Foundations of Machine Learning

Module 4:
Part A: Probability Basics

Sudeshna Sarkar
IIT Kharagpur

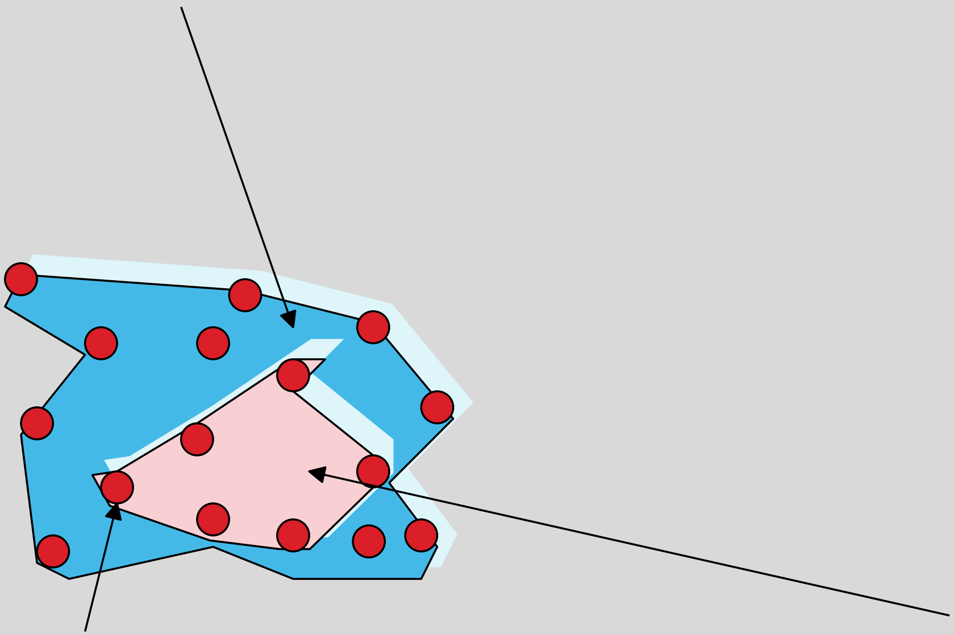
- *Probability* is the study of randomness and uncertainty.
- A *random* experiment is a process whose outcome is uncertain.

  Examples:
  - Tossing a coin once or several times
  - Tossing a die
  - Tossing a coin until one gets Heads
  - …

Events and Sample Spaces

Sample Space

The sample space is the set of all possible outcomes.



Simple Events

The individual outcomes are called simple events.

Event

An event is any collection of one or more simple events

Sample Space

- Sample space Ω : the set of all the possible outcomes of the experiment
  - If the experiment is a roll of a six-sided die, then the natural sample space is {1, 2, 3, 4, 5, 6}
  - Suppose the experiment consists of tossing a coin three times.

    $$\Omega = \{(hhh, hht, hth, htt, thh, tht, tth, ttt\}$$
  - the experiment is the number of customers that arrive at a service desk during a fixed time period, the sample space should be the set of nonnegative integers: $\Omega = Z^+ = \{0, 1, 2, 3, \ldots\}$

Events

- Events are subsets of the sample space
  - A= {the outcome that the die is even} ={2,4,6}
  - B = {exactly two tosses come out tails}=(htt, tht, tth}
  - C = {at least two heads} = {hhh, hht, hth, thh}

Probability

- A Probability is a number assigned to each event in the sample space.
- Axioms of Probability:
  - For any event *A*, 0 ◉ *P*(*A*) ◉ 1.
  - P(⌖) =1 and $P\left(\phi\right)=0$
  - If A_1, A_2, … A_n is a partition of *A*, then

    P(A) = P(A_1) + P(A_2) + ...+ P(A_n)

Properties of Probability

- For any event A, $P(A^c) = 1 - P(A)$.
- If $A \subset B$, then $P(A) \leq P(B)$.
- For any two events A and B,

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

  For three events, A, B, and C,

  $P(A \cup B \cup C) =$

  $P(A) + P(B) + P(C)$

  $- P(A \cap B) - P(A \cap C) - P(B \cap C)$

  $+ P(A \cap B \cap C)$

Intuitive Development (agrees with axioms)

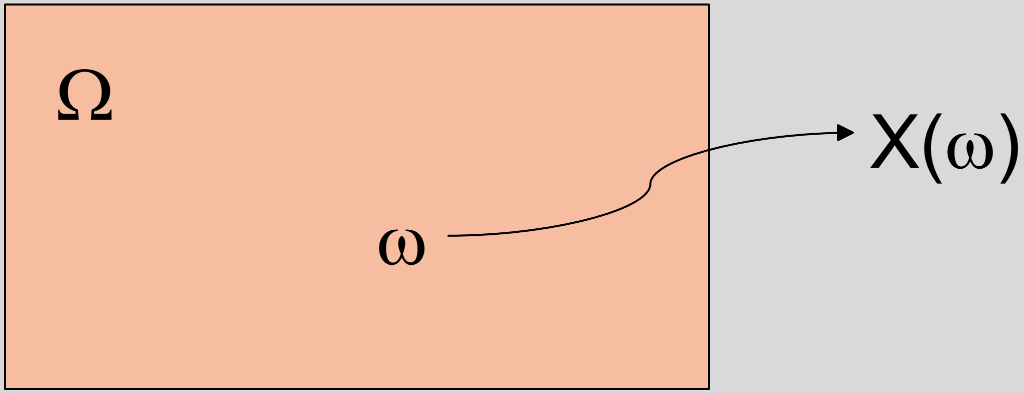
- Intuitively, the probability of an event **a** could be defined as:

$$P(a) = \lim_{n \to \infty} \frac{N(a)}{n}$$

Where N(a) is the number that event a happens in n trials

Random Variable

- A *random variable* is a function defined on the sample space Ω
  - maps the outcome of a random event into real scalar values



Discrete Random Variables

- Random variables (RVs) which may take on only a countable number of distinct values
  - e.g., the sum of the value of two dies

- X is a RV with arity *k* if it can take on exactly one value out of k values,
  - e.g., the possible values that X can take on are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Probability of Discrete RV

- Probability mass function (pmf): $\mathrm{P}\left(\mathrm{X} = x_i\right)$
- Simple facts about pmf
  - $\sum_i \mathrm{P}\left(\mathrm{X} = x_i\right) = 1$
  - $\mathrm{P}\left(\mathrm{X} = x_i \cap \mathrm{X} = x_j\right) = 0$ if $i \neq j$
  - $\mathrm{P}\left(\mathrm{X} = x_i \cup \mathrm{X} = x_j\right) = \mathrm{P}\left(\mathrm{X} = x_i\right) + \mathrm{P}\left(\mathrm{X} = x_j\right)$ if $i \neq j$
  - $\mathrm{P}\left(\mathrm{X} = x_1 \cup \mathrm{X} = x_2 \cup \ldots \cup \mathrm{X} = x_k\right) = 1$

Common Distributions

- Uniform $X \sim U[1, \cdots, N]$
  - X takes values 1, 2, …, *N*
  - $\mathrm{P}(\mathrm{X} = i) = 1/N$
  - E.g. picking balls of different colors from a box
- Binomial $X \sim Bin(n, p)$
  - X takes values 0, 1, …, *n*
  - $\mathrm{P}(\mathrm{X} = i) = \binom{n}{i} p^i (1 - p)^{n-i}$
  - E.g. coin flips

Joint Distribution

- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together
  - e.g.
    you and your friend each toss a coin 10 times
    P(You get 5 heads AND you friend get 7 heads)
- $\sum_x \sum_y \mathrm{P}\left(\mathrm{X} = x \cap \mathrm{Y} = y\right) = 1$

$$\sum_{i=0}^{50} \sum_{j=0}^{100} \mathrm{P}\left(\text{You get } i \text{ heads AND your friend get } j \text{ heads}\right) = 1$$

Conditional Probability

- $\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Y}=y\right)$ is the probability of $X = x$, given the occurrence of $Y = y$
  - E.g. you get 0 heads, given that your friend gets 3 heads

- $$\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Y}=y\right)=\frac{\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)}{\mathrm{P}\left(\mathrm{Y}=y\right)}$$

Law of Total Probability

- Given two discrete RVs X and Y, which take values in $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$, We have

$$
\begin{aligned}
\mathrm{P}(\mathrm{X} = x_i) &= \sum_j \mathrm{P}(\mathrm{X} = x_i \cap \mathrm{Y} = y_j) \\
&= \sum_j \mathrm{P}(\mathrm{X} = x_i \mid \mathrm{Y} = y_j)\,\mathrm{P}(\mathrm{Y} = y_j)
\end{aligned}
$$

Marginalization

Marginal Probability

Joint Probability

$$\mathrm{P}\left(\mathrm{X} = x_i\right) = \sum_j \mathrm{P}\left(\mathrm{X} = x_i \cap \mathrm{Y} = y_j\right)$$

$$= \sum_j \mathrm{P}\left(\mathrm{X} = x_i \middle| \mathrm{Y} = y_j\right)\mathrm{P}\left(\mathrm{Y} = y_j\right)$$

Conditional Probability

Marginal Probability

Bayes Rule

- X and Y are discrete RVs…

$$\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Y}=y\right)=\frac{\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)}{\mathrm{P}\left(\mathrm{Y}=y\right)}$$

$$\Downarrow$$

$$\mathrm{P}\left(\mathrm{X}=x_i \middle| \mathrm{Y}=y_j\right)=\frac{\mathrm{P}\left(\mathrm{Y}=y_j \middle| \mathrm{X}=x_i\right)\mathrm{P}\left(\mathrm{X}=x_i\right)}{\sum_k \mathrm{P}\left(\mathrm{Y}=y_j \middle| \mathrm{X}=x_k\right)\mathrm{P}\left(\mathrm{X}=x_k\right)}$$

Independent RVs

- X and Y are independent means that $X = x$ does not affect the probability of $Y = y$

- Definition: X and Y are independent iff
  - P(XY) = P(X)P(Y)
  - $$\mathrm{P}\left(\mathrm{X} = x \cap \mathrm{Y} = y\right) = \mathrm{P}\left(\mathrm{X} = x\right)\mathrm{P}\left(\mathrm{Y} = y\right)$$

More on Independence

$$\begin{tikzcd}
& \mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)=\mathrm{P}\left(\mathrm{X}=x\right)\mathrm{P}\left(\mathrm{Y}=y\right) \arrow[dl] \arrow[dr] & \\
\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Y}=y\right)=\mathrm{P}\left(\mathrm{X}=x\right) & & \mathrm{P}\left(\mathrm{Y}=y \middle| \mathrm{X}=x\right)=\mathrm{P}\left(\mathrm{Y}=y\right)
\end{tikzcd}$$

- E.g. no matter how many heads you get, your friend will not be affected, and vice versa

Conditionally Independent RVs

- Intuition: X and Y are conditionally independent given Z means that once Z is **known**, the value of X does not add any **additional** information about Y
- Definition: X and Y are conditionally independent given Z iff

$$\mathrm{P}\left(\mathrm{X}=x\cap\mathrm{Y}=y\middle|\mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{X}=x\middle|\mathrm{Z}=z\right)\mathrm{P}\left(\mathrm{Y}=y\middle|\mathrm{Z}=z\right)$$

More on Conditional Independence

$$\mathrm{P}\left(\mathrm{X}=x\cap\mathrm{Y}=y\,\middle|\,\mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{X}=x\,\middle|\,\mathrm{Z}=z\right)\mathrm{P}\left(\mathrm{Y}=y\,\middle|\,\mathrm{Z}=z\right)$$

$$\mathrm{P}\left(\mathrm{X}=x\,\middle|\,\mathrm{Y}=y,\mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{X}=x\,\middle|\,\mathrm{Z}=z\right)$$

$$\mathrm{P}\left(\mathrm{Y}=y\,\middle|\,\mathrm{X}=x,\mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{Y}=y\,\middle|\,\mathrm{Z}=z\right)$$

Continuous Random Variables

- What if X is continuous?
- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable *x*.

PDF

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
  - $f(x) \leq 1$???
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

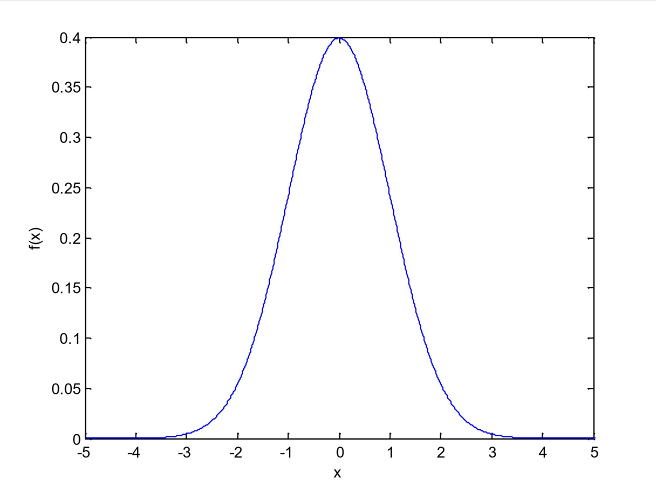
$$\mathrm{P}(0 \leq X \leq 1) = \int_0^1 f(x)dx$$

Cumulative Distribution Function

- $F_{\mathrm{X}}(v) = \mathrm{P}(\mathrm{X} \le v)$
- Discrete RVs
  - $F_{\mathrm{X}}(v) = \sum_{v_i} \mathrm{P}(\mathrm{X} = v_i)$
- Continuous RVs
  - $F_{\mathrm{X}}(v) = \int_{-\infty}^{v} f(x)\,dx$
  - $\frac{d}{dx} F_{\mathrm{X}}(x) = f(x)$

Common Distributions

- Normal $X \sim N(\mu, \sigma^2)$
  - $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R}$
  - E.g. the height of the entire population



Multivariate Normal

- Generalization to higher dimensions of the one-dimensional normal

Covariance Matrix

- $$f_{\vec{X}}\left(x_1,\ldots,x_d\right)=\frac{1}{\left(2\pi\right)^{d/2}\left|\Sigma\right|^{1/2}}$$
$$\cdot\exp\left\{-\frac{1}{2}\left(\vec{x}-\mu\right)^T\Sigma^{-1}\left(\vec{x}-\mu\right)\right\}$$

Mean

Mean and Variance

- Mean (Expectation): $\mu = E(\mathrm{X})$
  - Discrete RVs: $E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x) dx$
- Variance:

$$V(\mathrm{X}) = E(\mathrm{X} - \mu)^2$$

  - Discrete RVs:
  - Continuous RVs: $V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \mathrm{P}(\mathrm{X} = v_i)$

$$V(\mathrm{X}) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Mean Estimation from Samples

- Given a set of N samples from a distribution, we can estimate the mean of the distribution by:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Variance Estimation from Samples

- Given a set of N samples from a distribution, we can estimate the variance of the distribution by:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2$$

Thank You