## CATEGORIES:

→ Active Learning

Ask / Seek questions or labels during learning.

Goal: Reduce the number of queries.

## HOW TO MODEL SUPERVISED LEARNING:

INPUT: Some domain $X$

$X$

eg: images, documents, browsing profiles, credit card history.

LABELS: cats / dogs / foxes

$L$

$\{0, 1\}$

$\mathbb{R}$

DATASET: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$x_1 \to X$, $y_1 \to L$

GOAL: Given a new $x \in X$, want to predict its label!

ASSUMPTION 1: There is an underlying ground truth distribution on training and test examples.

i.e. Both training, test similar cases. if we train on dog, cat and test on polar bear, makes no sense.

Distribution, D on X: $x_1, \dots x_n$ are i.i.d samples from D.

$x$ is also from D.

ASSUMPTION 2: Labels cannot be arbitrary.

There is a class of functions $H : X \to L$.

ASSUMPTION 3: Prediction cannot be accurate all the time.

eg: We say model is 60% accurate.

$\rightarrow \Pr_{x \sim D}[y \neq f(x)] \leq \varepsilon \rightarrow$ "error"

$f(x)$ ↳ my prediction

we need to relax "$\neq$"

depending on application.

**Assumption 4:** Cannot expect to succeed always!

Sometimes, the ~~dataset~~ training data when chosen comes out very bad → like all dogs! Now there is no way it can predict cats.

→ "Probability of failure" $\delta$.

This is the chance of utter failure

→ Probability of being able to build a model with 60% accuracy.

## PAC MODEL:

(PROBABLY APPROXIMATELY CORRECT)

→ Otherwise approximately correct.

→ Some chances of utter failure

An algorithm $(\varepsilon, \delta)$ - PAC learns a hypothesis class $H$ with sample complexity $n(\varepsilon, \delta, H)$ if

INPUT: $(x_1, f^*(x_1)), \ldots, (x_n, f^*(x_n))$ where $x_i \leftarrow D$, $f^* \in H$.

OUTPUT: Some predictor $h$

$$\boxed{\begin{array}{l} \forall \text{ with probability } 1-\delta \text{ over } x_1, \ldots x_n \\ \Pr_{x \leftarrow D}[h(x) = f^*(x)] \geq 1-\varepsilon \end{array}}$$

PAC is a very strong requirement! Very difficult to find $h$ even with 51% accuracy. So use, may be, the structure in the distribution to simplify.

→ OFTEN BECOMES INTRACTABLE.

## LEARNING AS OPTIMIZATION:

→ We have an underlying $D$ on $X$. Labels $L$.

→ Dataset $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

→ Loss function $\ell : L \times L \to \mathbb{R}$

true label $y$ ↙ ↘ predicted label $h(x)$

**GOAL:** Find a hypothesis $h$ such that

$$\boxed{\sum_{i=1}^{n} \ell(h(x_i), y_i) \text{ is small}}$$

**PROBLEM:** Just memorize all $(x_i, y_i)$ and this will satisfy.

**PARAMETERIZED HYPOTHESIS CLASS:** $H$

Restrict, say memory available, this will prevents memorization. This restricted space (subspace) is $H$, parameterized by $\theta$.

$$\boxed{\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x_i), y_i)}$$

is the predictor as specified by the "parameters" $\theta$.

THIS IS "Empirical Risk Minimization" (ERM).

We are using loss on training data.

IDEAL: $\mathbb{E}_{x \sim D} \left[ \ell [h_\theta(x), y] \right]$

(testing data)

EXAMPLE 1 (ERM): $X \equiv \mathbb{R}^d$ : $d \equiv \mathbb{R}$

d-dimensional vector.

Parametric family : Linear predictors (predictors are linear functions)

$(\Theta \equiv \mathbb{R}^d)$

$$h_\theta(x) = \sum_{i=1}^{d} \theta_i x_i = \underbrace{\langle \theta, x \rangle}_{\text{inner product}}$$

$$= \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d.$$

(More generally, $h_{\theta, t}(x) = \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d + t$).

## LEAST SQUARES REGRESSION:

$$h_\theta(x) = \langle \theta, x \rangle$$

$$\ell : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$$

$$\ell(a,b) = (a-b)^2$$

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(x_i), y_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} ( \langle \theta, x_i \rangle - y_i )^2$$

$$\boxed{\text{ERM : Find } \theta \text{ to minimize } L(\theta)}$$

EXAMPLE 2: $\ell(a,b) = |a-b|$ "L1 - REGRESSION"

ERM $\min L(\theta) = \frac{1}{n} \sum_{i=1}^{n} | \langle \theta, x_i \rangle - y_i |$

EXAMPLE 3 : What if labels are discrete?

$L \in \{0,1\}$

PARAMETRIC FAMILY $h_\theta(x) = \begin{cases} 1 & \text{if } \langle \theta, x \rangle > 0 \\ 0 & \text{if } \langle \theta, x \rangle < 0 \end{cases}$

$$\ell(a,b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{else} \end{cases}$$

LINEAR THRESHOLD FUNCTIONS

(HALFSPACES)

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\text{sign}(\langle \theta, x_i \rangle) \neq y_i)$$

0 if sign matches $y_i$

1 else.