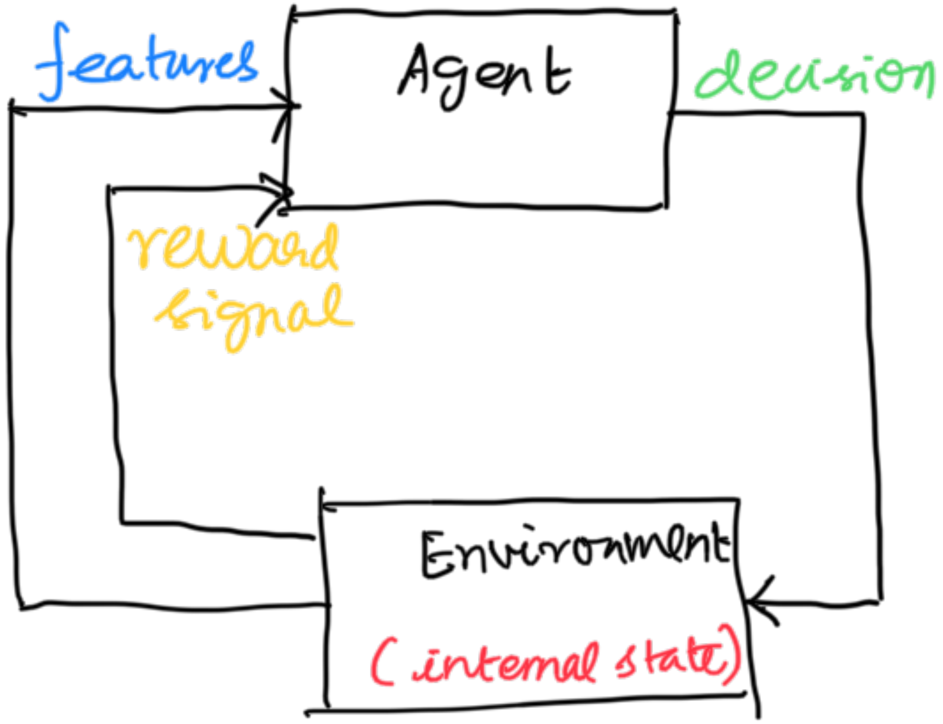
Statistical Decision Theory (Birds eye view)



max (over decision) total reward (Decision Theoretic Goal)

Statistical → probabilistic / frequency of occurrence

Statistical → probabilities / frequency of occurances
and co-occurances

Probability Model (Motivating example)

Say I would like to model an environment / world which has 360 days and 20% days it rains.

Environment

(internal state)

1 = rain

0 = no rain

y_t : internal state (rain / no-rain)

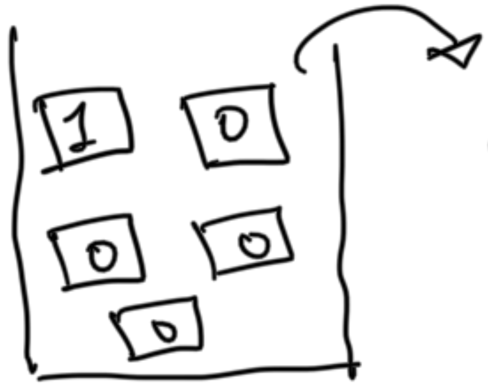
Deterministic 1) model

$$y_t = 1, \underbrace{1, \ldots, 1}_{72 \text{ days}}, 0, 0, \ldots, \underset{\substack{\uparrow \\ \text{day } 360}}{0}$$

Deterministic Model 2)

$$y_t = 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, \ldots$$

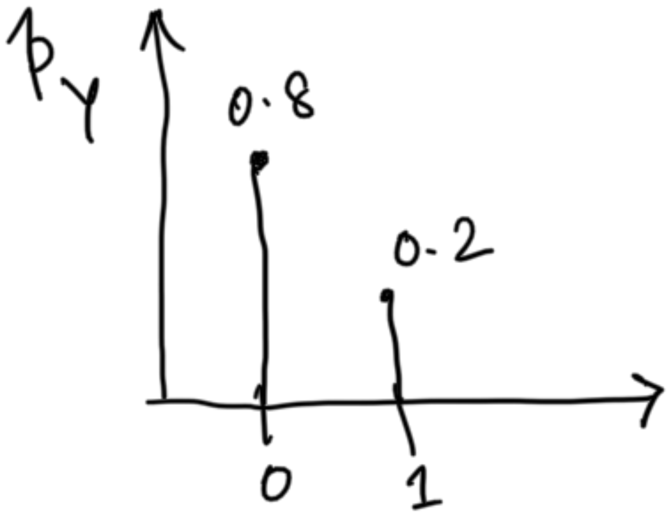
Stochastic (Probabilistic) Model



at time t

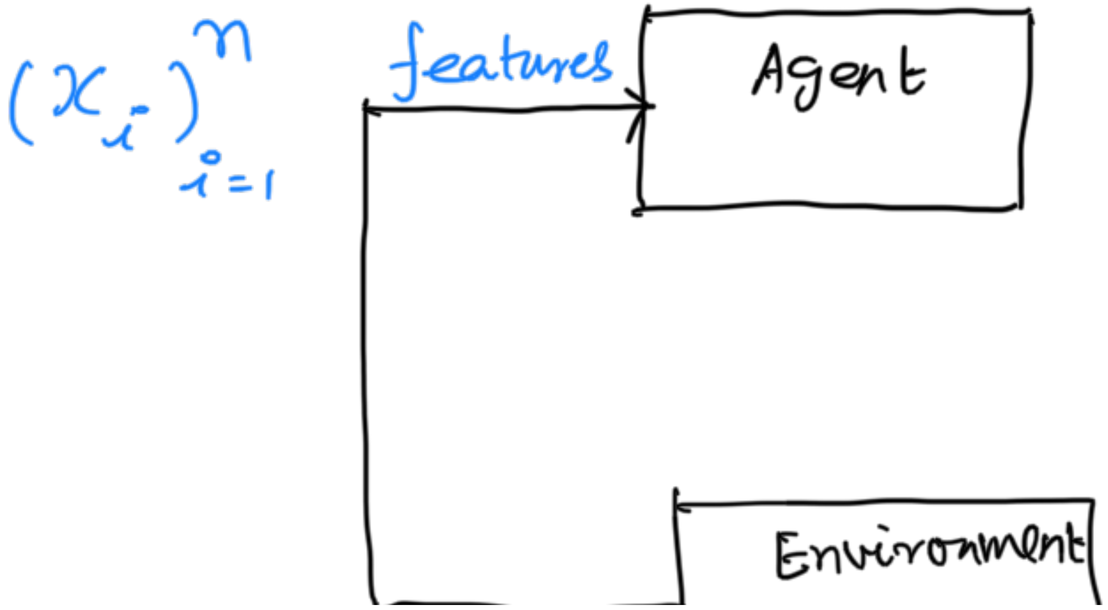
- * shuffle well
- * pick a chit and read out

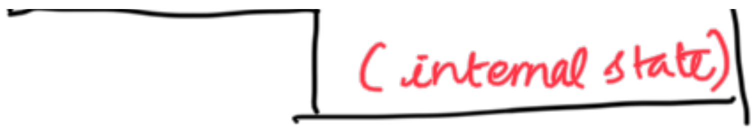
$$y_t \overset{i.i.d}{\sim} p_Y(\cdot) \,, \quad \text{where}$$



Descriptive Task

(Data Representation)





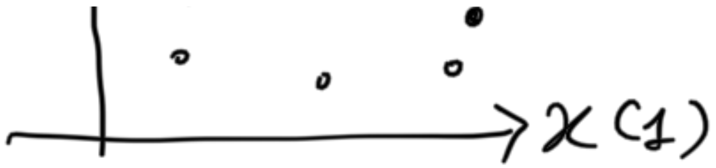
Agent does not make any decision

- Features : $(x_i)_{i=1}^{n} \in \mathbb{R}^d$

Eg 1) clustering : Goal is to say group articles in a newspaper by topic

n: articles , $x_i \in \mathbb{R}^d$ (d = size of vocabulary)

$x_i(j)$ = # word j occured in article i.

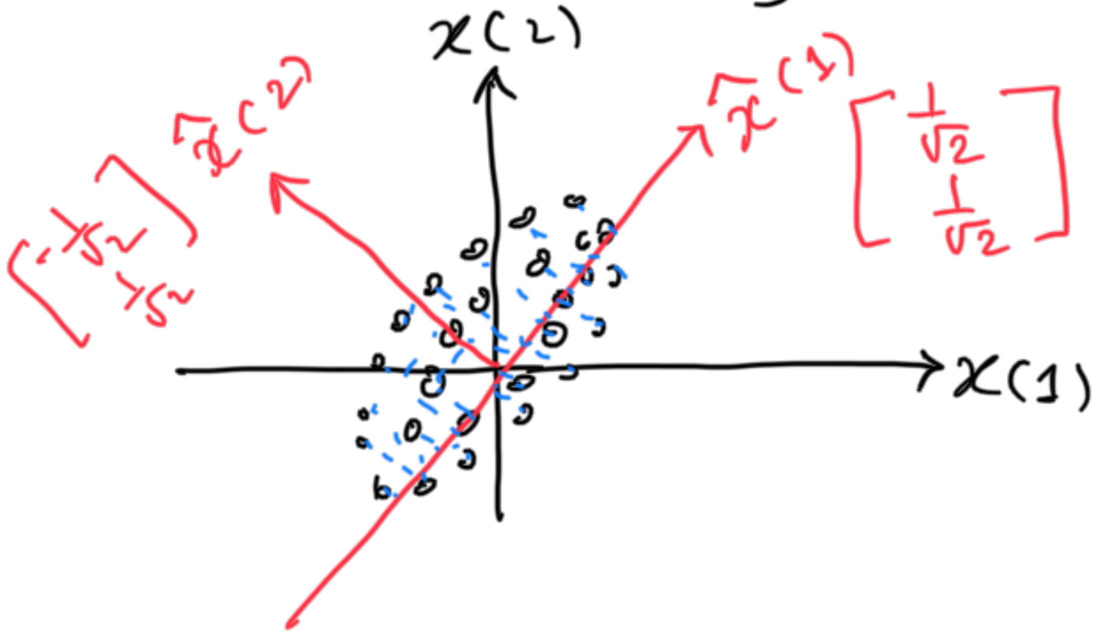


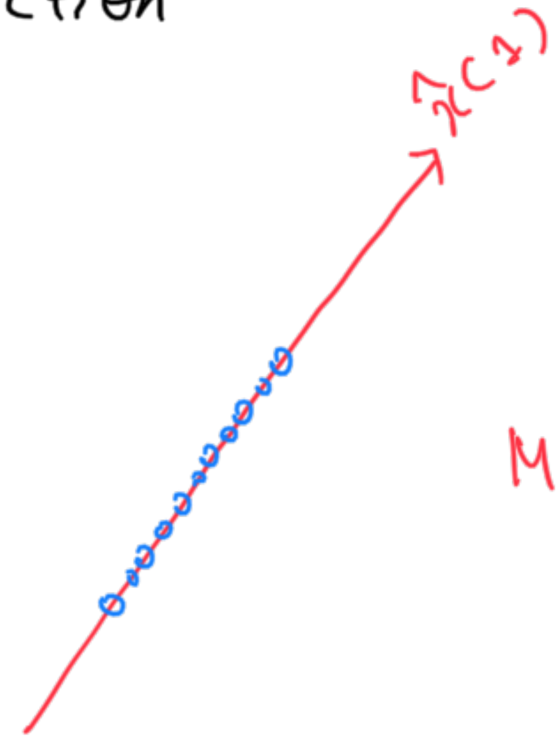
clustering $\mathcal{C} : \{1, \dots, n\} \longrightarrow \{1, \dots, k\}$

total clusters

$x_i \in$ cluster $\mathcal{C}(i)$

Eg (2) Dimensionality Reduction





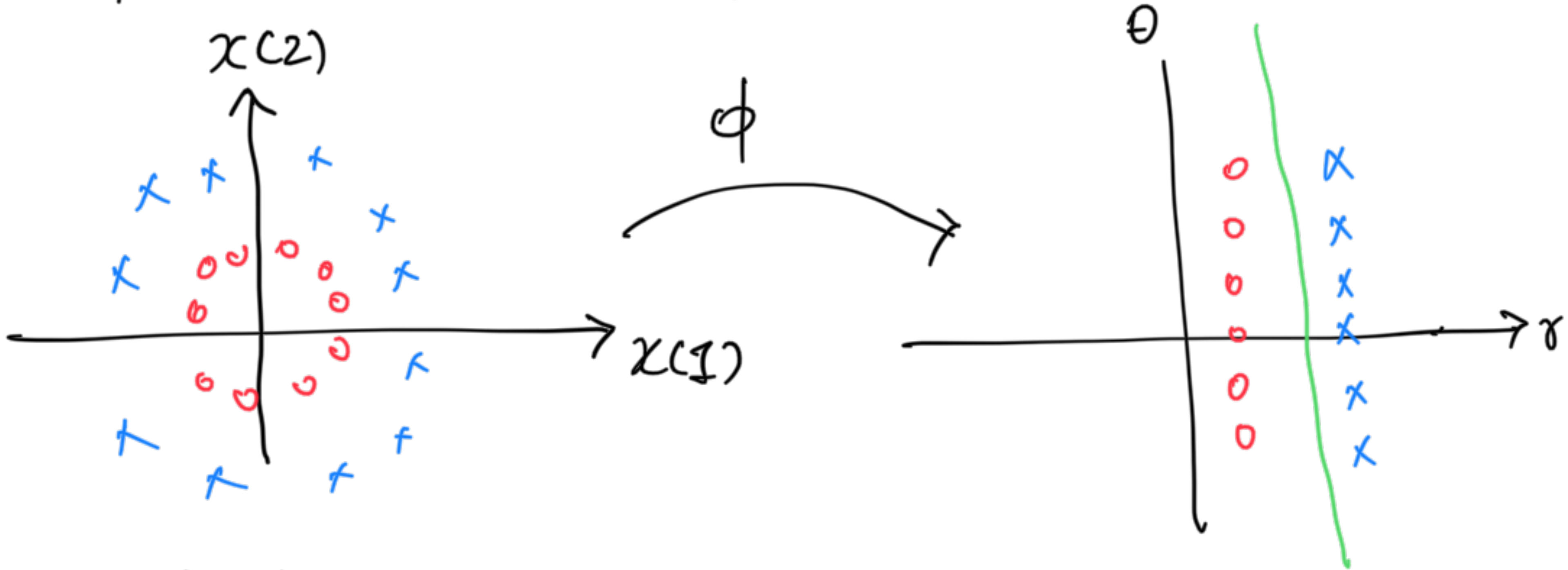
$M = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ ✓

$$(x_i)_{i=1}^{n} \in \mathbb{R}^{d} \xrightarrow{M} (\hat{x}_i)_{i=1}^{n} \in \mathbb{R}^{d'} \quad (d' << d)$$

$$\hat{x}_i = M x_i$$

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Eg (3) : Representation Learning



Cartesian

Polar

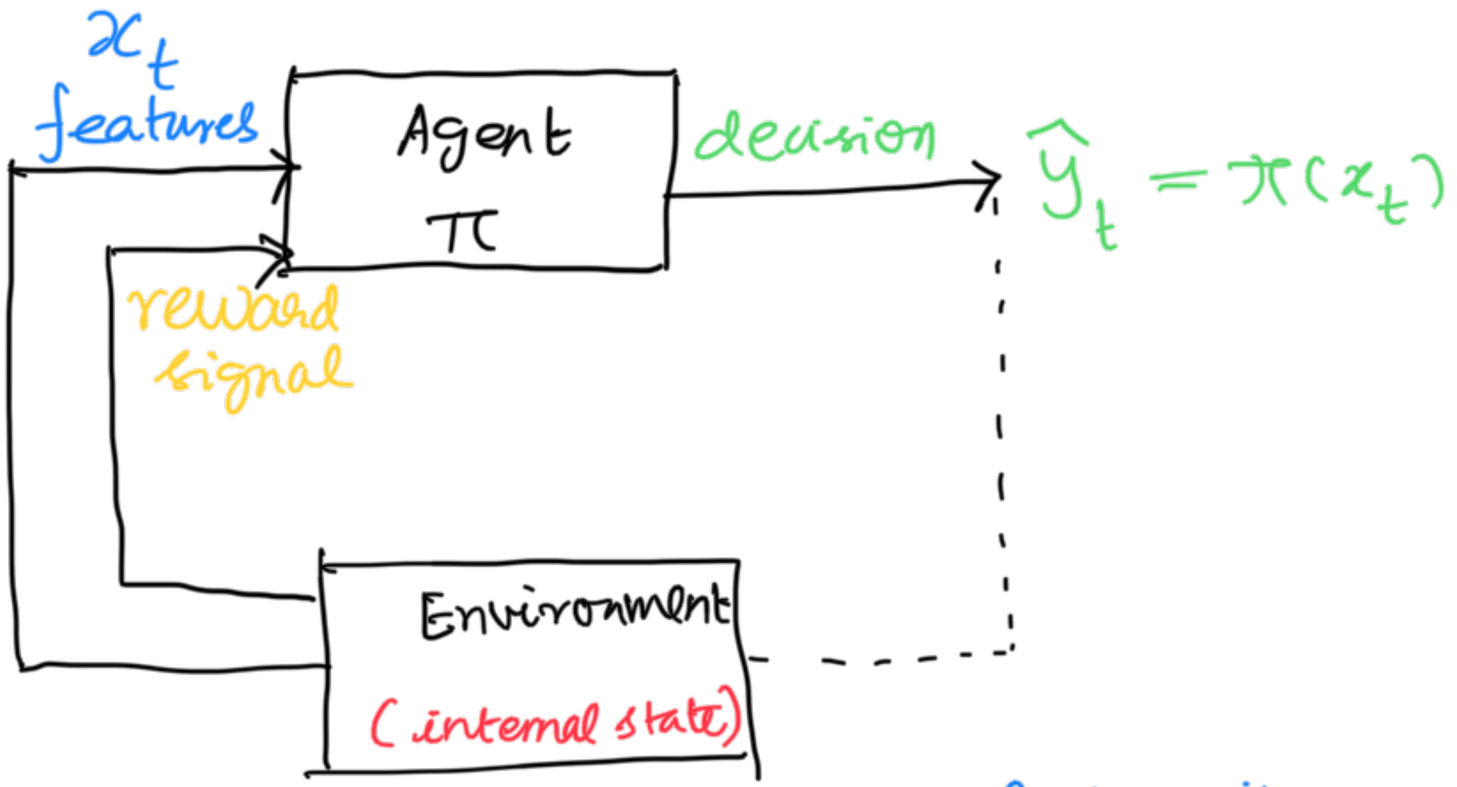
$$\phi : \mathbb{R}^{d} \rightarrow \mathbb{R}^{m}$$

$\phi : \mathbb{R} \rightarrow \mathbb{R}$

$$\hat{x}_i = \phi(x_i)$$

Predictive Task

Static Prediction



$\hat{y}_t = \pi(x_t)$

$y_t \sim p_y(\cdot)$ (distribution of categories)

$x_t \sim p_{X|Y}(\cdot|y_t)$ (distribution of image given category)

Multi-class classification (object classification)

- $y_t \in \mathcal{Y} = \{1, \dots, \overset{\text{categories}}{c}\}$ (label space)

For instance , object classification

$$\mathcal{Y} = \{\text{house, person, dog, cat}, \dots, \text{car, plane} \dots\}$$

- $x_t \in \mathcal{X}$ (feature space) $\subseteq \mathbb{R}^d$

In object classification x_t is an image

$$x_t \in \mathbb{R}^{\#\text{ pixels}}$$

- $\pi : \mathcal{X} \longrightarrow \mathcal{Y}$

- $\ell_t = L(\text{state, decision})$

  $= L(\text{label, predicted label})$

- $L(y_t, \hat{y}_t) = 0, \quad \hat{y}_t = y_t$

  $= 1, \quad \hat{y}_t \neq y_t$

$p_Y(\cdot)$

$p_{X|Y}(\cdot \mid y_t = \text{dog})$

Goal: $\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right]$

Model is known , i.e., p , p

$$\hat{y}_t = \pi_*(x_t) = \underset{y \in Y}{\arg\max}\ p_Y(y) \cdot p_{X|Y}(x_t | y)$$

(Bayesian Decision Theory)

Regression

- $X \subseteq \mathbb{R}^d$ (feature)
- $Y \subseteq \mathbb{R}^m$
- $(x_t, y_t) \overset{iid}{\sim} p_{XY}(\cdot,\cdot)$
- $\ell_t = L(y_t, \hat{y}_t)$

$L(y, \hat{y}) = \| y - \hat{y} \|^2$

$L(y_t, y_t) = \| y_t \quad y_t \|_2$

Goal : $\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right]$

$$\hat{y}_t = \pi_*(x_t) = \mathbb{E}[Y | X = x_t]$$

(Bayesian Decision theory)

$$(x_t, y_t)_{t=1}^{T}$$