Lecture 8 : 12th Jan 2026:

Vector valued RVs:

Here the Range Set of the func^c (RV) , is $\mathbb{R}^d$, where "d" is a Scalar

$$X : \Omega \to \mathbb{R}^d$$

Joint distribution:

Let Ω be a Sample space

Define 2 fun's , x_1, x_2

$$\left.\begin{aligned} x_1 &: \Omega \to \mathbb{R} \\ x_2 &: \Omega \to \mathbb{R} \end{aligned}\right] \text{ multiple RV on Same } \Omega$$

Corresponding $\mathbb{P}_{x_1}$, $\mathbb{P}_{x_2}$ be the respective DF.

define Joint Prob distributions as

$$P_{x_1,x_2}(a,b) = \mathbb{P}[\, E : \text{Intersection of inverse images of } (-\infty, a] \text{ \& } (-\infty, b] \text{ under } x_1, x_2 \text{ respectively }]$$

The above idea Can be extended to "d" Scalar RV.

Define Conditional Probabilities / Distributions.

if $A \& B \in \mathcal{F}$

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Suppose X & Y are 2 RVs defined on same Ω

(Range can be diff)

$$\left.\begin{array}{l} X : \Omega \to \mathbb{R}^d \\ Y : \Omega \to \mathbb{R}^k \end{array}\right] \; d \neq k$$

$$\mathbb{P}_{X|Y}(x \mid Y=y) = \frac{\mathbb{P}_{XY}}{\mathbb{P}_Y}$$

Joint / Marginal

Marginal distribution : If X & Y are 2 RVs

marginal of X is defined as.

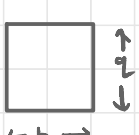
$$\mathbb{P}_X = \int_{\substack{Y \\ \text{range}}} \mathbb{P}_{XY} \, dy$$

→ Assuming that integration is possible.

$$\mathbb{P}_Y = \int_{\substack{X \\ \text{range}}} \mathbb{P}_{XY} \, dx$$

→ depending on the dim it can be a multidimensional integral

Example : Consider an image of dim $p \times q$



can be viewed as a $\mathbb{R}^{pq}$ dim Vector.

* In general any datapoint is a "d" dim vector.

* every datapoint is an element in the range space of a random variable

* The distribution function indicates @ quantifies the likelihood of observing a datapoint under X.

* Distribution func^{n} completely specifies the sample space.

* Typically a "label" is taken to be another additional RV, defined on the same sample space.

* In this scenario, we have 2 RVs, X & Y

$X \in \mathbb{R}^d$

$Y \in \mathbb{R}^k$

* The convention is as follows:

"Learning" starts with a dataset

$D = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \} \sim iid\ P_{xy}$

X: Feature @ Data Space

Y: Target @ Label space.

* i.i.d : Independent & Identically distributed.

$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$: statistically indpt.

* in a vector valued RV, we are not assuming independence among the scalars of that vector RV.
* our assumption is 2 data points $\in \mathbb{R}^d$ are independent.
* Independence is accross data points & not data dimensions.
* All problems in machine learning can be seen as Given D n unknown $\mathbb{P}_x$, estimate $\mathbb{P}_x$ (or) conditionals on $\mathbb{P}_x$ & Sample from it.

Lecture 4 : 14th Jan 2026

Recall we have given a dataset

$$D = \{(x_i, y_i)\}_{i=1}^{n} \sim \text{iid } P_{XY} : \text{Unknown.}$$

$$x_i \in \mathbb{R}^d$$

$$y_i \in \mathbb{R}^k \text{ or } \{1, 2, \ldots, K\}$$

Typically X : input / features / data

Y : Label / output

X & Y are random variables on a Sample space

The Fundamental problem of ML :

Given $D \sim$ Unknown distribution.

1) Estimate the distribution.

2) Sample from the distribution.

Lets consider the Distribution Estimation.

Given $D = \{(x_i, y_i)\}_{i=1}^{n} \sim \text{iid } P_{XY}$

estimate the distribution or the moments of the distribution.

Examples :

1) Estimate $\mathbb{P}_{Y|X}$ Classification $Y \in \{1, 2, \dots K\}$

Regression $Y \in \mathbb{R}^d$

(Bounding box)

2) Estimate

$\mathbb{P}_X$

$\mathbb{P}_Y$

$\mathbb{P}_{XY}$

$\mathbb{P}_{X|Y}$

* Estimating joint & marginal, does not mean its generative.

* In all conditional $\mathbb{P}_{X|Y}$ $(X \mid Y = y)$ (dummy, fixed)

Conditional distribution is for a specific "y"

Probability density func:

Given a CRV with a dist func $\mathbb{P}_X$,

density func $p_X : X \to \mathbb{R}^+$

$$P_x(x) = \int_{-\infty}^{x} p_x(x)\, dx.$$

* $p_x(x)$ does not corresponds to probability
* Not all RV has density func^c
* Our assumption is all distribution func^c that we consider has well behaved density func^c
* we will estimating density from now on.

The challenge with ML

* Given D, i.i.d from unknown p_x, estimate p_x
* challenge is P_x is completely unknown
* Given the samples from P_x can we estimate P_x.

Question:

* How to estimate a density func^c given its samples.
* Consider we have a dataset D,

$$D = \{x_i\}_{i=1}^{N} \sim \text{i.i.d } p_x$$

Mathematically how do we solve it?

(1) Assume a parametric functional form on p_x.

& lets call its a p_θ

eg: a) $p_\theta^a(x) = w^T x + b$

$x \in \mathbb{R}^d$
$w \in \mathbb{R}^d$
$b \in \mathbb{R}$

b) $p_\theta^b(x) = \mathcal{N}(x; \mu, \Sigma)$

$\theta = \{\mu, \Sigma\}$

$\mu \in \mathbb{R}^d$

$\Sigma \in \mathbb{R}^{d \times d}$

These are different model choices

* This assumption on p_x is a huge step.
* We also have universal fun^l approx.
* Once we make an assumption on parametric assumption on p_θ, our algorithm can say in that parametric form which one to use.

(2) Define & compute a distance metric between p_x & p_θ

we need say how good our assumption is.

* Let D_M denote the "distance metric" b/w p_x & p_θ

* $D_M(p_x \| p_\theta) : p_x \times p_\theta \longrightarrow \mathbb{R}^+$

③ Find ⓐ Estimate the parameter θ by solving the following optimization problem.

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, D_M(p_x \| p_\theta)$$

This is the Training of model.