



Intelligenza Artificiale

Università di Verona
Imbriani Paolo -VR500437
Professor Alessandro Farinelli

January 15, 2026

Contents

1	Introduzione	4
1.1	Machine Learning	4
1.2	Agenti intelligenti	5
2	Risolvere problemi con la ricerca	5
2.1	Agents and environments	5
2.1.1	Multi-robot Patrolling	6
2.2	Tipi di ambiente	7
2.3	Problem Solving Agents	8
2.3.1	Tree Search Algorithm	8
2.4	Strategie di ricerca	9
2.4.1	Stati ripetuti	10
2.5	Ricerca non informata	10
2.5.1	Breadth-first search	10
2.5.2	Uniform-cost search	11
2.5.3	Depth-first search	11
2.5.4	Iterative deepening search	12
2.6	Ricerca informata	14
2.6.1	Best-first search	14
2.6.2	Greedy best-first search	14
2.6.3	A* search	15
2.7	Ricerca locale	16
2.7.1	Simulated annealing	17
2.7.2	Local beam search	18
2.8	Ricerca locale in spazio continuo	18
2.8.1	Algoritmo di Newton-Raphson	19
2.9	Constrained Satisfaction Problem	20
2.9.1	Grafo dei vincoli	21
2.9.2	Ipergrafo e Grafi duali	22
2.9.3	Problemi combinatori	22
2.9.4	Tree Decomposition	23
3	Logical Agents	23
3.1	Logica in generale	25
3.1.1	Entailment	25
3.1.2	Modelli	26
3.2	Inferenza	26

3.3	Logica proposizionale	26
3.3.1	Sintassi	26
3.3.2	Semantica	27
3.4	Metodi di dimostrazioni	28
3.5	Sistema di inferenza	29
3.5.1	Proprietà dei sistemi di inferenza	30
3.5.2	Conversione in CNF	30
3.5.3	Algoritmo di Resolution	31
3.6	Forward and Backward Chaining	32
3.6.1	Horn Clauses	32
4	Rappresentare l'incertezza	34
4.1	Probabilità	35
5	Markov Decision Process	35
5.1	POMDP (Partially Observable Markov Decision Process)	35
6	Machine Learning: introduzione e regressione lineare	35
6.1	Concetti di base e terminologie	37
6.1.1	Perché voglio stimare $f()$: Predizione	38
6.1.2	Perché voglio stimare $f()$: Inference	38
6.1.3	Termini parametrici	39
6.1.4	Metodi non parametrici	40
7	Machine Learning	41
7.1	Transformer: neural network + attention	41
7.2	Reinforcement Learning	41
7.2.1	Relazioni con gli MDP	41
7.2.2	Come usare un modello?	42
7.2.3	Metodi Model-based	43
7.2.4	Model-free Reinforcement Learning	44
7.2.5	Q-Learning	45
7.2.6	Proprietà del Q-learning	45
7.2.7	Exploration vs Exploitation	47
7.2.8	Funzioni di esplorazione	47
7.3	Deep Reinforcement Learning	47
7.3.1	Gradient Q-Learning	48

1 Introduzione

Alle origini dell'intelligenza artificiale vi è un bisogno diverso da quello che abbiamo oggi. Alan Turing, negli anni 50 si era chiesto se le macchine potessero pensare, creando un test famoso ancora ora come "test di Turing" dove un interrogatore umano si deve interfacciare con un umano e una macchina e doveva capire chi dei due fosse chi. Nel 1956 ci fu uno studio fatto da il progetto di ricerca di Dartmouth, che aveva l'intento di risolvere compiti che richiedeva l'intelligenza di una persona attraverso una macchina, comprendendo che le *anche le macchine possono imparare*. La definizione più "accettata" di Intelligenza Artificiale è quella dove viene vista come una complessa e affascinante *disciplina* che studia come simulare l'intelligenza in scenari complessi usando come strumenti agenti autonomi per delle task ripetitive, sporche e pericolose che sfruttano l'analisi dei dati (predizione e classificazione).

☞ Definizione 1.1

L'intelligenza artificiale è una disciplina che studia come **simulare** l'intelligenza umana in scenari complessi.

Bisogna distinguere machine learning e programmazione:

- **Programmazione:** macchine programmate per ogni task che devono eseguire (il concetto chiave è **il programma**)
- **Machine Learning:** insegnare alla macchina (attraverso esempi) come risolvere task più complesse (il concetto chiave è **il modello**)

1.1 Machine Learning

L'idea di far apprendere una macchina si possono dividere in tre paradigmi contraddistinti:

- Unsupervised learning
- Supervised learning
- Reinforcement learning

Esistono poi i trasformatori, che sono modelli di machine learning probabilistici che si basano sul concetto di attenzione, che sono alla base di modelli come GPT. Il concetto dell'attenzione è quello di dare più importanza ad alcune parole rispetto ad altre in un contesto, per esempio in una frase. La potenza di questi trasformatori è che riescono a fare un'analisi del contesto molto più profonda rispetto ai modelli precedenti, permettendo di fare analisi di immagini come per esempio riconoscere oggetti in un'immagine o riconoscere dove è presente l'acqua all'interno di una foto.

1.2 Agenti intelligenti

Un agente intelligente è un'entità che percepisce il suo ambiente attraverso dei sensori e agisce su di esso attraverso degli attuatori.

- Percepisce l'ambiente attraverso dei **sensori**
- Agisce sull'ambiente attraverso degli **attuatori**
- Ha un **obiettivo** da raggiungere

Come dovrebbe comportarsi un agente intelligente?

- **Razionale:** agisce per massimizzare il raggiungimento dell'obiettivo
- **Performance measure:** misura di quanto bene l'agente sta raggiungendo l'obiettivo

Quando vogliamo ragionare sul Reinforcement Learning, è utile usare il *Markov Decision Process*.

Definizione 1.2

Un **Markov Decision Process (MDP)** è una tupla (S, A, P, R) dove:

- S è un insieme di stati
- A è un insieme di azioni
- $P(s'|s, a)$ è la probabilità di transizione dallo stato s allo stato s' eseguendo l'azione a
- $R(s, a, s')$ è la ricompensa ottenuta eseguendo l'azione a nello stato s e transizionando nello stato s'

Poi si ha la *policy* che è una funzione che mappa uno stato in un'azione.

2 Risolvere problemi con la ricerca

2.1 Agents and environments

Gli agenti includono umani, robot, softbot, termostati, ecc. La funzione agente mappa la storia delle percezioni in azioni.

$$f : \mathcal{P}^* \mapsto A$$

Il *programma dell'agente* viene eseguito su un'architettura fisica che produce f .

Esempio 2.1

Immaginiamo di avere un agente aspirapolvere che percepisce il luogo e i suoi contenuti.

- **Percezioni:** bump, Dirty e location (A o B)
- **Azioni:** left, right, suck, noOp

un esempio di sequenza percepita potrebbe essere:

$(A, \text{Dirty}), \text{Suck}, (A, \text{Dirty}), \text{Suck}, (A, \text{Clean}), \text{Right}$

$(B, \text{Dirty}), \text{Suck}, (B, \text{Clean}), \text{Left}, (A, \text{Clean}), \text{NoOp}$

Cosa fa la funzione *Right*? Può essere implementata in un piccolo programma agente? Se un agente ha $|\mathcal{P}|$ possibili percezioni, quante "entries" avrà la tabella della funzione agente dopo T time steps?

$$\sum_{t=1}^T |\mathcal{P}|^t$$

L'obiettivo dell'IA è quello di progettare **piccoli** programmi agenti che permettono di rappresentare grandi funzioni agenti.

```
function Reflex-Vacuum-Agent([location,status]) returns an action
    if status = dirty then return suck
    else if location = A then return right
    else if location = B then return left
```

2.1.1 Multi-robot Patrolling

Esempio 2.2

Considerate il seguente ambiente:

- Tre stanze (A,B,C) e due robot (r_1, r_2)
- r_1 può pattugliare A e B , r_2 può pattugliare B e C
- r_1 inizia da A e r_2 inizia da C
- Il tempo di viaggio tra le stanze è 0
- Performance Measure: minimizzare il tempo medio di inattività tra le stanze
- Media di inattività: somma degli intervalli nella quale la stanza non è stata visitata da nessun robot

- Quale potrebbe essere un comportamento razionale di questo ambiente?

Quello che succede in maniera ragionevole è la seguente, dove S è la tupla in cui i robot sono posizionati: TODO

Nei diversi casi si ha che il miglior modo per fare girare i robot è quello di farli muovere alternando chi entra nella stanza B minimizzando anche la varianza nelle varie stanze perché dobbiamo stare attenti a non penalizzare troppo una stanza.

2.2 Tipi di ambiente

Il tipo di ambiente determina la progettazione di un agente? Nel mondo reale è ovviamente parzialmente visibile, stocastico, sequenziale, dinamico, continuo, multi-agente.

- **Completamente osservabile vs parzialmente osservabile:** un agente ha accesso completo allo stato dell'ambiente in ogni istante di tempo?
- **Deterministico vs stocastico:** il prossimo stato dell'ambiente è completamente determinato dallo stato corrente e dall'azione eseguita dall'agente?
- **Episodico vs sequenziale:** l'esperienza dell'agente è divisa in episodi indipendenti?
- **Statico vs dinamico:** l'ambiente può cambiare mentre l'agente sta pensando?
- **Discreto vs continuo:** il numero di stati, percezioni e azioni è finito o infinito?
- **Singolo agente vs multi-agente:** l'agente agisce da solo o ci sono altri agenti che possono influenzare l'ambiente?

	Crosswords	Robo-selector	Poker	Taxi
Osservabile	Sì	Parziale	Parziale	Parziale
Deterministico	Sì	No	No	No
Episodico	No	Sì	No	No
Statico	Sì	No	Sì	No
Discreto	Sì	No	Sì	No
Singolo agente	Sì	Sì	No	No

- Se il problema è deterministico e completamente osservabile, è un **single-state problem**
- Se il problema non è osservabile, è un **conformant problem**
- Se il problema è non deterministico o parzialmente osservabile, è un **contingency problem**
- Quando non conosco lo spazio degli stati è un **exploration problem**

2.3 Problem Solving Agents

Una forma ristretta di agente generale sono i: **Goal Based Agent**

- Formula un goal e un problema partendo dallo stato corrente
- Cerco una soluzione a questo problema
- Eseguo la soluzione ignorando le percezioni

Notiamo che questo si chiama anche offline problem; la soluzione viene eseguita ad "occhi chiusi".

```
function Simple-Problem-Solving-Agent(percept) returns an action
    static: solution, state, problem, action
    state <- Update-State(state, percept)
    if seq is empty then
        goal <- Formulate-Goal(state)
        problem <- Formulate-Problem(state, goal)
        seq <- Search(problem)
    action <- First(seq)
    seq <- Rest(seq)
    return action
```

Esempio 2.3 (Vacanze in Romania)

In viaggio in Romania, se attualmente ad Arad. Il viaggio parte domani da Bucharest.

- **Formulate Goal:** essere a Bucharest
- **Formulate Problem:** stati: varie città, azioni: guidare tra le città
- **Search:** trovare una sequenza di azioni che portano da Arad a Bucharest
- **Esempio di Soluzione:** Arad, Sibiu, Fagaras, Bucharest

2.3.1 Tree Search Algorithm

Idea base: offline, esplorazione simulata di spazio di stati, generando successori di stati già esplorati.

```
function Tree-Search(problem) returns a solution, or failure
    initialize the frontier using the initial state of problem
    loop do
        if the frontier is empty then return failure
        node <- Pop an element from the frontier
        if problem.GOAL-TEST(node.STATE) then return SOLUTION(node)
```

```
    expand node, adding the resulting nodes to the frontier  
end
```

☞ Definizione 2.1

Uno **stato** è una rappresentazione di una configurazione fisica.

☞ Definizione 2.2

Un **nodo** è una struttura dati che contiene:

- uno stato
- un puntatore al nodo genitore
- l'azione che ha generato lo stato
- il costo del cammino dal nodo radice a questo nodo

Gli stati non hanno parenti, azioni, figli, costi e profondità!

```
function Expand(node, problem) returns a set of nodes  
    successors <- an empty list  
    for each action in problem.ACTIONS(node.STATE) do  
        child <- CHILD-NODE(problem, node, action)  
        add child to successors  
    return successors
```

2.4 Strategie di ricerca

Una strategia è definita dal scegliere l'ordine dei nodi di espansione. Strategie vengono valutate insieme alle seguenti metriche:

- Completezza: la strategia trova una soluzione se esiste
- Tempo: tempo di esecuzione della strategia
- Spazio: memoria usata dalla strategia
- Optimalità: la strategia trova la soluzione ottima?

Tempo e spazio sono misurati in termini di:

- b branching factor (numero massimo di figli per nodo)
- d profondità della soluzione più superficiale
- m profondità massima dell'albero di ricerca (potrebbe essere infinito)

2.4.1 Stati ripetuti

Fallire nel riconoscere stati ripetuti può trasformare un problema lineare in un problema esponenziale. Bisogna quindi mantenere una lista di stati già visitati e non espandere nodi che portano a stati già visitati:

```
1 function Graph-Search( problem, frontier) returns a solution , or failure
2   explored <- an empty set
3   frontier <- Insert(Make-Node(problem.Initial-State))
4   while not IsEmpty(frontier) do
5     node <- Pop(frontier)
6     if problem.Goal-Test(node.State) then return node
7     if node.State is not in explored then
8       add node.State to explored
9       frontier <- InsertAll(Expand(node, problem))
10      end if
11    end loop
12  return failure
```

2.5 Ricerca non informata

Gli algoritmi di ricerca non informata utilizzano soltanto i dati disponibili nella definizione del problema e i principali sono:

- Breadth-first search
- Uniform-cost search (Dijkstra)
- Depth-first search
- Depth-limited search
- Iterative deepening search

2.5.1 Breadth-first search

Questo algoritmo espande il nodo non esplorato più superficiale, cioè il nodo più vicino alla radice. Utilizza una coda FIFO per la frontiera e i nuovi successori vengono aggiunti alla fine della coda.

```
1 function BFS( problem) returns a solution , or failure
2   node <- node with State=problem.Initial-State,Path-Cost=0
3   if problem.Goal-Test(node.State) then return node
4   explored <- empty set frontier <- FIFO queue with node as the only element
5   loop do
6     if frontier is empty then return failure
7     node <- Pop(frontier)
8     add node.State to explored
9     for each action in problem.Actions(node.State) do
```

```

10     child <- Child-Node(problem, node, action)
11     if child.State is not in (explored or frontier) then
12       if problem.Goal-Test(child.State) then return child
13       frontier <- Insert(child)
14     end if
15   end for
16 end loop

```

Questo tipo di ricerca è:

- **Completa:** Sì, soltanto se b è finito, cioè se il branching factor è limitato
- **Complessità di tempo:** $b + b^2 + b^3 + \dots + b^d = O(b^d)$
- **Complessità di spazio:** $O(b^d)$, perché bisogna memorizzare tutti i nodi generati
- **Ottimale:** Sì, soltanto se il costo delle azioni è uniforme

2.5.2 Uniform-cost search

Questo algoritmo espande il nodo non esplorato con il **costo del percorso più basso**. La frontiera è una coda di priorità ordinata in base al costo del percorso. Questo tipo di ricerca è:

- **Completa:** Sì, se il costo minimo delle azioni $\geq \varepsilon$ (con piccola ma $\varepsilon > 0$)
- **Complessità di tempo:** Numero di nodi $g \leq$ del costo del percorso ottimale C^* . $O(b^{1+\lfloor C^*/\varepsilon \rfloor})$
- **Complessità di spazio:** $O(b^{1+\lfloor C^*/\varepsilon \rfloor})$
- **Ottimale:** Sì perchè i nodi vengono espansi in ordine di costo del percorso

Ci sono due modifiche principali rispetto alla BFS che garantiscono l'ottimalità:

1. Il goal test viene fatto quando il nodo viene estratto dalla frontiera, non quando viene generato. (Questo elemento spiega il $+1$ nella complessità)
2. Controllare se un nodo generato è già presente nella frontiera con un costo più alto e in tal caso sostituirlo con il nuovo nodo a costo più basso

2.5.3 Depth-first search

Questo algoritmo espande il nodo non esplorato più profondo, cioè il nodo più lontano dalla radice. Utilizza una pila LIFO per la frontiera e i nuovi successori vengono aggiunti all'inizio. Questo tipo di ricerca è:

- **Completa:** No, perchè può rimanere bloccata in un ramo infinito, a meno che l'albero di ricerca non abbia una profondità limitata. Si potrebbero evitare loop modificando l'algoritmo per evitare stati ripetuti sul percorso corrente

- **Complessità di tempo:** $O(b^m)$, dove m è la profondità massima dell'albero di ricerca
- **Complessità di spazio:** $O(bm)$, bisogna memorizzare soltanto il percorso corrente e i nodi fratelli
- **Ottimale:** No, perché non garantisce di trovare la soluzione migliore

2.5.4 Iterative deepening search

Questo algoritmo combina i vantaggi della BFS e della DFS. Esegue una serie di ricerche in profondità limitata, aumentando progressivamente il limite di profondità fino a trovare una soluzione.

```

1 # Depth-Limited Search
2 function DLS(problem, limit) returns soln/fail/cutoff
3   R-DLS(Make-Node(problem.Initial-State), problem, limit)
4
5
6 function R-DLS(node, problem, limit) returns soln/fail/cutoff
7   if problem.Goal-Test(node.State) then return node
8   else if limit = 0 then return cutoff # raggiunta la profondita' massima
9   else
10     # flag: c'e' stato un cutoff in uno dei sottoalberi?
11     cutoff-occurred? <- false
12     for each action in problem.Actions(node.State) do
13       child <- Child-Node(problem, node, action)
14       result <- R-DLS(child, problem, limit-1)
15       if result = cutoff then cutoff-occurred? <- true
16       else if result 6 = failure then return result
17     end for
18     if cutoff-occurred? then return cutoff else return failure
19   end else
20
21 # Iterative Deepening Search
22 function IDS(problem) returns a solution
23   inputs: problem, a problem
24   for depth <- 0 to infinity do
25     result <- DLS(problem, depth)
26     if result 6 = cutoff then return result
27   end

```

Questo tipo di ricerca è:

- **Completa:** Sì
- **Complessità di tempo:** $db^1 + (d - 1)b^2 + \dots + b^d = O(b^d)$
- **Complessità di spazio:** $O(bd)$
- **Ottimale:** Sì, se il costo delle azioni è uniforme

Esempio 2.4

Assumi:

1. Un albero di ricerca ben bilanciato, tutti i nodi hanno lo stesso numero di figli
2. Il goal state è l'ultimo che viene espanso nel suo livello (il più a destra)
3. Se il branching factor è 3, la soluzione più superficiale è a profondità 3 (la radice è a profondità 0) e si utilizza la ricerca in ampiezza quanti nodi vengono generati?
4. Se il branching factor è 3, la soluzione più superficiale è a profondità 3 (la radice è a profondità 0) e si utilizza la iterative deepening quanti nodi vengono generati?

Esempio 2.5

Un uomo ha un lupo, una pecora e un cavolo. L'uomo è sulla riva di un fiume con una barca che può trasportare solo lui e un altro oggetto. Il lupo mangia la pecora e la pecora mangia il cavolo, quindi non può lasciarli insieme da soli.

1. Formalizza il problema come un problema di ricerca
2. Usa BFS per risolvere il problema

Soluzione:

Formalizziamo gli stati come una tupla:

$$< W, S, C, M, B >$$

dove:

- W : posizione del lupo
- S : posizione della pecora
- C : posizione del cavolo
- M : posizione dell'uomo
- B : stato della barca

La posizione può essere 0 (left) o 1 (right).

Lo stato iniziale è:

$$< 0, 0, 0, 0, 0 >$$

Lo stato obiettivo è:

$$< 1, 1, 1, 1, 1 >$$

Le azioni possibili sono:

- Porta il lupo (CW)
- Porta la pecora (CS)
- Porta il cavolo (CC)
- Porta niente (CN)

Operatore	Precondizione	Funzione
CW	$M = B, M = W, S \neq C$	$\langle W, S, C, M, B \rangle \mapsto \langle \bar{W}, S, C, \bar{M}, \bar{B} \rangle$
CS	$M = B, M = S$	$\langle W, S, C, M, B \rangle \mapsto \langle W, \bar{S}, C, \bar{M}, \bar{B} \rangle$
CC	$M = B, M = C, W \neq S$	$\langle W, S, C, M, B \rangle \mapsto \langle W, S, \bar{C}, \bar{M}, \bar{B} \rangle$
CN	$M = B$	$\langle W, S, C, M, B \rangle \mapsto \langle W, S, C, \bar{M}, \bar{B} \rangle$

Notiamo che in tutte le precondizioni c'è $M = B$ perchè l'uomo deve essere sempre con la barca, quindi si possono unire i due stati in uno solo M .

2.6 Ricerca informata

Gli algoritmi di ricerca informata utilizzano informazioni aggiuntive (euristiche) per guidare la ricerca verso la soluzione in modo più efficiente.

2.6.1 Best-first search

Questo algoritmo usa una **funzione di valutazione** per ogni nodo che stima la "desiderabilità". La frontiera è una coda ordinata in ordine decrescente di desiderabilità. A seconda di come viene definita la desiderabilità si ottengono diversi algoritmi:

- Greedy best-first search
- A*

2.6.2 Greedy best-first search

Questo algoritmo espande il nodo che sembra essere il più vicino alla soluzione secondo una funzione di valutazione euristica $h(n)$ che stima il costo rimanente per raggiungere l'obiettivo da un nodo n .

Esempio 2.6

In una mappa di una città, la funzione di valutazione potrebbe essere la distanza in linea d'aria dal nodo corrente alla destinazione. In questo modo, l'algoritmo esplora prima

i nodi che sembrano più vicini alla destinazione, riducendo il numero di nodi esplorati rispetto a una ricerca non informata.

Questo tipo di ricerca è:

- **Completa:** No, perché può rimanere bloccata in un ciclo infinito. È completo se lo spazio di ricerca è finito e ci sono controlli per evitare stati ripetuti
- **Complessità di tempo:** $O(b^m)$ nel peggiore dei casi, ma può essere molto più veloce con una buona euristica
- **Complessità di spazio:** $O(b^m)$, bisogna memorizzare tutti i nodi generati
- **Ottimale:** No

2.6.3 A* search

Questo algoritmo evita di espandere cammini che sono già molto costosi e ha come funzione di valutazione:

$$f(n) = g(n) + h(n)$$

dove:

- $g(n)$: costo del percorso dal nodo iniziale a n
- $h(n)$: stima del costo rimanente per raggiungere l'obiettivo da n
- $f(n)$: stima del costo totale del percorso passando per n

L'euristica, per poter garantire l'ottimalità, deve essere **ammissibile**, cioè per ogni nodo la stima di quel nodo deve essere minore o uguale del vero costo per arrivare all'obiettivo, quindi non deve **sovrestimare** il costo rimanente:

$$h(n) \leq h^*(n) \quad h(n) \geq 0 \rightarrow h(G) = 0$$

dove $h^*(n)$ è il costo effettivo del percorso da n .

Teorema 2.6.1

Per A* l'euristica ammissibile implica l'ottimalità

Questo tipo di ricerca è:

- **Completa:** Sì, tranne se ci sono nodi infiniti con $f \leq f(G)$
- **Complessità di tempo:** Esponenziale in errore relativo in $h \times$ lunghezza del numero di passi della soluzione ottimale. (Se l'euristica è buona, la complessità sarà molto più bassa)

- **Complessità di spazio:** $O(b^d)$, bisogna memorizzare tutti i nodi generati
- **Ottimale:** Sì, se l'euristica è ammissibile e consistente

2.7 Ricerca locale

In molte problemi di ottimizzazione il "path" è irrilevante, il traguardo è importante. In questi casi, allora lo spazio degli stati è un insieme di configurazioni:

- Trovare la configurazione ottimale (TSP (Travelling Salesperson Problem), etc...)
- Trovare una configurazione che soddisfi dei vincoli (n-Queens, per esempio, dove ci sono 8 regine su una scacchiera e per trovare la configurazione dove nessuna delle 8 è sotto attacco, parto da una configurazione "base" e sposto le regine finché non trovo la configurazione traguardo, etc...)

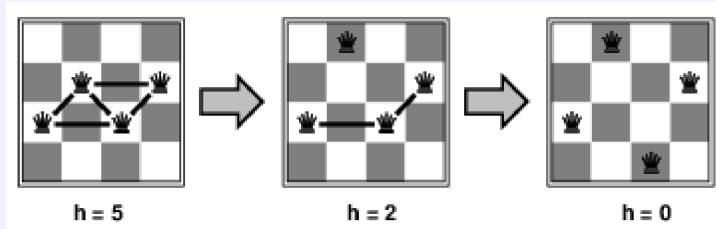
Si possono usare algoritmi di "iterative improvement":

- Mantenere un singolo stato corrente
- Cercare di migliorarlo

Spazio costante, fatto apposta per online e offline search. Varianti di questo approccio arrivano fino a 1% di soluzione ottimali.

Esempio 2.7 (Problema delle n regine)

- Inserire n regine su una scacchiera $n \times n$ in modo che nessuna regina possa attaccarne un'altra (quindi due regine non devono essere sulla stessa riga, colonna o diagonale).
- Muovi una regina per volta, cercando di ridurre il numero di conflitti.



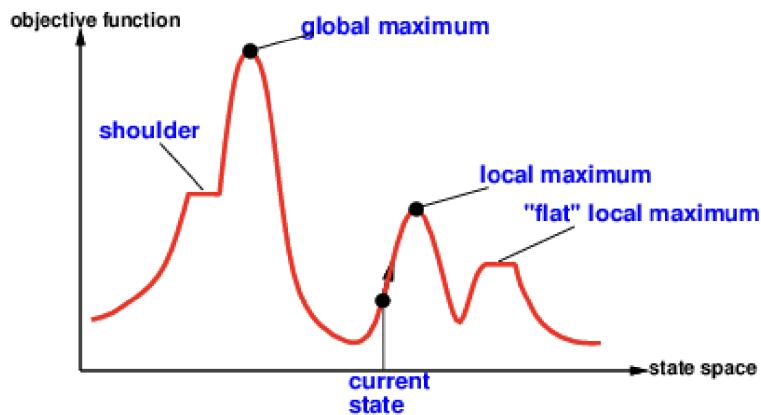
Quasi sempre si solve una problema di questo tipo in pochi passi, anche per $n = 1$ milione.

Ecco ora l'algoritmo di "hill-climbing" (come scalare il monte everest in una fitta nebbia con amnesia):

```

function Hill-Climbing(problem) returns a state that is a local maximum
    inputs: problem, a problem
    local variables: current, a node
                neighbor, a node
    current <- MAKE-NODE(problem.INITIAL-STATE)
    loop do
        neighbor <- a highest-value neighbor of current
        if neighbor.VALUE <= current.VALUE then return
            current.STATE
        current <- neighbor
    
```

Utile per considerare lo *state space landscape*:



Ci sono varianti di questo algoritmo:

- **Random-restart hill climbing** è una variante che supera il massimo locale, trivialmente completo.
- **Random sideways moves** è buono perché esce dalle *shoulder* ma non completamente perché può rimanere bloccato in un ciclo infinito su "flat local maxima".

2.7.1 Simulated annealing

Simulated annealing è un algoritmo di ottimizzazione ispirato al processo di raffreddamento dei metalli. L'idea è di permettere occasionalmente mosse che peggiorano la soluzione corrente per evitare di rimanere bloccati in massimi locali.

- Inizia con una temperatura alta che permette molte mosse peggiorative
- La temperatura diminuisce gradualmente, riducendo la probabilità di accettare mosse peggiorative

- Alla fine, la temperatura raggiunge zero e l'algoritmo si comporta come hill-climbing
- La scelta della schedule di raffreddamento è cruciale per le prestazioni dell'algoritmo

```

1 function Simulated-Annealing(problem, schedule) returns a solution state
2   inputs: problem, a problem
3   schedule, a mapping from time to "temperature"
4   local variables: current, a node
5           next, a node
6           T, a "temperature" controlling prob. of downward steps
7   current <- Make-Node(problem.Initial-State)
8   for t <- 1 to infinity do
9     T <- schedule(t)
10    if T = 0 then return current
11    next <- a randomly selected successor of current
12    deltaE <- next.Value - current.Value
13    if deltaE > 0 then current <- next
14    else current <- next only with probability  $e^{-\Delta E/T}$ 
```

A temperatura fissata T , la probabilità di accettare una mossa che peggiora la soluzione di ΔE è $e^{\Delta E/T}$.

$$p(x) = \alpha e^{\frac{E(x)}{kT}}$$

Decrescendo T abbastanza, si può garantire la convergenza alla soluzione ottimale. Perché

$$e^{\frac{E(x^*)}{kT}} / e^{\frac{E(x)}{kT}} = e^{\frac{E(x^*) - E(x)}{kT}} \gg 1 \quad \text{per } T \rightarrow 0$$

2.7.2 Local beam search

Local Beam Search è un algoritmo di ricerca locale che mantiene k stati invece di uno solo. Inizia con k stati casuali e ad ogni iterazione:

- Genera tutti i successori di tutti i k stati correnti
- Seleziona randomicamente i k migliori successori tra tutti quelli generati
- Ripete fino a quando non viene trovata una soluzione o non ci sono più miglioramenti
- Se tutti i k stati convergono allo stesso punto, si può introdurre diversità sostituendo alcuni stati con nuovi stati casuali

2.8 Ricerca locale in spazio continuo

La ricerca locale può essere estesa a spazi di stato continui. Per risolvere questi problemi si possono utilizzare tecniche come:

- **Discretizzazione:** suddividere lo spazio continuo in una griglia di punti discreti e applicare algoritmi di ricerca locale su questi punti

- **Randomiche Perturbazioni:** introdurre piccole perturbazioni casuali alle soluzioni correnti per esplorare lo spazio delle soluzioni con metodi come il simulated annealing (il prossimo stato è scelto randomicamente)
- **Gradiente:** utilizzare il gradiente della funzione obiettivo per guidare la ricerca verso direzioni di miglioramento (il prossimo stato è scelto in base alla direzione del gradiente). Il metodo del gradiente calcola:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)$$

Per trovare la direzione di massima crescita della funzione obiettivo si pone il gradiente uguale a zero:

$$\nabla f(x) = 0$$

A volte però non riusciamo a risolvere $\nabla f(x) = 0$ analiticamente, quindi possiamo migliorarla localmente:

- Si performa un update nella direzione della salita per ogni coordinata
- Più la funzione è ripida più si fanno passi grandi

Aggiornare una coordinata viene effettuato tramite una funzione generale $g(x_1, x_2)$

$$x_1 \leftarrow x_1 + \alpha \frac{\partial g(x_1, x_2)}{\partial x_1} \quad x_2 \leftarrow x_2 + \alpha \frac{\partial g(x_1, x_2)}{\partial x_2}$$

Oppure in forma vettoriale:

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad nablag(X) = \begin{bmatrix} \frac{\partial g(x_1, x_2)}{\partial x_1} \\ \frac{\partial g(x_1, x_2)}{\partial x_2} \end{bmatrix}$$

$$X \leftarrow X + \alpha \nabla g(X)$$

Dove α è lo step size, cioè la dimensione del passo da fare:

- Se è troppo grande si rischia di saltare soluzioni
- Se è troppo piccolo la convergenza sarà molto lenta

2.8.1 Algoritmo di Newton-Raphson

È una tecnica generale per trovare le radici di una funzione cioè risolvere un'equazione $f(x) = 0$. Per farlo si trova un'approssimazione iniziale \bar{x}_0 della soluzione e iterativamente si aggiorna l'approssimazione usando la formula:

$$\bar{x}_{n+1} = \bar{x}_n - \frac{f(\bar{x}_n)}{f'(\bar{x}_n)}$$

dove:

$$g'(x) = \frac{d}{dx}g(x)$$

✍ Esempio 2.8

Consideriamo la funzione $f(x) = x^2 - a$.

- Mostrare che il metodo di Newton conduce a:

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{a}{x_n} \right)$$

- Fissato $a = 4$, $x_0 = 1$ calcolare le prime tre iterazioni. ($x_i = \{1, 2, 3\}$)

Quindi abbiamo

$$\begin{aligned} f(x) &= x^2 - a \\ x_{n+1} &= \bar{x}_n - \frac{f(\bar{x}_n)}{f'(\bar{x}_n)} \\ x_{n+1} &= \bar{x}_n - \frac{\bar{x}_n - a}{2\bar{x}_n} \end{aligned}$$

2.9 Constrained Satisfaction Problem

Un **Constrained Satisfaction Problem (CSP)** è un problema definito da:

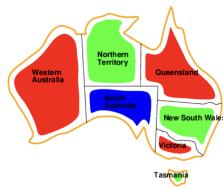
- Un insieme di variabili $X = \{X_1, X_2, \dots, X_n\}$
- Un insieme di domini $D = \{D_1, D_2, \dots, D_n\}$ dove ogni D_i è l'insieme dei valori possibili per la variabile X_i
- Un insieme di vincoli $C = \{C_1, C_2, \dots, C_m\}$ che specificano le relazioni tra le variabili

Assunzioni: single agent, azioni deterministiche, stato completamente osservabile

✍ Esempio 2.9 (Map-Coloring)

Il Map-coloring è un problema specifico di Graph coloring. Dato un insieme di regioni geografiche e un insieme di colori, assegnare un colore a ogni regione in modo che regioni adiacenti non abbiano lo stesso colore.

- Variabili: regioni geografiche (es. WA, NT, Q, NSW, V, SA, T)
- Domini: colori (es. rosso, verde, blu)
- Vincoli: regioni adiacenti non possono avere lo stesso colore



Solutions are assignments satisfying all constraints, e.g.,
 $\{WA = \text{red}, NT = \text{green}, Q = \text{red}, NSW = \text{green}, V = \text{red}, SA = \text{blue}, T = \text{green}\}$

Esempio 2.10 (N-Queens Problem)

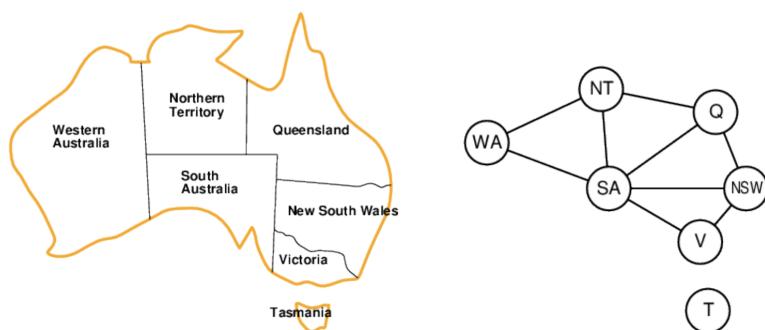
Il N-Queens Problem è un problema di posizionamento di N regine su una scacchiera $N \times N$ in modo che nessuna regina possa attaccarne un'altra.

- Variabili: posizioni delle regine sulle righe della scacchiera
- Domini: colonne della scacchiera (es. 1, 2, ..., N)
- Vincoli: questa volta formuliamo i vincoli in maniera più complessa. Un vincolo per ogni coppia di variabili specificando le posizioni "permesse" PER OGNI ogni due regine.

Questa formulazione rende alcuni vincoli impliciti, per esempio, non è possibile assegnare due regine la stessa colonna quindi non ci sta bisogno di controllare.

2.9.1 Grafo dei vincoli

Un **grafo dei vincoli** chiamato anche grafo primale consiste nel costruire un nodo per ogni variabile e un arco per ogni vincolo tra due variabili.



2.9.2 Ipergrafi e Grafi duali

Le relazioni tra ipergrafi e grafi binari:

- Si può sempre convertire un ipergrafo in un grafo binario
- Ogni variabile ha un dominio esponenzialmente grande

2.9.3 Problemi combinatori

Dato un insieme di possibili soluzioni bisogna trovare quella migliore che soddisfa i vincoli. Alcuni esempi:

- Decisionali: colorare un grafo con k colori
- Ottimizzazione: trovare la colorazione con il minor di conflitti
- Ottimizzazione Multi-obiettivo: portfolio investment, minimizzare il rischio e massimizzare il guadagno
- Modelli grafici:
 - Insieme di variabili, domini e funzioni locali (vincoli)
 - Funzioni globali è un aggregazione di funzioni locali
 - Soluzioni: l'assegnamento di variabili che ottimizza la funzione globale

☞ Definizione 2.3: Rete a vincoli

Una tupla di tre elementi $CN = (X, D, C)$ dove:

- $X = \{X_1, X_2, \dots, X_n\}$ è un insieme di variabili
- $D = \{D_1, D_2, \dots, D_n\}$ è un insieme di domini associati alle variabili
- $C = \{C_1, C_2, \dots, C_m\}$ è un insieme di vincoli che specificano le relazioni tra le variabili
- Ogni vincolo C_i è una tupla (S_i, R_i) dove:
 - $S_i \subseteq X$ è lo scopo, l'insieme delle variabili coinvolte in R_i
 - R_i sottoinsieme del prodotto cartesiano delle variabili in S_i
 - R_i specifica le tuple permesse su S_i
- **Soluzione:** assegnamento di tutte le variabili che soddisfano tutti i vincoli.
- Obiettivo: consistency check, trovare una o tutte le soluzioni, ottimizzare una funzione obiettivo.

Ci si può avvicinare alla soluzione attraverso una soluzione parziale:

- Soluzione parziale consistente: soluzione parziale che soddisfa tutti i vincoli di cui lo scopo non contiene variabili non assegnate
- Una soluzione parziale consistente non è necessariamente estendibile a una soluzione completa

2.9.4 Tree Decomposition

☞ Definizione 2.4: Cycle Cutset

Dato un grafo non orientato, un sottoinsieme di nodi nel grafo è un cycle cutset se la rimozione di questi nodi rende il grafo aciclico.

Il concetto è:

- Una volta che una variabile viene assegnata può essere rimossa dal grafo
- Se rimoviamo un cycle cutset allora il grafo rimanente è un albero
- Si può usare arc-consistency per risolvere l'albero rimanente
- Dobbiamo controllare ogni possibile assegnazione delle variabili del cycle cutset e fare propagazione negli archi
- La complessità è comunque esponenziale ma nella dimensione del cycle cutset.

3 Logical Agents

Perché costruire un agente basato sulla logica? Solitamente hanno due componenti:

- **Knowledge base (KB):** insieme di proposizioni che rappresentano ciò che l'agente sa sul mondo
- **Inference engine:** meccanismo per dedurre nuove proposizioni

La knowledge base è un insieme di proposizioni in un linguaggio formale. Un approccio dichiarativo per costruire un agente:

- Dire cosa sa l'agente
- Poi può chiedersi da solo cosa fare e le risposte dovrebbero seguire la knowledge base.

Gli agenti possono essere visti al livello di conoscenza i.e cosa sanno, non come sono implementati. O a livello di implementazione, cioè come sono costruiti.

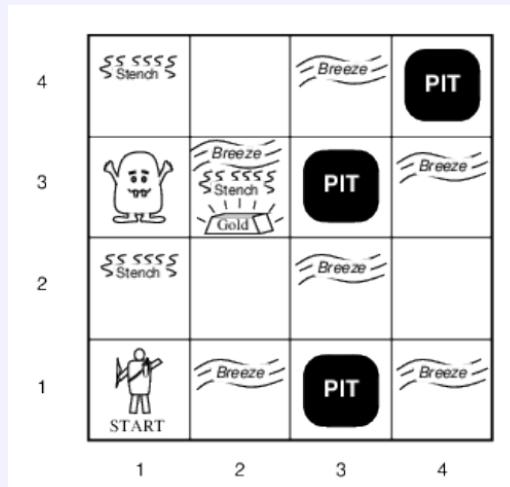
```

function KB-Agent(percept) returns an action
    inputs: percept, a percept
    static: KB, a knowledge base, initially empty
        action, an action, initially null
        t, a counter initially 0
    KB <- Tell(KB, Percept-To-Sentence(percept))
    action <- Ask(KB, Action-Sentence())
    KB <- Tell(KB, Action-To-Sentence(action))
    t <- t + 1
    return action

```

Esempio 3.1 (Wumpus World PEAS)

Un esempio famoso di agente logico è l'agente Wumpus World.



- **Performance measure:** +1000 per uscita, -1 per ogni azione, -1000 per essere mangiati dal Wumpus o cadere in un buco
- **Environment:**
 - Celle adiacenti al wumpus sono maleodoranti
 - Celle adiacenti a un buco sono ventose
 - C'è scintillio se l'oro è nella stessa cella
 - Sparare uccide il wumpus se si è rivolti verso di lui
 - Sparare consuma la sola freccia
 - Prendere raccoglie l'oro se si è nella stessa cella

- Rilasciare lascia cadere l'oro nella stessa cella
- **Actuators:** muovi, gira a sinistra, gira a destra, spara, prendi oro, esci
- **Sensors:** brivido, puzza, scintilla, bump, urlo, sensore di glitter, sensore di impatto

Questo problema è:

- Osservabile? No—solo percezione locale
- Deterministico? Sì—i risultati sono esattamente specificati
- Episodico? No—sequenziale a livello di azioni
- Statico? Sì—Wumpus e buchi non si muovono
- Discreto? Sì
- Single-agent? Sì—Wumpus è essenzialmente una caratteristica naturale

3.1 Logica in generale

La logiche sono linguaggi formali per rappresentare informazioni come per trarre conclusioni. Come sappiamo la logica è divisa in:

- **Sintassi:** definisce le frasi nel linguaggio
- **Semantica:** definisce il "significato" delle frasi; cioè definisce la verità di una frase in un mondo

Un esempio è il linguaggio dell'aritmetica:

- $x + 2 \geq y$ è una frase; $x2 + y >$ non è una frase
- $x + 2 \geq y$ è vera se il numero $x + 2$ non è minore del numero y
- $x + 2 \geq y$ è vera in un mondo dove $x = 7, y = 1$
- $x + 2 \geq y$ è falsa in un mondo dove $x = 0, y = 6$

3.1.1 Entailment

L'entailment è una relazione tra frasi in un linguaggio logico e ha a che fare con i modelli non con la prova formale.

$$KB \models \alpha$$

Knowledge base KB entaila la frase α se e solo se α è vera in ogni mondo dove KB è vera.

Esempio 3.2

Per esempio se KB contiene "La Juventus ha vinto" e "Roma ha vinto" allora KB entaila "O la Juventus o Roma ha vinto". Oppure se $x + y = 4$ allora $4 = x + y$. Entailmente è una relazione tra frasi (cioè sintassi) basata sulla semantica. I computer sono molto bravi a processare regole sintattiche.

3.1.2 Modelli

I logici solitamente ragionano in termini di modelli, che sono formalmente strutturati in mondi rispetto alla verità che deve essere valutata. Diciamo che m è un modello per una frase α se α è vera in m . $M(\alpha)$ è l'insieme di tutti i modelli per α . Allora $KB \models \alpha$ se e solo se $M(KB) \subseteq M(\alpha)$.

Esempio 3.3

Prendendo l'esempio di prima se KB contiene "La Juventus ha vinto" e "La Roma ha vinto" allora α può essere "La Juventus ha vinto".

3.2 Inferenza

$KB \vdash_i \alpha$ vuol dire che α può essere derivata da KB con una procedura i . Le conseguenze di KB sono un pagliaio; α è un ago. Entailmente = ago nel pagliaio; inferenza = trovarlo.

- **Soundness:** i è corretto (sound) se dove $KB \vdash_i \alpha$, è anche vero che $KB \models \alpha$
- **Completeness:** i è completo se dove $KB \models \alpha$, è anche vero che $KB \vdash_i \alpha$

Preview: la logica del primo ordine è abbastanza espressiva da poter dire quasi tutto ciò che ci interessa, ed esiste una procedura di inferenza corretta e completa. Quindi, la procedura risponderà a qualsiasi domanda la cui risposta segua da ciò che è noto dalla KB.

3.3 Logica proposizionale

3.3.1 Sintassi

La logica proposizionale è il linguaggio logico più semplice. I simboli P_1, P_2 sono proposizioni atomiche.

- Se S è una proposizione, allora $\neg S$ è una proposizione
- Se S_1 e S_2 sono proposizioni, allora $S_1 \wedge S_2$, $S_1 \vee S_2$, $S_1 \Rightarrow S_2$, $S_1 \Leftrightarrow S_2$ sono proposizioni

3.3.2 Semantica

Ogni modello specifica se qualcosa è vero o falso per ogni simbolo proposizionale. Le regole per valutare rispetto ad un modello M sono:

- $\neq S$ è vero in M se e solo se S è falso in M
- $S_1 \wedge S_2$ è vero in M se e solo se sia S_1 che S_2 sono veri in M
- $S_1 \vee S_2$ è vero in M se e solo se almeno uno tra S_1 e S_2 è vero in M
- $S_1 \Rightarrow S_2$ è vero in M se e solo se S_1 è falso in M o S_2 è vero in M
- $S_1 \iff S_2$ è vero in M se e solo se entrambi sono veri in M

 **Esempio 3.4** (Wumpus World Sentences)

Consideriamo $P_{i,j}$ sia vero se ci sta un pozzo nella cella (i,j) e $B_{i,j}$ sia vero se c'è del vento nella cella (i,j) .

- $R_1 : \neg P_{1,1}$ (non c'è un pozzo nella cella (1,1))
- $R_2 : \neg B_{1,1}$
- $R_3 : B_{1,2}$

”Il pozzo è in una cella adiacente alla cella con vento”:

$$R_4 : B_{1,1} \Leftrightarrow (P_{1,2} \vee P_{2,1})$$

$$R_5 : B_{1,2} \Leftrightarrow (P_{1,1} \vee P_{1,3} \vee P_{2,2})$$

z Una cella è ventosa **se e solo se** c'è un pozzo in una cella adiacente.

```
function TT-Entails(KB,a) returns true or false
    inputs: KB, the knowledge base, a sentence in prop. logic
    a, the query, a sentence in prop. logic
    symbols <- a list of the proposition symbols in KB and a
    return TT-Check-All(KB,a,symbols,[])
```

Quello che fa questa funzione è quello di fare una DFS di ogni possibile assegnazione, controllando se in ogni modello dove KB è vero anche α è vero. Ritnerà vero o falso in base a questo controllo. Un altro algoritmo è:

```
function TT-Check-All (KB, alpha, symbols, model) return true or false
    if symbols is empty then
        if MODEL-IS-TRUE(KB, model) then
```

```

    return MODEL-IS-TRUE(alpha, model)
else
    return true
else
    P <- FIRST(symbols)
    rest <- REST(symbols)
    return TT-Check-All(KB, alpha, rest,
        EXTEND-MODEL(model, P, true)) and
        TT-Check-All(KB, alpha, rest,
        EXTEND-MODEL(model, P, false))

```

Questo algoritmo ci permette di verificare se $KB \models \alpha$ attraverso una ricerca in profondità di tutte le possibili assegnazioni delle variabili proposizionali. La complessità di questo algoritmo è

$$O(2^n)$$

dove n è il numero di simboli proposizionali ed è co-NP completo.

3.4 Metodi di dimostrazioni

I metodi di dimostrazione si dividono in due categorie:

- **Model checking:**
 - Enumerazione delle tabelle di verità (sempre esponenziale per n)
 - Migliora il backtracking (e.g DPLL)
 - Ricerca con euristiche nello spazio dei modelli (sound ma non completo)
- **Applicazione delle regole d'inferenza:**
 - Legittime (sound) generazioni di nuove formule da quelle vecchie
 - Dimostrazione: una sequenza di applicazione di regole di inferenza (Può essere regole di inferenza come operatori).
 - Tipicamente richiedono traduzioni delle formule in forme normali

Due formule sono logicamente equivalenti quando sono vere negli stessi modelli.

$$\alpha \equiv \beta \iff \alpha \models \beta \wedge \beta \models \alpha$$

Ci sono diverse formule equivalenti utili:

- $\neg(\alpha \wedge \beta) \equiv \neg\alpha \vee \neg\beta$ (Legge di De Morgan 1)
- $\neg(\alpha \vee \beta) \equiv \neg\alpha \wedge \neg\beta$ (Legge di De Morgan 2)

- $\alpha \Rightarrow \beta \equiv \neg\alpha \vee \beta$
- $\alpha \Leftrightarrow \beta \equiv (\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha)$
- $\neg(\alpha \Rightarrow \beta) \equiv \alpha \wedge \neg\beta$
- $\neg(\alpha \Leftrightarrow \beta) \equiv (\alpha \wedge \neg\beta) \vee (\neg\alpha \wedge \beta)$
- $\alpha \vee (\beta \wedge \gamma) \equiv (\alpha \vee \beta) \wedge (\alpha \vee \gamma)$ (Distributività)
- $\alpha \wedge (\beta \vee \gamma) \equiv (\alpha \wedge \beta) \vee (\alpha \wedge \gamma)$ (Distributività)

Una proposizione è valida se è vera in ogni modello.

$$\top \quad \alpha \vee \neg\alpha \quad A \implies A \quad (A \wedge (A \implies B)) \implies B$$

La validità è connessa all'inferenza attraverso il teorema della deduzione.

$$KB \vDash \alpha \iff KB \implies \alpha \text{ è valida}$$

Una proposizione è soddisfacibile se è vera in almeno un modello.

$$A \vee B \quad C$$

Una proposizione è insoddisfacibile se non è vera in nessun modello.

$$\perp \quad A \wedge \neg A$$

La soddisficiabilità è connessa all'inferenza attraverso il seguente fatto:

$$KB \vDash \alpha \iff KB \wedge \neg\alpha \text{ è insoddisfacibile}$$

i.e dimostrare α attraverso la Reductio ad Absurdum.

3.5 Sistema di inferenza

È un insieme di regole di inferenza: Le regole di inferenza sono scritte come:

$$\frac{\text{premesse}}{\text{conclusione}}$$

$$\frac{A_1 \dots A_k}{A}$$

Una derivazione A è derivato da un insieme di formule Γ se con il mio sistema di inferenza R' esiste una sequenza A_1, \dots, A_n tale che:

- A è A_n

- $\forall i \in \{1 \dots n\}$ uno dei seguenti è vero:
 - $A_i \in \Gamma$
 - A_i è derivazione diretta di A_j con $j < i$ usando una regola di inferenza in R'

La sequenza A_1, \dots, A_n è chiamata dimostrazione di A da Γ . Γ sono le premesse di A .

3.5.1 Proprietà dei sistemi di inferenza

La **correttezza** delle regole di inferenza significa che le conclusioni sono conseguenze logiche delle premesse $\psi_1, \dots, \psi_n \models \psi$. La **completezza** ci dice che se $H \models \psi$ allora esiste una derivazione di ψ da H . Si può arrivare alla completezza dicendo: esiste una derivazione della contraddizione se $H \cup \{\neg\psi\}$ non è soddisfacibile. L'algoritmo che andremo a vedere è basato su questo principio. Dimostrare che $\Gamma \models A$ per:

- Dimostrazione per refutazione (reductio ad absurdum): provare che $\Gamma \wedge \neg A$ non è soddisfacibile. L'algoritmo si chiama **Resolution** che ha bisogno di CNF (Conjunctive Normal Form) ed è completo.
- Esiste anche il **Forward/Backward reasoning** sound, completo e anche polinomiale ma SOLO per una parte ristretta di logica proposizionale chiamata Horn clauses.

La forma normale congiuntiva (CNF) è una congiunzione di clausole, dove una clausola è una disgiunzione di letterali e.g. $(A \vee \neg B \vee C) \wedge (\neg A \vee D)$. La **resoluzione** è una regola di inferenza per CNF completa per la logica proposizionale.

$$\frac{l_1 \vee \dots \vee l_k \quad m_1 \vee \dots \vee m_n}{l_1 \vee \dots \vee l_{i-1} \vee l_{i+1} \vee \dots \vee l_k \vee m_1 \vee \dots \vee m_{j-1} \vee m_{j+1} \vee \dots \vee m_n}$$

dove l_i e m_j sono letterali complementari. La resoluzione è corretta e completa per la logica proposizionale.

3.5.2 Conversione in CNF

Partiamo da una formula qualsiasi e la convertiamo in CNF seguendo questi passi:

$$B_{1,1} \iff P_{1,2} \vee P_{2,1}$$

- Elimina \iff , rimpiazza $\alpha \iff \beta$ con $(\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha)$:

$$(B_{1,1} \Rightarrow P_{1,2} \vee P_{2,1}) \wedge (P_{1,2} \vee P_{2,1} \Rightarrow B_{1,1})$$

- Elimina \Rightarrow , rimpiazza $\alpha \Rightarrow \beta$ con $\neg \alpha \vee \beta$:

$$(\neg B_{1,1} \vee P_{1,2} \vee P_{2,1}) \wedge (\neg(P_{1,2} \vee P_{2,1}) \vee B_{1,1})$$

- Muovi \neg verso le proposizioni atomiche usando le leggi di De Morgan e la doppia negazione:

$$(\neg B_{1,1} \vee P_{1,2} \vee P_{2,1}) \wedge ((\neg P_{1,2} \wedge \neg P_{2,1}) \vee B_{1,1})$$

- Applica la distributività e appiattisci:

$$(\neg B_{1,1} \vee P_{1,2} \vee P_{2,1}) \wedge (\neg P_{1,2} \vee B_{1,1}) \wedge (\neg P_{2,1} \vee B_{1,1})$$

Come possiamo notare alla fine abbiamo ottenuto una formula in CNF, ovvero una congiunzione di disgiunzioni.

3.5.3 Algoritmo di Resolution

Dimostrazione per assurdo ovvero mostrare che $KB \wedge \neg\alpha$ è insoddisfacibile.

```
function PL-Resolution (KB, alpha) returns true or false
    inputs: KB, the knowledge base, a sentence in prop. logic
            alpha, the query, a sentence in prop. logic
    clauses <- CNF(KB and not alpha)
    new <- {}
    repeat
        for each pair of clauses (Ci, Cj) in clauses do
            resolvents <- PL-Resolve(Ci, Cj)
            if resolvents contains the empty clause then
                return true
            new <- new union resolvents
            if new contains clauses then
                return false
            clauses <- clauses union new
```

Per alcuni casi dove potremmo ottenere la clausola vuota, devo decidere quale risolvente generare:

$$\frac{P \vee \neg Q \quad \neg P \vee Q}{\neg Q \vee Q} \quad \text{oppure} \quad \frac{P \vee Q \quad \neg P \vee \neg Q}{P \vee \neg P}$$

Questo algoritmo però se gli viene data una formula che non è soddisfacibile, nel momento in cui genera tutti i possibili risolventi deve poter generare una contraddizione per potersi fermare.

 **Esempio 3.5** (Esempio di risoluzione dove la risoluzione $KB \not\models \alpha$)

$$\alpha = P_{1,2} \quad KB = B_{1,1} \iff (P_{1,2} \vee P_{2,1}) \wedge \neg B_{1,1}$$

Traduciamo KB in CNF:

$$(\neg B_{1,1} \vee P_{1,2} \vee P_{2,1}) \wedge (\neg P_{1,2} \vee B_{1,1}) \wedge (\neg P_{2,1} \vee B_{1,1}) \wedge \neg B_{1,1} \wedge P_{1,2}$$

A queste assunzioni aggiungiamo $\neg\alpha$ ovvero $\neg P_{1,2}$. Generiamo le seguenti clausole:

- $\neg B_{1,1} \vee P_{1,2} \vee B_{1,1}$
- $P_{1,2} \vee \neg P_{1,2} \vee P_{2,1}$
- $\neg B_{1,1} \vee P_{2,1} \vee B_{1,1}$
- $\neg P_{2,1} \vee P_{2,1} \vee P_{1,2}$
- $\neg P_{2,1}$

Alcune clausole vengono generate mettendo insieme le clausole originali con quelle generate.

- $\neg B_{1,1} \vee P_{1,2}$
- $\neg P_{2,1} \vee P_{1,2}$

A questo punto non possiamo più generare nuove clausole e non abbiamo generato la contraddizione. L'algoritmo quindi non riesce a terminare e nota che abbiamo generato le stesse formule quindi $KB \not\models \alpha$.

3.6 Forward and Backward Chaining

3.6.1 Horn Clauses

Una **Horn clause** è:

- Una proposizione atomica, oppure
- Una disgiunzione di letterali con al massimo un letterale positivo

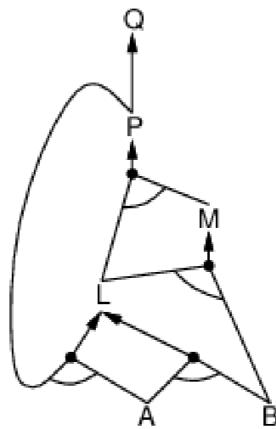
Per esempio: $C \wedge (B \implies A) \wedge (C \wedge D \implies B)$. Possiamo ancora fare il modus ponens, completo per Horn clauses.

$$\frac{\alpha_1, \dots, \alpha_n \quad \alpha_1, \dots, \alpha_n \implies B}{\beta}$$

Può essere usato per il **forward chaining** e il **backward chaining**. Questi algoritmi sono molto naturali e possono essere eseguiti in tempo lineare. Idea: Usa qualsiasi regola di cui le premesse sono soddisfatte in KB e aggiungi la sua conclusione a KB , fino a quando la query è

trovata.

$$\begin{aligned}
 P &\implies Q \\
 L \wedge M &\implies P \\
 B \wedge L &\implies M \\
 A \wedge P &\implies L \\
 A \wedge B &\implies L \\
 A \\
 B
 \end{aligned}$$



```

function PL-FC-Entails(KB, q) return true or false
    inputs: KB, the knowledge base, a set of Horn clauses
        q, the query, a proposition symbol
    count[p] = number of premises in rules with conclusion p
    inferred[p] = false for each proposition symbol p
    agenda = all proposition symbols known to be true in KB
    while agenda is not empty do
        p = POP(agenda)
        if p == q then return true
        if inferred[p] == false then
            inferred[p] = true
            for each horne_clause in KB where p is one of the pi do
                count[r] = count[r] - 1
                if count[r] == 0 then
                    PUSH(r, agenda)
    return false

```

1. FC arriva ad un punto fisso dove non vengono derivate nuove formule atomiche
2. Consideriamo lo stato finale come un modello m , assegnando vero/falso ai simboli
3. Ogni clausola nella KB originale è vera in m , per assurdo supponiamo che una clausola $a_1 \wedge \dots \wedge a_k \implies b$ sia falsa in m . Allora $a_1 \wedge \dots \wedge a_k$ è vera in m e b è falsa in m . Quindi l'algoritmo non ha raggiunto un punto fisso!
4. Quindi m è un modello di KB
5. Per ogni a (formula atomica), se $KB \models a$, a è vera in ogni modello di KB, incluso m

Idea generale: costruire un qualsiasi modello di KB attraverso inferenza corretta, controllare α . Per il back chaining: lavora al contrario dalla query q . Per provare q per BC, controlla se q è noto in KB o prova per BC le premesse di qualche regola includendo q . Evita loop: cerca se il goal è già nella pila dei goal attivi. Evita lavoro ripetuto: controlla se il nuovo subgoal:

- È stato già provato
- È già fallito

FC è *data driven* mentre BC è *goal driven*.

4 Rappresentare l'incertezza

Esempio 4.1

Consideriamo un'azione $A_t =$ lasciare l'aeroporto t minuti prima del volo. A_t mi permetterà di arrivare in tempo? Problemi:

- Osservabilità parziale (non so se c'è traffico, incidenti, ecc)
- Sensori rumorosi (report del traffico)
- Incertezza sulle conseguenze delle azioni (tempo di viaggio variabile)
- Complessità immensa per modellare e predire il traffico

Quindi un approccio puramente logico potrebbe:

- Rischia della falsità: " A_{25} mi porterà in tempo"
- Porta alla conclusione che sono troppo debole per la scelta delle decisioni

Ci sono dei metodi per gestire l'incertezza:

- Logica non monotonica:

- Assumo che la mia macchina non ha problemi alle gomme
- Assumo A_{25} funziona se non ho contraddizioni nelle prove
- Problemi: Quali assunzioni sono ragionevoli? Come gestire la contraddizione?
- Fuzzy logic: gestisce gradi di verità NON incertezza come:
 - WETGRASS è vera al 70%
- Probabilità:
 - Data una prova disponibile, A_{25} arriveremo in tempo con una probabilità 0.04.

4.1 Probabilità

La probabilità misura il grado di credenza in una proposizione. Questa riassume effetti come:

- Laziness: fallire ad enumerare l'eccezioni
- Ignoranza: non so se c'è traffico, condizioni iniziali

La probabilità **Soggettiva** o **bayesiana**: Le probabilità mettano in relazione delle proposizioni con lo stato di conoscenza di un'altra proposizione.

5 Markov Decision Process

5.1 POMDP (Partially Observable Markov Decision Process)

Un POMDP è una generalizzazione di un MDP per ambienti parzialmente osservabili. POMDP ha un modello di osservazione $O(s, e)$ definisce la probabilità che l'agente ottenga l'evidenza e in uno stato s . Se l'agente non sa in quale stato si trova, allora non ha senso parlare di policy $\pi(s)$.

Teorema 5.1.1 Teorema di Astrom

La policy ottimale in una POMDP è una funzione $\pi(b)$ dove b è la **belief state** dell'agente, cioè una distribuzione di probabilità.

Posso convertire una POMDP in un MDP con stati che sono belief states, dove $T(b, a, b')$ è la probabilità che il nuovo belief state sia b' dato il vecchio belief state b e l'azione a .

6 Machine Learning: introduzione e regressione lineare

Automaticamente costruisce modelli secondo alcuni dati ed esistono diversi tipi di apprendimento:

- **Supervised learning:** ($y = f(x)$) dato (x,y) impara $f()$ (approssimazione di funzione)

 **Esempio 6.1** (Supervised Learning: riconoscere cifre scritte a mano)

L'idea è di avere un training set:

- Immagini conosciute che rappresentano le cifre. (x)
- labels che rappresentano il valore della cifra. (t)

Il **training** costruisce il modello che mappa le immagini a cifre $y(x)$. Il **test set** vengono date nuove mai viste prime immagini senza etichette. E il **testing** applica il modello al testing set.

- **Unsupervised learning:** dato x impara $f()$ come rappresentazione compatta di x (Clustering)

 **Esempio 6.2** (Unsupervised Learning: gruppe di celle basati sull'espressione del gene)



Il clustering è un esempio di unsupervised learning dove abbiamo una matrice dove:

- Righe: geni
- Colonne: celle
- Entry (gene, cella) = livello di espressione del gene nella cella

- nessuna etichetta
- obiettivo: trovare gruppi di geni e celle con modelli simili
 - * Per celle (colonne)
 - * Per geni simili
 - * Per entrambi (bi-clustering)

- **Reinforcement learning:** dato x, z impara $f()$ per generare y (x è lo stato, z è la ricompensa)

 **Esempio 6.3** (Pianificare traiettorie con un braccio robotico)

- Controllo del problema (decision making sequenziale)
- Segnale di ricompensa che guida le azioni del robot (alta ricompensa quando l'obiettivo viene raggiunto)
- Ha bisogno di interazione con l'ambiente
- Usa tecniche statiche orientati ai dati ma si concentra sul controllo non sull'analisi dei dati.

6.1 Concetti di base e terminologie

Variabili di input:

- Varaibili indipendenti, predictors, features
- X quando ho più variabili X_1, X_2, \dots, X_n

Variabili di output:

- Risposta, variabile dipendente
- Y

Variabili quantitative: (valori numerici)

- Temperatura, altezza, guadagno, etc.

Variabili qualitativi: (categorie)

- Genere, brand del prodotto etc.

Regressione: output quantitativi

Classificazione: output qualitativi

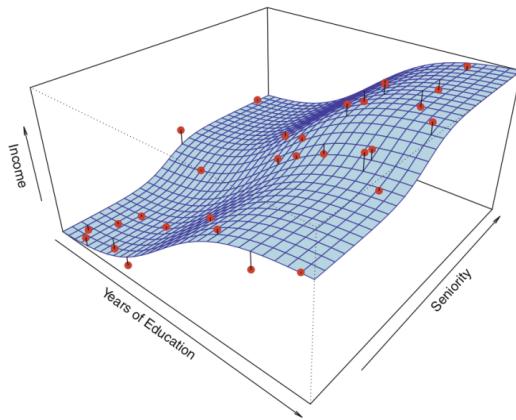
6.1.1 Perché voglio stimare $f()$: Predizione

Dato un set di input facili da ottenere $X = X_1, X_2, \dots, X_n$, predire l'output Y che è costoso da misurare $Y = f(X) + \varepsilon$ dove ε è il rumore (errore casuale indipendente da X) e $f(\cdot)$ rappresenta informazione sistematica tra X e Y .

Definizione 6.1: Predizione

La predizione costruisce un modello $\hat{f}(\cdot)$ per calcolare $\hat{Y} = \hat{f}(X)$ dato X .

L'obiettivo è minimizzare l'errore di predizione.



L'accuratezza di una \hat{Y} dipende da due quantità:

- **Errore riducibile:** errore nella costruzione del modello
- **Errore irriducibile:** variabilità di ε (errore randomico indipendente da X).

Assumere $\hat{f}(\bullet)$ e X fissati:

$$E[(Y - \hat{Y})^2] = E[\underbrace{f(X) + \varepsilon - \hat{f}(X)}_{\text{Errore riducibile}}]^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Errore riducibile}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Errore irriducibile}}$$

- $E(X)$ con X come variabile aleatoria, è il valore atteso o media di X .
- $\text{Var}(X) = E[(X - E(X))^2]$ è la varianza di X .

6.1.2 Perché voglio stimare $f()$: Inferenza

Obiettivo: capire come Y varia quando varia X (non predire). Dobbiamo conoscere la forma di $f(\bullet)$ non una **black box**. Possibili domande per l'inferenza:

- Quali variabili indipendenti sono associati con la risposta? (il punto chiave è l'interpretabilità)

- Quale relazione tra la risposta e ogni variabile indipendente? (effetti positivi o negativi)
- Come possiamo codificare la relazione tra X e Y con modello lineare? (modelli lineari sono semplici e facili da usare ma hanno limitazioni sul potere rappresentativo)

 **Esempio 6.4** (Come stimare f ?)

Abbiamo:

- I dati di training: $n = 30$ data points
- numero di predittori: $p = 2$

su anni di educazione e anzianità.

- x_{ij} è il valore della osservazione i del predittore j dove $i = 1, \dots, n$ e $j = 1, \dots, p$.
- y_{ij} è il valore della risposta per osservazione i .

Il training set: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ dove $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$.

Obiettivo: trovare la funzione $\hat{f}(\bullet)$ tale che $Y \approx \hat{f}(X)$ per ogni osservazione nel training set.

6.1.3 Termini parametrici

Esiste una procedura a due passi:

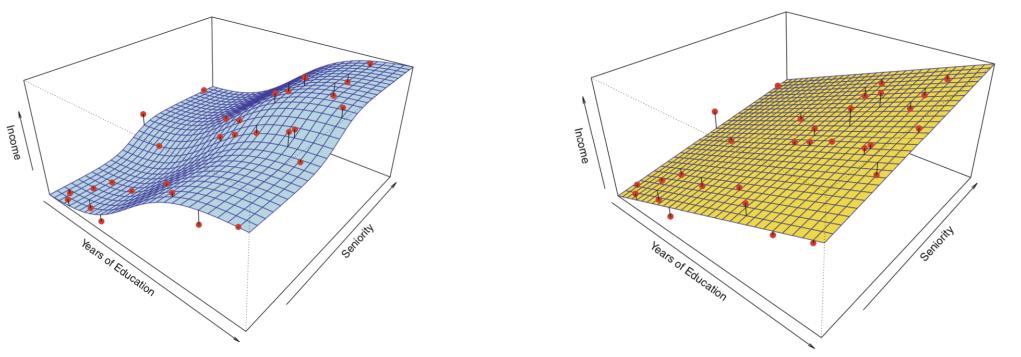
- Fare un assunzione sulla forma di f (modello parametrico)
 - esempio: assumi $f(\bullet)$ sia lineare
 - $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$
 - Abbiamo $p + 1$ parametri $\beta_0, \beta_1, \dots, \beta_p$
 - Punto chiave: un modello è completamente identificato da parametri.
- Trovare una procedura per capire quali sono i valori di β .
 - Trovare un modello lineare per stimare i parametri β .
 - Trovare i valori per cui i parametri:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

– **Least Square** è l'approccio più comune per allenare modelli lineari.

 **Esempio 6.5**

Modello **guadagno** $\approx \beta_0 + \beta_1 \text{educazione} + \beta_2 \text{anzianità}$. I minimi quadrati stimano i parametri $\beta_0, \beta_1, \beta_2$.



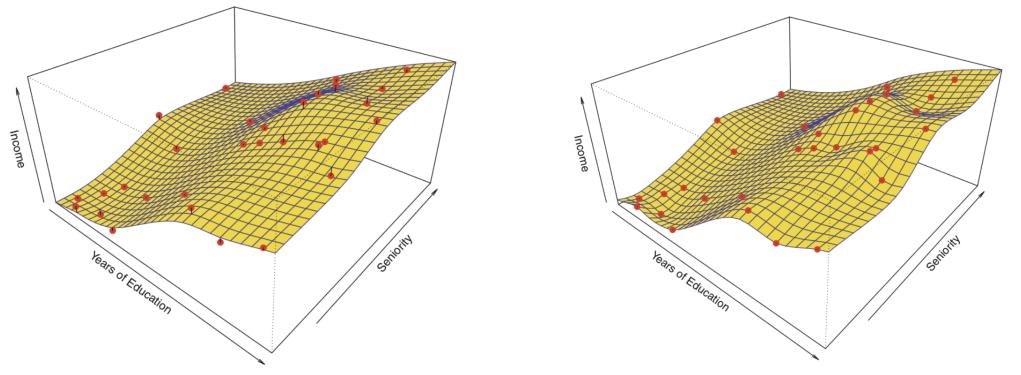
6.1.4 Metodi non parametrici

L'idea è di non fare assunzioni su $f(\bullet)$. Si cerca di fissare $f(\bullet)$ il più possibile sui dati di allenamento.

- **Pro:** posso approssimare i dati di allenamento molto bene (nessuna assunzione sul modello)
- **Contro:** rischio di overfitting (non generalizza bene su dati mai visti prima)

Esempio 6.6

Piano "sottile" per estimare f usando *smooth spline* e *rough spline*.



Quella migliore è la *smooth spline* perché riesce a catturare il trend generale senza overfittare i dati di training.

7 Machine Learning

7.1 Transformer: neural network + attention

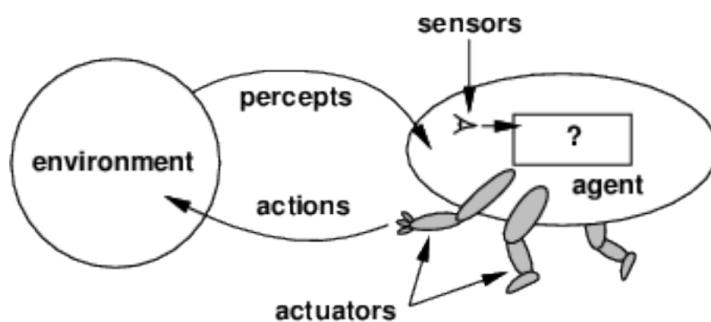
L'attention è un meccanismo che permette al modello di focalizzarsi su parti specifiche dell'input quando genera l'output.

7.2 Reinforcement Learning

Reinforcement learning: impara come mappare situazioni ad azioni così da poter massimizzare una ricompensa numerica cumulativa. I concetti chiave dei RL sono:

- Trial and error mentre interagisce con l'ambiente
- Reward "in ritardo" (le azioni hanno effetto nel futuro)

Essenzialmente, dobbiamo stimare il valore a lungo termine di $V(s)$ e trovare $\pi(s)$.



7.2.1 Relazioni con gli MDP

Guidare una MDP senza sapere le dinamiche:

- Non conosce quali sono gli stati buoni/cattivi (no $R(s, a, s')$)
- Non sa a cosa portano le azioni (no $T(s, a, s')$)
- Quindi dobbiamo provare azioni/stati e ottenere le ricompense

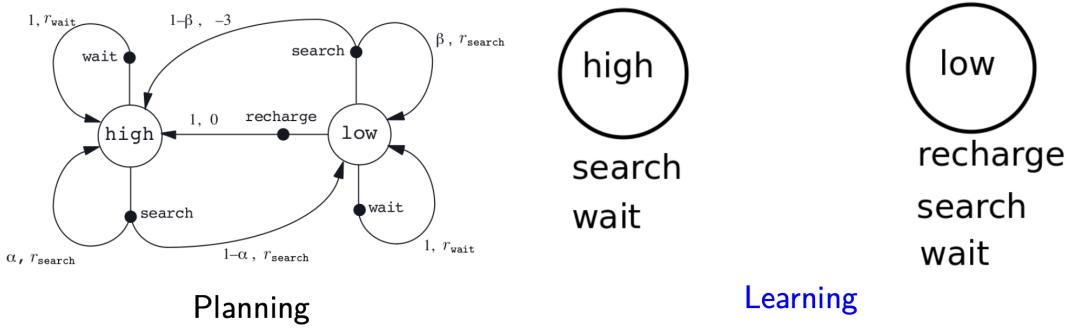


Figure 1: Esempio di un robot che impara a riciclare

7.2.2 Come usare un modello?

Ci sono due approcci:

- **Metodi Model-based** metodi per provare ad imparare il modello:
 - evitare di ripetere stati e azioni "cattivi"
 - meno passi di esecuzione
 - uso efficiente dei dati
- **Metodi Model-free** per imparare la Q -funzione e la policy direttamente:
 - Semplicità, nessun bisogno di costruire ed usare un modello
 - nessun bias nella progettazione del modello

Esempio 7.1 (Età attesa, Model Based vs Model Free)

Obiettivo: calcolare l'età attesa per questa classe. Data la distribuzione di probabilità dell'età:

$$E[A] = \sum_a a \cdot P(a)$$

- **Model based:** stima $\hat{P}(a)$.
 - $\hat{P}(a) = \frac{\text{num}(a)}{N}$
 - $E[A] = \sum_a a \cdot \hat{P}(a)$
 - dove $\text{num}(a)$ è il numero di studenti con età a e N è il numero totale di studenti
 - Funziona perché abbiamo imparato il modello corretto
- **Model free:** nessuna stima
 - $E[A] \approx \frac{1}{N} \sum_{i=1}^N a_i$

- dove a_i è l'età dello studente i
- funziona perché ogni campione appare con la giusta frequenza

L'idea generale:

- Stima $P(x)$ dai campioni.
 - Ottieni campioni $x_i \tilde{P}(x)$
 - Stima $\hat{P}(x) = \text{count}(x)/k$
- Stima $\hat{T}(s, a, s')$ dai campioni:
 - Ottieni i campioni $s_0, a_0, s_1, a_1, \dots, s_k$
 - Stima $\hat{T}(s, a, s') = \text{count}(s, a, s')/\text{count}(s, a)$
 - Funziona perché i campioni appaiono con la giusta frequenza

Esempio 7.2 (Imparare il modello per il robot che ricicla)

Guardate l'immagine del robot che ricicla e dati i seguenti episodi di apprendimento:

$$E_1 : (L, R, H, 0), (H, S, H, 10), (H, S, L, 10)$$

$$E_2 : (L, R, H, 0), (H, S, L, 10), (L, R, H, 10)$$

$$E_3 : (H, S, L, 0), (L, R, H, 10), (H, S, L, 10)$$

Stima $T(s, a, s')$ e $R(s, a, s')$. Se voglio stimare $T(L, R, H)$:

$$T(L, R, H) = \frac{\text{count}(L, R, H)}{\text{count}(L, R)} = \frac{3}{4} = 0.75$$

Mentre se voglio stimare i reward $R(L, R, H)$:

$$R(L, R, H) = \frac{r_1 + r_2 + r_3}{\text{count}(L, R, H)} = \frac{0 + 10 + 10}{3} = \frac{20}{3} \approx 6.67$$

7.2.3 Metodi Model-based

Il seguente, è un algoritmo model-based per RL:

Algorithm 1 Model Based approach to RL

Require: A, S, S_0 **Ensure:** $\hat{T}, \hat{R}, \hat{\pi}$

- 1: Initialize $\hat{T}, \hat{R}, \hat{\pi}$
 - 2: **repeat**
 - 3: Execute $\hat{\pi}$ for a learning episode
 - 4: Acquire a sequence of tuples $\langle s, a, s', r \rangle$
 - 5: Update \hat{T} and \hat{R} according to tuples $\langle s, a, s', r \rangle$
 - 6: Given current dynamics, compute a policy $\hat{\pi}$ (e.g., VI or PI)
 - 7: **until** termination condition is met
-

7.2.4 Model-free Reinforcement Learning

- Vogliamo calcolare il peso atteso da $P(x)$

$$E[f(x)] = \sum_x f(x) \cdot P(x)$$

- Stima Model-based $P(x)$ dai campioni e poi calcolare:

$$x_i \tilde{P}(x), \hat{P}(x) = \text{num}(x)/N, E[f(x)] \approx \sum_x f(x) \cdot \hat{P}(x)$$

- Stima model-free lo fa direttamente sui campioni:

$$x_i \tilde{P}(x), E[f(x)] \approx \frac{1}{N} \sum_{i=1}^N f(x_i)$$

Obiettivo: calcolare il valore della funzione data una policy π

- Media di tutti i campioni osservati
- Esegui π per alcuni episodi di apprendimento
- Calcola la somma delle ricompense (scontate) ogni volta che uno stato viene visitato
- Calcola la media sui campioni raccolti

TD learning for control:

- TD da una valutazione basata sui campioni data una policy π
- Vogliamo calcolare la policy basata su $V(s)$
- Non posso direttamente usare V per calcolare π :

$$-\pi(s) = \arg \max_a Q(s, a)$$

$$- Q(s, a) = \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V(s')]$$

- Idea chiave: possiamo imparare i Q -values direttamente!

7.2.5 Q-Learning

Q-Learning è un metodo di RL a modello libero per imparare la funzione $Q(s, a)$.

- Q-learning: Iterazione Q-value basata su campioni
- Value iteration:

$$V_{k+1} = \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V_k(s')]$$

$$Q_{k+1} = \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma \max_{a'} Q_k(s', a')]$$

Questa equazione trova iterativamente il valore ottimo di Q .

Ma come funziona questa iterazione?

- Calcola l'aspettativa basata sui campioni: $\mathbb{E}(f(x)) = \frac{1}{N} \sum_{i=1}^N f(x_i)$
- I nostri campioni: $R(s, a, s') + \gamma \max_{a'} Q_k(s', a')$
- Imparo $Q(s, a)$ valori durante l'esecuzione:
 - Ricevi il campione: (s, a, s', r)
 - Considero il mio vecchio stimato $Q(s, a)$
 - Considero il nuovo campione: $R(s, a, s') + \gamma \max_{a'} Q(s', a')$
 - Incoporo il nuovo stimato nella media usando un **tasso di apprendimento** α :

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a')]$$

Il tasso di apprendimento α controlla quanto peso dare al nuovo campione e quanto al vecchio stimato.

7.2.6 Proprietà del Q-learning

Il Q-learning converge ad una policy ottimale se:

- Esplori abbastanza
- Se rendi il tasso di apprendimento α abbastanza piccolo col tempo (ma che non decresca troppo velocemente)
- $\alpha = \frac{1}{n(s, a)}$ dove $n(s, a)$ è il numero di visite per (s, a)
- La selezione delle scelte non impatta sulla convergenza: **off-policy learning** impara la policy ottimale senza seguirla.

- **MA** per garantire la convergenza devi visitare ogni coppia (s, a) un numero infinito di volte.

Algorithm 2 Q-learning

```

INITIALIZE  $Q(s, a)$  arbitrarily for all  $s \in S, a \in A(s)$ 
2: repeat
    INITIALIZE  $s$ 
4:   repeat
    CHOOSE  $a$  from  $s$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)
6:   TAKE action  $a$ , OBSERVE  $r, s'$ 
    UPDATE  $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a')]$ 
8:    $s \leftarrow s'$ 
    until  $s$  is terminal
10:  until per ogni episodio

```

Scegliamo l'azione A in base a diversi algoritmi tra cui ϵ -greedy: Scegli la migliore azione la maggior parte delle volte, ma ogni tanto (con probabilità ϵ) scegli un'azione a caso per esplorare nuove azioni (con uguale probabilità).

Algorithm 3 Q-learning con SARSA

```

INITIALIZE  $Q(s, a)$  arbitrarily for all  $s \in S, a \in A(s)$ 
repeat
3:   INITIALIZE  $s$ 
    CHOOSE  $a$  from  $s$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)
    repeat
6:      TAKE action  $a$ , OBSERVE  $r, s'$ 
        CHOOSE  $a'$  from  $s'$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)
        UPDATE  $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a')]$ 
9:       $s \leftarrow s'$ 
       $a \leftarrow a'$ 
    until  $s$  is terminal
12:  until per ogni episodio

```

SARSA deriva dalla tupla (s, a, r, s', a') :

- Caratterizzata dal fatto che calcoliamo la prossima azione basata sulla policy
- Se la policy converge (nel limite) con la policy greedy (ed ogni coppia è stata visitata abbastanza) *SARSA* converge alla policy ottimale.

Q-learning impara la policy ottimale ma occasionalmente fallisce a causa della selezione delle azioni di ϵ -greedy. SARSA, essendo *on – policy* ha prestazioni *on – line* migliori.

7.2.7 Exploration vs Exploitation

Un problema fondamentale nel reinforcement learning è il trade-off tra

- **Exploration:** provare nuove azioni per scoprire il loro valore
- **Exploitation:** usare le azioni che si pensa abbiano il valore più alto

La scelta dipende dal contesto e dall'obiettivo dell'agente. Ci sono due approcci principali:

- **Epsilon-greedy:** con probabilità ϵ scegli un'azione a caso (exploration), altrimenti scegli l'azione con il valore più alto (exploitation).
- **Softmax:** assegna una probabilità a ogni azione basata sui suoi valori stimati, quindi scegli un'azione in base a queste probabilità. La probabilità di scegliere un'azione a in stato s è data da:

$$P(a|s) = \frac{e^{Q(s,a)/\tau}}{\sum_{a'} e^{Q(s,a')/\tau}}$$

dove τ è la temperatura che controlla il livello di exploration. Se τ è alta le azioni hanno probabilità simili (più exploration), se τ è bassa l'azione con il valore più alto ha probabilità maggiore (più exploitation).

7.2.8 Funzioni di esplorazione

L'idea è quella di aggiungere un bonus di esplorazione al valore stimato di un'azione per incentivare l'esplorazione di azioni meno visitate. Considera un estimato u e visita n volte e calcola $f(u, n) = u + \frac{k}{n}$:

- Aggiornamento regolare: $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a')]$
- Aggiornamento modificato: $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_{a'} f(Q(s', a'), N(s, a))]$
- Dove $N(s, a)$ è il numero di volte che l'azione a è stata scelta nello stato s .
- k è un parametro fissato

7.3 Deep Reinforcement Learning

Per molti problemi reali non possiamo esplicitamente rappresentare funzioni chiave con reinforcement learning (es. $Q(s, a)$, $\pi(s)$, $V(s)$) Per risolvere questo problema possiamo usare:

- Approssimazione lineare
- Approssimazione tramite rete neurale (Deep RL)
- Approssimazione con Deep Q-Networks (DQN)

7.3.1 Gradient Q-Learning

Per approssimare $Q(s, a)$ con una rete parametrica $Q_w(s, a)$ si può usare il gradiente per aggiornare i pesi w :

- Stima $Q(s, a)$ con $Q_w(s, a)$
- Target per Q-learning: $y = r(s, a, s') + \gamma \max_{a'} Q_w(s', a')$
- Errore quadrato:

$$E(w) = (Q_w(s, a) - r(s, a, s') - \gamma \max_{a'} Q_w(s', a'))^2$$

- Gradiente:

$$\frac{\partial E(w)}{\partial w} = 2(Q_w(s, a) - r(s, a, s') - \gamma \max_{a'} Q_w(s', a')) \cdot \frac{\partial Q_w(s, a)}{\partial w}$$

Il valore scalare 2 è un fattore costante e non è importante per l'update.

Algorithm 4 Gradient Q-Learning

```

Initiliaze weights  $w$  arbitrarily
INITIALIZE  $s$  (osserva stato corrente)
loop
  4:   SELECT and EXECUTE actiona  $a$ 
        OBSERVE new state  $s'$  receive immediate reward  $r$ 
        COMPUTE target:  $y = r + \gamma \max_{a'} Q_w(s', a')$ 
        UPDATE weights:  $w \leftarrow w - \alpha(Q_w(s, a) - y) \cdot \nabla_w Q_w(s, a)$ 
  8:    $s \leftarrow s'$ 
end loop

```
