

Product space, joint distribution

Consider two measurable spaces $(\Omega_1, \mathcal{A}_1)$, $(\Omega_2, \mathcal{A}_2)$.

Define the product space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ with

$$\Omega_1 \times \Omega_2 = \{ (\omega_1, \omega_2) \mid \omega_1 \in \Omega_1, \omega_2 \in \Omega_2 \}$$

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \{ A_1 \times A_2 \mid A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2 \}.$$

Consider two RVs. $X_1: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_1, \mathcal{A}_1)$

$$X_2: (\Omega, \mathcal{A}, P) \rightarrow (\Omega_2, \mathcal{A}_2)$$

$$X := (X_1, X_2): (\Omega, \mathcal{A}, P) \rightarrow (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$$

$$(X_1, X_2)(\omega) = (X_1(\omega), X_2(\omega))$$

The distribution $P_{(X_1, X_2)}$ on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ is called the joint distribution of X_1 and X_2 .

Example in ML: (X, Y) where X is the input data, Y is the label.

Product measure: Let $(\Omega_1, \mathcal{A}_1, P_1), (\Omega_2, \mathcal{A}_2, P_2)$ be two probability spaces. We define the product measure $P_1 \otimes P_2$ on the product space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ as

$$(P_1 \otimes P_2)(A_1 \times A_2) := P_1(A_1) \cdot P_2(A_2)$$

Theorem: Two RVs X_1, X_2 are independent if and only if their joint distribution coincides with the product distributions:

$$P_{(X_1, X_2)} = P_1 \otimes P_2$$

Marginal Distributions

Consider the joint distribution $P_{(x_1, x_2)}$ of two RVs. $X := (x_1, x_2)$. The marginal distribution of X w.r.t. x_1 is the original distribution of x_1 on $(\Omega_1, \mathcal{A}_1)$, namely P_{x_1} . Similarly for x_2 as well.

Example in the discrete case:

$Y \setminus X$	x_1	x_2	x_3	Σ
y_1	p_{11}	p_{12}	p_{13}	$p_{11} + p_{12} + p_{13} = P(Y=y_1)$
y_2	p_{21}	p_{22}	p_{23}	$p_{21} + p_{22} + p_{23} = P(Y=y_2)$
Σ	$p_{11} + p_{21}$ $= P(X=x_1)$	$p_{12} + p_{22}$ $= P(X=x_2)$	$p_{13} + p_{23}$ $= P(X=x_3)$	marginal dist. w.r.t. Y .

Marginal distributions in case of densities

$X, Y : (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. $Z := (X, Y)$.

Assume that the joint distribution of Z has a density f on \mathbb{R}^2 . Then we have the following statements:

(1) Both X and Y have densities on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by.

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

↑ joint dist. → sum over y

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

↑ joint dist. → sum over x

(2) X and Y are independent iff

$$\underbrace{f(x, y)}_{\text{joint}} = \underbrace{f_x(x)}_{\text{marginals}} \cdot \underbrace{f_y(y)}_{\text{marginals}} \quad \text{a.s.}$$

almost surely.

Mixed cases

For example, consider X is a continuous RV. with density and Y a discrete RV.

Say, $X =$ image
(2d-continuous signal)

$Y =$ "cat", "dog" discrete

Special case: marginals of multivariate Normal.

2 dim Consider a 2-dim normal RV $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ with mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^2$ and cov. $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$

Then the marginal dist. of X w.r.t. x_1 is again a normal distribution with mean μ_1 and var σ_1^2 .

sum up the
y-direction
↳ marginal w.r.t. x
is normal




n-dim: $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$. Group the

Variables $S = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \in \mathbb{R}^k$, $T = \begin{pmatrix} x_{k+1} \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n-k}$

Want to look at the marginal of X

w.r.t. S . $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$ mean, $\mu_S := \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$, $\mu_T := \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix}$

$$\Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{ST} \\ \Sigma_{TS} & \Sigma_{TT} \end{pmatrix} \in \mathbb{R}^{n \times n}$$



Now the marginal of X w.r.t. S is a normal dist. on \mathbb{R}^k with mean μ_S and cov Σ_{SS} .

In summary: Marginals of normal dist. are again normal dist.

Conditional Distributions

Discrete case:

Know conditional probabilities: $P(A|B)$ defined for events $A, B \in \mathcal{A}$, and $P(B) > 0$.

Let $X, Y: (\Omega, \mathcal{A}, P) \rightarrow \mathbb{R}$ be discrete RV, $y \in \mathbb{R}$ such that $P(Y=y) > 0$. Then we can define the conditional probability measure

$$P_{X|Y=y}: \mathcal{A} \mapsto P(X \in A | Y=y).$$

This is a probability measure.

For general RV this is very complicated!

(skipping)

Conditional distributions in case of densities:

Assume $Z := (X, Y)$ has a joint density $f: \mathbb{R}^2 \rightarrow \mathbb{R}$, and marginal densities $f_X, f_Y: \mathbb{R} \rightarrow \mathbb{R}$.

Then the function,

$$f_{X|Y=y}(x) := \frac{f(x, y)}{f_Y(y)}$$

is then also a density on \mathbb{R} , called the conditional density of X given $Y=y$.

Example: normal distribution

$$\mu = \begin{pmatrix} \mu_s \\ \vdots \\ \mu_T \end{pmatrix} \begin{matrix} \} \mu_s \\ \\ \} \mu_T \end{matrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{ST} \\ \Sigma_{TS} & \Sigma_{TT} \end{pmatrix}$$

If $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \sim N(\mu, \Sigma)$, then the

conditional distributions of $X_S = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$

conditioned on $X_T = \begin{pmatrix} x_{p+1} \\ \vdots \\ x_n \end{pmatrix}$ is given by

$$P_{X_S|X_T} \sim N \left(\mu_T + \Sigma_{ST} \Sigma_{TT}^{-1} (X_S - \mu_T), \right. \\ \left. \Sigma_{TT} - \Sigma_{ST}^T \Sigma_{SS}^{-1} \Sigma_{ST} \right)$$



marginal
(collapsing)



conditional
(slicing)

Conditional Expectation

Def (discrete case): $X, Y : (\Omega, \mathcal{A}, P) \rightarrow \mathbb{R}$. Assume X takes finitely (countably) many values $x_1, x_2, \dots, x_n \in \mathbb{R}$, Y takes finitely (countably) many values $y_1, \dots, y_m \in \mathbb{R}$. Always with positive probability.

$$E(Y | X = x_i) := \sum_{j=1}^m y_j \underbrace{P(Y = y_j | X = x_i)}_{\text{well-defined}}$$

Example: two dice, $X =$ value of die 1, $Y =$ value of die 2, independent dice.

$$E(\text{sum} | X = 1) = \sum_{i=1}^{12} i \cdot P(\text{sum} = i | X = 1)$$

$$= \sum_{k=1}^6 (1+k) \cdot P(Y = k | X = 1)$$

$$= \sum_{k=1}^6 (1+k) \cdot P(Y = k) = \sum_{k=1}^6 (1+k) \cdot \frac{1}{6} = 4.5$$

So far we defined $E(Y|X=x_i)$, but often we want to consider the "function" $E(Y|X)(\omega)$.

This is a RV: $E(Y|X):(\Omega, \mathcal{A}, \mathcal{P}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

This leads to the following:

Def (discrete case): X, Y as before. Then the conditional expectation is defined as follows:

$$E(Y|X) := f(X) \text{ with}$$

$$f(x) = \begin{cases} E(Y|X=x) & \text{if } P(X=x) > 0 \\ \text{arbitrary, say } 0 & \text{otherwise} \end{cases}$$

 $E(Y|X)$ is only defined a.s.

Now we want to consider the more general case.

Sketch: X continuous RV

Y discrete RV $\sim Y_1, Y_2, \dots, Y_s$

We want to look at $E(X|Y)$.

