Learning

- Bait shyness of rats: Rats eat poisoned food, get sick, and afterwards avoid the same food.
- Pidgeon superstition: Pidgeons in a cage get food at regular intervals of time, afterwards repeat the same actions to get more food.

Setting

- Supervised: We learn from labeled training data with a target value and try to derive a rule how to come up with the target value
- passive learner: The learner has no influence on the environment and its actions don't influence it
- Teacher is passive
- Batch learning: We get some labeled training data set and some time to learn from it and need to predict new labels afterwards.

How to formalize this?

- Domain set $\mathcal{X}$: The set of all possible examples, often $\mathcal{X} \subseteq \mathbb{R}^n$
- Label set $\mathcal{Y}$: The set of all possible target values, often $\mathcal{Y} = \{+1, -1\}$

Additional assumptions:

- We have a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is unknown but fixed.

We have a measure of success:

$$\underbrace{L_{\mathcal{D}}(h)}_{\text{learner}} := \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

Learning Task:

Input: A training sequence $S = ((x_1, y_1), \dots, (x_m, y_m))$ drawn independently and identically distributed from $\mathcal{D}$

Output: a prediction rule $A(S)$ that can, given $x \in \mathcal{X}$ produce a value $A(S)(x) \in \mathcal{Y}$

s.t. $L_{\mathcal{D}}(A(S))$ is minimized.

First learning rule: Empirical Risk Minimization (ERM)

Given a training sequence S, return a prediction rule that minimizes the empirical error (on the training sequence)

$$L_S(A(S)) := \frac{|\{i \in [m] : A(S)(x_i) \neq y_i\}|}{m}$$

„How to fix the problem of memorizing / overfitting"

We introduce a restricted search space $\mathcal{H}$, called hypothesis class.

Then ERM becomes the following:

Given: S and some finite representation of $\mathcal{H}$

return: $A(S) := ERM_{\mathcal{H}}(S) \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$

Intuition

More restricted $\mathcal{H}$ "$\Rightarrow$" a better safeguard against overfitting

"$\Rightarrow$" stronger inductive bias

PAC & APAC

Which hypothesis classes are learnable?

"we can come up with an algorithm that for any distribution $\mathcal{D}$ and training sequence S minimizes $L_{\mathcal{D}}(A(S))$ with some guarantees."

DEF A hypothesis class $\mathcal{H}$ (for some $\mathcal{X}, \mathcal{Y}$) is APAC learnable if there exists

- a function $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$
- a learning algorithm A with:

$\forall\, \varepsilon, \delta \in (0,1)$, $\forall$ distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$

~~$\forall$ labelling functions $f : \mathcal{X} \longrightarrow \mathcal{Y}$~~

~~with realizable: $\exists h \in \mathcal{H}$ s.t. $L_{\mathcal{D}}(h) = 0$~~

there holds: when running A on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d examples from drawn from $\mathcal{D}$ labeled by f. the algorithm returns $A(S) \in \mathcal{H}$ s.t. with prob. $1-\delta$ $L_{\mathcal{D}}(A(S)) \leq \varepsilon + \min_{h' \in H} L_{\mathcal{D}}(h')$

Corr: Every realizable finite hypothesis class $\mathcal{H}$ is APAC learnable using ERM with sample complexity

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{2\log(|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil$$

No Free Lunch

Let $\mathcal{H} \subseteq \{f : f : \mathcal{X} \to \{+1, -1\}\}$

Let A be any algorithm for binary classification w.r.t. to 0-1 loss over $\mathcal{X}$. Let $m \leq \frac{|\mathcal{X}|}{2}$ (finite)

then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ s.t.

1) $\exists f : \mathcal{X} \to \mathcal{Y}$ with $L_{\mathcal{D}}(f) = 0$
2) With prob. at least $\frac{1}{7}$ over $S \sim \mathcal{D}^m$ we have $L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$

DEF: Shattering

A hyp. class $\mathcal{H}$ shatters some finite set $C \subseteq \mathcal{X}$ if $\mathcal{H}|_C$ is the set of all functions from C to $\{1, -1\}$

DEF VC-Dimension

The VC-dimension of of a hyp. class $\mathcal{H}$ is the cardinality of the largest set $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.