#Code

Artificial Intelligence and Machine Learning (AIML)

2023–24

- **Module so far**:
  - Symbolic AI
  - Logical reasoning
  - Machine Learning
    - Unsupervised technique (clustering)
    - Supervised techniques
      - Regression and classification
      - Neural networks and deep learning

- **This lecture**: statistical machine learning
  - Probability and probabilistic AI, probabilistic graphical models

Quantifying uncertainty in probabilistic AI and ML

#Code

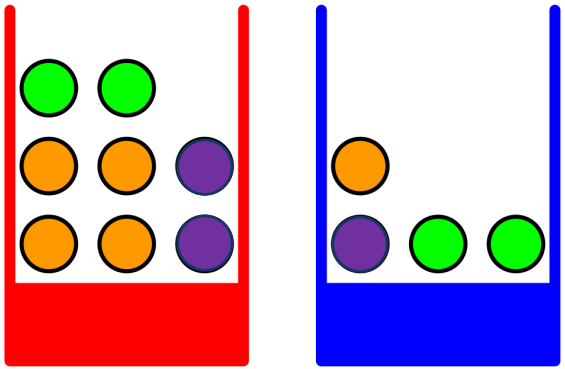
- Most of real-world problems need to handle **uncertainty**
  - Noise on measurements
  - Partial observability

- In these situations, the optimal decision will be taken under uncertainty and depends on both the relative importance of various goals and the likelihood that it can be achieved.
  - **Example:** uncertain logical reasoning: toothache diagnosis
    - toothache ⇒ cavity 口腔 ✗ *not all patients with toothaches have cavities*
    - toothache ⇒ (cavity) ∨ (gum problem) ∨ (abscess) ✗ *what is the size of the list?*
    - cavity ⇒ tootache 脓肿 ✗ *not all cavities cause pain*
    - Makes more sense to provide a **degree of belief** to a given proposition

Probability Theory

#Code

- Consistent framework for the quantification and manipulation of uncertainty
- Probability provides a calculus for **random events**, conveniently indexed using **random variables** (RVs)



Identity of the box is a random variable: B

Two possible values (outcomes): red ($B = r$) or blue ($B = b$)

Suppose we are presented with the red box 40% of the time:

$$P(B = r) = \frac{40}{100} = 0.4$$

If the two boxes are the only possibilities we have, what would be the chances of being presented with the blue box? 1-0.4=0.6

What would be the chances of being presented with either the red or the blue boxes? 1

Probability Theory

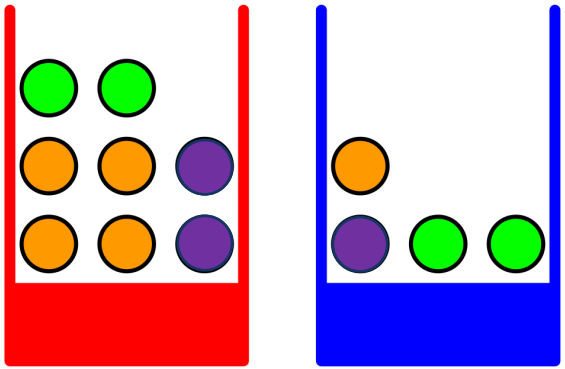
#Code

- **Rules (axioms) of probability**

(i) $P(A) \geq 0$

(ii) $P(A \cup B) = P(A) + P(B)$, if events are non-overlapping (i.e., $A \cap B = \emptyset$) ∅ = empty set

(iii) if Ω represents all possible events (e.g., $\Omega = \{A, B, \ldots, N\}$), then $P(\Omega) = 1$



Identity of the box is a random variable: B

Two possible values (outcomes): red ($B = r$) or blue ($B = b$)

Suppose we are presented with the red box 40% of the time:

$$P(B = r) = \frac{40}{100} = 0.4$$

If the two boxes are the only possibilities we have, what would be the chances of being presented with the blue box?

What would be the chances of being presented with either the red or the blue boxes?

Probability distribution functions

#Code

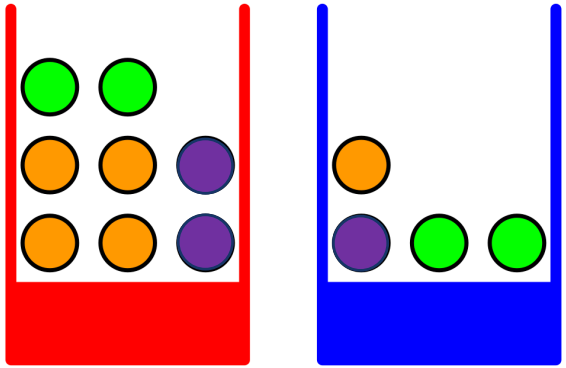
- Informs how are the “chances” of obtaining outcomes for a random variable, which can be **discrete** or **continuous**.

  - **Discrete probability mass function (PMF)** gives the probability that the RV X takes on the value x (i.e., $P(X = x)$)
    - Each possibility lies in the range $[0,1]$ and must have $\sum_{x \in \Omega} P(X = x) = 1$ (normalization).
  - **Continuous probability density function (PDF)** gives the amount of **probability per unit** (probability density)
    - must satisfy $p(x) \geq 0$ for all $x \in \Omega_X$;
    - volume under this function gives the probability of the event represented by that volume, $\int_A p(x)\, dx = \mathrm{P}(A)$, and it must be normalized, $\int_{\Omega X} p(x)\, dx = 1$.

The probability of every event x is greater or equal to 0

Probability distribution functions

#Code



Identity of the box is a random variable: B

Two possible values (outcomes): red ($B = r$) or blue ($B = b$)

$$P(B = r) = 0.4$$

$$P(B = b) = 0.6$$

Bernoulli Distribution

- Binary distribution
- **Sample space**: $\Omega = \{r,b\}$
- **Mass function (PMF)**:

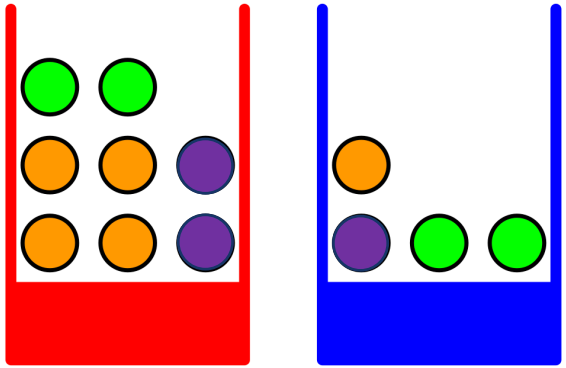
$$P(B = r) = p \qquad p \in [0,1]$$
$$P(B = b) = 1 - p$$

- **Normalization**:

$$\sum_{x \in \Omega} P(X = x) = P(X = r) + P(X = b)$$
$$= p + 1 - p = 1$$

Probability distribution functions

#Code



Colour of the ball as a random variable: C

Three possible values (outcomes):

- green ($C = g$)
- orange ($C = o$)
- purple ($C = p$)

Categorical Distribution

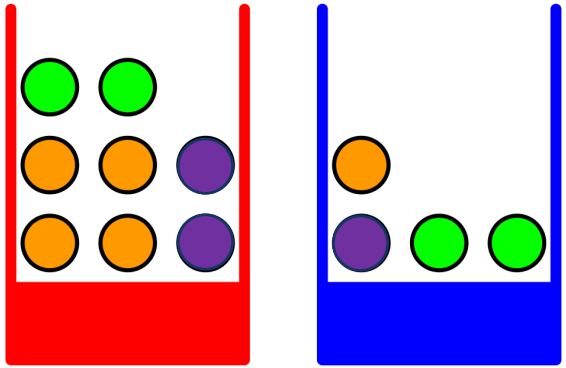
- Discrete distribution
- **Sample space**: $\Omega = \{1, 2, \ldots, N\}$
- **Mass function (PMF)**: each outcome can have a different probability, so $P(X = x) = p_x$ (i.e., parameter vector with each $p_x \in [0,1]$)
- **Normalization**:

$$\sum_{x \in \{1,2,\ldots,N\}} p_x = 1$$

The sum of the probability of each color category

Probability distribution functions

#Code



Colour of the ball as a random variable: C

Three possible values (outcomes):

- green ($C = g$)
- orange ($C = o$)
- purple ($C = p$)

Uniform Distribution

- Special case of the categorical distribution where each outcome is equally likely
- **Sample space**: $\Omega = \{1, 2, \ldots, N\}$
- **Mass function (PMF)**: each outcome can have a different probability, so

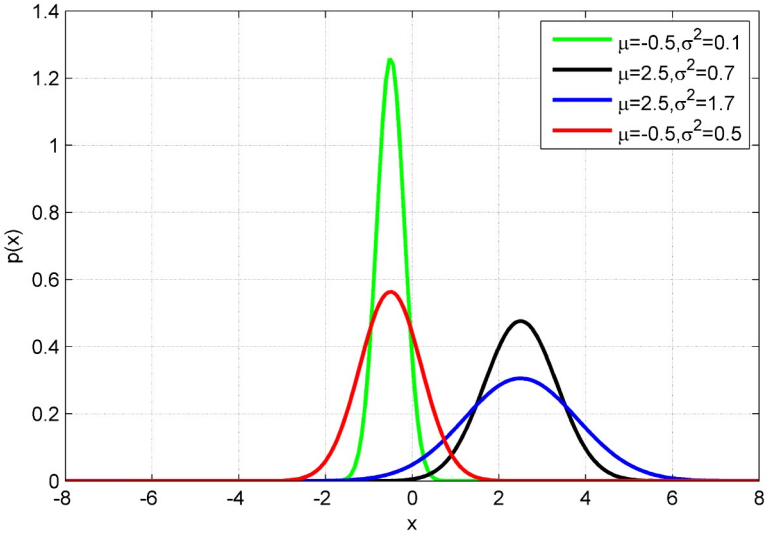
$$P(X = x) = p = \frac{1}{N}, \qquad p \in [0,1]$$

- **Normalization**:

$$\sum_{x \in \{1,2,\ldots,N\}} \frac{1}{N} = N \times \frac{1}{N} = 1$$

Probability distribution functions

#Code



Gaussian (Normal) distribution

- Most ubiquitous continuous distribution
- **Sample space**: $\Omega = \mathbb{R}$
- **Density function (PDF)**:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

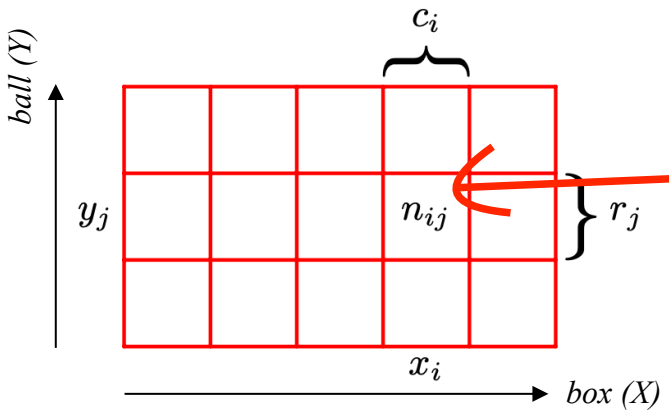
with **mean** $\mu \in \mathbb{R}$ and **variance** $\sigma^2 > 0$.

- **Notes:** μ is both **mean**, **median** and **mode** of the density

Other probabilities intuition

#Code

- Extension to *M* boxes and *L* colors for each ball



Consider *N* trials in which we sample a ball within a box.

What is the probability that we will take a box $X = x_i$ and ball $Y = y_i$?

Joint Probability

- Two or more events occurring simultaneously are represented by **joint RVs**, and corresponding **joint PMFs and/or PDFs**
- For this case, $P(X = x_i, Y = y_i) = \frac{n_{ij}}{N}$
- The joint probability should also be normalized. For the discrete case,

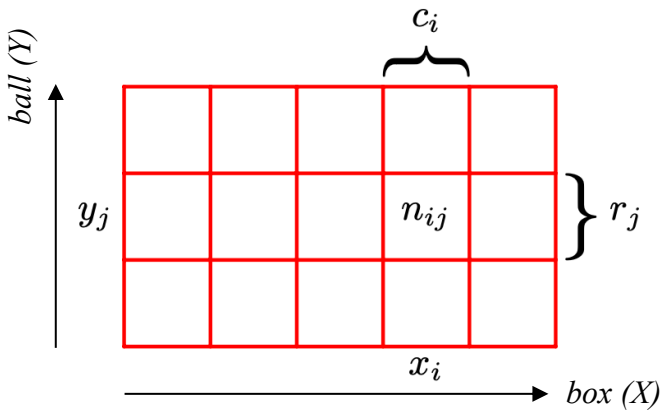
$$\sum_{x \in \Omega_X, y \in \Omega_Y} P(X = x, Y = y) = 1$$

- We often use the simplified notation $P(X, Y)$ to indicate $P(X = x, Y = y)$, where there is no ambiguity.

Other probabilities intuition

#Code

- Extension to *M* boxes and *L* colors for each ball



Consider *N* trials in which we sample a ball within a box.

What is the probability that we will take a box $X = x_i$, irrespective of the ball $Y = y_i$ we take?

Marginal Probability

- A joint distribution contains all information about the RVs; we can find the distribution of one of the RVs from the joint, by **marginalizing** (i.e., summing out) the other RVs
- For this case,

$$P(X = x_i) = \frac{c_i}{N}$$

- Since $c_i = \sum_j n_{ij}$, we can also write

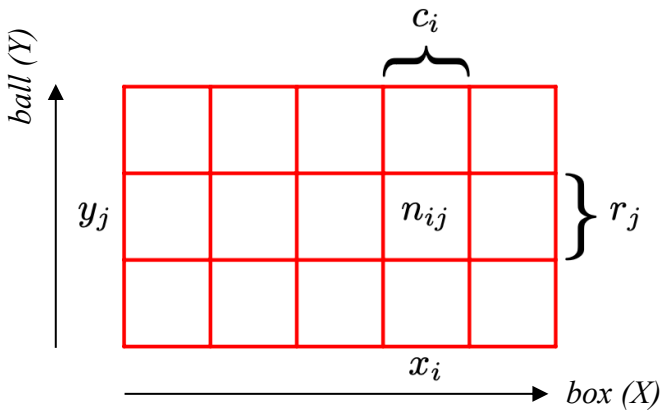
$$P(X = x) = \sum_{y \in \Omega_Y} P(X = x, Y = y)$$

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

Other probabilities intuition

#Code

- Extension to *M* boxes and *L* colors for each ball



Consider *N* trials in which we sample a ball within a box.

What is the probability that we will take a ball $Y = y_i$, given we took box $X = x_i$ already?

Conditional Probability

- Properly normalized, fixing one variable allows the joint to behave as a new distribution called the **conditional**.
- For this case,

$$P(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i}$$

- Since $p(X = x, Y = y) = \frac{n_{ij}}{N}$,

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

$$P(X, Y) = p(Y|X)p(X) \text{ or } P(X, Y) = p(X|Y)p(Y)$$

Marginals and conditionals: another example (Table 15.1)

#Code

Joint $P(X,Y)$	$y=0$	$y=1$
$x=0$	$\frac{3}{7}$	$\frac{1}{7}$
$x=1$	$\frac{3}{15}$	$\frac{8}{35}$

Marginal $P(X)$	
$x=0$	$P(X=0,Y=0)+P(X=0,Y=1)=\frac{4}{7}$
$x=1$	$P(X=1,Y=0)+P(X=1,Y=1)=\frac{3}{7}$

Marginal $P(Y)$	
$y=0$	$P(X=0,Y=0)+P(X=1,Y=0)=\frac{22}{35}$
$y=1$	$P(X=0,Y=1)+P(X=1,Y=1)=\frac{13}{35}$

Conditional $P(X\|Y)$	$y=0$	$y=1$
$x=0$	$\frac{P(X=0,Y=0)}{P(Y=0)}=\frac{15}{22}$	$\frac{P(X=0,Y=1)}{P(Y=1)}=\frac{5}{13}$
$x=1$	$\frac{P(X=1,Y=0)}{P(Y=0)}=\frac{7}{22}$	$\frac{P(X=1,Y=1)}{P(Y=1)}=\frac{8}{13}$

Conditional $P(Y\|X)$	$y=0$	$y=1$
$x=0$	$\frac{P(X=0,Y=0)}{P(X=0)}=\frac{3}{4}$	$\frac{P(X=0,Y=1)}{P(X=0)}=\frac{1}{4}$
$x=1$	$\frac{P(X=1,Y=0)}{P(X=1)}=\frac{7}{15}$	$\frac{P(X=1,Y=1)}{P(X=1)}=\frac{8}{15}$

Probabilistic graphical models

#Code

- Given a probability distribution with multiple RVs, we can represent their mutual conditional dependence graphically using a **probabilistic graphical model** (PGM)
- Consider an arbitrary joint distribution over 3 RVs: X, Y, Z

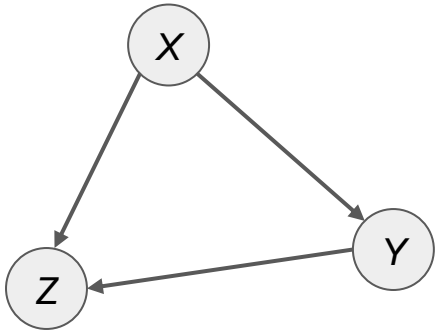
$$P(X, Y, Z) = P(Z|X, Y)P(X, Y)$$

- Applying the product rule again on $P(X, Y) = P(Y|X)P(X)$, we have

$$P(X, Y, Z) = P(Z|X, Y)P(Y|X)P(X)$$

- **Note:** While the left-hand side is symmetrical w.r.t. the 3 RVs, the right-hand side is not; we have chosen a particular ordering that yielded a specific decomposition.

There are as many different forms of the chain rule, as there are **permutations** of the variables



- One node for each RV
- Associate each node with the corresponding conditional distribution

- X is **parent** of Y, Z
- Y is **child** of X

AIML

Independence of RVs

#Code

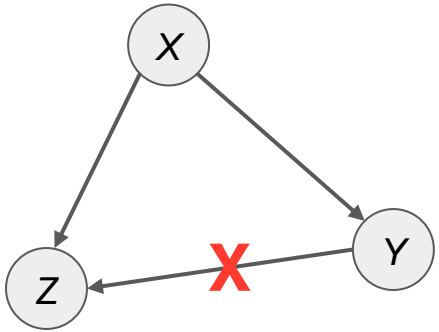
- Two RVs are **independent** if and only if their joint distribution factors into a product of marginals:

$$P(X = x, Y = y) = P(X = x)P(Y = y), \qquad \forall x \in \Omega_X, y \in \Omega_Y$$

- Since $P(X, Y) = p(Y|X)p(X)$, the condition above implies that

$$P(Y|X) = p(Y), \qquad \forall y \in \Omega_Y$$

- We can read the above as "X does NOT add any information about Y", or "knowing X does not change our belief in Y"



- If Z is **conditionally independent** on Y in the example before, then

$$P(Z|X, Y) = P(Z|X)$$

PGMs in Probabilistic AI and ML

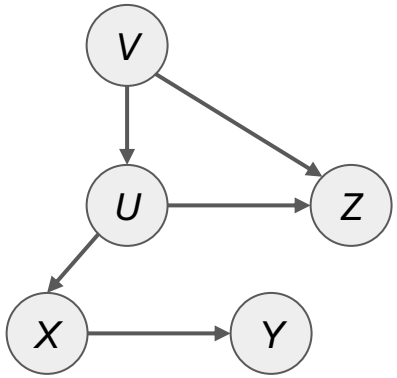
#Code

- Independence/conditional independence such as this, is critical important in probabilistic AI and ML, since it allows joint RVs to be modelled by subsets of the data; basis of **probabilistic graphical models**
    - Special graph structures for applications such as time or spatially ordered data
- The graph must have **no cycles** (loops), a **directed acyclic graph** (DAG)
- Every PGM has a corresponding **Markov factorization**, which is a form of the chain rule compatible with a **topological ordering** of the DAG, taking into account the conditional independencies due to the absence of edges

PGMs: Markov factorization

#Code

- Markov factorization of the PGM below, considering the specific topological order $[V, U, X, Y, Z]$



$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X,U,V,Z)$$

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(U,V,Z)$$

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U,V)$$

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U|V)P(V)$$

Simplify using conditional independence:

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U|V)P(V)$$

$$P(U,V,X,Y,Z) = P(V)P(U|V)P(X|U)P(Y|X)P(Z|U,V)$$

#Code

To recap

- We touched on key probabilistic concepts to build a **probabilistic graphical model (PGM)**
- Learned how to find a **Markov factorization** of a given PGM, considering a topological ordering

- **Next**: Bayesian Models and probabilistic classification using naïve Bayes

Further Reading

- **PRML**, Section 1.2, Section 8.1-8.2
- **MLSP**, Section 1.4, Section 1.6, Section 5.1-5.2
- **R&N**, Section 14.1-14.2