels"; $\mathcal{C}$: concept class; $\mathcal{H}$: hypo. set; D: unknown dist. (sample (random)); c: unknown concept; $c \in \mathcal{C}$: target.

(2) The accuracy of a hypo. h is formalized by the next two concepts, generalization error (or risk) and empirical error. (or empirical risk).

Def 2.1 Given: $h \in \mathcal{H}$, $c \in \mathcal{C}$, D.

The generalization error or risk of h is defined by

$$R(h) = \mathbb{P}_{x \sim D}[h(x) \neq c(x)] = \mathbb{E}_{x \sim D}[\mathbb{1}\{h(x) \neq c(x)\}]$$

error in an ideal setup depends on c and D.

indicator fn returning 1 if the condition holds, 0 otherwise.

Def 2.2 Given: $h \in \mathcal{H}$, $c \in \mathcal{C}$, D, $S = (x_1, \dots, x_m) \sim D^m$ (m iid draws from D).

The empirical error or empirical risk of h is defined by

$$\widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq c(x_i)\} = \mathbb{E}_{x \sim \widehat{D}}[\mathbb{1}\{h(x) \neq c(x)\}]$$

error that can be measured in practice. depends on c, D

uniform distribution over $(x_1, \dots, x_m)$.

Lemma. $\mathbb{E}_{S\sim D^m}[\widehat{R}_S(h)] = R(h)$. That is, $\widehat{R}_S(h)$ is an unbiased estimate of $R(h)$.

Proof.
$$\mathbb{E}_{S\sim D^m}[\widehat{R}_S(h)] = \mathbb{E}_{S\sim D^m}\left[\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h(x_i)\neq c(x_i)\}}\right]$$
$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{x_i\sim D}[\mathbb{1}_{\{h(x_i)\neq c(x_i)\}}] = \frac{1}{m}\sum_{i=1}^{m} R(h) = R(h) \quad \square.$$

Let n be the size of the representation of each $x \in \mathcal{X}$.

Def 2.3. [PAC learning] $\mathcal{X}, \mathcal{Y}$: given.

A concept class $\mathcal{C}$ is PAC-learnable if $\exists$ an algo. $\mathcal{A}$ and a polynomial $\text{poly}(\cdot,\cdot,\cdot,\cdot)$ s.t.

$$\forall c \in \mathcal{C}\ \forall D \in \Pr(\mathcal{X})\quad \forall \varepsilon > 0\ \forall \delta > 0\quad \forall m \in \mathbb{N},$$

(D is a prob. distribution on $\mathcal{X}$)

if $m \geq \text{poly}(1/\varepsilon, 1/\delta, n, \text{size}(c))$, then

$$\mathbb{P}_{S\sim D^m}[R(h_S) \leq \varepsilon] \geq 1-\delta.$$

- $\text{size}(c)$: the size of the representation of c.
- h_S: the result of running the alg. $\mathcal{A}$ on $\{(x_i, c(x_i)) \mid x_i \in S\}$.

$\mathcal{C}$ is efficiently PAC-learnable if $\mathcal{A}$ runs in $\text{poly}'(1/\varepsilon, 1/\delta, n, \text{size}(c))$ for some polynomial poly'.

In that case, $\mathcal{A}$ is called a PAC-learning algo.

- *Note 1 : Polynomial sample complexity is the key requirement of PAC learnability.
- *Note 2 : Efficient PAC learnability additionally requires polynomial time (& so space) complexity.
- *Note 3 : No assumption on the distribution D.

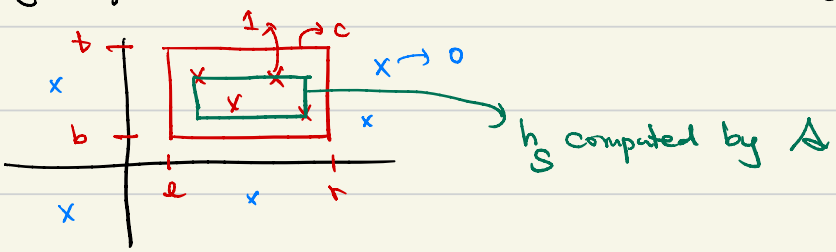
* Note 4: The definition is applicable usually when $\exists\ h$ s.t. $\hat{R}(h) = 0$. $\in \mathcal{H}$.

* Note 5: size(c) will not play any significant role. We will ignore it (in the course)

(3) Example : Learning axis-aligned rectangles.

$\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0,1\}$.

$\mathcal{H} = C = \{ (a,b) \mapsto \mathbb{1}\{(a,b) \in [l,r] \times [b,t]\} \mid l \le r \wedge b \le t \}$.



Q Prove that C is PAC-learnable.

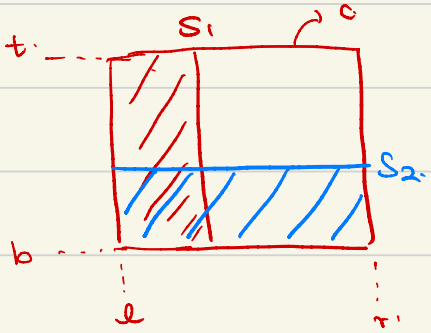
A $A((x_1,y_1), \cdots, (x_m,y_m))$ = the smallest rectangle containing all x_i's with $y_i = 1$.

$\text{poly}\left(\frac{1}{\varepsilon}, \frac{1}{\delta}\right) = \frac{4}{\varepsilon}\log\frac{4}{\delta}$.

Fix $D \in \mathcal{P}(\mathcal{X})$, $c \in C$. , $\varepsilon, \delta > 0$, $m \ge \frac{4}{\varepsilon}\log\frac{4}{\delta}$.

We will show that:

$$\mathbb{P}\left[R(h_S) > \varepsilon \right] \le \delta.$$



Pick S_1, S_2, S_3, S_4 s.t.

$$\mathbb{P}_{x \sim D}\left[x \in [l, s_1] \times [b,t] \right] \ge \varepsilon/4$$

$$\mathbb{P}_{x \sim D}\left[x \in [l, s_1) \times [b,t] \right] \le \varepsilon/4.$$

S_2, S_3, S_4 satisfy similar properties for the other three sides of the rectangle.

Let: $r_1 = [l, s_1] \times [b,t]$

$r_2 = [l,r] \times [b, s_2]$ and r_3, r_4 be similar rectangles.

Then, $R(h_S) > \varepsilon$ implies that $\bigcup_{i=1}^{4} \left(h_S^{-1}(1) \cap r_i = \phi \right)$.

Thus, $\underset{S\sim D^m}{\mathbb{P}}[R(h_S) > \varepsilon] \le \underset{S\sim D^m}{\mathbb{P}}[\bigvee_{i=1}^{4} h_S^{-1}(1) \cap r_i = \phi]$

$\le \sum_{i=1}^{4} \mathbb{P}[h_S^{-1}(1) \cap r_i = \phi] \le \sum_{i=1}^{4} (1-\frac{\varepsilon}{4})^m \le 4\exp(-\frac{m\varepsilon}{4})$

(∵ union-bound) (∵ $1-x \le e^{-x}$)

$\le 4\exp(-\frac{\varepsilon}{4}\cdot\frac{4}{\varepsilon}\log\frac{4}{\delta}) = \delta$.

□

3. PAC learning for finite consistent hypothesis sets.

C1)

Assumption : $\mathcal{H}$ is finite and $C \subseteq \mathcal{H}$.

Thm 2.5 A ... algo. s.t. for any target concept $c \in \mathcal{H}$ and any $S = (x_1, \dots, x_m) \in \mathcal{X}^m$, it returns a consistent h_S:

$\hat{R}_S(h_S) = 0$.

Then, $\forall D \in \Pr(\mathcal{X})\ \forall \varepsilon, \delta > 0\ \forall m > 0$,

if $m \ge \frac{1}{\varepsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta})$, then $\underset{S\sim D^m}{\mathbb{P}}[R(h_S) \le \varepsilon] \ge 1-\delta$.

□

* Note 1 : Equivalent statement that expresses generalisation bound:

$\forall D\ \forall \delta > 0\ \forall m > 0$

$\underset{S\sim D^m}{\mathbb{P}}[R(h_S) \le \frac{1}{m}(\log|\mathcal{H}| + \log\frac{1}{\delta})] \ge 1-\delta$.

* Note 2: Thm 2.5 implies PAC-learnability (if $\log|\mathcal{H}|$ is polynomial in n (i.e., the size of input)), although it doesn't imply efficient PAC-learnability.

algo.-independent. uniform bound.

Proof We will show that:

$\underset{S\sim D^m}{\mathbb{P}}[\exists h \in \mathcal{H} \text{ s.t. } \hat{R}_S(h) = 0 \wedge R(h) > \varepsilon] \le \delta$.

Since A returns a consistent hypo. h_S, the above inequality implies the thm..

$$\mathbb{P}_{S\sim D^m}\left[\exists h \in \mathcal{H} \text{ s.t. } \hat{R}_S(h)=0 \wedge R(h)>\varepsilon\right]$$

$$\leq \sum_{\substack{h\in\mathcal{H}\\ R(h)>\varepsilon}} \mathbb{P}_{S\sim D^m}\left[\hat{R}_S(h)=0\right] \quad \text{by union bound}$$

$$= \sum_{\substack{h\in\mathcal{H}\\ R(h)>\varepsilon}} \prod_{i=1}^{m} \mathbb{P}_{x\sim D}\left[h(x)=c(x)\right] \leq \sum_{h\in\mathcal{H}} (1-\varepsilon)^m = |\mathcal{H}|\,(1-\varepsilon)^m \leq |\mathcal{H}| e^{-\varepsilon m}$$

uniform bound. — exponential decay due to repeated trials.

$$= |\mathcal{H}|\, e^{-\varepsilon \times \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)} = \delta. \qquad \square$$

↑ assumption on m.

* Note 1. : $\frac{1}{m}\log|\mathcal{H}|$ comes from the use of uniform bound, and $\frac{1}{m}\log\frac{1}{\delta}$ comes from the estimation error for a single h.

* Note 2: $O(\frac{1}{m})$ convergence rate in the generalization bound. This is a good news. In a more difficult inconsistent case, we will get $O(\frac{1}{\sqrt{m}})$ convergence rate.

* Note 3: $\log|\mathcal{H}|$ corresponds to the # of bits needed to encode each hypothesis in $\mathcal{H}$.

(2) Examples / Applications.

① Conjunctions of boolean literals.

- n boolean variables $v_1, \cdots, v_n$
- each example specifies the values of those variables. $\mathcal{X} = 2^n$
  $x = (v_1=1, v_2=1, v_3=0, \cdots, v_n=1)$ (denoted by $(110\cdots1)$)
- $\mathcal{C} = \mathcal{H} = \{ v_1 \wedge \bar{v}_2 \;,\; v_1 \wedge v_3 \wedge \bar{v}_4 \;, \cdots \}$ = the set of conjunctions of literals
- $\log|\mathcal{H}| = \log 3^n = n \log 3$.

- consistent-hypothesis-producing A.

$A((x_1,y_1),\cdots,(x_m,y_m)) = \wedge\{v_i \mid \forall j.\ y_j=1 \Rightarrow x_{j,i}=1\}$

$\wedge\ \wedge\{\overline{v_i} \mid \forall j.\ y_j=1 \Rightarrow x_{j,i}=0\}$.

- PAC-learnable by Thm 2.5. Sample complexity bound is

$m \geq \frac{1}{\varepsilon}(n\log 3 + \log\frac{1}{\delta})$.

($\delta=0.02$, $\varepsilon=0.1$, $n=10$, $m \geq 149$)

- Efficiently PAC-learnable because A runs in poly time.

② Universal concept class

- $\mathcal{U}_n = [\mathcal{X} \to \mathcal{Y}]$, $\mathcal{X}=2^n$, $\mathcal{Y}=\{0,1\}$.

(i.e., the set of all subsets of $\mathcal{X}$).

$|\mathcal{H}| \geq |\mathcal{U}_n|$.

- Sample complexity bound from Thm 2.5:

$$m \geq \frac{1}{\varepsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta}) \geq \frac{1}{\varepsilon}(\log|\mathcal{U}_n| + \log\frac{1}{\delta})$$
$$= \frac{1}{\varepsilon}(\log 2^{2^n} + \log\frac{1}{\delta})$$
$$= \frac{1}{\varepsilon}(2^n\log 2 + \log\frac{1}{\delta})$$

exponential in n.

- In fact, not PAC-learnable.

③ k-term DNF formulas.

- $\mathcal{H} = \mathcal{C} = \{\bigvee_{j\in\underbrace{[u]}_{\{1,2,\cdots,u\}}} c_j \mid$ c_j is a conjunction of literals and $u \leq k\}$.

$$|\mathcal{H}| = |\mathcal{C}| = 3^{n\times 0} + 3^{n\times 1} + 3^{n\times 2} + \cdots + 3^{n\times k} \leq (k+1)3^{n\times k}$$

- Sample complexity bound from Thm 2.5:

$$\frac{1}{\varepsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta}) \leq \frac{1}{\varepsilon}(n\times k\times\log 3 + \log(k+1) + \log\frac{1}{\delta}) \cdots \text{poly}.$$

$\therefore$ PAC-learnable.

- But not efficiently PAC-learnable if RP ≠ NP.

④ k-CNF

$\mathcal{C} = \mathcal{H} = \{ \bigwedge_{j=1}^{u} c_j \mid c_i \neq c_j \text{ for } i \neq j, \ c_i \text{ is a disjunction of at most } k \text{ literals} \}$.

variable v or its negation $\bar{v}$

- Reduction to the case ① by replacing all disjunctions of at most k literals, $(l_1 \vee l_2 \vee \cdots \vee l_u)$, with new variables $v_{l_1, l_2, \cdots, l_u}$. Then, the # of these new variables $\leq (2n+1)^k$.

∴ The sample complexity bound from Thm 2.4 is

$$m \geq \frac{1}{\varepsilon}\left((2n+1)^k \log 3 + \log \frac{1}{\delta} \right).$$

∴ PAC-learnable. In fact, efficiently PAC-learnable.

4. Guarantee for finite inconsistent-hypothesis sets.

(1) We now consider the case that the hypo. set $\mathcal{H}$ is finite but $\mathcal{C} \notin \mathcal{H}$, more accurately, $\exists\, c \in \mathcal{C}$ and $D \in P(\mathcal{X})$ s.t. $R(h) > 0$ for any $h \in \mathcal{H}$ when R is defined with c and D

(2) In such a case, PAC-learning is not possible usually, because sometime we cannot reduce $R(h)$ enough, no matter how many examples we use in $S = (x_1, \cdots, x_m)$

(3) We instead try to find a high-prob. uniform bound on generalization error which has the form:

$$\mathbb{P}[\, \forall h \in \mathcal{H}.\ R(h) \leq \hat{R}_S(h) + f(m, \delta, |\mathcal{H}|) \,] \geq 1 - \delta.$$

Such a bound lets us talk about the quality of a learnt classifier measured by $R(h_S)$ using the estimated $\hat{R}_S(h_S)$.

Thm 2.13. $\mathcal{H}$ is finite.

$\Rightarrow \forall c \in \mathcal{C}\ \forall D \in \mathcal{P}(\mathcal{X})\ \forall \varepsilon > 0\ \forall \delta > 0.$

$$\mathbb{P}_{S \sim D^m}\left[\forall h \in \mathcal{H}.\ R(h) \le \hat{R}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta.$$

uniform bound.

$O\left(\sqrt{\frac{1}{m}}\right)$ instead of $O\left(\frac{1}{m}\right)$

Equivalent statement based on sample complexity bound.

$$\forall c\ \forall D\ \forall \varepsilon, \delta > 0\ \forall m \ge \frac{1}{\varepsilon^2}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right),$$

$$\mathbb{P}\left[\forall h \in \mathcal{H}.\ R(h) \le \hat{R}_S(h) + \varepsilon\right] \ge 1-\delta.$$

(+) Proof of thm 2.13:

① Hoeffding's inequality. ... summing ind. random vars leads to concentration around the mean.

Thm D.2 $X_1, \dots, X_m$ ... independent random variables s.t.

$$X_i \in [a_i, b_i].\ \text{for all } i \in [m].$$

$$S_m = \sum_{i=1}^{m} X_i \qquad \varepsilon > 0.$$

$\Rightarrow$

$$\mathbb{P}\left[S_m - \mathbb{E}[S_m] \ge \varepsilon\right] \le \exp\left(-2\varepsilon^2 \Big/ \sum_{i=1}^{m}(b_i - a_i)^2\right)$$

and $$\mathbb{P}\left[S_m - \mathbb{E}[S_m] \le -\varepsilon\right] \le \exp\left(-2\varepsilon^2 \Big/ \sum_{i=1}^{m}(b_i - a_i)^2\right)$$

Cor 2.11 Let $h \in \mathcal{H}$. and $\delta > 0$. Fix c, D, m. Then,

$$\mathbb{P}_{S \sim D^m}\left[R(h) \le \hat{R}_S(h) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta.$$

Proof. It suffices to show that

$$\mathbb{P}_{S\sim D^m}\left[R(h) \geq \hat{R}_S(h) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \right] \leq \delta.$$

"

$$= \mathbb{P}\left[\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq c(x_i)\} - \mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq c(x_i)\}\right] \geq \sqrt{\frac{m}{2}\log\frac{1}{\delta}} \right].$$

Hoeffding $\geq$

$$\leq \exp\left(- \frac{2\times\left(\sqrt{\frac{m}{2}\log\frac{1}{\delta}}\right)^2}{\sum_{i=1}^{m}(1-0)^2} \right) = \exp\left(\frac{-2\times\frac{m}{2}\log\frac{1}{\delta}}{m} \right)$$

$$= \delta. \qquad \square$$

Equivalent version of "

for all $\varepsilon > 0$,

$$\mathbb{P}_{S\sim D^m}[R(h) \geq \hat{R}_S(h) + \varepsilon] \leq \exp(-2m\varepsilon^2).$$

② Proof of thm 2.13. Let $\varepsilon = \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$.

We will show that

$$\mathbb{P}[\exists h \in \mathcal{H} \text{ s.t. } R(h) \geq \hat{R}_S(h) + \varepsilon] \leq \delta.$$

"

$$\leq \sum_{h\in\mathcal{H}} \mathbb{P}[R(h) \geq \hat{R}_S(h) + \varepsilon].$$

$$\leq \sum_{h\in\mathcal{H}} \exp(-2m\varepsilon^2) = |\mathcal{H}| \exp\left(-2m \times \left(\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}\right)\right)$$

$$= \delta. \qquad \square$$

(5) Notes on thm 2.13.

① $$R(h) \leq \hat{R}_S(h) + O\left(\sqrt{\frac{\log_2|\mathcal{H}|}{m}}\right).$$

Contrast it with the generalization bound of thm 2.5

$$R(h) \leq O\left(\frac{\log_2|\mathcal{H}|}{m}\right).$$

Thus. worse. bound.

② Simpler $\mathcal{H}$ and larger m tighten the bound.

related to Occam's razor principle. ... the simplest explanation is best.

5. Stochastic scenarios

(1) A natural generalisation of the setting of a distribution D on the inputs in $\mathcal{X}$ and an unknown concept c is to assume simply a distribution D' on the input-output pairs. Note that the previous setting is a special case of this stochastic general setting where $D'(x,y)$ is defined to be $D(x) \times \mathbb{1}\{y = c(x)\}$. (More formally, D' is the push-forward measure of D along the function $x \longmapsto (x, c(x))$.)

(2) In this new stochastic setting, we cannot avoid a certain level of error, and this is not compatible with the PAC framework that usually works when we can minimize error as much as we can. So, we have to modify PAC.

(3) New setup and agnostic PAC learning.

$$D \in Pr(\mathcal{X} \times \mathcal{Y})$$

$$S = ((x_1, y_1), \dots, (x_m, y_m)) \sim D^m$$

$$R(h) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y] = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}\{h(x) \neq y\}].$$

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}.$$

Def 2.14 $\mathcal{H}$ -- hypo. set. A ... a learning algorithm.

A is an agnostic PAC-learning algo. if $\exists$ a polynomial $poly(\cdot,\cdot,\cdot)$ s.t. $\forall D \in P(\mathcal{X}\times\mathcal{Y})$ $\forall \varepsilon, \delta > 0$ $\forall m \in \mathbb{N}$.

if $m \geq poly\left(\frac{1}{\varepsilon}, \frac{1}{\delta}, n\right)$, (n: size of each input.) then

$$\mathbb{P}_{S\sim D^m}\left[R(h_S) - \min_{h\in\mathcal{H}} R(h) \leq \varepsilon \right] \geq 1-\delta.$$

($\mathcal{H}$ is finite and: error that cannot be avoided.)

(4) Suppose that $\mathcal{H} = [\mathcal{X} \to \mathcal{Y}]$, the set of all functions.

What is the solution of the following opt. pb?

$$\min_{h\in\mathcal{H}} R(h) = \min_{h\in\mathcal{H}} \mathbb{E}[\mathbb{1}\{h(x)\neq y\}].$$

- Optimal solution h_{Bayes} (called Bayes hypothesis, Bayes classifier)

$$h_{Bayes}(x) = \underset{y\in\{0,1\}}{\operatorname{argmax}} \mathbb{P}[y \mid x].$$

- Optimal value $R^* = R^*(h_{Bayes}) = \mathbb{E}[noise(x)]$ Bayes error.

where $noise(x) = \min\{\mathbb{P}[1|x], \mathbb{P}[0|x]\}$.

called noise at x.

- For each $D \in P(\mathcal{X}\times\mathcal{Y})$, the Bayes error describes the difficulty of the learning problem wrt. D.