Probabilistic Machine Learning (1): Some Basics of Probability

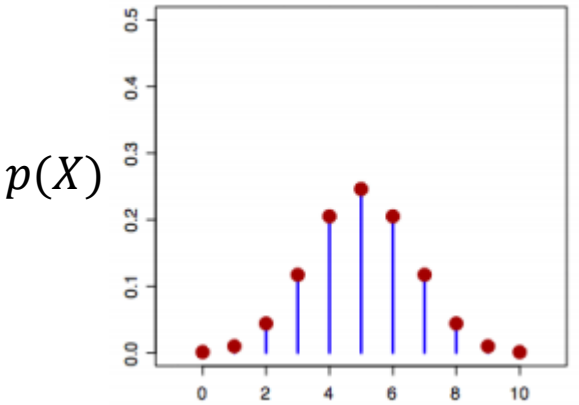
CS771: Introduction to Machine Learning

Piyush Rai

Some Probability Basics

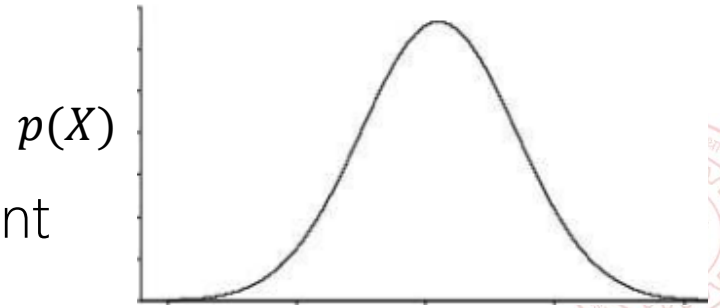
Random Variables

- 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



X (a discrete r.v.)

- 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

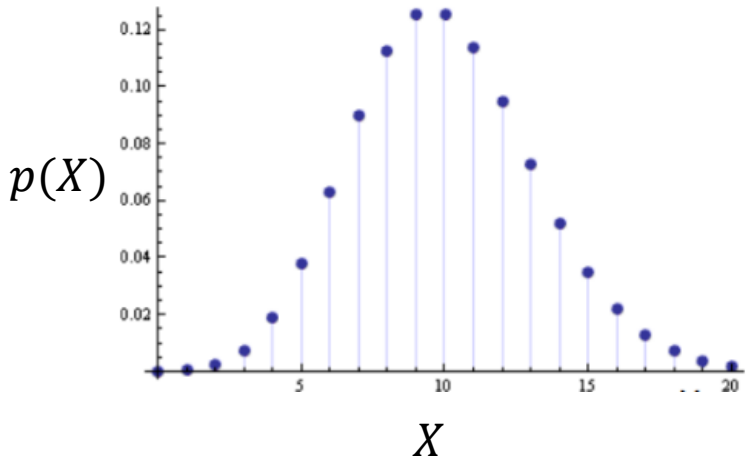


X (a continuous r.v.)

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

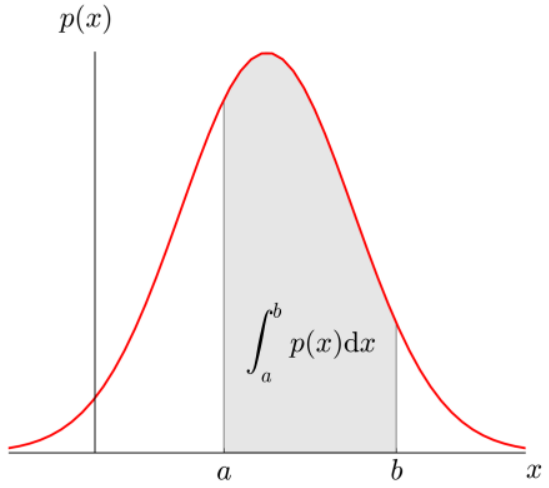
- For cont. r.v., we talk in terms of prob. within an interval $X \in (x, x + \delta x)$
  - $p(x)\delta x$ is the prob. that $X \in (x, x + \delta x)$ as $\delta x \to 0$
  - $p(x)$ is the probability density at $X = x$

Yes, probability density at a point x can very well be larger than 1. The integral however must be equal to 1

$$p(x) \geq 0$$

~~$p(x) \leq 1$~~

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



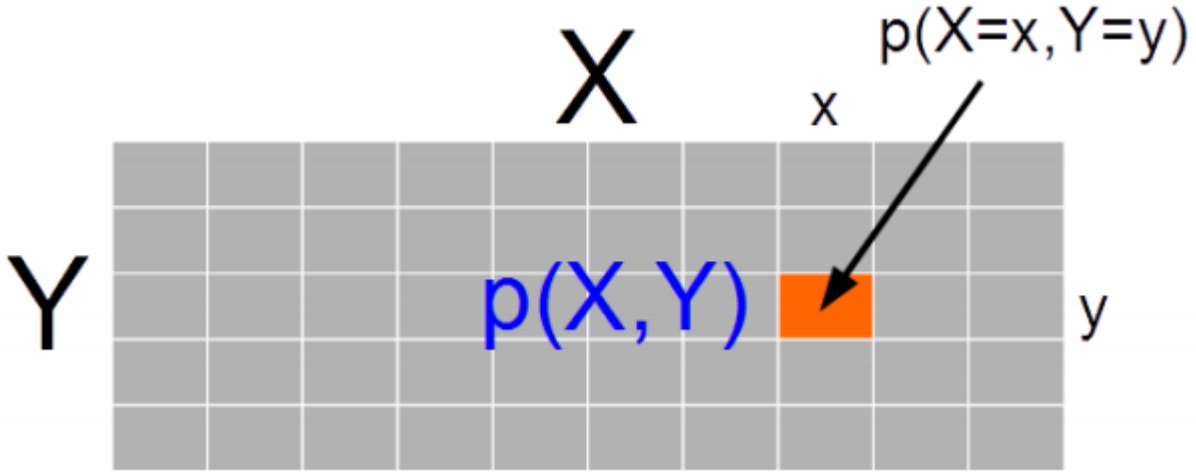
A word about notation

- $p(.)$ 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 $p(X)$

$$x \sim p(X)$$

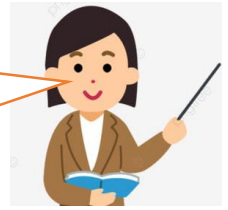
Joint Probability Distribution

- 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)



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

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



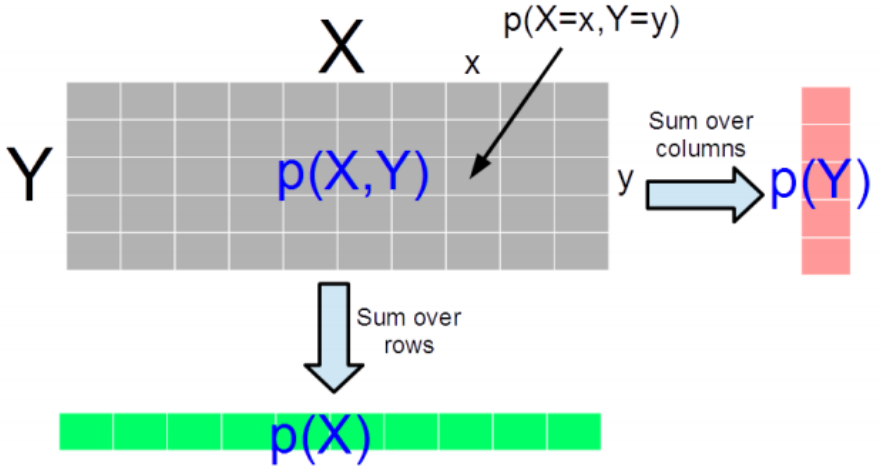
- For two continuous r.v.'s X and Y, we have joint PDF $p(X, Y)$

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

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

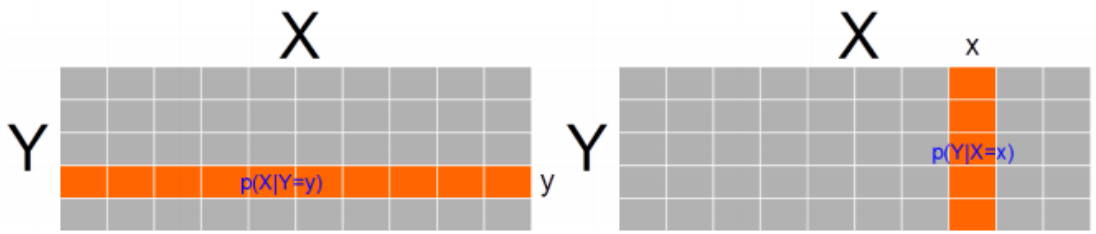
For discrete r.v.'s, marginalization is called summing over, for continuous r.v.'s, it is called "integrating out"

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

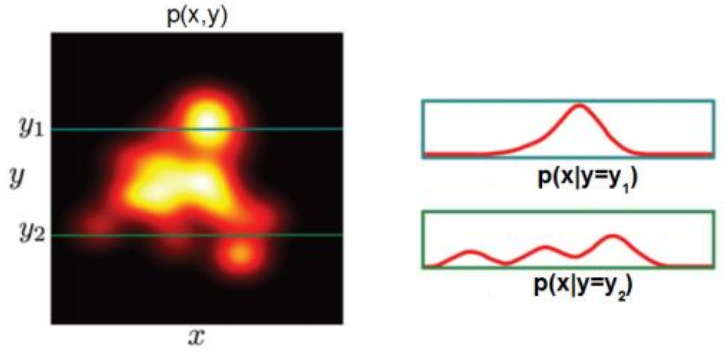
Conditional Probability Distribution

- 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 y given weights w (r.v.) and features X written as $p(y|w, X)$

Some Basic Rules

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

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

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

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

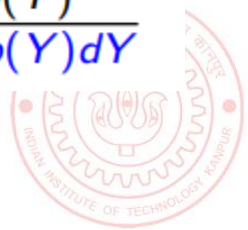
- **Bayes' rule:** Gives conditional probability distribution (can derive it from product rule)

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

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

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

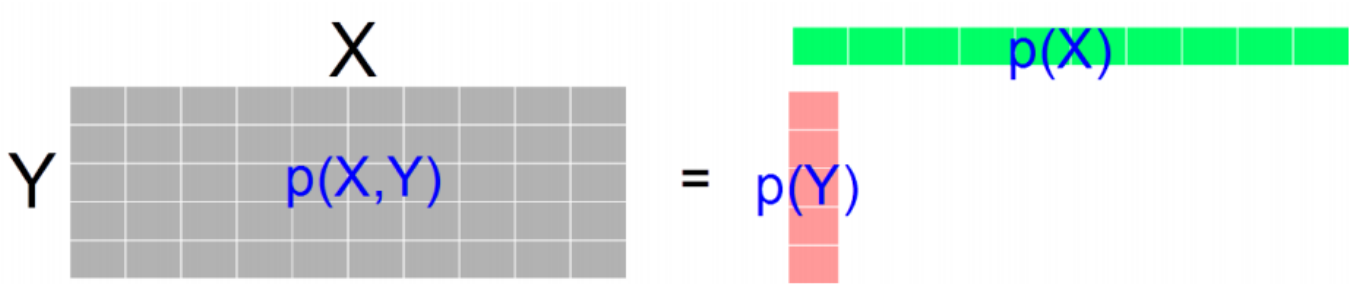
- Chain Rule: $p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2|X_1)\ldots p(X_N|X_1, \ldots, X_{N-1})$



Independence

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

$$p(X|Y=y) = p(X)$$
$$p(Y|X=x) = p(Y)$$
$$p(X,Y) = p(X)p(Y)$$



- 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

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

Coming up next

- Some other basic concepts from probability and statistics
- Probabilistic models and parameter estimation in probabilistic models
  - MLE, MAP, Bayesian approaches