强化学习

与环境交互

智能体、环境、状态、动作、策略、奖励(现实中奖励和惩罚的统合)

基于评估、交互性、序列决策过程

离散马尔可夫链

t+1时刻状态仅与t时刻状态相关

离散马尔可夫链：引入奖励机制：$R: S\times S \mapsto \mathbb{R}$，$R(S_t, S_{t+1})$描述t(状态)→t+1所获得奖励，

引入回报：反映该时刻的累加奖励，

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots\cdots$

$\gamma \in [0,1]$：折扣因子

R_{t+k}：t+k时刻获得奖励

设计的奖励机制对当前时刻及附近时刻能够带来的奖励更为关注

λ

马尔可夫奖励过程：MRP（S,P,R,\lambda）

马尔可夫决策过程：MDP=（S,A,P,R,\lambda）

状态集合S，动作集合A，状态转移概率 $P(S_{t+1}|S_t, A_t)$

奖励函数，$R(S_t, A_t, S_{t+1})$ 在S_t执行A_t后到达S_{t+1}时，Agent得到的奖励

折扣因子 $\gamma \in [0,1]$

与环境交互：$S_0 \xrightarrow[A_0]{R_1} S_1 \xrightarrow[A_1]{R_2} S_2 \to \cdots$

轨迹：状态序列 $(S_0, S_1, \cdots)$

{包含终止状态：分段问题 → 一个从初态到终态的完整轨迹一片段

不包含，……：持续问题。

策略函数：刻画了智能体选择动作的机制。$\pi: S\times A \to [0,1]$，$\pi(s|a)$：状态S下采取动作a的概率

若只有一个动作a，记$a=\pi(s)$。

一个好的策略应该：最大化回报值G_t。

价值函数：衡量某个状态的好坏，反映智能体从当前状态转移到该状态时能够为目标完成带来多大好处

$V: S\to\mathbb{R}$，$V_\pi(s) = E_\pi[G_t|S_t=s]$，t时刻处于状态s时，按$\pi$采取行动所获回报的期望

动作-价值函数 $q: S\times A\to\mathbb{R}$，$q_\pi(s,a) = E_\pi[G_t|S_t=s, A_t=a]$。

强化学习可转化为一个策略学习问题，给定一个马尔可夫决策过程MDP，学习个最优策略\pi, 对任意s，使得V

$\pi(s)$最大。

贝尔曼方程：{价值函数 $V_\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} + \cdots | S_t=s]$

动作-价值函数：$q_\pi(s,a) = E_\pi[R_{t+1} + \cdots | S_t=s, A_t=a]$

$$V_\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t=s]$$
$$= E_{a\sim\pi(s,\cdot)}[E_\pi[\cdots | S_t=s, A_t=a]]$$
$$= \sum_{a\in A}\pi(s,a)\times q_\pi(s,a)$$

$$q_\pi(s,a) = \sum_{s'\in S} P(s'|s,a)[R(s,a,s') + \gamma V_\pi(s')]$$

$$V_\pi(s) = \sum \pi(s,a)\times q_\pi(s,a)$$

$$q_\pi(s,a) = \sum_{s'\in S} P(s'|s,a)[R(s,a,s') + \gamma V_\pi(s')]$$

通过策略计算价值函数的过程叫做策略评估，通过价值函数优化策略的过程叫做策略优化

策略评估和策略优化交替进行的强化学习求解方法叫做通用策略迭代

二、基于价值的强化学习

什么是更好的策略？

$\forall s\in S$ 若 $q_\pi(s,\pi'(s)) \geq q_\pi(s,\pi(s))$ 则 $V_{\pi'}(s) \geq V_\pi(s)$（策略优化定理）

策略评估：根据π来计算相应价值函数V_π或Q_π

①动态规划：初始化V_π.

for

$s \in S$

$$V_\pi(s) \leftarrow \sum_{a\in A}\pi(s,a)\sum_{s'\in S}P_r(s'|s,a)[R(s,a,s')+\gamma V_\pi(s')]$$

缺点：智能体需要事先知道状态转移概率，无法处理状态集合大小无限的情况。

②蒙特卡洛采样 选择不同起始状态，按当前π采样若干轨迹，D.

for $s\in S$：计算D中s每次出现时对应的反馈 $G_1,\cdots,G_k$.

$$V_\pi(s) \leftarrow \frac{1}{k}\sum_{i=1}^{k}G_i$$

蒙特卡洛方法的核心思想：期望可以通过样本均值来估计。给定状态s，从该状态触发不断采样后续状态，得到不同的采样序列。通过这些采样序列来分别计算状态s的回报值，取均值作为s价值函数的估计，避免对状态转移概率的依赖

③时序差分. 初始化V_π.

for-

初始化s为初状态.

循环

$a\sim\pi(s)$，执行a，观察R和s'，

$V_\pi(s)\leftarrow V_\pi(s)+\alpha[R(s,a,s')+\gamma V_\pi(s')-V_\pi(s)]$

$s\leftarrow s'$. $= (1-\alpha)V_\pi(s)+\alpha[R(s,a,s')+\gamma V_\pi(s')]$

（$(1-\alpha)V_\pi(s)$：过去；$\alpha[R(s,a,s')+\gamma V_\pi(s')]$：学习得到）

直至s终止

直至V_π收敛

时序差分法可以看作蒙特卡洛方法和动态规划方法的有机结合。时序差分方法从实际经验中获取信息,无需提前获知环境模型的全部信息(类似于蒙特卡洛树搜索），与动态规划法相似之处：利用前序已知信息进行在线实时学习。
时序差分法根据下一个状态的价值函数来估计，客服了采样轨迹的稀疏可能带来样本方差较大的不足问题，同时缩短了反库周期。

④价值迭代算法（每次迭代只对一个状态进行评估和优化）

$$\pi(s)\leftarrow \arg\max_a Q_\pi(s,a)$$

$$V_\pi(s)\leftarrow Q_\pi(s,\pi(s))$$

⑤Q学习算法. 只有Q_π参与

初始化Q_π

for 初始s为初态

for $a\leftarrow \arg\max_a Q_\pi(s,a)$，执行a，

$Q_\pi(s,a)\leftarrow Q_\pi(s,a)+\alpha[R+\gamma\max_{a'}Q_\pi(s',a')-Q_\pi(s,a)]$

$s\leftarrow s'$

s为终

Q_π收敛

⑥探索与利用 ε贪心策略

$$\begin{cases}\arg\max_a Q_\pi(s,a) & 1-\varepsilon\\ \text{random } a\in A & \varepsilon\end{cases}$$

⑦参数化与深度强化学习

$$L(\theta)=\frac{1}{2}[R+\gamma\max_a Q_\pi(s',a';\theta)-Q_\pi(s,a;\theta)]^2$$

$$\theta\leftarrow\theta-\eta\frac{\partial L}{\partial\theta}$$

⑪ DQN. (经验重现)

每探索到一个新的四元组，便将该四元组添加到经验重现表，在对参数进行更新时，则从表里随机选取一批样本计算损失函数。
一方面样本之间相关性显著减弱，另一方面过去的经验可以被重复利用，提高信息利用的效率。

$$L(\theta)=\frac{1}{2}\left[R+\gamma \max_{a'} q_{\pi}(S',a';\theta^{-})-q_{\pi}(S,a;\theta)\right]^{2}$$

θ^{-} 相对稳定，较低频率更新 $\theta^{-}\leftarrow\theta$.

---

⑫ 策略梯度法

通过直接参数化策略函数的方法求解强化学习问题。
函数取值表示在状态s下选择动作a的概率。选择一个动作的概率是随着参数的改变而光滑变化的，光滑性对算法收敛有更好保证。$\pi_{\theta}(S,a)$.

最大化目标：$J(\theta)=V_{\pi_\theta}(S_0)$

梯度上升：$\nabla_\theta J(\theta)=\nabla_\theta \sum_S \underbrace{\mu_{\pi_\theta}(S)}_{\text{策略分布}} \sum_a q_{\pi_\theta}(S,a)\pi_\theta(S,a)$

$$\nabla_\theta J(\theta) \propto \sum_S \mu_{\pi_\theta}(S)\sum_a q_{\pi_\theta}(S,a)\nabla_\theta \pi_\theta(S,a)$$

$$=E_{S,a,\sim(S,a)\sim\pi}\left[G_t \nabla_\theta \ln \pi_\theta(S,a)\right]$$

⑬ Actor-Critic.

机器学习

① 监督学习，从假设空间学习得到最优映射 f（决策函数）

训练集，将训练集中一部分数据作为验证集。
经验风险：映射函数在训练集上产生的损失
期望风险：所有数据中计算模型的损失。(真实风险，真实误差)
机器学习中模型优化目标一般为经验风险最小化。

? 模型度量方法？

参数优化 { 频率学派：最大似然估计
贝叶斯学派：最大后验估计

② 回归分析（刻画不同变量间的关系）
└ 要从标注数据中学习得到

一元线性回归 $L(a,b)=\sum_{i=1}^{n}(y_i-ax_i-b)^2$

$\frac{\partial L(a,b)}{\partial b}=\sum_{i=1}^{n}2(y_i-ax_i-b)(-1)=0 \Rightarrow \sum_{i=1}^{n}(y_i)-a\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}b=0$

即 $n\bar{y}-na\bar{x}-nb=0$

$\frac{\partial L(a,b)}{\partial a}=\cdots\cdots \quad a=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$

③ 线性回归问题：outlier 非常敏感

logistic 回归：$y=\frac{1}{1+e^{-z}}=\frac{1}{1+e^{-(w^Tx+b)}}$

④ 决策树

建立决策树的过程，就是不断选择属性值对样本集进行划分，直至每个子样本为同一个类别。
构建决策树时划分属性的顺序选择是重要的。性能好->决策树的分支结点样本集的纯度越来越高
信息熵：衡量样本集合纯度的指标，信息熵越大说明该集合的不确定性越大，纯度越低。
选择属性划分样本集前后信息熵的减少量称为信息增益，信息增益被用来衡量样本集合复杂度所减少的程度。

K个集合，集合样本D. $Ent(D)=-\sum_{k=1}^{K}p_k\log p_k$

信息增益：$Gain(D,A)=Ent(D)-\sum_{i=1}^{n}\frac{|D_i|}{|D|}Ent(D_i)$ ✓

⑪ K均值聚类：

将 n个d维数据划分为K个聚簇，使簇内方差最小化。
⇒ 找到一个局部最优，即没有任何其它聚类结果让簇内方差最小化。
* 易受初值影响

1. 初始化聚类质心
2. 对数据进行聚类，$dist(x_i,c_j)=\sqrt{\sum_{a=1}^{d}(x_{ia}-c_{ja})^2}$ $\quad \arg\min_{c\in C} mindist(x_i,c_j)$
3. 更新聚类质心 $c_j=\frac{1}{|c_j|}\sum x_i$
4. 迭代

另一种解释：$\arg\min_{G}\sum_{i=1}^{k}\sum_{x\in G_i}\|x-G_i\|^2=\arg\min_{G}\sum_{i=1}^{k}|G_i|Var G_i$

kmeans → 最小化聚簇内的数据方差

⑪线性判别分析(LDA)

基于监督学习的降维方法. (Fisher LDA)

对于一组具有标签信息的高维数据样本，LDA利用其类别信息，将其线性投影到一个低维空间，在低维空间中同一类别样本尽可能靠近，不同类别样本尽可能彼此原理

样本集 $D=\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\}$, $y_i\in\{C_1,C_2,\dots,C_k\}$. k类样本.

第i类样本协方差阵 $\Sigma_i=\sum_{x\in X_i}(x-m_i)(x-m_i)^T$, m_i为均值.

对于二分类问题 $y(x)=w^Tx$ 投影.

$$S_1=\sum_{x\in C_1}(w^Tx-w^Tm_1)^2=w^T\sum_{x\in C_1}[(x-m_1)(x-m_1)^T]w$$

$S_1=w^T\Sigma_1 w$, $S_2=w^T\Sigma_2 w$: S_1,S_2衡量同一类别样本间分散程度.

$\min(S_1+S_2)$ $\max\|m_1-m_2\|_2^2$, $m_1=w^Tm_1$, $m_2=w^Tm_2$

$$\Rightarrow \max J(w)=\frac{\|m_2-m_1\|_2^2}{S_1+S_2}=\frac{\|w^T(m_2-m_1)\|_2^2}{w^T\Sigma_1 w+w^T\Sigma_2 w}=\frac{w^TS_bw}{w^TS_ww}$$

S_b: 类间散度矩阵

S_w: 类内散度矩阵

可令 $w^TS_ww=1$ $L(w)=w^TS_bw-\lambda(w^TS_ww-1)$

$\Rightarrow$ 求导 $S_w^{-1}S_bw=\lambda w$

$S_b=(m_2-m_1)(m_2-m_1)^T$ 令 $\lambda w=(m_2-m_1)^Tw$

$S_bw=(m_2-m_1)(m_2-m_1)^Tw=\lambda w(m_2-m_1)$

$\because S_w^{-1}S_bw=S_w^{-1}(m_2-m_1)\times\lambda w=\lambda w$

$\therefore w=S_w^{-1}(m_2-m_1)$

最大维度 $\min(k-1,d)$.

⑫PCA. (KL变换. 霍林特变换. 本征正交分解)

样本方差 $Var(X)=\frac{1}{n-1}\sum_{i=1}^n(x_i-\mu)^2$

协方差 $Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^n(x_i-EX)(y_i-EY)$

皮尔逊相关系数 $Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var X\cdot Var Y}}$

$|Corr(X,Y)|\le 1$. 刻画线性相关程度

PCA: d维$\to e$, d维映射到e维. 去除冗余性.

1. 中心化处理
2. 协方差矩阵
3. Σ特征分解, $\lambda_1\ge\lambda_2\ge\dots\ge\lambda_e$
4. 对应的 $w_1,w_2,\dots,w_e$ 组成映射矩阵.

⑬特征人脸方法

本质是使用一组特征向量的线性组合来表示人脸
人脸具有一定的拓扑结构，可以用低维向量表达
原始图像大部分信息。可以用PCA降维，但是在特征维度较高是，可以用奇异值分解来实现PCA

$A=UDV^T$

↓ ↓ ↓ ↓

nxd nxn nxd dxd

$AA^T=UDV^TVD^TU^T$

$=UD^2U^T$

$\Rightarrow (AA^T)U=UD^2$

⑭演化算法.

突变重组 + 自然选择. 模拟自然演化过程.

1.初始化具有若干规模数目的群体。当前进化代数Generation=0;
2.采用评估函数对群体所有的染色体进行评价，分别计算每个染色体的适应值，保存适应值最大的染色体Best
3.采用轮盘赌选择算法对群体染色体进行选择操作，产生规模同样的种群。
4.按照概率从种群中选择染色体进行交配。
5.按照概率对新种群的基因进行变异操作。
6.变异后的新群体取代原有群体，重新计算群体中各个染色体的适应值。若大于best则取代best
7.当前进化数加1

知识表达与推理
一、命题逻辑
逻辑与推理是人工智能的核心问题
二、谓词逻辑
三、知识图谱推理
知识图谱由有向图构成，被用来描述显示世界中实体及实体之间的关系。每个节点表示客观世界中的一个实体，两个节点之间的连线表示节点具有某一关系，知识图谱中存在连线的两个实体可以表达为三元组形式。三元组也可以表示为一阶逻辑的形式。
从数据到知识(推导规则)：归纳， 从知识到数据(实例) ：演绎
归纳逻辑程序设计是机器学习和逻辑程序设计交叉领域的研究内容。
FOIL通过序贯覆盖学习推理。
FOIL算法：
目标谓词是需要推断规则的结论，也称为规则头。在给定推理结论后，FOIL算法学习得到使得结论满足的前提条件，即目标谓词作为结论的推理规则。
正例
反例可以从知识图谱中构造出来，只有两个实体之间存在的关系且确定相悖时才能用来做反例。
背景知识
算法思路：从一般到特殊，逐步添加目标谓词的前提约束谓词，直到所构成的推理规则不覆盖任何反例。

FOIL信息增益值：$FOIL\_Gain = m_+ \cdot \left(\log_2 \frac{m_+}{m_+ + m_-} - \log_2 \frac{n_+}{n_+ + n_-}\right)$

m_+：添加后，n_+：原推理

添加前提约束谓词后所得推理规则的质量好坏由信息增益值来判断

路径排序推理算法：
- 特征提取：生产并选择路径特征集合
特征计算：计算每个训练样例的特征值。$P(s \to t; \pi_j)$ 从s出发通过π_j到达t的概率
分类器训练，分类器可以用于推理两个实体之间是否存在目标关系

概率图谱推理：{贝叶斯网络（DAG）→局部马尔可夫性：联合分布 $= \prod P(\text{节点} \mid \text{父节点})$
马尔可夫网络（无向图）

因果推理　混淆关联 $T \leftarrow X \rightarrow Y$.
选择关联 $T \rightarrow S \leftarrow Y$

⑩因果图：DAG，$P(x_1, \dots, x_d) = \prod_{j=1}^{d} P(x_j \mid x_{pa(j)})$.
⑩干预：
改变明确存在关联关系的某变量取值，研究变量取值改变对结果变量的影响。
do算子：$do(X=x)$，$P(Y=y \mid do(X=x))$
因果效应差/平均因果效应 $P(Y=1 \mid do(X=1)) - P(Y=1 \mid do(X=0))$.

对用药情况X进行干预并固定其值为x时，可将所有指向X的边移除

操纵概率P_m：{边缘概率 $P(Z=z)$不变. $P_m(Z=z) = P(Z=z)$
条件概率 $P(Y=y \mid X=x, Z=z)$不变. $P_m(Y=y \mid X=x, Z=z) = P(Y=y \mid X=x, Z=z)$

$\therefore P(Y=y \mid do(X=x)) = P_m(Y=y \mid X=x)$
$= \sum_z P(Y=y \mid X=x, Z=z) \cdot P(Z=z)$.（调整公式）Z-调整/Z-控制

⑪反事实推理

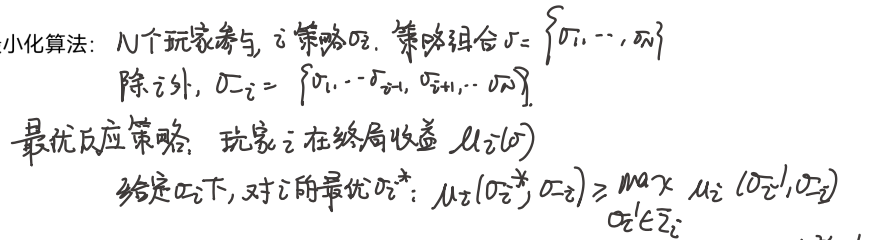
人工智能博弈

推动机器学习从数据拟合过程中以求取最优解为核心向博弈对抗过程中以求去均衡解为核心的转变
博弈论主要研究博弈行为中最优的对抗策略及其稳定局势
相关概念： 参与者或玩家，策略，某个参与者可采纳策略的全体组合形成了策略集，所有参与者鸽子采取行动后形成的状态被称为局势。如果参与者可以通过一定概率分布来选择若干个不同的策略，称为混合策略，有确定的策略称为纯策略
收益，混合策略下的收益为期望收益
博弈论研究的范式：建模者对参与者规定可采区的策略集和取得的收益，观察当参与者选择若干策略以最大化其收益时会产生什么结果

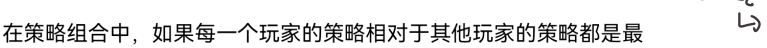
分类：合作博弈与非合作博弈，
静态博弈(所有参与者同时决策，或参与者互相不知道对方的决策）和动态博弈（参与者所采取的行为先后顺序由规则决定，且后行动者知道先行动者所采取的行为）
完全信息博弈与不完全信息博弈 完全信息：参与者均了解其他参与者的策略集，收益等信息

纳什均衡：博弈的稳定局势：参与者作出的这样一种策略组合，在该策略组合上，任何参与者单独改变策略都不会得到好处。即当所有其他人不改变策略时，没有人会改变自己的策略。
Nash定理：若参与者有限，每位参与者的策略集有限，收益函数为实值函数，则博弈必定存在混合策略意义下的纳什均衡。
策梅洛定理：对于任意一个有限步的双人完全信息零和动态博弈，一定存在先手必胜策略或后手必胜策略或保平策略。

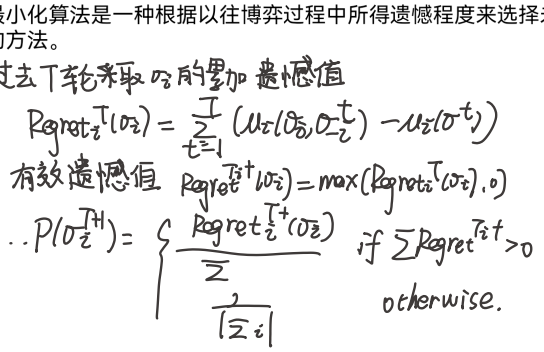
虚拟遗憾最小化算法：



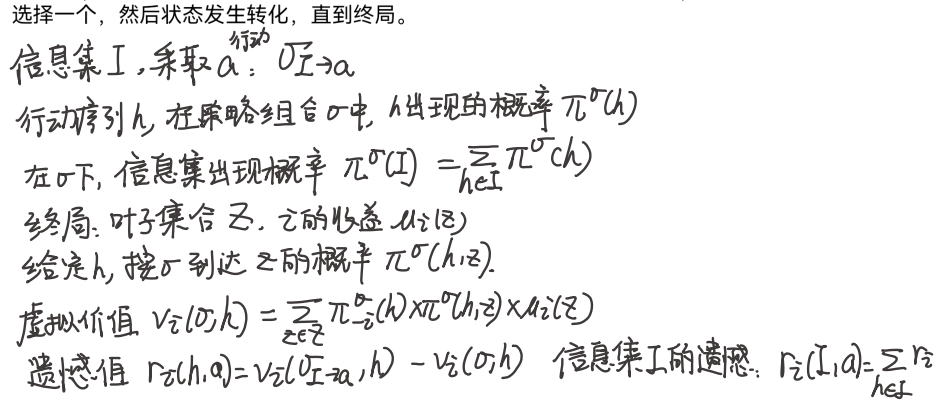
在策略组合中，如果每一个玩家的策略相对于其他玩家的策略都是最佳反应策略，则该策略组合就是一个纳什均衡策略。在有限对手，有限策略情况下，纳什均衡一定存在。



遗憾最小化算法是一种根据以往博弈过程中所得遗憾程度来选择未来行为的方法。



对于任意序贯决策的博弈对抗，可将博弈过程表示成一棵博弈树，博弈树中的每一个中间节点都是一个信息集I，信息集中包含了博弈中当前的状态。给定博弈树的每一个节点，玩家都可以从一系列的动作中选择一个，然后状态发生转化，直到终局。



深度学习

神经元. 感知机.

在感知机模型中增加若干隐藏层，增强神经网络的非线性表达能力，就会让神经网络具有更强的拟合能力。
多个隐藏层构成的多层感知机，也被称为前馈神经网络。
层层递进，逐层抽象；非线性映射；误差反馈调优

$\frac{1}{1+e^{-x}}$

激活函数：sigmoid: 概率形式输出，单调递增，非线性变化；梯度消失问题。
ReLU: 有效缓解梯度消失问题，x<0，稀疏性一定程度上克服过拟合现象。但是也导致神经元死亡。

$Softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}}$

神经网络参数优化：监督学习的过程，让神经网络对数据进行拟合。
模型利用反向传播算法将损失函数计算所得误差从输出端触发，由后向前传递给神经网络中每个单元，通过梯度下降对神经网络中参数进行更新。
损失函数（代价函数），计算模型预测值与真实值之间的误差。
均方损失误差
交叉熵损失误差:

$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$

熵用来表示热力学系统所呈现的无序程度，信息熵，通过对数函数来测量信息的不确定性。
交叉熵用来度量两个概率分布间的差异，刻画了两个概率分布之间的距离，旨在描绘通过概率分布 来表达的困难程度。

梯度下降：

$f(x+\Delta x) - f(x) \approx (\nabla f(x))^T \Delta x < 0$

$= \|\nabla f(x)\| \|\Delta x\| \cos\theta$

$\theta = \pi$ 时 下降最快

误差反向传播.

利用损失函数来计算模型预测结果与真实结果之间的误差以优化调整模型参数，从输出端向输入端，由后向前递进进行。

批量梯度下降：在整个训练集上计算损失误差
随机梯度下降：使用训练集中每个训练样本计算损失函数，梯度方向有很大的波动，收敛慢，但是有助于跳出局部最优解
小批量梯度下降：选取训练集中小批量样本计算，根据每一批量样本所得到的累加误差来更新参数，保证训练过程更稳定，利用矩阵计算优势。

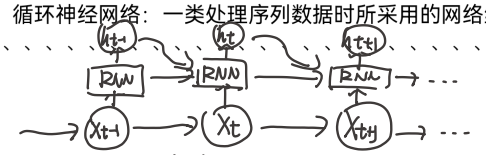
卷积神经网络:
对于图像这样的数据，不能直接将所构成的像素点向量与前馈神经网络中的每个神经元相连
卷积神经网络的前身：级联方式（逐层滤波）实现一种满足平移不变性的网络
卷积滤波结果：特征图
滤波可以视为在给定卷积核权重前提下，记住了领域像素点之间的若干特定空间模式，忽略了某些靠近空间模式
卷积的思想利用了图像中像素点存在空间依赖度的特点，对图像进行了下采样操作

图像平滑操作：中心位置的权重系数小，且与其它卷积权重系数差吧娇小
填充和步长
感受野是卷积神经网络每一层输出的特征图上的像素点在输入图像上映射区域的大小。
局部感知、参数共享

$\frac{W - F + 2P}{S} + 1$

池化：最大/平均/k-max 池化

循环神经网络：一类处理序列数据时所采用的网络结构。



$h_t = \phi(Ux_t + Wh_{t-1})$

前序时刻信息影响后续

得到句子的向量编码 { 最后一个单词隐式编码：整个句子向量编码表示
每个单词隐式编码加权平均 → 句子向量编码

容易出现梯度消失

长短时记忆模型

⑩内部记忆单元门

⑪注意力机制　　自注意力→单词之间概率关联

查询向量 $q_i = W^q \times w_i$

键 … : $k_i = W^k \times w_i$

值 … : $v_i = W^v \times w_i$

$q_i = W^q \times w_i$

$q \cdot k \rightarrow a_{3i}'$

$a_{3i} = q_3 \cdot k_i \xrightarrow{Softmax} a_{3i}'$

$\sum_j a_{3i}' \times v_i$ 为 w_3 注意到与其他单词的关联程度

⑫正则化.

*Dropout. 随机丢掉一部分神经元

批归一化

L1 正则化：稀疏规则算子

搜索

状态，动作，状态转移，路径和代价，目标测试
搜索过程可视为搜索树的构建
评测标准：完备性，最优性，时间空间复杂度

贪婪最佳优先搜索：
评价函数：从当前节点n出发，根据评价函数来选择后续节点：下一个节点是谁？
启发函数：计算从节点n到目标节点之间所形成路径的最小代价值：完成任务还需要多少代价？

贪婪最佳优先搜索：启发函数等于评价函数，不一定最优
A*搜索算法：在评价函数中考虑从起始节点到当前节点的路径代价 可溶性，一致性

可溶性. $\forall n, h(n) \le \underbrace{h^*(n)}_{\text{最小代价}}$ 启发函数不会过高估计代价

一致性. $h(n) \le c(n, a, n') + h(n')$

一致性必然导致可容性

minmax搜索：对抗搜索/博弈搜索
一方想最大化自身的利益，另一方想最小化对手的利益

Alpha-Beta 剪枝.

对于MAX节点，若子孩子的收益大于 α，则 α = 收益.

$\alpha = -\infty$, $\beta = +\infty$ 若 $\alpha > \beta$ 则剪枝

(1) 蒙特卡洛树搜索.

惩值函数 $\rho_T = T\mu^* - \sum_{t=1}^{T} r_t$.

ε-贪心. $r_t = \begin{cases} \arg\max \bar{x}_{i,T(i,t-1)} & 1-\varepsilon \\ \text{random.} & \varepsilon \end{cases}$

$C\sqrt{\frac{2\ln t}{T(i,t-1)}}$

上限置信区间 $P(\mu_i - \bar{x}_{i,T(i,t-1)} > s) \le e^{-2T(i,t-1)s^2}$

认为 $\bar{x}_{i,T(i,t-1)} + s$ 是 μ_i 的上界.

$C\sqrt{\frac{2\ln t}{T(i,t-1)}}$

选择、扩展、模拟、反向传播