Probability Basics

CS771: Introduction to Machine Learning

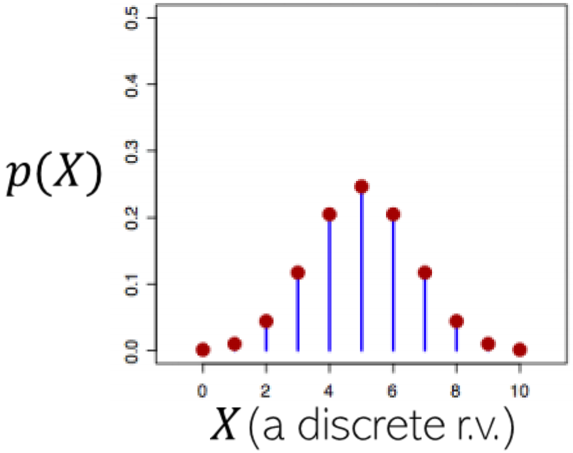
Nisheeth

Random Variables

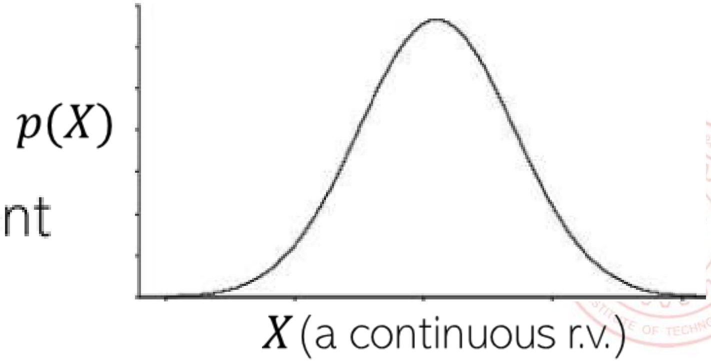
- Informally, a random variable (r.v.) X denotes possible outcomes of an event

- Can be discrete (i.e., finite many possible outcomes) or continuous

- Some examples of discrete r.v.
  - $X \in \{0,1\}$ denoting outcomes of a coin-toss
  - $X \in \{1,2,\ldots,6\}$ denoting outcome of a dice roll



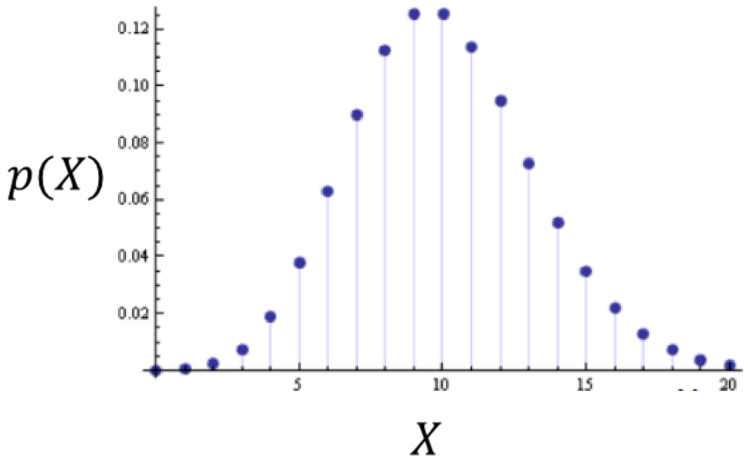
- Some examples of continuous r.v.
  - $X \in (0,1)$ denoting the bias of a coin
  - $X \in \mathbb{R}$ denoting heights of students in CS771
  - $X \in \mathbb{R}$ denoting time to get to your hall from the department



Discrete Random Variables

- For a discrete r.v. X, $p(x)$ denotes $p(X = x)$ - probability that $X = x$
- $p(X)$ is called the probability mass function (PMF) of r.v. X
    - $p(x)$ or $p(X = x)$ is the value of the PMF at x

$$p(x) \geq 0$$
$$p(x) \leq 1$$
$$\sum_x p(x) = 1$$



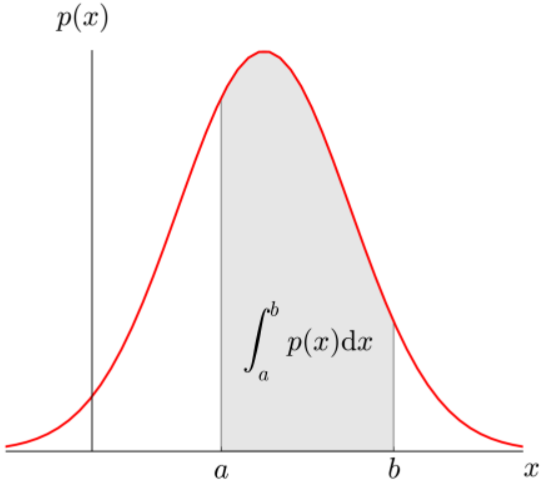
Continuous Random Variables

- For a continuous r.v. X, a *probability* $p(X = x)$ or $p(x)$ is meaningless
- For cont. r.v., we talk in terms of prob. within an interval $X \in (x, x + \delta x)$
  - $p(x)\delta x$ is the prob. that $X \in (x, x + \delta x)$ as $\delta x \to 0$
  - $p(x)$ is the probability density at $X = x$

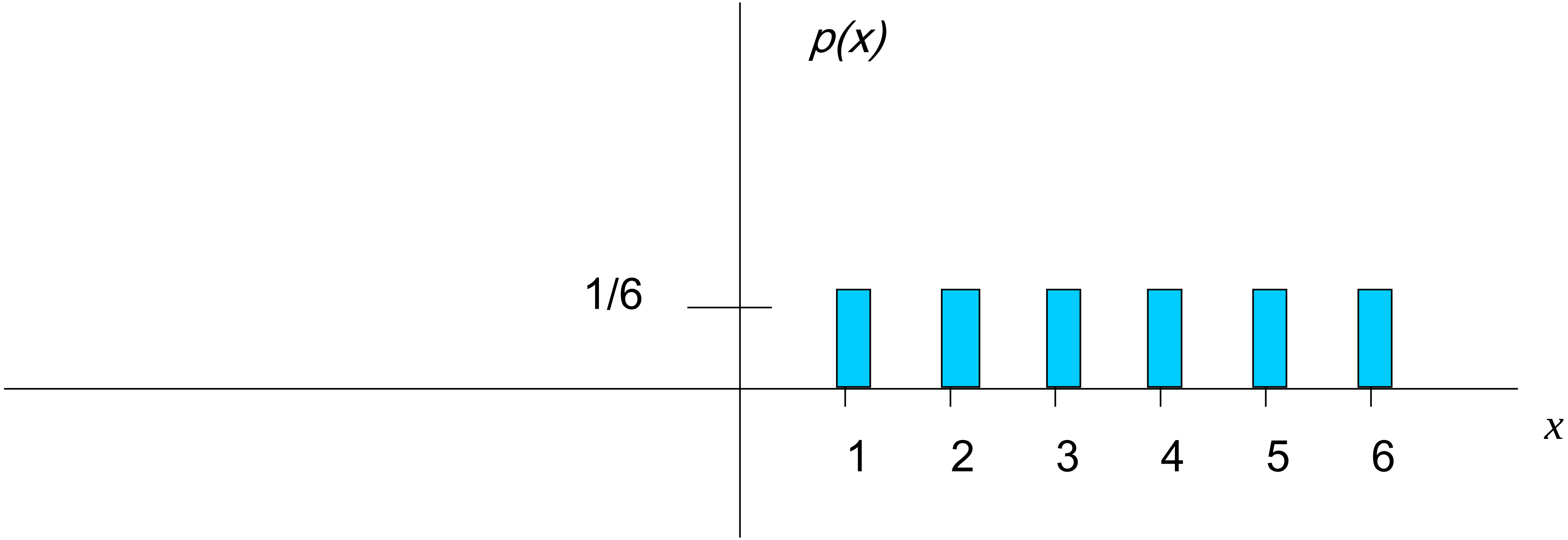
$$p(x) \geq 0$$

$$\cancel{p(x) \leq 1}$$

$$\int p(x) dx = 1$$



Discrete example: roll of a die



$$\sum_{\text{all } x} P(x) = 1$$

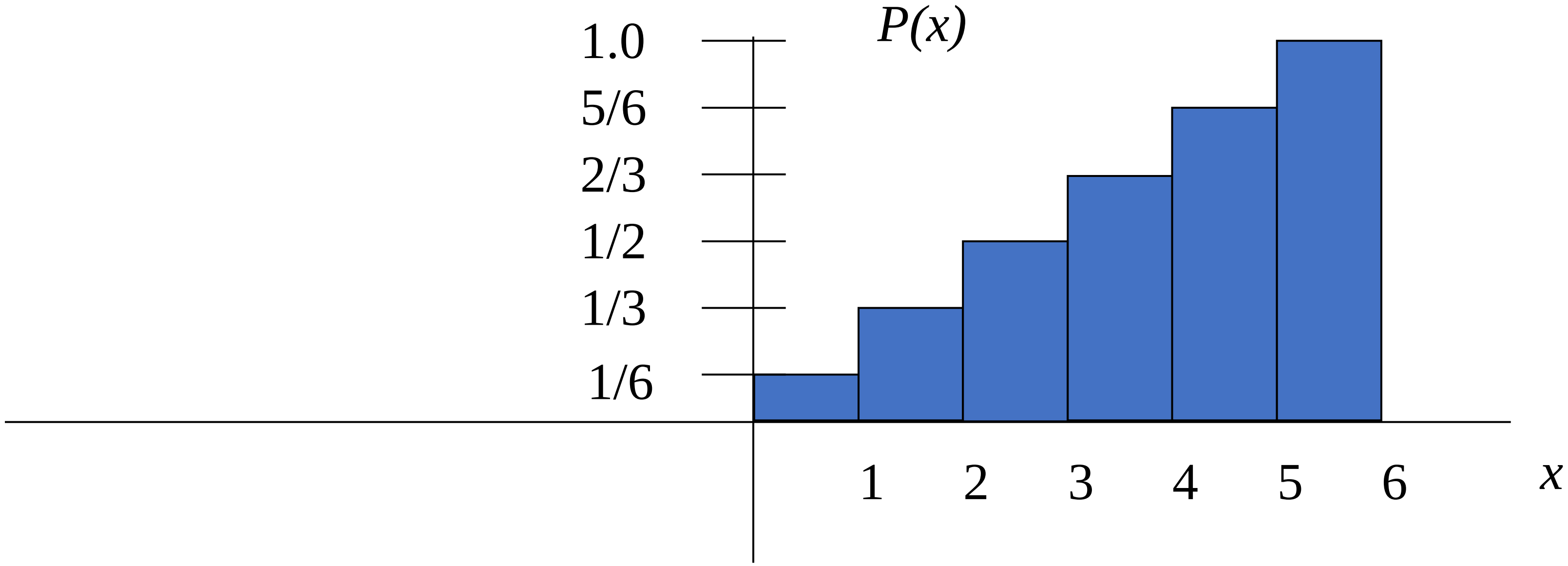
Probability mass function (pmf)

x	$p(x)$
1	$p(x=1)$=1/6
2	$p(x=2)$=1/6
3	$p(x=3)$=1/6
4	$p(x=4)$=1/6
5	$p(x=5)$=1/6
6	$p(x=6)$=1/6

Cumulative distribution function

x	$P(x \leq A)$
1	$P(x \leq 1) = 1/6$
2	$P(x \leq 2) = 2/6$
3	$P(x \leq 3) = 3/6$
4	$P(x \leq 4) = 4/6$
5	$P(x \leq 5) = 5/6$
6	$P(x \leq 6) = 6/6$

Cumulative distribution function (CDF)



Practice Problem

- The number of patients seen in a clinic in any given hour is a random variable represented by *x*. The probability distribution for *x* is:

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

Find the probability that in a given hour:

a. exactly 14 patients arrive $p(x=14)= .1$

b. At least 12 patients arrive $p(x\geq 12)= (.2 + .1 +.1) = .4$

c. At most 11 patients arrive $p(x\leq 11)= (.4 +.2) = .6$

Continuous case

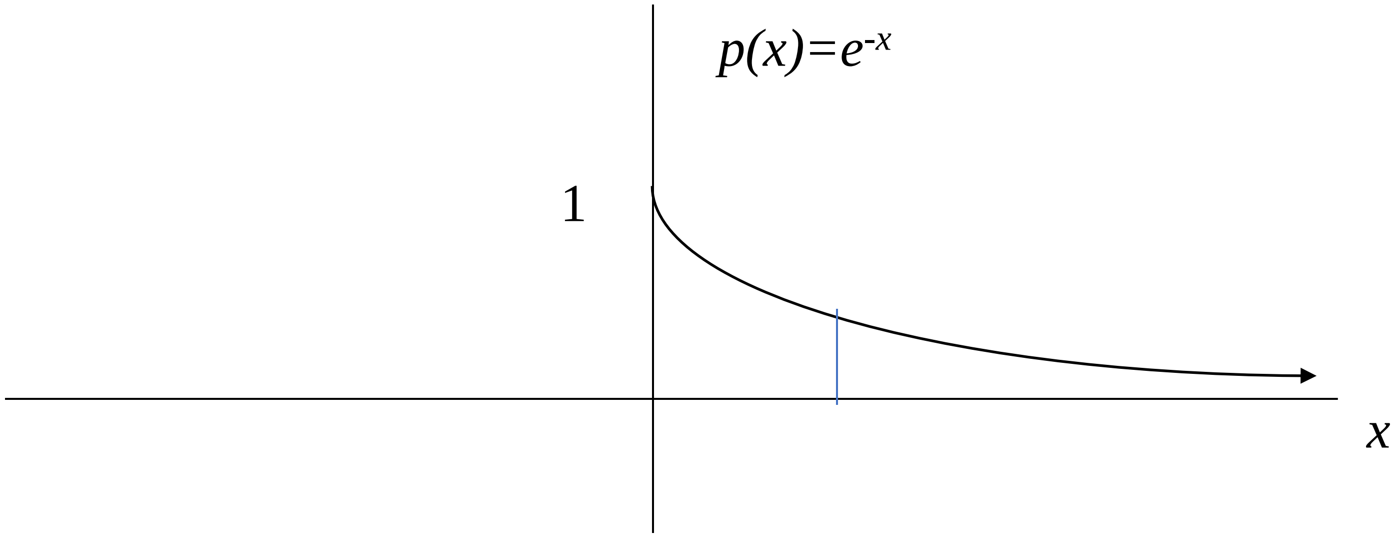
- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
  - For example, recall the negative exponential function (in probability, this is called an “exponential distribution”):

$$f(x) = e^{-x}$$

  - This function integrates to 1:

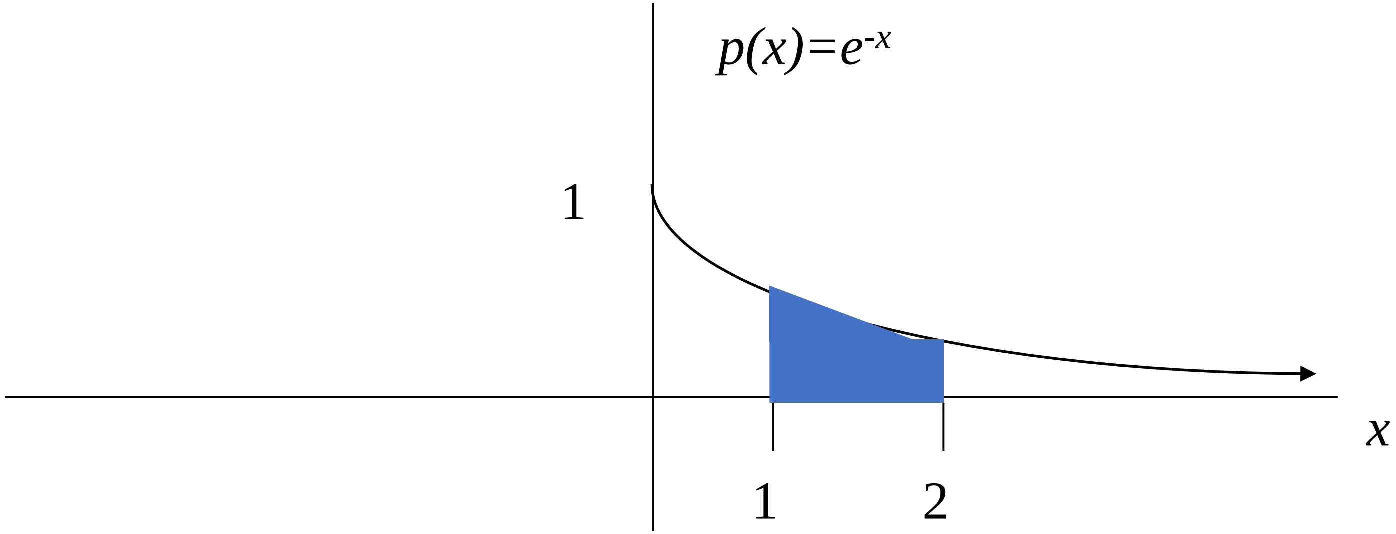
$$\int_{0}^{+\infty} e^{-x} = -e^{-x} \Big|_{0}^{+\infty} = 0 + 1 = 1$$

Continuous case: “probability density function” (pdf)



The probability that *x* is any exact particular value (such as 1.9976) is 0; we can only assign probabilities to possible ranges of x.

For example, the probability of *x* falling within 1 to 2:

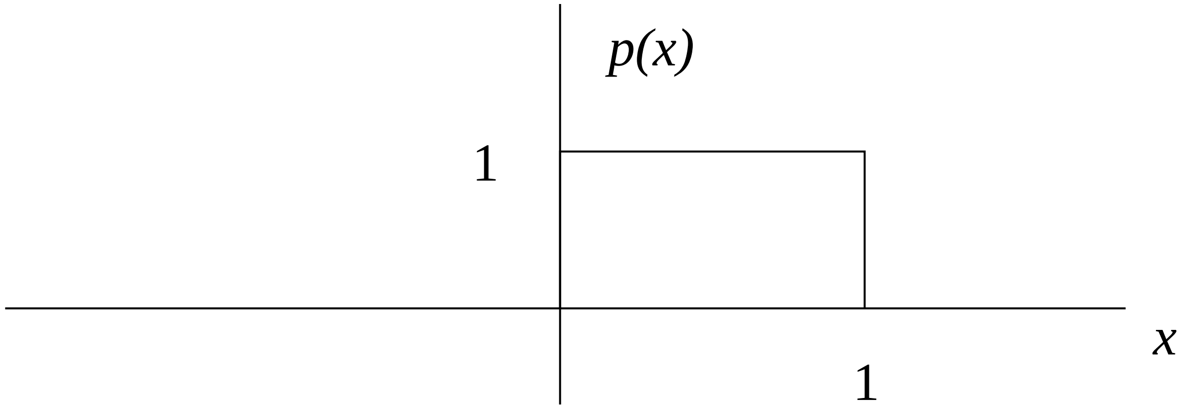


$$P(1 \le x \le 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - -e^{-1} = -.135 + .368 = .23$$

Example 2: Uniform distribution

The uniform distribution: all values are equally likely.

$f(x) = 1$, for $1 \geq x \geq 0$

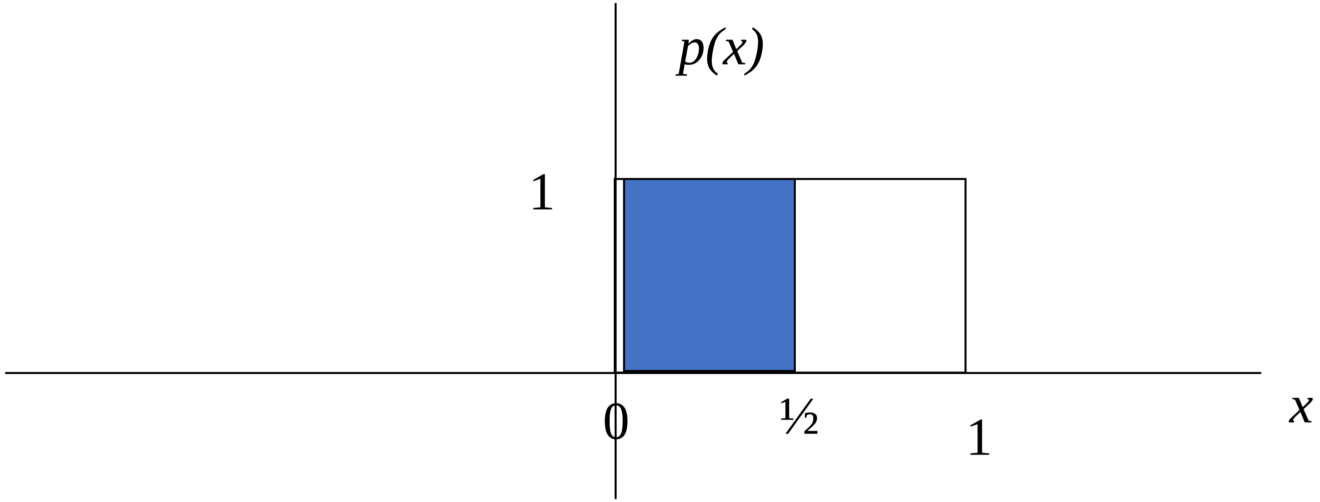


We can see it's a probability distribution because it integrates to 1 (the area under the curve is 1):

$$\int_0^1 1 = x \Big|_0^1 = 1 - 0 = 1$$

Example: Uniform distribution

What’s the probability that *x* is between 0 and ½?



P(½ ≥x≥ 0)= ½

A word about notation

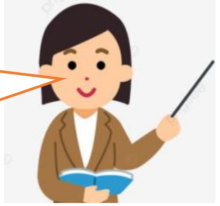
- $p(.)$ can mean different things depending on the context
- $p(X)$ denotes the distribution (PMF/PDF) of an r.v. X
- $p(X = x)$ or $p_X(x)$ or simply $p(x)$ denotes the prob. or prob. density at value x
  - Actual meaning should be clear from the context (but be careful)
- Exercise same care when $p(.)$ is a specific distribution (Bernoulli, Gaussian, etc.)
- The following means generating a random sample from the distribution $p(X)$

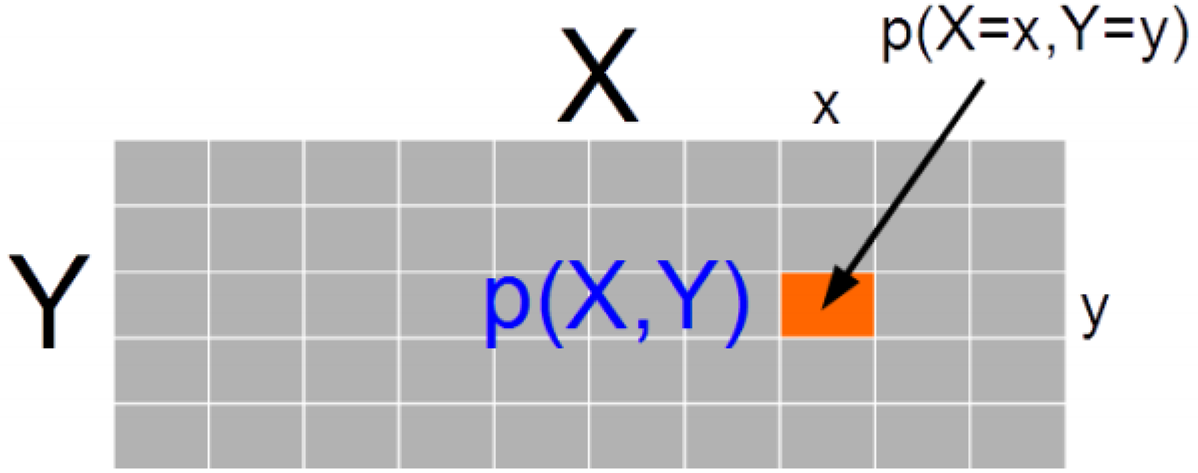
$$x \sim p(X)$$

Joint Probability Distribution

- Joint prob. dist. $p(X, Y)$ models probability of co-occurrence of two r.v. X, Y
- For discrete r.v., the joint PMF $p(X, Y\ )$ is like a table (that sums to 1)

For 3 r.v.'s, we will likewise have a "cube" for the PMF. For more than 3 r.v.'s too, similar analogy holds



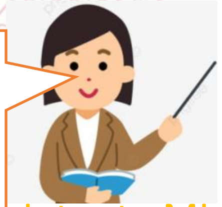


$$\sum_x \sum_y p(X = x, Y = y) = 1$$

- For two continuous r.v.'s X and Y, we have joint PDF $p(X, Y)$

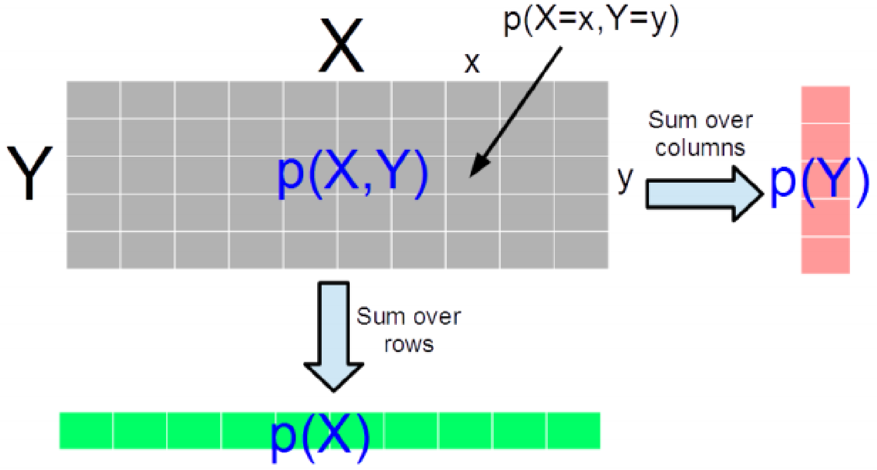
$$\int_x \int_y p(X = x, Y = y)dxdy = 1$$

For more than two r.v.'s, we will likewise have a multi-dim integral for this property



Marginal Probability Distribution

- Consider two r.v.'s X and Y (discrete/continuous – both need not of same type)
- Marg. Prob. is PMF/PDF of one r.v. accounting for all possibilities of the other r.v.
- For discrete r.v.'s, $p(X) = \sum_y p(X, Y = y)$ and $p(Y) = \sum_x p(X = x, Y)$
- For discrete r.v. it is the sum of the PMF table along the rows/columns



The definition also applied for two sets of r.v.'s and marginal of one set of r.v.'s is obtained by summing over all possibilities of the second set of r.v.'s

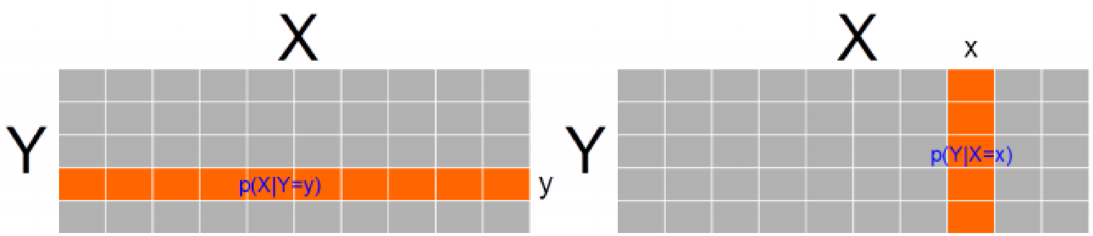
For discrete r.v.'s, marginalization is called summing over, for continuous r.v.'s, it is called "integrating out"

- For continuous r.v.'s, $p(X) = \int_y p(X, Y = y)dy$, $p(Y) = \int_x p(X = x, Y)dx$

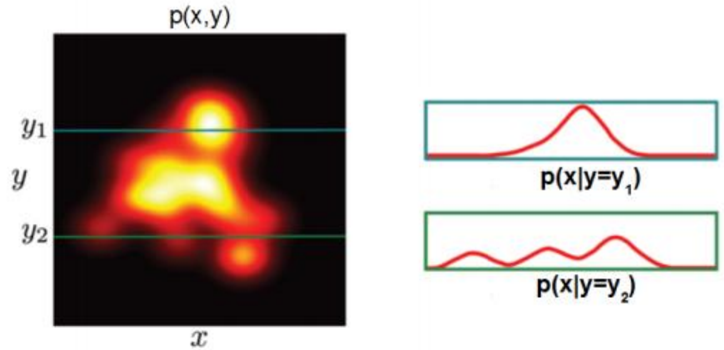
Conditional Probability Distribution

- Consider two r.v.'s X and Y (discrete/continuous – both need not of same type)
- Conditional PMF/PDF $p(X|Y)$ is the prob. dist. of one r.v. X, fixing other r.v. Y
- $p(X|Y = y)$ or $p(Y|X = x)$ like taking a slice of the joint dist. $p(X,Y)$

Discrete Random Variables



Continuous Random Variables



- Note: A conditional PMF/PDF may also be conditioned on something that is not the value of an r.v. but some fixed quantity in general

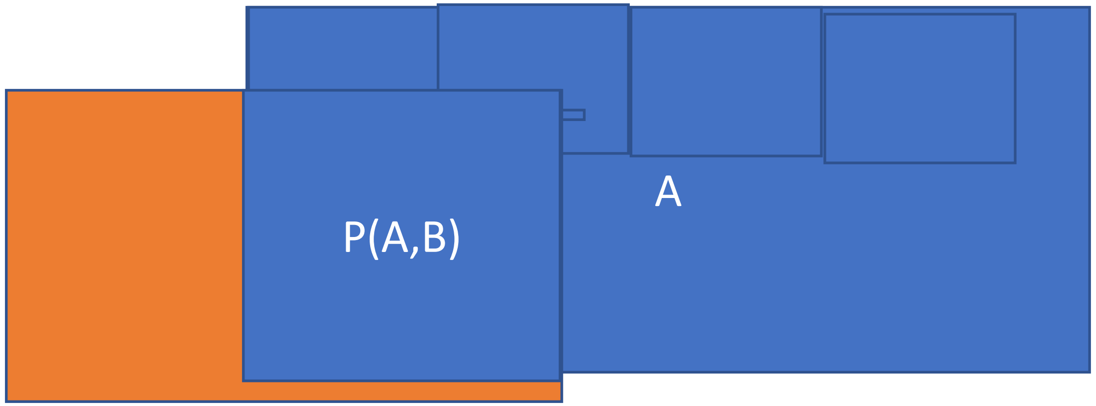
We will see cond. dist. of output $\boldsymbol{y}$ given weights $\boldsymbol{w}$ (r.v.) and features $\boldsymbol{X}$ written as $p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X})$

An example

X/Y	raining	sunny
Umbrella	0.5	0.1
No umbrella	0.2	0.2

P (B|A)p(A) = p(A|B)p(B)

P(x) = {0.6, 0.4}
P(y) = {0.7, 0.3}



P(X|Y = raining) = {0.5, 0.2}

P(X = umbrella|Y = raining)

Some Basic Rules

- **Sum Rule:** Gives the marginal probability distribution from joint probability distribution

For discrete r.v.: $p(X) = \sum_Y p(X, Y)$

For continuous r.v.: $p(X) = \int_Y p(X, Y) dY$

- **Product Rule:** $p(X,Y) = p(Y|X)p(X) = p(X|Y)p(Y)$

- **Bayes' rule:** Gives conditional probability distribution (can derive it from product rule)

$$\boxed{p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}}$$

For discrete r.v.: $p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$

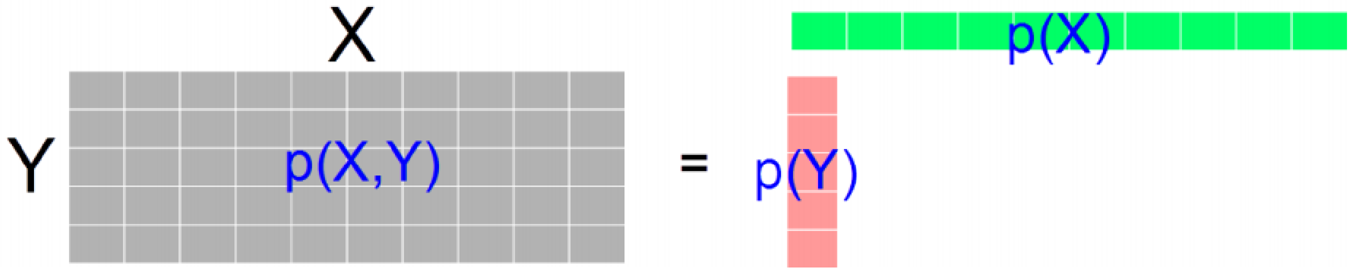
For continuous r.v.: $p(Y|X) = \frac{p(X|Y)p(Y)}{\int_Y p(X|Y)p(Y)dY}$

- **Chain Rule:** $p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2|X_1)\ldots p(X_N|X_1, \ldots, X_{N-1})$

Independence

- X and Y are independent when knowing one tells nothing about the other

$$\begin{aligned} p(X|Y=y) &= p(X) \\ p(Y|X=x) &= p(Y) \\ p(X,Y) &= p(X)p(Y) \end{aligned}$$



- The above is the marginal independence $(X \perp\!\!\!\perp Y)$

- Two r.v.'s X and Y may not be marginally indep but may be given the value of another r.v. Z

$$p(X,Y|Z=z) = p(X|Z=z)p(Y|Z=z) \qquad X \perp\!\!\!\perp Y|Z$$

Expectation

- Expectation of a random variable tells the expected or average value it takes
- Expectation of a discrete random variable $X \in S_X$ having PMF $p(X)$

$$\mathbb{E}[X] = \sum_{x \in S_X} x p(x)$$

Probability that $X = x$

- Expectation of a continuous random variable $X \in S_X$ having PDF $p(X)$

$$\mathbb{E}[X] = \int_{x \in S_X} x p(x) dx$$

Probability density at $X = x$

Note that this exp. is w.r.t. the distribution $p(f(X))$ of the r.v. $f(X)$

- The definition applies to functions of r.v. too (e.g.., $\mathbb{E}[f(X)]$)

Often the subscript is omitted but do keep in mind the underlying distribution

- Exp. is always w.r.t. the prob. dist. $p(X)$ of the r.v. and often written as $\mathbb{E}_p[X]$

Expectation: A Few Rules

X and Y need not be even independent. Can be discrete or continuous

- Expectation of sum of two r.v.'s: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Proof is as follows
  - Define $Z = X + Y$

$$\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot p(Z = z) \qquad \text{s.t. } z = x + y \text{ where } x \in S_X \text{ and } y \in S_Y$$

$$= \sum_{x \in S_X} \sum_{y \in S_Y} (x+y) \cdot p(X = x, Y = y)$$

$$= \sum_x \sum_y x \cdot p(X = x, Y = y) + \sum_x \sum_y y \cdot p(X = x, Y = y)$$

$$= \sum_x x \sum_y p(X = x, Y = y) + \sum_y y \sum_x p(X = x, Y = y)$$

$$= \sum_x x \cdot p(X = x) + \sum_y y \cdot p(Y = y)$$

Used the rule of marginalization of joint dist. of two r.v.'s

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Expectation: A Few Rules (Contd)

α is a real-valued scalar

α and β are real-valued scalars

f and g are arbitrary functions.

- Expectation of a scaled r.v.: $\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$
- Linearity of expectation: $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$
- (More General) Lin. of exp.: $\mathbb{E}[\alpha f(X) + \beta g(Y)] = \alpha \mathbb{E}[f(X)] + \beta \mathbb{E}[g(Y)]$
- Exp. of product of two independent r.v.'s: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Law of the Unconscious Statistician (LOTUS): Given an r.v. X with a known prob. dist. $p(X)$ and another random variable $Y = g(X)$ for some function g

Requires finding $p(Y)$

Requires only $p(X)$ which we already have

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{y \in S_Y} y p(y) \quad = \sum_{x \in S_X} g(x) p(x)$$

LOTUS also applicable for continuous r.v.'s

- Rule of iterated expectation: $\mathbb{E}_{p(X)}[X] = \mathbb{E}_{p(Y)}[\mathbb{E}_{p(X|Y)}[X|Y]]$

Variance and Covariance

- Variance of a scalar r.v. tells us about its spread around its mean value $\mathbb{E}[X] = \mu$

$$\text{var}[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2$$

- Standard deviation is simply the square root is variance
- For two scalar r.v.'s X and Y, the covariance is defined by

$$\text{cov}[X,Y] = \mathbb{E}[\{X-\mathbb{E}[X]\}\{Y-\mathbb{E}[Y]\}] = \mathbb{E}[XY] - \mathbb{E}[\text{X}]\mathbb{E}[\text{Y}]$$

- For two vector r.v.'s X and Y (assume column vec), the covariance matrix is defined by

$$\text{cov}[X,Y] = \mathbb{E}[\{X-\mathbb{E}[X]\}\{Y^\top-\mathbb{E}[Y^\top]\}] = \mathbb{E}[XY^\top] - \mathbb{E}[\text{X}]\mathbb{E}[Y^\top]$$

- Cov. of components of a vector r.v. X: $\text{cov}[X] = \text{cov}[X,X]$
- Note: The definitions apply to functions of r.v. too (e.g., $\text{var}[f(X)]$)
- Note: Variance of sum of independent r.v.'s: $\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$

Important result

Transformation of Random Variables

- Suppose $Y = f(X) = AX + b$ be a linear function of a vector-valued r.v. X (A is a matrix and b is a vector, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\mathbf{cov}[X] = \Sigma$, then for the vector-valued r.v. Y

$$\mathbb{E}[Y] = \mathbb{E}[AX + b] = A\mu + b$$

$$\text{cov}[Y] = \text{cov}[AX + b] = A\Sigma A^\top$$

- Likewise, if $Y = f(X) = a^\top X + b$ be a linear function of a vector-valued r.v. X (a is a vector and b is a scalar, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\mathbf{cov}[X] = \Sigma$, then for the scalar-valued r.v. Y

$$\mathbb{E}[Y] = \mathbb{E}[a^\top X + b] = a^\top \mu + b$$

$$\text{var}[Y] = \text{var}[a^\top X + b] = a^\top \Sigma a$$

Common Probability Distributions

Important: We will use these extensively to model data as well as parameters of models

- Some common discrete distributions and what they can model
  - Bernoulli: Binary numbers, e.g., outcome (head/tail, 0/1) of a coin toss
  - Binomial: Bounded non-negative integers, e.g., # of heads in n coin tosses
  - Multinomial/multinoulli: One of K (>2) possibilities, e.g., outcome of a dice roll
  - Poisson: Non-negative integers, e.g., # of words in a document
- Some common continuous distributions and what they can model
  - Uniform: numbers defined over a fixed range
  - Beta: numbers between 0 and 1, e.g., probability of head for a biased coin
  - Gamma: Positive unbounded real numbers
  - Dirichlet: vectors that sum of 1 (fraction of data points in different clusters)
  - Gaussian: real-valued numbers or real-valued vectors