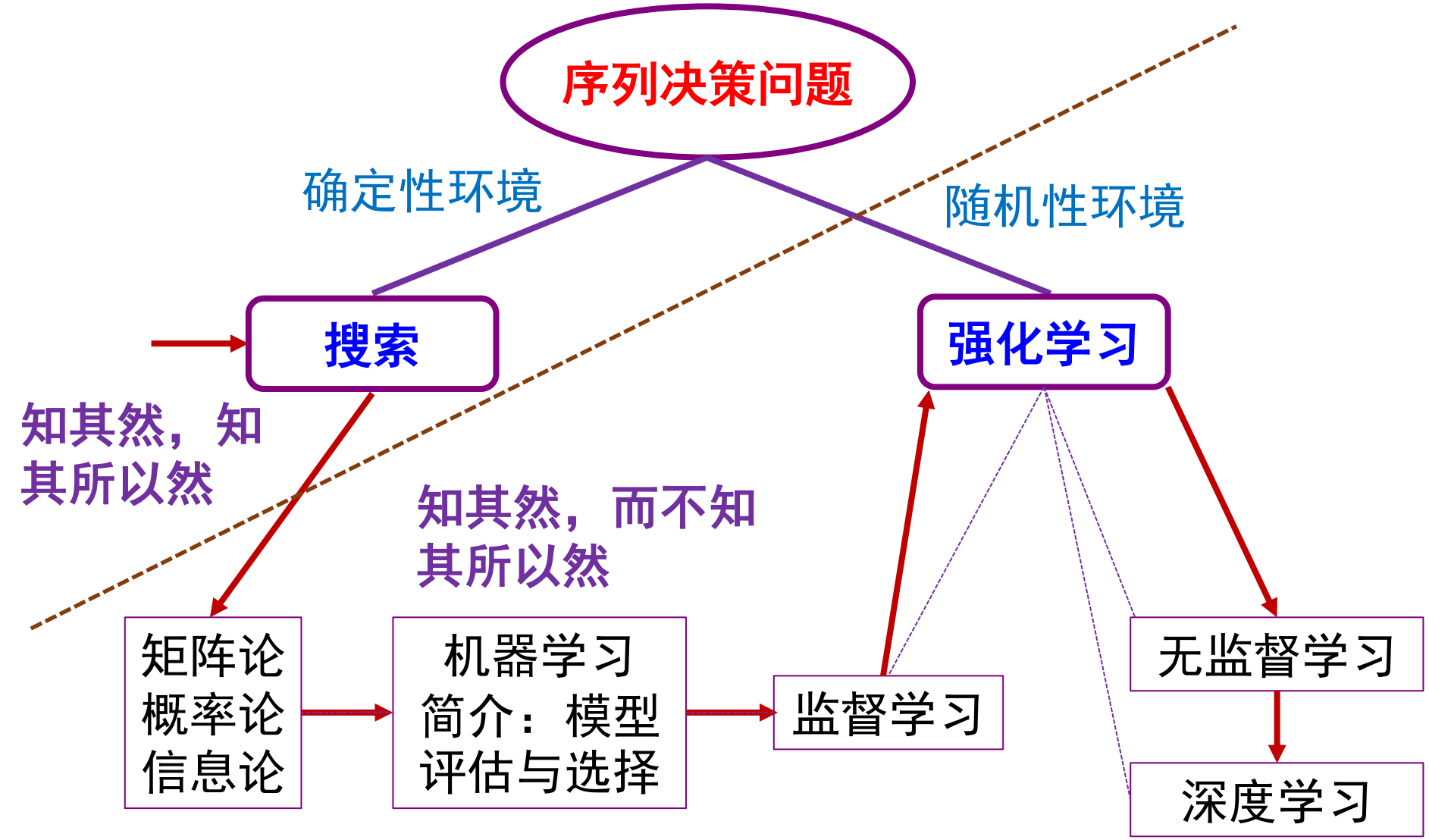
# 人工智能 I

# 本周四实验课安排

| | | |
|---|---|---|
| **周次** | 7 | |
| **日期** | 2025/11/06 | |
| **星期** | 星期四 | |
| **时间** | 16:40-18:15 | 19:00-20:35 |
| **机房地点** | 中心楼实验室<br>D区 | 五五楼435<br>软实七 |
| **实验任务** | 对抗搜索 | |

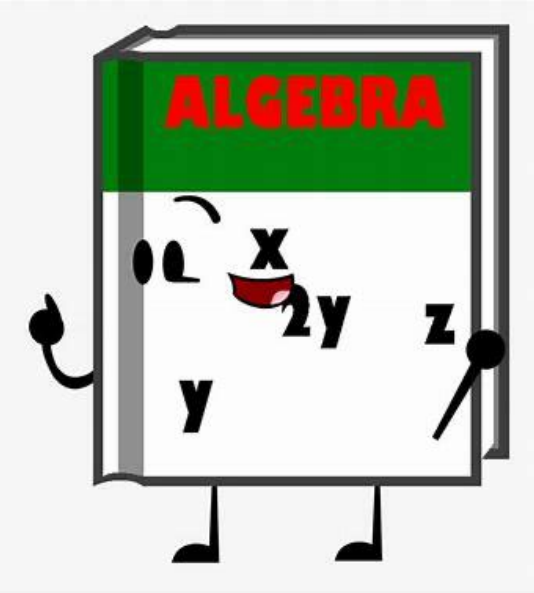
**注意两个时间段是不同机房！请勿跑错！**
**注意18:15机房电脑自动关机！请及时保存！**

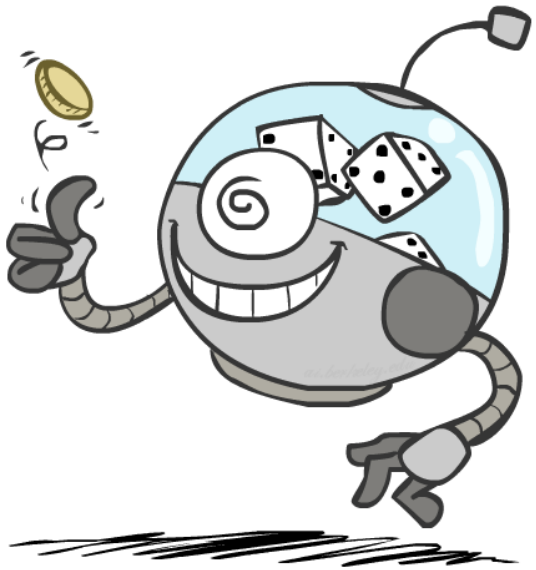
# 课程内容


序列决策问题
确定性环境
随机性环境
搜索
强化学习
知其然，知其所以然
知其然，而不知其所以然
矩阵论
概率论
信息论
机器学习
简介：模型
评估与选择
监督学习
无监督学习
深度学习

# 本节课安排

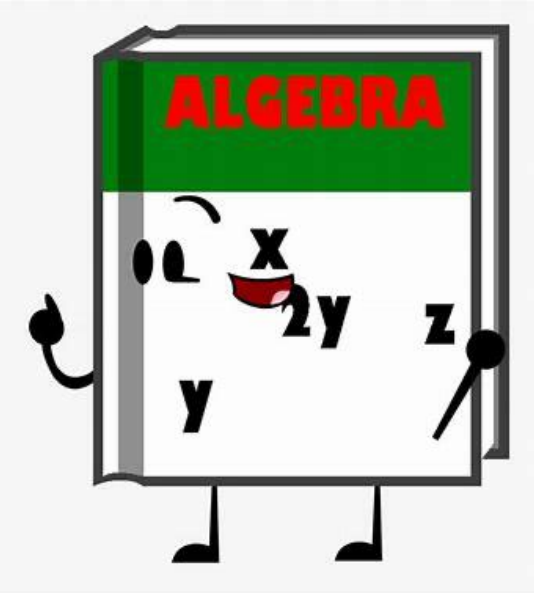
- 回顾：基于搜索的问题求解
- 机器学习的数学基础
  - 矩阵论
  - 概率论
  - 信息论

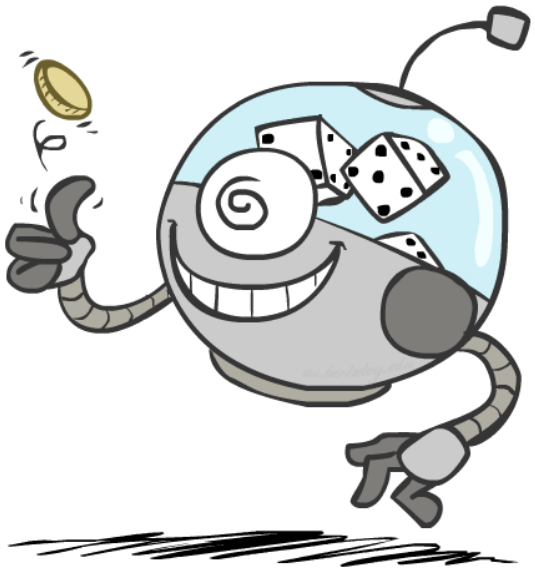


# 本节课安排

- 回顾：基于搜索的问题求解
- 机器学习的数学基础
  - 矩阵论
  - 概率论
  - 信息论



# 基于搜索的问题求解

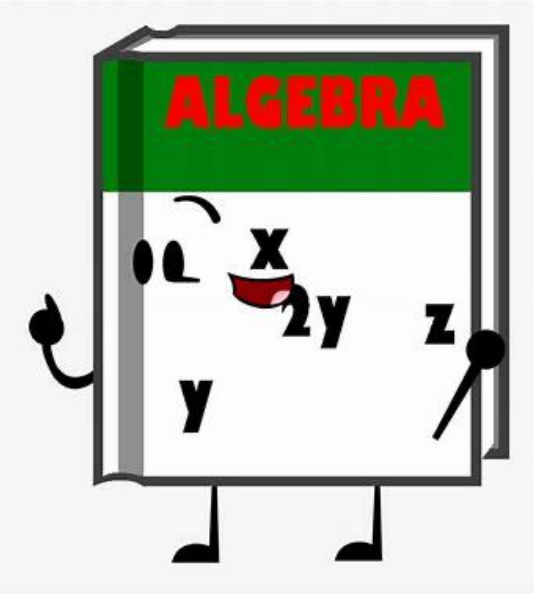
① **搜索问题（路径规划问题）**
- 无信息搜索
  - 广度优先、深度优先、代价一致
- 有信息搜索
  - 贪婪算法、A*算法

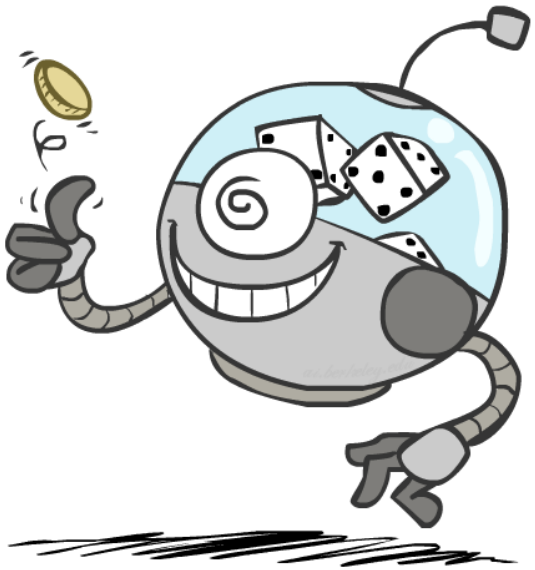
② **博弈问题**
- 两位玩家、零和博弈
  - 极小极大算法
  - $\alpha-\beta$剪枝

③ **约束满足问题**
- 回溯搜索、及其改进方法：
  - 预处理（静态相容性检查）：节点相容、弧相容（AC-3算法）
  - 排序：
    - 变量排序：最少剩余值、度启发式
    - 值的排序：最少约束值
  - 推理（动态相容性检查）：前向检验、弧相容保持
  - 结构：独立子问题、树结构

# 本节课安排

- 回顾：基于搜索的问题求解
- 机器学习的数学基础
  - 矩阵论
  - 概率论
  - 信息论



# 矩阵与向量

- 矩阵
  - m * n个数排成m行n列，简记为A

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

  - 当m=n时，则称A是n阶方阵
- 向量
  - 一种特殊的矩阵

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

# 矩阵运算

- 矩阵加法
  - 矩阵$A$、$B$均为m * n矩阵，则m * n矩阵$C = A + B$，称为矩阵$A$与$B$的和
- 矩阵数乘
  - 矩阵$A$为m * n矩阵，k为常数，k与矩阵$A$每个元素相乘，称为数k与矩阵$A$的数乘
- 矩阵乘法
  - 矩阵$A$是m * n矩阵，
  - 矩阵$B$是n * s矩阵，
  - m * s矩阵$C = A \times B$（or $C = AB$）称为矩阵$A$与矩阵$B$的乘法，其中，

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

# 矩阵运算

- 矩阵的乘法不满足交换律：

$$A \times B \neq B \times A$$

- 矩阵的乘法满足结合律：

$$A \times (B \times C) = (A \times B) \times C$$

# 逆、转置

- 单位矩阵
  - 从左上角到右下角的对角线（称为主对角线）上的元素均为1，其余全都为0，一般用I表示
  - 性质

$$AA^{-1} = A^{-1}A = I$$

- 逆
  - 如矩阵A是一个m * m 矩阵（方阵），如果A可逆，则：

$$AA^{-1} = A^{-1}A = I$$

# 逆、转置

- m * n 矩阵A，其转置矩阵大小为 n * m，其中的元素为：

$$(A^T)_{ij} = A_{ji}$$

- 基本性质

$$(A \pm B)^T = A^T \pm B^T$$

$$(A \times B)^T = B^T \times A^T \qquad (AB)^T = B^T A^T$$

$$\left(A^T\right)^T = A$$

# 范数（Norm）

- 在机器学习中，数据通常表示为向量或矩阵，范数是描述数据性质的重要指标
- 向量范数
  - 0-伪范数，向量非0元素的个数：

$$\| \mathbf{x} \|_0 = \#\{i: x_i \neq 0\}$$

    - 稀疏性
  - 1-范数，向量元素绝对值之和

$$\| \mathbf{x} \|_1 = \sum_i |x_i|$$

    - 通常用于替换0-伪范数。
  - 2-范数，即向量元素的平方和再开方

$$\| \mathbf{x} \|_2 = \sqrt{\sum_i x_i^2}$$

# 范数（Norm）

- 矩阵范数
  - 把矩阵$A \in \mathbb{R}^{m\times n}$看成$mn$维的向量
    - F−范数，即矩阵元素的平方和再开方

$$\| A \|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

    - 0−伪范数，即矩阵非0元素的个数，稀疏性

$$\| A \|_0 = \#\{(i,j): a_{ij} \neq 0\}$$

    - 1−范数，即矩阵元素的绝对值求和，替换0−伪范数

$$\| A \|_1 = \sum_{i,j} |a_{ij}|$$

    - ∞−范数，即矩阵元素的绝对值的最大值

$$\| A \|_\infty = \max_{i,j} |a_{ij}|$$

# 线性空间

- 所有$m$维向量构成的集合记为$\mathbb{R}^m$，称为欧式空间，为特殊的线性空间
- **定义**：设$\mathcal{V}$是非空集合，$\mathbb{Q}$是数域。如果$\mathcal{V}$满足如下10条性质，则称$\mathcal{V}$是数域$\mathbb{Q}$上的线性空间
  - 加法：对于$\mathbf{x} \in \mathcal{V}, \mathbf{y} \in \mathcal{V}$，有唯一和$\mathbf{x}+\mathbf{y} \in \mathcal{V}$（封闭性）
    - 结合律：$\mathbf{x}+(\mathbf{y}+\mathbf{z})=(\mathbf{x}+\mathbf{y})+\mathbf{z}$
    - 交换律：$\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$
    - 零元律：存在零元素$\mathbf{0}$，使得$\mathbf{x}+\mathbf{0}=\mathbf{x}$
    - 负元律：存在负元素$-\mathbf{x}$，使得$\mathbf{x}+(-\mathbf{x})=\mathbf{0}$
  - 乘法：对于$\mathbf{x} \in \mathcal{V}, \alpha \in \mathbb{Q}$，有唯一积$\alpha\mathbf{x} \in \mathcal{V}$（封闭性）
    - 数因子分配率：$\alpha(\mathbf{x}+\mathbf{y})=\alpha\mathbf{x}+\alpha\mathbf{y}$
    - 分配率：$(\alpha+\beta)\mathbf{x}=\alpha\mathbf{x}+\beta\mathbf{x}$
    - 结合律：$\alpha(\beta\mathbf{x})=(\alpha\beta)\mathbf{x}$
    - 恒等率：$1\mathbf{x}=\mathbf{x}$

# 线性空间

- **性质**：零元素是唯一的，负元素也是唯一的

$$0\mathbf{x}=\mathbf{0},(-1)\mathbf{x}=-\mathbf{x}$$

- 例1：$\mathbb{R}$是实数域$\mathbb{R}$上的线性空间
- 例2：$\mathbb{R}^m$是实数域$\mathbb{R}$上的线性空间

加法

$$\begin{bmatrix}x_1\\x_2\\\vdots\\x_m\end{bmatrix}+\begin{bmatrix}y_1\\y_2\\\vdots\\y_m\end{bmatrix}=\begin{bmatrix}x_1+y_1\\x_2+y_2\\\vdots\\x_m+y_m\end{bmatrix},$$

乘法

$$\alpha\begin{bmatrix}x_1\\x_2\\\vdots\\x_m\end{bmatrix}=\begin{bmatrix}\alpha x_1\\\alpha x_2\\\vdots\\\alpha x_m\end{bmatrix}$$

- 例3：$\mathcal{V}=\{$全体正实数$\}$，其加法与乘法定义分别为

$$\mathbf{x}\oplus\mathbf{y}=\mathbf{xy},\alpha\otimes\mathbf{x}=\mathbf{x}^{\alpha}$$

则$\mathcal{V}$是实数域$\mathbb{R}$上的线性空间

- 例4：$\mathbb{R}^{m\times n}$是实数域$\mathbb{R}$上的线性空间

# 线性相关性

- 线性组合：$\forall \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathcal{V}, \alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{Q}$,则$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_n\mathbf{x}_n$为元组$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$的线性组合。
- 线性表示：$\mathcal{V}$中的某个元素$\mathbf{x}$可表示为某元组的线性组合，则称$\mathbf{x}$可由该元组线性表示
- 线性相关：如果存在不全为0的数$\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{Q}$，使得

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_n\mathbf{x}_n = 0$$

则称元组$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$线性相关，否则称其线性无关。

例1：线性空间$\mathbb{R}, \mathbf{x}_1 = 1, \mathbf{x}_2 = 2$线性相关

例2：线性空间$\mathbb{R}^2$

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

# 基与坐标

- 设$\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r$是线性空间$\mathcal{V}$上的线性无关组，且$\mathcal{V}$中任意元素可由$\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r$线性表示，则$\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r$是$\mathcal{V}$的一组基，且维数$\dim(\mathcal{V}) = r$.
- 若元素$\mathbf{x}$在基$\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r$下的线性表示为

$$\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a_2} + \cdots + x_r\mathbf{a}_r$$

则称$(x_1, x_2, \cdots, x_r)$为$\mathbf{x}$在该基（坐标系）下的坐标。

- $\mathbb{R}^2$中的向量$\mathbf{x} = \begin{bmatrix}1\\1\end{bmatrix}$

$$\mathbf{a}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \mathbf{a}_2 = \begin{bmatrix}0\\1\end{bmatrix} \quad (1,1)$$

$$\mathbf{a}_1 = \begin{bmatrix}1\\1\end{bmatrix}, \mathbf{a}_2 = \begin{bmatrix}0\\1\end{bmatrix} \quad (1,0)$$

例：$\mathbb{R}^{m\times n}$是实数域$\mathbb{R}$上的线性空间，它的一组基$\{E_{ij}, i = 1, \cdots, m, j = 1, \cdots, n\}$

$$E_{ij} = \mathbf{e}_i\mathbf{e}_j^T$$

# 标准正交基

- 设$\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r$是线性空间$\mathcal{V}$的一组基，且

$$\langle \mathbf{a}_i, \mathbf{a}_j \rangle = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$$

则$\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r$是线性空间$\mathcal{V}$的一组标准正交基

例：$\mathbb{R}^n$的一组标准正交基：

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

# 线性子空间

- 定义：设$\mathcal{V}_1$是数域$\mathbb{Q}$上线性空间$\mathcal{V}$的非空子集，且对$\mathcal{V}$上的加法与数乘封闭：

  ① 如果$\mathbf{x} \in \mathcal{V}_1, \mathbf{y} \in \mathcal{V}_1$，那么$\mathbf{x}+\mathbf{y} \in \mathcal{V}_1$

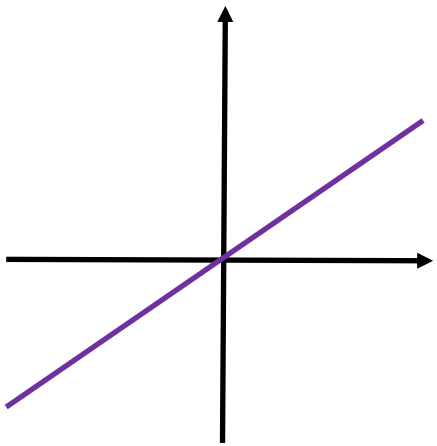
  ② 如果$\mathbf{x} \in \mathcal{V}_1, \alpha \in \mathbb{Q}$，那么$\alpha\mathbf{x} \in \mathcal{V}_1$

  则称$\mathcal{V}_1$是$\mathcal{V}$的（线性）子空间。

- 子空间的生成：设$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$是线性空间$\mathcal{V}$上的$n$个元素，令$\mathcal{V}_1$为$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$所有可能的线性组合的集合，$\mathcal{V}_1 = \{\sum_{i=1}^{n} \alpha_i \mathbf{x}_i\}$。那么$\mathcal{V}_1$是$\mathcal{V}$的子空间，称为由$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$张成的子空间，记为$\mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$



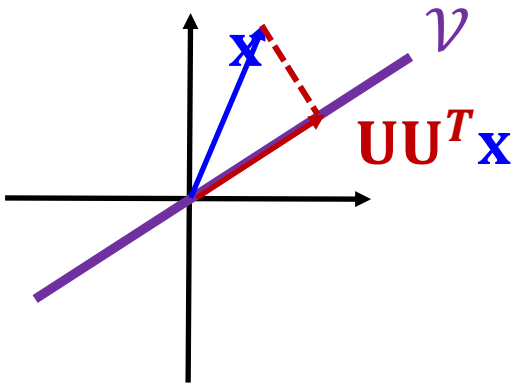
# 投影矩阵

- **问题：**
  - $\mathbf{x} = [x_1; x_2; \cdots; x_d] \in \mathbb{R}^d$，$\mathcal{V}$为$\mathbb{R}^d$的$r$维子空间，$\mathbf{u}_1, \cdots, \mathbf{u}_r$为$\mathcal{V}$的一组标准正交基。基矩阵：

$$\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_r] \in \mathbb{R}^{d \times r}, \mathbf{U}^T\mathbf{U} = I$$

  - 寻找$\mathcal{V}$中与$\mathbf{x}$距离最小的点

$$\min_{\mathbf{y} \in \mathcal{V}} \| \mathbf{y} - \mathbf{x} \|_2$$




- **解：**设 $\mathbf{y} = \sum_{i=1}^{r} w_i \mathbf{u}_i = \mathbf{U}\mathbf{w}$，可得

$$\mathbf{w}^* = \mathbf{U}^T\mathbf{x},\ \mathbf{y}^* = \mathbf{U}\mathbf{U}^T\mathbf{x}$$

  - $\mathbf{UU}^T\mathbf{x}$为$\mathbf{x}$到子空间$\mathcal{V}$上的正交投影在原坐标系$\mathbf{e}_1, \cdots, \mathbf{e}_d$下的坐标
  - $\mathbf{UU}^T$为子空间$\mathcal{V}$上的正交投影矩阵
  - $\mathbf{U}^T\mathbf{x}$为$\mathbf{x}$的正交投影在坐标系$\mathbf{u}_1, \cdots, \mathbf{u}_r$下的坐标

# 特征值

- $\mathbf{A} \in \mathbb{R}^{n \times n}$，如果存在常数$\lambda$和非零向量$\mathbf{x} \in \mathbb{R}^{n}$，使得：

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

则称$\lambda$为矩阵**A**的一个特征值(Eigenvalue)，**x**为矩阵**A**对应于特征值$\lambda$的一个特征向量(Eigenvector)。

- 如果**A**有$n$个线性无关的特征向量，那么可对**A**做特征值分解(Eigenvalue Decomposition / ED)

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$$

  - **Q**是特征向量组成的可逆矩阵
  - **Λ**是特征值组成的对角矩阵

$$\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

  - 不是所有矩阵都存在特征值分解，即便存在，**Q**与**Λ**中可能有复数
  - 机器学习中，**A**通常是对称矩阵($\mathbf{A}^T = \mathbf{A}$)，一定存在特征值分解，且**Q**与**Λ**一定取实数值

# 奇异值

- $\mathbf{A} \in \mathbb{R}^{m\times n}$, 则$\mathbf{AA}^T \in \mathbb{R}^{m\times m}$与$\mathbf{A}^T\mathbf{A} \in \mathbb{R}^{n\times n}$均为半正定矩阵，具有非负特征值。
- $\mathbf{AA}^T$的特征值的开方为A的左奇异值(Singular Value)
- $\mathbf{AA}^T$的特征向量为A的左奇异向量(Singular Vector)
- $\mathbf{A}^T\mathbf{A}$的特征值的开方为A的右奇异值
- $\mathbf{A}^T\mathbf{A}$的特征向量为A的右奇异向量
  - 左奇异值=右奇异值
- 任何情况下，可对A做奇异值分解(Singular Value Decomposition / SVD)

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

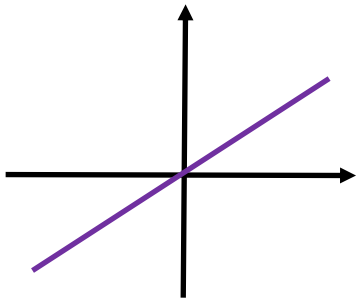
  - $\mathbf{U} \in \mathbb{R}^{m\times m}$是左奇异向量组成的正交矩阵
  - $\mathbf{V} \in \mathbb{R}^{n\times n}$是右奇异向量组成的正交矩阵
  - $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$是奇异值组成的准对角矩阵

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \end{bmatrix}_{m\times n}$$

# 关于SVD

- 非零奇异值的数目为矩阵的秩(rank)，也是矩阵行或列所张成的子空间的维度
  - 行子空间
  - 列子空间

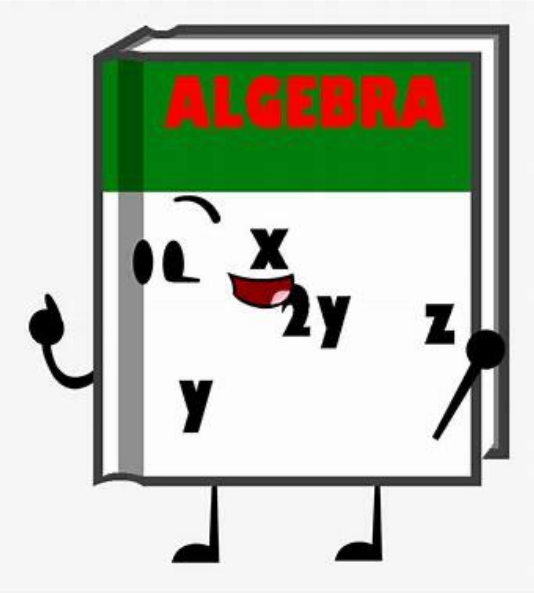
$$\mathbf{A}=\begin{bmatrix}1 & 2 & 3 & 4\\ 1 & 2 & 3 & 4\end{bmatrix}$$

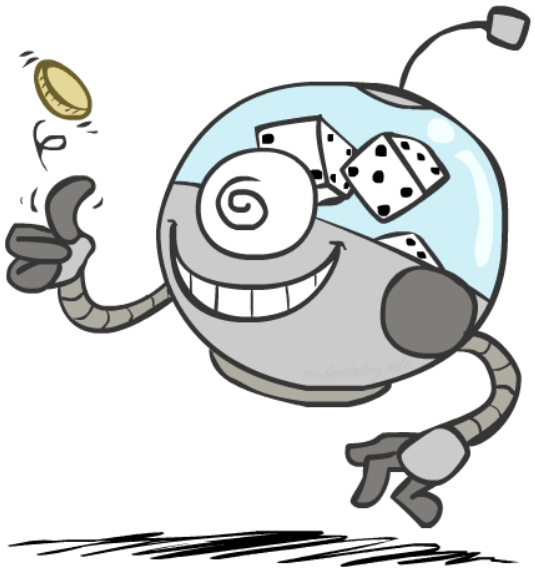


- $\mathbf{A}\in\mathbb{R}^{m\times n},\mathbf{A}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,\mathbf{U}^T\mathbf{U}=\mathbf{I},\mathbf{V}^T\mathbf{V}=\mathbf{I}$
  - SVD：$\mathbf{U}\in\mathbb{R}^{m\times m},\mathbf{U}\mathbf{U}^T=\mathbf{I};\mathbf{V}\in\mathbb{R}^{n\times n},\mathbf{V}\mathbf{V}^T=\mathbf{I};\mathbf{\Sigma}\in\mathbb{R}^{m\times n}$为准对角矩阵
  - thin SVD（设$m<n$）：$\mathbf{U}\in\mathbb{R}^{m\times m},\mathbf{U}\mathbf{U}^T=\mathbf{I};V\in\mathbb{R}^{n\times m},\mathbf{V}\mathbf{V}^T\neq\mathbf{I};\mathbf{\Sigma}\in\mathbb{R}^{m\times m}$为对角矩阵
  - skinny SVD（设$r=rank(\mathbf{A})$）：$\mathbf{U}\in\mathbb{R}^{m\times r},\mathbf{U}\mathbf{U}^T\neq\mathbf{I};\mathbf{V}\in\mathbb{R}^{n\times r},\mathbf{V}\mathbf{V}^T\neq\mathbf{I};\mathbf{\Sigma}\in\mathbb{R}^{r\times r}$为对角矩阵
    - $\mathbf{U}\mathbf{U}^T$为列子空间的正交投影矩阵
    - $\mathbf{V}\mathbf{V}^T$为行子空间的正交投影矩阵

# 本节课安排

- 回顾：基于搜索的问题求解
- 机器学习的数学基础
  - 矩阵论
  - 概率论
  - 信息论



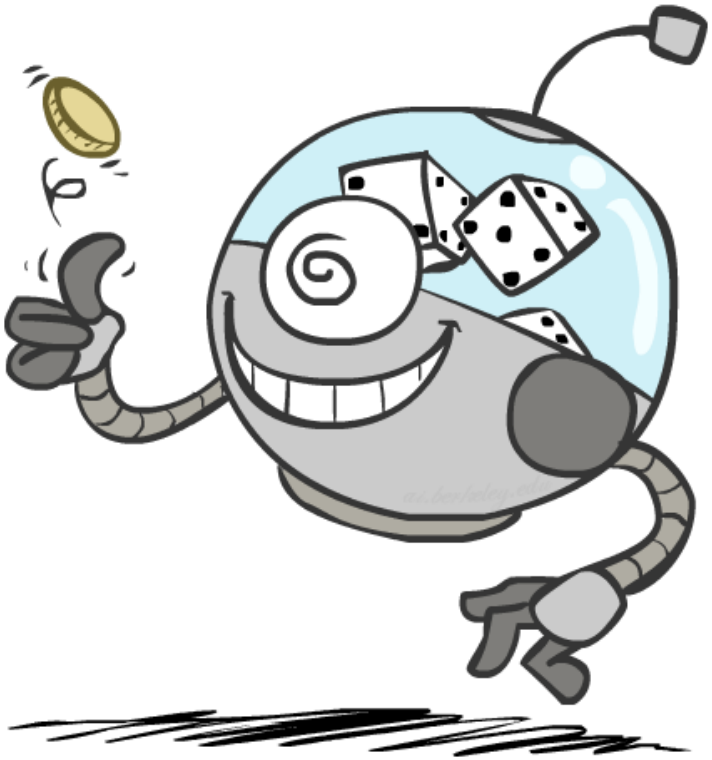



# 为什么需要概率?

- 计算机学科的大多数分支主要处理确定性问题
  - 程序员通常可以放心地假设CPU将完美地执行每一条指令
- 但是机器学习总是需应对不确定性(Uncertainty)
  - 随机性(Stochasticity, Nondeterministic)
  - 未知性(Unknown)
- 导致不确定性的因素
  - 本质上的随机性
    - 抛一枚硬币，正面朝上还是背面朝上?
  - 观测不完整
    - 纸牌游戏
  - 建模不完整
    - 下围棋

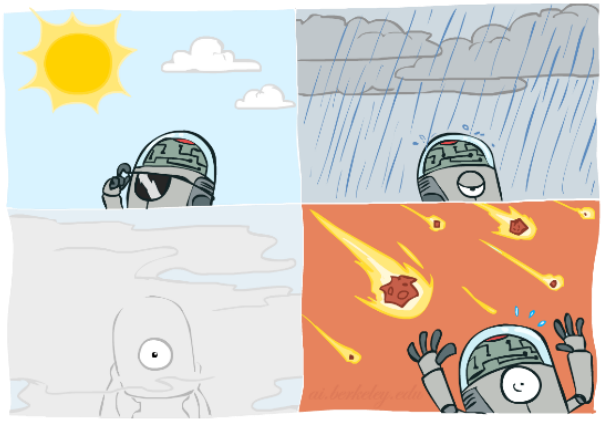
# 随机变量

- 随机变量(Random Variable)是能随机取不同值的变量
  - $X =$ 硬币哪面朝上
  - $R =$ 明天是否会下雨
  - $T =$ 开车上班需要的时间
- 通常使用大写字母表示随机变量
- 随机变量也有值域
  - $X \in$ {正面，背面}
  - $R \in$ {是，否}
  - $T \in [20, 60]$



# 概率分布

- 概率分布(Probability Distribution)是对随机变量呈现其每种可能状态（取值）的可能性的描述



| *T* | *P* |
|---|---|
| 正面 | 0.5 |
| 背面 | 0.5 |

| *W* | *P* |
|---|---|
| 晴天 | 0.6 |
| 雨天 | 0.1 |
| 雾 | 0.3 |

# 概率质量函数(PMF)

- 概率质量函数(Probability Mass Function, PMF)
  - 设随机变量$X$的值域为$\mathcal{D}$
  - PMF：$P(X = x)$表示$X$取值为$x$的概率，简记为$P(x)$
    - $\sum_{x\in\mathcal{D}} P(x) = 1$
    - $\forall x \in \mathcal{D}, 0 \leq P(x) \leq 1$
  - $X \sim P$表示$X$服从分布$P$，即$X$的PMF为$P(x)$
  - E. g.， 均匀分布（Uniform Distribution）
    - 随机变量X的值域为$\mathcal{D} = \{x_1, x_2, \cdots, x_k\}$

$$P(x_i) = \frac{1}{k}, i = 1, \cdots, k$$

| $T$ | $P$ |
|---|---|
| 正面 | 0.5 |
| 背面 | 0.5 |

# 联合分布

- 多个随机变量的概率分布称为联合分布（Joint Probability Distribution)
  - $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$表示$X_1$取值为$x_1$、且$X_2$取值为$x_2$、……、且$X_n$取值为$x_n$的概率，简记为$P(x_1, x_2, \cdots, x_n)$

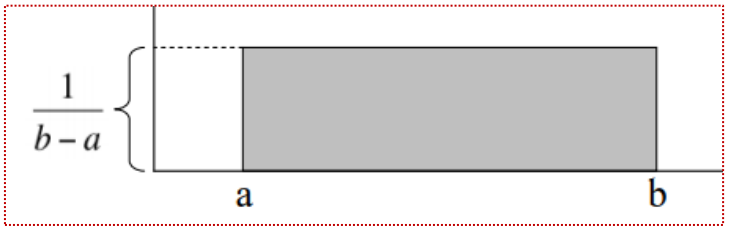
| T | W | P |
|---|---|---|
| 炎热 | 晴天 | 0.4 |
| 炎热 | 雨天 | 0.1 |
| 寒冷 | 晴天 | 0.2 |
| 寒冷 | 雨天 | 0.3 |

  - $(X_1, \cdots, X_n)$可看做是一个多元随机变量$\boldsymbol{X}$

# 概率密度函数(PDF)

- PMF是针对离散随机变量
- 概率密度函数(Probability Density Function, PDF)
  - 设连续随机变量$X$的值域为$\mathcal{D}$
  - $X$的概率分布由PDF描述
    - $\forall x \in \mathcal{D}, P(x) \geq 0$
      - 注意：$P(x)$可以大于1，不是$X$取值为$x$的概率.
    - $\int_{\mathcal{S}} P(x)\,dx$表示$X$的取值属于某集合$\mathcal{S}$的概率（$\mathcal{S} \subseteq \mathcal{D}$）
      - $P(X = x) = \int_x^x P(t)dt = 0.$
    - $\int_{\mathcal{D}} P(x)dx = 1$
  - e. g.， 均匀分布（Uniform Distribution）
    - $\mathcal{D} = [a, b], b > a$

$$P(x) = \frac{1}{b-a}$$



# 边缘分布

- 一组随机变量的子集的分布称为边缘分布（Marginal Probability Distribution）
  - 边缘分布也是一个概率分布
  - 比如，已知$P(x,y)$求$P(x)$

$$P(x)=\sum_y P(x,y) \quad \text{or} \quad P(x)=\int P(x,y)dy.$$

| T | W | P |
|---|---|---|
| 炎热 | 晴天 | 0.4 |
| 炎热 | 雨天 | 0.1 |
| 寒冷 | 晴天 | 0.2 |
| 寒冷 | 雨天 | 0.3 |

$$P(t)=\sum_s P(t,s)$$

| T | P |
|---|---|
| 炎热 | 0.5 |
| 寒冷 | 0.5 |

$$P(s)=\sum_t P(t,s)$$

| W | P |
|---|---|
| 晴天 | 0.6 |
| 雨天 | 0.4 |

# 练习

$P(X,Y)$

| X | Y | P |
|---|---|---|
| +x | +y | 0.2 |
| +x | -y | 0.3 |
| -x | +y | 0.4 |
| -x | -y | 0.1 |

$$P(x) = \sum_y P(x,y)$$

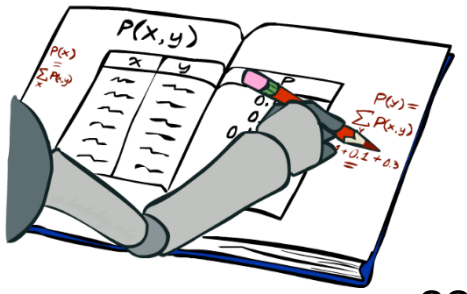
$P(X)$

| X | P |
|---|---|
| +x | |
| -x | |

$$P(y) = \sum_x P(x,y)$$

$P(Y)$

| Y | P |
|---|---|
| +y | |
| -y | |



# 事件（Event）

- 随机变量$X$的值域为$\mathcal{D}$
- 值域 $\mathcal{D}$ 的子集 $\mathcal{E}$ 称为一个事件，$X \in \mathcal{E}$表示事件$\mathcal{E}$发生
- 事件发生的概率
  - $P(\mathcal{E}) = \sum_{x \in \mathcal{E}} P(x)$
  - $\int_{\mathcal{E}} P(x)\, dx$
- 练习：通过联合分布，计算事件发生的概率
  - 炎热并且是晴天的概率
  - 炎热或者是晴天的概率
  - 炎热的概率

$P(T, W)$

| T | W | P |
|---|---|---|
| 炎热 | 晴天 | 0.4 |
| 炎热 | 雨天 | 0.1 |
| 寒冷 | 晴天 | 0.2 |
| 寒冷 | 雨天 | 0.3 |

# 条件概率

- 条件概率（Conditional Probability）：在事件$\mathcal{A}$发生的条件下，事件$\mathcal{B}$发生的概率，记为$P(\mathcal{B}|\mathcal{A})$
  - 条件概率公式

$$P(\mathcal{B}|\mathcal{A}) = \frac{P(\mathcal{A},\mathcal{B})}{P(\mathcal{A})}$$

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

$P(T,W)$

| T | W | P |
|---|---|---|
| 炎热 | 晴天 | 0.4 |
| 炎热 | 雨天 | 0.1 |
| 寒冷 | 晴天 | 0.2 |
| 寒冷 | 雨天 | 0.3 |

$$P(W = \text{晴天}|T = \text{寒冷}) = \frac{P(W = \text{晴天}, T = \text{寒冷})}{P(T = \text{寒冷})} = \frac{0.2}{0.5} = 0.4$$

$$= P(W = \text{晴天}, T = \text{寒冷}) + P(W = \text{雨天}, T = \text{寒冷})$$
$$= 0.2 + 0.3 = 0.5$$

# 练习

$P(X,Y)$

| X | Y | P |
|---|---|---|
| +x | +y | 0.2 |
| +x | -y | 0.3 |
| -x | +y | 0.4 |
| -x | -y | 0.1 |

- $P(+x|+y) = ?$
- $P(-x|+y) = ?$
- $P(-y|+x) = ?$

# 贝叶斯法则（Bayes’ Rule)

- 对两个变量的联合分布进行条件概率分解的两种方法：

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- 两边同除，得到：

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

- 有什么用？
  - 反向构建条件
  - 通常一个计算比较麻烦，但另一个计算相对简单。比如：假设已知$P(\text{抽烟})$与$P(\text{肺癌})$，求$P(\text{肺癌}|\text{抽烟})$

$$P(\text{肺癌}|\text{抽烟}) = \frac{P(\text{抽烟}|\text{肺癌})P(\text{肺癌})}{P(\text{抽烟})}$$

# 链式法则(Chain Rule)

- 三个随机变量$A, B, C$:

$$P(a, b, c) = P(a|b, c)P(b, c)$$

$$P(b, c) = P(b|c)P(c)$$

$$P(a, b, c) = P(a|b, c)P(b|c)P(c)$$

- $n$个随机变量$X_1, \cdots, X_n$

$$P(x_1, x_2, \cdots, x_n) = P(x_1)\prod_{i=2}^{n} P(x_i|x_1, \cdots, x_{i-1})$$

# 独立与条件独立

- 事件$\mathcal{A}$与$\mathcal{B}$独立(Independence)，记为$\mathcal{A} \perp \mathcal{B}$

$$P(\mathcal{A}, \mathcal{B}) = P(\mathcal{A})P(\mathcal{B})$$

$$P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A}), P(\mathcal{B}|\mathcal{A}) = P(\mathcal{B})$$

  - 连续随机变量$X$与$Y$独立，记为$X \perp Y$

$$P(x, y) = P(x)P(y), \forall x, y$$

- 事件$\mathcal{A}$与$\mathcal{B}$在事件$\mathcal{C}$发生的条件下独立(Conditional Independence)，记为$\mathcal{A} \perp \mathcal{B}|\mathcal{C}$

$$P(\mathcal{A}, \mathcal{B}|\mathcal{C}) = P(\mathcal{A}|\mathcal{C})P(\mathcal{B}|\mathcal{C})$$

$$P(\mathcal{A}|\mathcal{B}, \mathcal{C}) = P(\mathcal{A}|\mathcal{C}), P(\mathcal{B}|\mathcal{A}, \mathcal{C}) = P(\mathcal{B}|\mathcal{C})$$

  - 连续随机变量$X$与$Y$在$Z$给定的条件下独立，记为$X \perp Y|Z$

$$P(x, y|z) = P(x|z)P(y|z), \forall x, y, z$$

- 独立与条件独立之间的关系：
  - 独立不意味着条件独立，e. g.， 抛股子
  - 条件独立不意味着独立，e. g.， 马尔科夫链

# 期望、方差、协方差

- 随机变量$X$的函数$f(X)$, e. g. , $f(X)=X$。如何描述$f(X)$?
  - 期望(Expectation)：

    $$\mathbf{E}_{X\sim P}[f(X)]=\sum_x f(x)P(x) \quad \text{or} \quad \int f(x)P(x)dx$$

    - 简记：$\mathbf{E}_X[f(X)], \mathbf{E}[f(X)]$
  - 方差(Variance)：

    $$\mathbf{Var}[f(X)]=\mathbf{E}\left[(f(X)-\mathbf{E}[f(X)])^2\right]$$
- 随机变量$X$的函数$f(X)$, 随机变量$Y$的函数g$(Y)$
  - 协方差(Covariance)

$$\mathbf{Cov}[f(X),g(Y)]=\mathbf{E}[(f(X)-\mathbf{E}[f(X)])(g(Y)-\mathbf{E}[g(Y)])]$$

  - $\mathbf{Var}[f(X)]=\mathbf{Cov}[f(X),f(X)]$

# 练习

- 计算$\mathbf{E}[T], \mathbf{E}[W]$
- 计算$\mathbf{Var}[T], \mathbf{Var}[W]$
- 计算$\mathbf{Cov}[T, W]$

$P(T, W)$

| T | W | P |
|---|---|---|
| 1 | 1 | 0.4 |
| 1 | 0 | 0.1 |
| 0 | 1 | 0.2 |
| 0 | 0 | 0.3 |

$$\mathbf{E}[f(X)] = \sum_x f(x)P(x)$$

$$\mathbf{Var}[f(X)] = \mathbf{E}\left[(f(X) - \mathbf{E}[f(X)])^2\right]$$

$$\mathbf{Cov}[f(X), g(Y)] = \mathbf{E}[(f(X) - \mathbf{E}[f(X)])(g(Y) - \mathbf{E}[g(Y)])]$$

# 常用概率分布

- 伯努利分布(Bernoulli Distribution)
  - $\mathcal{D} = \{0, 1\}$
  - PMF:

$$P(X = 1) = \rho \in [0,1]$$

$$P(X = 0) = 1 - \rho$$

$$P(x) = \rho^x (1 - \rho)^{1-x}$$

$$\mathrm{E}(X) = \rho$$

$$\mathrm{Var}(X) = \rho(1 - \rho)$$

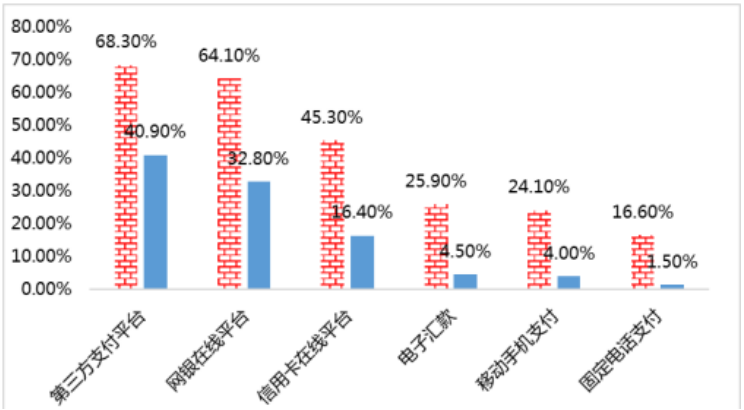
# 常用概率分布

- 多努利分布(Multinoulli Distribution)，也称分类分布(Categorical Distribution)
  - $\mathcal{D} = \{x_1, \cdots, x_k\}$
  - PMF：

$$P(X = x_i) = \rho_i \geq 0, i = 1, \cdots, k$$

$$\sum_{i=1}^{k} \rho_i = 1.$$

  - 主要用于类别建模，无需关注$x_i$的真实取值，无需计算期望与方差



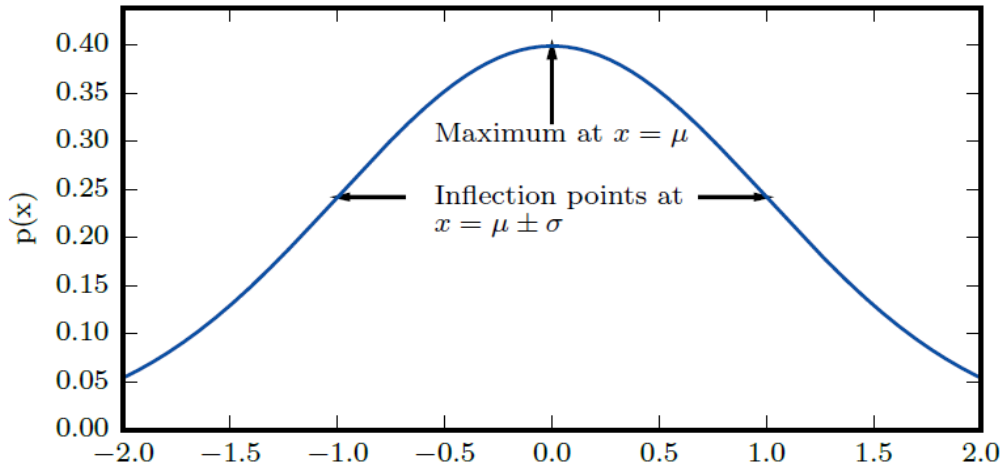
# 常用概率分布

- 正态分布(Normal Distribution)，也称高斯分布(Gaussian Distribution)
  - $\mathcal{D} = \mathbb{R}$
  - PDF:

简记为：$\mathcal{N}(\mu, \sigma)$

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

$$\mathbf{E}[X] = \mu, \mathbf{Var}[X] = \sigma^2.$$

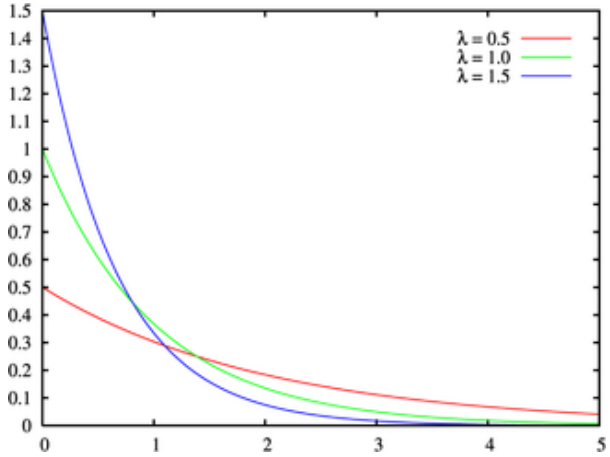



标准正态分布：

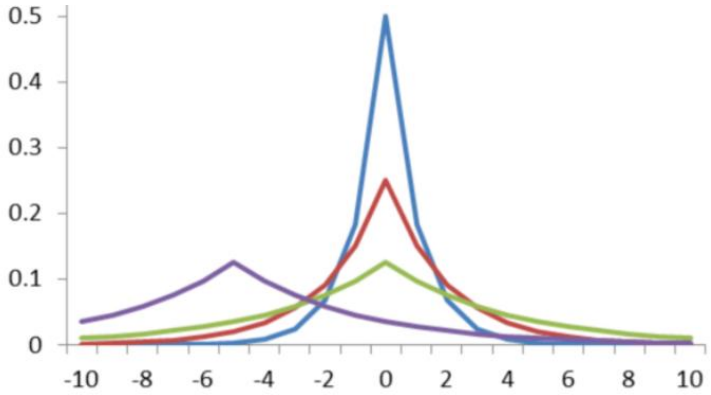
- $\mu = 0$
- $\sigma = 1$

# 常用概率分布

- 指数分布(Exponential Distribution)
  - $\mathcal{D} = \mathbb{R}^+$
  - PDF: $P(x) = \lambda \exp(-\lambda x), x \geq 0$
  - $\mathbf{E}[X] = 1/\lambda, \mathbf{Var}[X] = 1/\lambda^2$
  - 简记为：$\mathrm{Exp}(\lambda)$




- 拉普拉斯分布(Laplace Distribution)
  - $\mathcal{D} = \mathbb{R}$
  - PDF: $P(x) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$
  - $\mathbf{E}[X] = \mu, \mathbf{Var}[X] = 2\gamma^2$
  - 简记为：$\mathrm{Laplace}(\mu, \gamma)$



# 常用概率分布

- 经验分布(Empirical Distribution)
  - $\mathcal{D} = \mathbb{R}^d$
  - 把数据看做随机变量$\boldsymbol{X} \in \mathbb{R}^d$
  - 设有$n$个数据样本$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathbb{R}^d$
  - PDF:

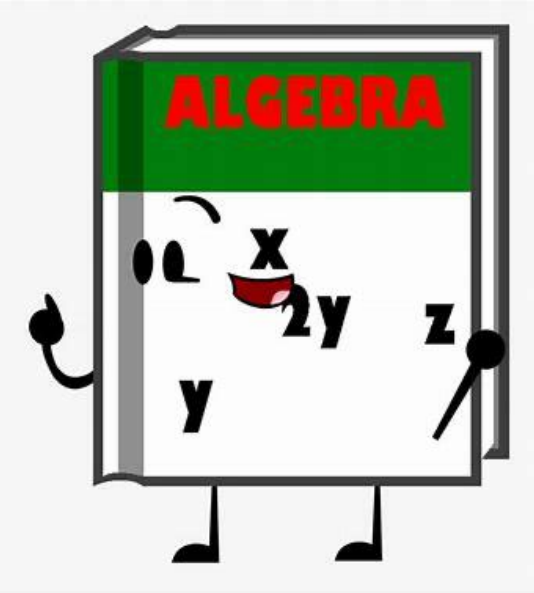
$$P(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\mathbf{x} - \mathbf{x}_i),$$

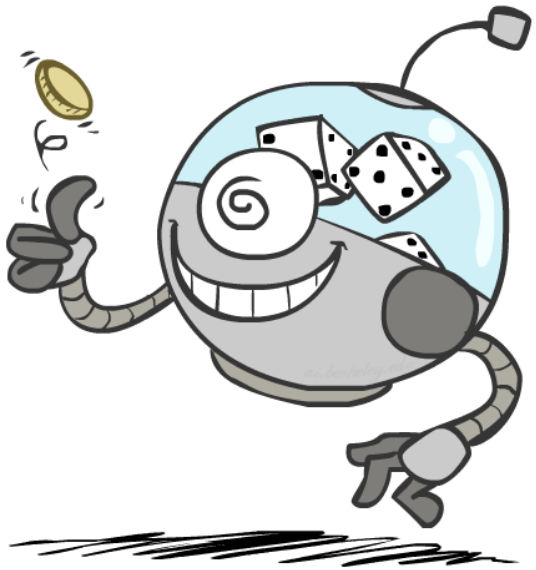
  - 狄拉克$\delta$函数(Dirac Delta Function):

$$\delta(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{0}, \\ 0, & \text{otherwise.} \end{cases}$$

# 本节课安排

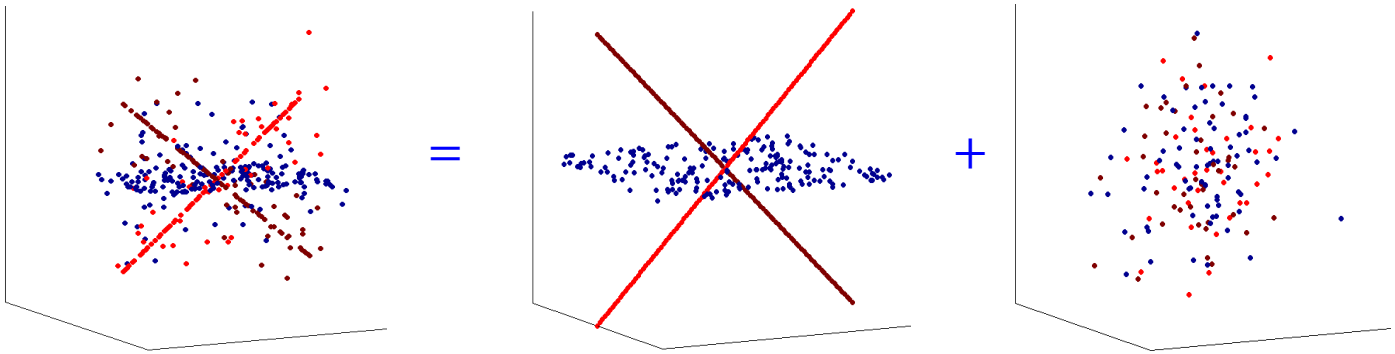
- 回顾：基于搜索的问题求解
- 机器学习的数学基础
  - 矩阵论
  - 概率论
  - 信息论



# 信息论（Information Theory）

- 应用数学的分支
- 主要关注如何量化信号（数据）中的信息量
- 信息论有什么用处？
  - （定量地）刻画概率分布的性质
  - 刻画概率分布之间的相似性
- 在机器学习中：
  - 矩阵论是为了对数据的确定性部分进行描述（主体）
  - 概率论与信息论是对数据的不确定性部分进行描述（重要补充）



# 熵（Entropy）

- 如何量化信号（数据）中的信息量?
  - 基本准则：
    - ✓ 事件发生的概率越小，包含的信息量越大（不确定量）
      - 今天早上太阳出来了
      - 今天早上发生日全食
    - ✓ 独立事件的信息量可累加
- 自信息量(Self-Information)：一个事件$X = x$的信息量为：

$$I(x) = \log \frac{1}{P(x)} = -\log P(x)$$

- 熵(Entropy)：随机变量 $X$ 的概率分布为 $P$，其信息量（不确定量）为：

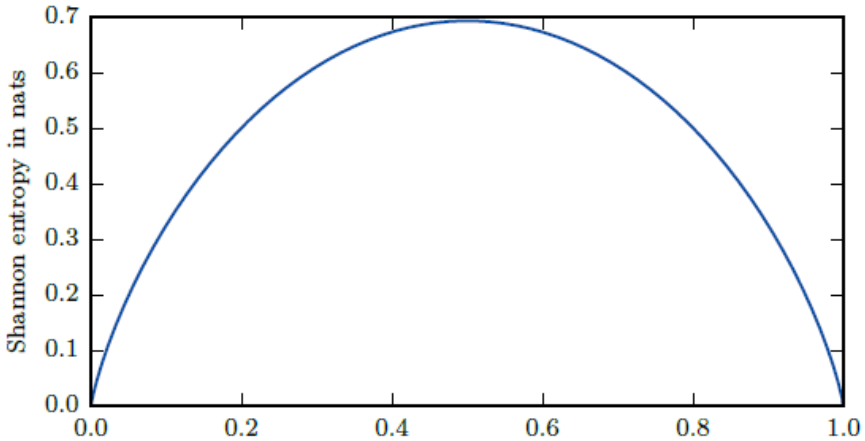
$$H(X) = \mathbf{E}[I(x)] = -\mathbf{E}[\log P(x)] = -\sum P(x) \log P(x)$$

  - 也记为$H(P)$，描述概率分布的性质

# 概率分布的熵

- 伯努利分布：

$$H(X) = -\mathbf{E}[\log P(x)] = -\rho \log \rho - (1-\rho)\log(1-\rho)$$




- 正态分布：

$$H(X) = -\mathbf{E}[\log P(x)] =$$

$$-\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) dx$$

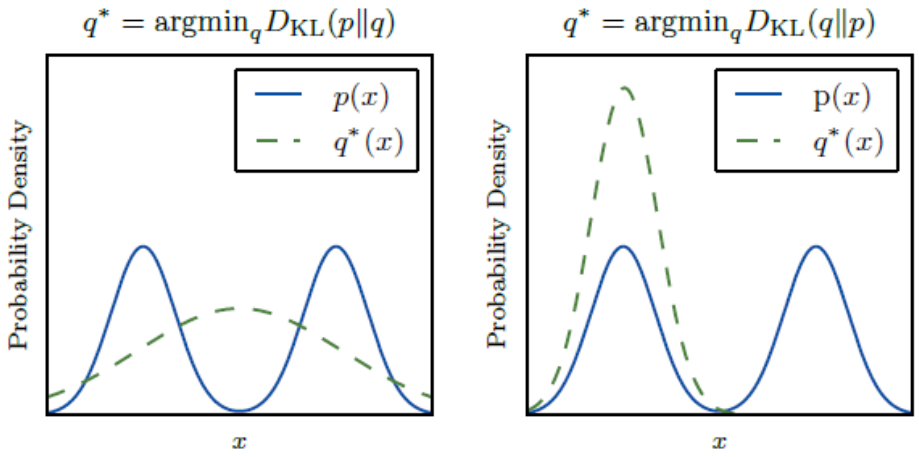
$$= \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2)$$

# KL散度

- 假设得到了随机变量$X$的两个分布$P(x)$与$Q(x)$, 如何衡量这两个分布之间的差异性?
- KL散度(Kullback-Leibler Divergence)：

$$D_{KL}(P||Q) = \mathbf{E}_{X\sim P}\log\left(\frac{P(x)}{Q(x)}\right) = \mathbf{E}_{X\sim P}[\log(P(x)) - \log(Q(x))]$$

  - $\mathrm{D}_{KL}(P||Q) \geq 0$, 取0当且仅当$P = Q$
  - $\mathrm{D}_{KL}(P||Q) \neq \mathrm{D}_{KL}(Q||P)$



# 交叉熵

- 交叉熵(Cross Entropy):

$$H(P,Q) = -\mathbf{E}_{X\sim P}\log(Q(x))$$

  - 离散变量:

$$H(P,Q) = -\sum_{x} P(x)\log(Q(x))$$

  - 连续变量:

$$H(P,Q) = -\int_{x} P(x)\log(Q(x))\,dx$$

  - $H(P,Q) = H(P) + D_{KL}(P||Q)$

- 寻找与已知分布$P$最接近的分布$Q$:

$$\min_{Q} D_{KL}(P||Q) \text{ 等价于 } \min_{Q} H(P,Q)$$

  - 机器学习中，$P$通常是真实标记的分布，$Q$为模型的预测标记分布

# 练习

1. 在人工智能问题中，经常会出现如下优化问题

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \| \mathbf{Ax} - \mathbf{b} \|_2^2 + \lambda \| \mathbf{x} \|_2^2$$

其中$\mathbf{A} \in \mathbb{R}^{m\times n}$、$\mathbf{b} \in \mathbb{R}^m$与$\lambda > 0$已知。

请写出求解**x**的推导步骤。

2. 小明所在的社区爆发了流感，社区中5万名人员共有1万人被感染。经统计得知，5万名人员中有1.2万人有发烧症状，1万名感染者中有8千人有发烧症状。某天，小明发烧了，非常想知道自己有没有感染流感。请运用概率论知识，估算出小明感染流感的概率。