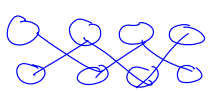


today: probability theory review

(aside: RBM )

why? → principled framework to model uncertainty
sources of uncertainty

- 1) intrinsic uncertainty → quantum mechanics
- 2) partial information / observation
(incomplete) e.g. rolling a dice → don't know exactly the initial conditions
- 3) incomplete modeling of complex phenomenon
(computation issues also important)
example: • "most birds can fly"
→ simplicity of rule is advantage but then yields uncertainty
- object recognition model

notation: X_1, X_2, X_3 or X, Y, Z

↳ random variables (usually real-valued)

x_1, x_2, x_3 x, y, z

→ their realizations

$\{X_1 = x_1\}$ represents event that ^{random variable} R.V. X_1 takes value x_1

example: rolling a dice $\Omega = \{1, 2, 3, 4, 5, 6\}$ sample space of "elementary events"

example: rolling a dice $\Omega = \{1, 2, 3, 4, 5, 6\}$ sample space of "elementary events"

R.V. $X = \mathbb{1}_{\{2, 4, 6\}}$

$\mathbb{1}_A$ \mapsto indicator function on set A $\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{o.w.} \end{cases}$

$\{X=1\}$ \mapsto event
that dice output was even here

Formally:

Ω is a sample space of "elementary events"
 $\omega_1, \omega_2, \omega_3, \dots$ (assume countable for now)

def: a random variable is a (measurable) mapping $X: \Omega \rightarrow \mathbb{R}$

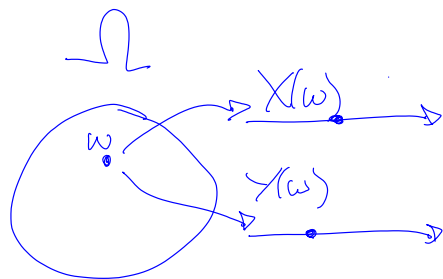
example of dice

[example 1]

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$X = \mathbb{1}_{\{2, 4, 6\}}$, "even"

$Y = \mathbb{1}_{\{1, 3, 5\}}$ "odd"



"world of elementary possibilities"

a probability distribution P

is a mapping $P: 2^\Omega \rightarrow [0, 1]$

$\mathcal{E} \triangleq$ set of all subsets of Ω

\uparrow set of "events"

("sigma-field" in measure theory
needed when Ω is uncountable)

"world of elementary possibilities"

a probability distribution P

is a mapping $P: 2^\Omega \rightarrow [0, 1]$

$\mathcal{E} \triangleq$ set of all subsets of Ω

\mathcal{E} set of "events"

(" σ -field" in measure theory needed when Ω is uncountable)

which satisfies the following properties

Kolmogorov axioms

$$\begin{cases} 1) P(E) \geq 0 \quad \forall E \in \mathcal{E} \\ 2) P(\Omega) = 1 \\ 3) P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \text{ when } E_i \text{'s are disjoint} \end{cases}$$

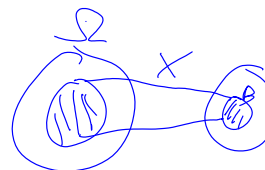
prob. on Ω induces a prob. dist. on image of X $\mathcal{Q}_X \triangleq X(\Omega)$

$$X(A) \triangleq \{y : \exists w \in A \text{ s.t. } X(w) = y\}$$

$$\text{set} = \{X(w) : w \in A\}$$

inverse image

$$X^{-1}(B) = \{w : X(w) \in B\}$$



$$X^{-1}(\{a\}) = \text{set of } w \text{'s s.t. } X(w) = a$$

(if X was invertible, this would be a singleton)

e.g. "event" $\{x\}$ in \mathcal{Q}_X , get $\mathcal{Q}_X(\{x\}) = P(X^{-1}(\{x\}))$

$$= P(\{w : X(w) = x\})$$

notation shortcut:

recap:

$\{X=x\}$ represents

both the event $\{x\}$ in Ω_X

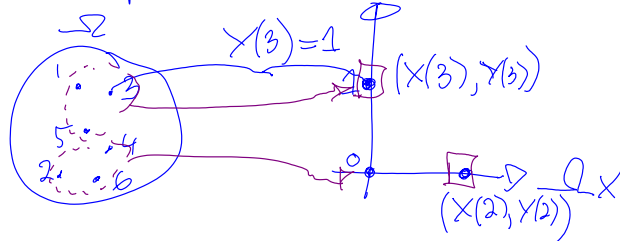
the event $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{E}$

notation shorthand:

$P(\{X=x\})$ (standard in stat)

$p(x)$ (standard in ML)

back to our example (*): Ω_X



joint distribution on $(X, Y) \in \Omega_X \times \Omega_Y$

$P_{X,Y} \{X=x, Y=y\}$

means "and" \Rightarrow intersection of events

$= P(X^{-1}(\{x\}) \cap Y^{-1}(\{y\}))$

can represent events of $\Omega_X \times \Omega_Y$ as table:

	$X=0$	$X=1$
$Y=0$	0	$\frac{1}{2}$
$Y=1$	$\frac{1}{2}$	0

$P(\{X=0, Y=1\}) = P(\{1, 3, 5\})$

$= \sum_{\omega \in \{1, 3, 5\}} p(\omega) = 3 \cdot \frac{1}{6} = \frac{1}{2}$

marginal distribution (in the context of the joint)

↳ distribution on the components of a random vector

$$P(\{X=x\}) = \sum_{y \in \Omega_Y} P(\{X=x, Y=y\})$$

"sum rule"

from $P_{X,Y}$ a distribution for (X,Y)

← this summation procedure is called "marginalization"

P_X "marginal distribution" ← distribution for X by itself

Other R.V. basics:

types of R.V.:

"discrete R.V." → Ω_X is countable

"continuous R.V." → Ω_X is uncountable + \exists a density

↳ its distribution P_X (on Ω_X) is fully defined

by probability mass fct. "pmf" $P(\{X=x\})$ for $x \in \Omega_X$

shorthand: $P_X(x)$

or even $p(x)$ (!)

for any (scalar) R.V., proba distribution P_X is fully characterized

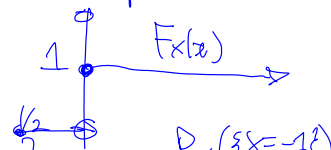
by its cumulative distribution function (cdf): $F_X(x) \triangleq P(\{X \leq x\})$

$$F_X: \mathbb{R} \rightarrow [0,1]$$

properties:

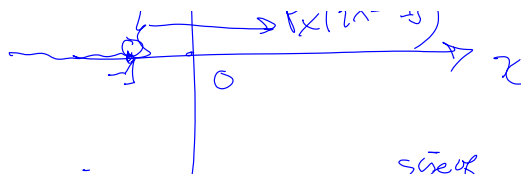
- F_X is non-decreasing
- $\lim_{x \rightarrow \infty} F_X(x) = 1$

example



$$\lim_{x \rightarrow +\infty} F_X(x) = 1$$

continuous from the right



(for a discrete R.V., the cdf is piecewise constant, and the jumps give value of pmf)

* for a continuous R.V., the cdf is "absolutely cts"

(\Leftrightarrow differentiable almost everywhere

and \exists function $f(x)$ s.t. $F_X(x) = \int_{-\infty}^x f(z) dz$

$$\frac{d}{dx} F_X(x) = f(x) \text{ when } x \text{ is a continuity pt. of } F$$

probability density function (pdf)

⊗ pdf for a cts. R.V. is the analog of the pmf for a discrete R.V.
with " \int " (for cts R.V.) replacing " \sum "

$$\sum_{x \in \mathcal{X}} p(x) = 1 \text{ for a discrete R.V.}$$

$$\int_{\mathcal{X}} f(x) dx = 1$$

call this $p(x)$ (∇) in ML

note: $p(x)$ ($f(x)$) can be bigger than 1 for a pdf

e.g. uniform dist. on $[0, \frac{1}{2}]$

$$\text{then } p(x) = \begin{cases} 2 & \text{for } x \in [0, \frac{1}{2}] \\ 0 & \text{o.w.} \end{cases}$$

"otherwise"