

Introduzione all’Intelligenza Artificiale

Federico Matteoni

A.A. 2019/20

Indice

1 Agenti Intelligenti	9
1.1 Intelligenza	9
1.2 Agenti	9
1.2.1 Caratteristiche	9
1.2.2 Percezioni e Azioni	10
1.2.3 Agente e ambiente	10
1.2.4 Agenti Razionali	11
1.2.5 Agenti Autonomi	11
1.3 Ambienti	12
1.3.1 PEAS	12
1.3.2 Simulatore di Ambienti	12
1.3.3 Proprietà dell'Ambiente-Problema	13
1.4 Struttura di un Agente	14
1.4.1 Strutture di Agenti Caratteristici	14
1.4.2 Tipi di rappresentazione	17
2 Problem Solving	19
2.1 Agenti Risolutori di Problemi	19
2.1.1 Processo di risoluzione	19
2.2 Algoritmi di Ricerca	20
2.2.1 Ricerca ad Albero	20
2.2.2 Breadth-First	21
2.2.3 Depth-First	23
2.2.4 Depth-First ricorsiva	23
2.2.5 Depth-Limited	23
2.2.6 Iterative-Deepening	24
2.3 Direzione della Ricerca	24
2.4 Problematiche	25
2.4.1 Tre soluzioni	26
2.5 Uniform-Cost	26
2.6 Confronto delle Strategie (albero)	27
2.7 Conclusioni	27
3 Ricerca Euristica	29
3.1 Funzione di Valutazione Euristica	29
3.2 Best-First	29
3.2.1 Algoritmo A	30
3.2.2 Algoritmo A*: La Stima Ideale	31
3.2.3 Perché A* è vantaggioso?	32
3.2.4 Conclusioni su A*	33
3.2.5 Casi speciali di A	33
3.3 Costruire le Euristiche di A*	34
3.3.1 Valutazione di funzioni euristiche	34
3.3.2 Confronto di euristiche ammissibili	34
3.3.3 Misura del potere euristico	34
3.3.4 Capacità di esplorazione	35
3.4 Come si inventa un'euristica?	35

3.4.1	Rilassamento del problema	35
3.4.2	Massimizzazione di euristiche	35
3.4.3	Pattern Disgiunti	36
3.4.4	Apprendere dall'esperienza	36
4	Algoritmi Evolutivi Basati su A*	37
4.1	Beam Search	37
4.2	IDA*	37
4.3	RBFS	37
4.4	A* con memoria limitata	37
4.5	Conclusioni	38
5	Oltre la Ricerca Classica	39
5.1	Verso ambienti più realistici	39
5.2	Ricerca Locale	39
5.2.1	Algoritmi di ricerca locale	39
5.2.2	Panorama dello spazio degli stati	40
5.2.3	Algoritmo Hill Climbing	40
5.2.4	Algoritmo di Tempra Simulata	42
5.2.5	Algoritmo Local Beam	42
5.2.6	Algoritmo Beam Search Stocastico	43
5.3	Algoritmi Genetici	43
5.3.1	Gradient	44
5.4	Ambienti più realistici	45
5.4.1	Azioni non deterministiche	45
5.4.2	Come si pianifica	45
6	I Giochi con Avversario	47
6.1	Giochi con Avversario	47
6.1.1	Ciclo <i>pianifica-agisci-percepisci</i>	47
6.2	Giochi come problemi di ricerca	48
6.2.1	Algoritmo MinMax	48
6.2.2	Algoritmo Min-Max Euristic (con orizzonte)	50
6.2.3	Potatura Alfa-Beta	51
7	Problemi di Soddisfacimento di Vincoli	53
7.1	Formulazione	53
7.2	Strategie per problemi CSP	53
7.2.1	Ricerca in problemi CSP	54
7.2.2	Backtracking	54
8	Agenti Basati su Conoscenza	57
8.1	Agenti Knowledge-Based	57
8.1.1	Il mondo del Wumpus	58
8.1.2	Knowledge-Base	58
8.1.3	Algoritmo TT-entails	59
8.2	Algoritmi per la soddisfabilità (SAT)	60
8.2.1	Algoritmo DPLL	61
8.2.2	Metodi locali per SAT	62
8.2.3	Algoritmo WalkSAT	62
8.3	Inferenza come Deduzione	63
8.3.1	Regola di risoluzione per PROP	64
8.3.2	Logica del Primo Ordine	65
8.3.3	Inferenza nella Logica del Prim'Ordine	65
8.3.4	Teorema di Herbrand	65
8.3.5	Regola di risoluzione per il FOL	65
8.4	Definizione e Confronto di Euristiche Ammissibili	68

9 Strategie di risoluzione	69
9.1 Strategia di risoluzione	69
9.1.1 Strategie di Cancellazione	69
9.1.2 Strategie di Restrizione	70
9.1.3 Strategie di Ordinamento	71
9.1.4 Sottoinsieme a regole del FOL	71
9.1.5 Sistemi a regole logici	71
9.1.6 Programmazione Logica	71
9.1.7 Risoluzione SLD	72
9.2 Sistemi a Regole in Avanti	73
9.2.1 Analisi di FOL-FC-Ask	74
9.2.2 FC Efficiente	74
10 Machine Learning	77
10.1 Introduzione al Machine Learning	77
10.2 Concept Learning	80
10.2.1 Supervised Learning	80
10.2.2 Regole congiuntive	80
10.2.3 Rappresentare le ipotesi	81
10.2.4 Numerare le istanze, concetti, ipotesi	82
10.2.5 Algoritmo Find-S	83
10.2.6 Algoritmo List-Then-Eliminate	83
10.2.7 Rappresentare i Version Spaces	84
10.2.8 Algoritmo Candidate Elimination	84
10.3 Bias Induttivo ed il suo ruolo	84
10.4 Sistemi Induttivi e Sistemi Deduttivi Equivalenti	85
10.5 Modelli Lineari	86
10.5.1 Problemi di Regressione	86
10.5.2 Gradient Descent Algorithm	89
10.6 Ridge Regression	91
10.7 Classificazione	91
10.7.1 Problema d'apprendimento per classificatori lineari	92
10.7.2 Algoritmo di Apprendimento	92
10.8 Conclusione sui modelli lineari	93
10.9 Alberi di Decisione	93
10.9.1 Algoritmo Top-Down Induction (ID3)	93
10.9.2 Ricerca nello spazio delle ipotesi dei Decision Trees	94
10.9.3 Problemi nell'apprendimento con DT	95
10.10 Validazione	96
10.11 SLT	98
10.12 Support Vector Machine	99
10.12.1 Maximum Margin Classifier	99
10.12.2 Kernel	101
10.12.3 Kernel notevoli	102
10.12.4 Sintesi su SVM	102
10.12.5 Uso pratico	102
10.13 K-Nearest Neighbors	103
10.13.1 1-Nearest Neighbor	103
10.13.2 K-Nearest Neighbors	104
10.13.3 Complessità	105
10.13.4 Un estremo	105
10.13.5 Qualche limite di K-NN	105
10.13.6 Considerazioni	106

11 Vari Modelli**107**

11.1 Overview sul supervised learning	107
11.2 Unsupervised learning	107
11.3 K-Means	108
11.4 Dimensionality Reduction	109
11.5 Altri task nell'ML	109
11.5.1 Reinforcement Learning	109
11.5.2 Semi-supervised learning	109
11.5.3 On-Line Learning	109
11.5.4 Apprendimento di domini strutturati e apprendimento relazionale	110
11.5.5 Reti Neurali	110
11.5.6 Deep Learning	111

12 Applicazioni e oltre IIA**113**

12.0.1 Pattern Recognition	113
12.0.2 Data Mining/Knowledge Discovery	113
12.0.3 Computational Intelligence	113
12.0.4 Deep Learning	113

Introduzione

Alessio Micheli, Maria Simi

elearning.di.unipi.it/course/view.php?id=174

Intelligenza Artificiale si occupa della **comprendione** e della **riproduzione** del comportamento *intelligente*.

Psicologia cognitiva: obiettivo comprensione intelligenza umana, costruendo modelli computazionali e verifica sperimentale.

Approccio costruttivo: costruire entità dotate di intelligenze e **razionalità**. Questo tramite codifica del pensiero razionale per risolvere problemi che richiedono intelligenza non necessariamente facendolo come lo fa l'uomo.

Definizioni di IA: pensiero-azione, umanamente-razionalmente.

Costruire macchine intelligenti sia che operino come l'uomo che diversamente.

formalizzaz conoscenze e meccanizzazione ragionemtno in tutti i settori dell'uomo

comprendione tramite modelli comp della psicologia e comportamento di uomini, animali ecc

rendere il lavoro con il calcolatore altrettanto facile e utile che del lavoro con persone capaci, abili e disponibili.

Poniamo definizione di IA: arte di creare macchine che svolgono funzioni che richiedono intelligenza quando svolte da esseri umani. Non definisce "Intelligenza", cosa significa "intelligente"?

Capitolo 1

Agenti Intelligenti

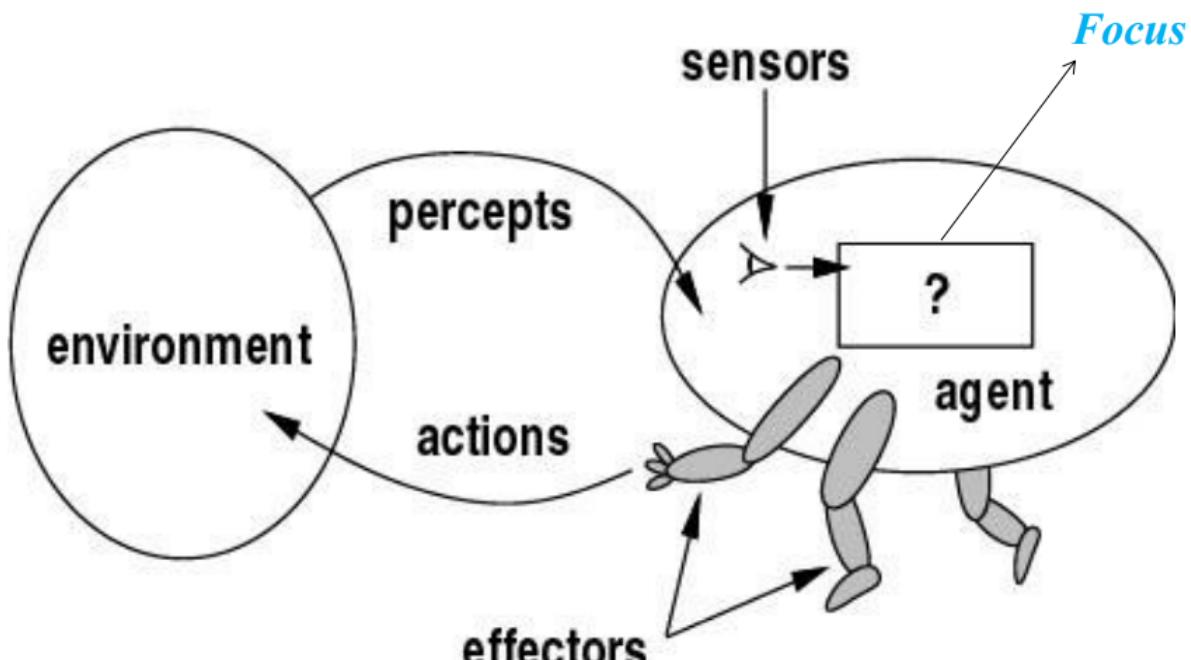
1.1 Intelligenza

Durante il progresso nell'area di ricerca, l'**intelligenza** è stata vista come l'**avere diverse capacità**: buon senso, interazione con un ambiente, acquisizione di esperienza, comunicazione, ragionamento logico...

L'intelligenza quindi non è una collezione di tecniche per risolvere problemi **specifici**, ma per l'informatica consiste nel fornire metodologie sistematiche per dotare le macchine di comportamenti intelligenti/*razionali* su problemi generali *difficili*.

1.2 Agenti

Iniziamo con inquadrare gli **agenti**. L'approccio moderno dell'IA consiste nella costruzione di agenti intelligenti. Questa visione ci offre un quadro di riferimento ed una prospettiva **diversa** all'analisi dei sistemi software. Il primo obiettivo sarà di costruire agenti per la risoluzione di problemi vista come una **ricerca in uno spazio di stati** (**problem solving**)



Ciclo percezione- azione

1.2.1 Caratteristiche

Sono qualcosa di più di un modulo software.

Situati Gli agenti sono situati in un ambiente da cui ricevono percezioni e su cui agiscono mediante azioni (attuatori).

Sociali Gli agenti hanno abilità sociali: comunicano, collaborano e si difendono da altri agenti.

Credenze, obiettivi, intenzioni...

Corpo Gli agenti hanno un corpo, sono embodied fino a considerare i meccanismi delle emozioni.

1.2.2 Percezioni e Azioni

Percezione Una percezione è un input da sensori.

Sequenza percettiva Storia completa delle percezioni

La scelta delle azioni è unicamente determinata dalla sequenza percettiva.

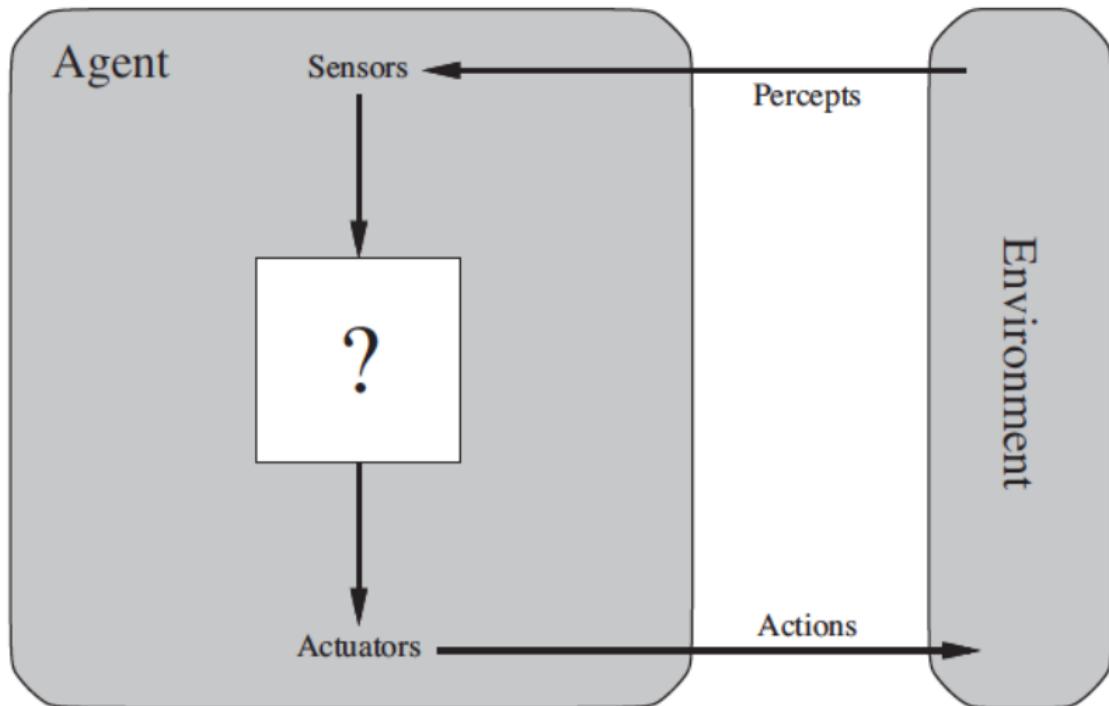
Funzione Agente Definisce l'azione da intraprendere per ogni sequenza percettiva e **descrive completamente** l'agente. Implementata da un **programma agente**.

$$\text{Sequenza Percettiva} \xrightarrow{f} \text{Azione}$$

Il compito dell'IA è progettare il programma agente.

1.2.3 Agente e ambiente

Architettura astratta



Esempi

Agente robotico Percepisce con camera, microfoni e sensori. Interagisce con motori, voce...

Agente finanziario Percepisce i tassi, le news. Interagisce con acquisti e scambi.

Agente di gioco Percepisce le mosse dell'avversario. Interagisce tramite le proprie mosse.

Agente diagnostico Percepisce i sintomi e le analisi dei pazienti. Interagisce fornendo la diagnosi.

Agente web Percepisce le query utente e le pagine web. Interagisce fornendo i risultati di ricerca.

1.2.4 Agenti Razionali

Agenti razionali Un agente razionale **interagisce con l'ambiente in maniera efficace**: *"fa la cosa giusta"*. L'agente razionale raggiunge l'obiettivo nella maniera più efficiente. Serve quindi una **misura di prestazione**, di *come vogliamo che il mondo evolva*, a seconda del problema e considerato l'ambiente.

Esterna, perché bisogna definirla *prima* di agire. Non si può definire l'obiettivo dopo aver iniziato ad agire, altrimenti non è significativo.

Esempio: la volpe che non arriva all'uva.

Scelta dal progettista a seconda del problema e considerando l'effetto che ha sull'ambiente.

Razionalità La razionalità è relativa/dipende da:

Misura delle prestazioni

Conoscenze pregresse dell'ambiente

Percezioni presenti e passate (sequenza percettiva)

Capacità dell'agente (le azioni possibili)

Definizione Un **agente razionale**, quindi, **esegue l'azione che massimizza il valore atteso della misura delle prestazioni per ogni sequenza di percezioni**, considerando le sue percezioni passate e la sua conoscenza pregressa.

Non si pretende perfezione e conoscenza del futuro, ma massimizzare il risultato *atteso*. Potrebbero essere necessarie azioni di acquisizione di informazioni o esplorative (**non onniscienza**).

Le capacità dell'agente possono essere limitate (**non onnipotenza**).

Razionalità e apprendimento Raramente il programmatore può fornire a priori tutta la conoscenza sull'ambiente. L'agente razionale, quindi, **deve essere in grado di modificare il proprio comportamento con l'esperienza**, cioè con le percezioni passate.

Può migliorarsi esplorando, **apprendendo**, aumentando la propria autonomia per operare in ambienti differenti o mutevoli.

1.2.5 Agenti Autonomi

Un agente è **autonomo quando il suo comportamento dipende dalla sua esperienza**. Se il suo comportamento fosse determinato solo dalla propria conoscenza *built-int* allora sarebbe **non autonomo** e poco flessibile.

1.3 Ambienti

Definire un problema per un agente significa **caratterizzare l'ambiente in cui lavora**, cioè l'**ambiente operativo**. L'agente razionale è la soluzione del problema.

1.3.1 PEAS

Performance, prestazioni

Environment, ambiente

Actuators, attuatori

Sensors, sensori

Esempio Autista di taxi

Prestazione	Ambiente	Attuatori	Sensori
Arrivare alla destinazione, sicuro, veloce, ligio alla legge, confortevole, consumo minimo di benzina, profitti massimi	Strada, altri veicoli, clienti	Sterzo, acceleratore, freni, frecce, clacson	Telecamere, sensori, GPS, contachilometri, accelerometro, sensori del motore...

Formulazione PEAS dei problemi

Problema	P	E	A	S
Diagnosi medica	Diagnosi corretta	Pazienti, ospedale	Domande, suggerimenti, test, diagnosi	Sintomi, test clinici, risposte del paziente
Analisi immagini	Numero di immagini/zona correttamente classificate	Collezione di fotografie	Etichettatore di zone nell'immagine	Array di pixel
Robot "selezionatore"	Numero delle parti correttamente classificate	Nastro trasportatore	Raccogliere le parti e metterle nei cestini	Telecamera (pixel di varia intensità)
Giocatore di calcio	Fare più goal dell'avversario	Altri giocatori, campo di calcio, porte	Dare calci al pallone, correre	Locazione del pallone, dei giocatori e delle porte

1.3.2 Simulatore di Ambienti

Uno **strumento software** con il compito di:

Generare gli stimoli per gli agenti

Raccogliere le azioni in risposta

Aggiornare lo stato dell'ambiente

Opzionalmente, attivare altri processi che influenzano l'ambiente

Valutare le prestazioni degli agenti

Gli esperimenti su classi di ambienti (variando le condizioni) sono essenziali per valutare la capacità di generalizzare. La valutazione delle prestazioni è fatta tramite la media su più istanze.

1.3.3 Proprietà dell'Ambiente-Problema

- Osservabilità

Completemente osservabile: l'apparato percettivo è in grado di dare una conoscenza completa dell'ambiente o almeno tutto quello che serve a decidere l'azione.

Parzialmente osservabile: sono presenti limiti o inaccuratezze nell'apparato sensoriale. (Es. la videocamera di un rover vede solo parte dell'ambiente in un dato istante).

- Singolo/Multi-Agente

Distinzione tra agente e non agente: il mondo può cambiare anche attraverso **eventi**, non necessariamente per le azioni di agenti.

Multi-Agente Competitivo, come gli scacchi: comportamento randomizzato ma razionale.

Multi-Agente Cooperativo, o benigno: stesso obiettivo e comunicazione.

- Predicibilità

Deterministico: lo stato successivo è completamente determinato dallo stato corrente e dall'azione.

Stocastico: esistono elementi di incertezza con probabilità associata. Es: guida, tiro in porta.

Non deterministico: si tiene traccia di più stati possibili che sono risultato dell'azione, ma non in base ad una probabilità.

- Episodico:

l'esperienza dell'agente è divisa in episodi atomici indipendenti. In ambienti episodici non c'è bisogno di pianificare.

Sequenziale: ogni decisione influenza le successive.

- Statico:

il mondo non cambia mentre l'agente decide l'azione.

Dinamico: l'ambiente cambia nel tempo, va osservata la contingenza. Tardare equivale a non agire.

Semi-dinamico: l'ambiente non cambia ma la valutazione dell'agente sì. Es: scacchi con timer, se non agisco prima dello scadere perdo.

- Discreto/Continuo

Lo stato, il tempo, le percezioni e le azioni sono tutti elementi che possono assumere valori discreti o continui. Combinatoriale (nel discreto) *vs* infinito (nel continuo).

- Noto/Ignoto

Distinzione riferita allo stato di conoscenza dell'agente sulle leggi fisiche dell'ambiente. **L'agente conosce l'ambiente o deve compiere azioni esplorative?**

Noto ≠ osservabile: posso giocare a carte coperte, ma con regole note.

Ambienti reali Parzialmente osservabili, stocastici, sequenziali, dinamici, continui, multi-agente e ignoti.

1.4 Struttura di un Agente

$$\text{Agente} = \text{Architettura} + \text{Programma}$$

$$\text{Ag: P} \rightarrow \text{Az}$$

L'Agente associa **Azioni** alle **Percezioni**. Il **programma dell'agente** implementa la funzione **Ag**.

Programma Agente Pseudocodice del programma agente.

```
function Skeleton-Agent(percept) returns action
    static: memory #la memoria del mondo posseduta dall'agente
    memory <- UpdateMemory(memory, percept)
    action <- Choose-Best-Action(memory) #Cuore dell'IA
    memory <- UpdateMemory(memory, action)
    return action
```

1.4.1 Strutture di Agenti Caratteristici

Agente basato su tabella La scelta dell'azione è un accesso ad una tabella che **associa un'azione ad ogni possibile sequenza di percezioni**.

Vari problemi:

Le **dimensioni** possono essere proibitive: per giocare a scacchi, la tabella dovrebbe contenere un numero di righe nell'ordine di $10^{120} >> 10^{80}$ numero di atomi nell'universo osservabile.

Difficile da costruire

Nessuna autonomia

Difficile da aggiornare, apprendimento complesso.

Con le IA vogliamo realizzare **automi razionali con un programma compatto**.

Agente Reattivo Semplice



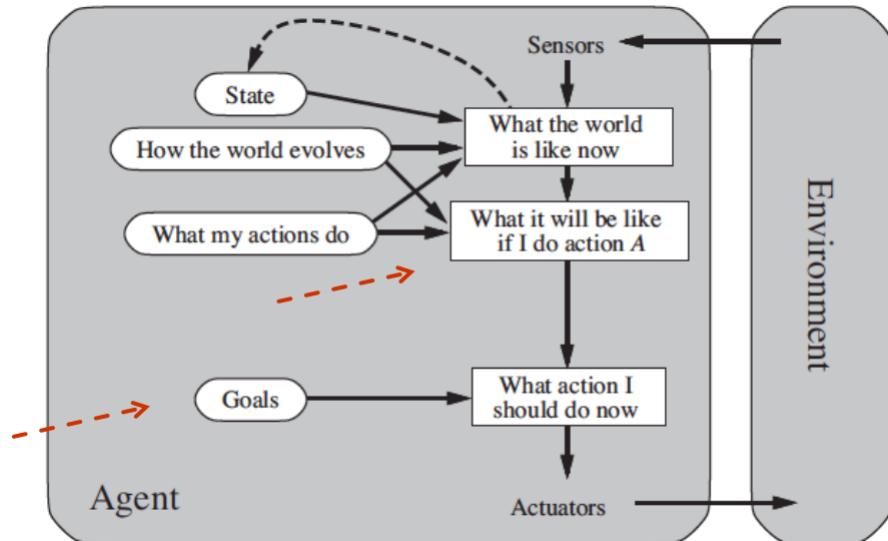
```
function Agente-Reattivo-Semplice(percezione) returns azione
    persistent: regole #insieme di regole condizione-azione (if-then)
    stato <- Interpreta-Input(percezione)
    regola <- Regola-Corrispondente(stato, regole)
    azione <- regola.Azione
    return azione
```

Agenti basati su modello



```
function Agente-Basato-su-Modello(percezione) returns azione
    persistent:      stato #descrizione dello stato corrente
                      modello #conoscenza del mondo
                      regole #insieme di regole condizione-azione
                      azione #azione piu recente
    stato <- Aggiorna-Stato(stato, azione, percezione, modello)
    regola <- Regola-Corrispondente(stato, regole)
    azione <- regola.Azione
    return azione
```

Agenti con obiettivo Bisogna pianificare una sequenza di azioni per raggiungere l'obiettivo. (In rosso sono indicate le parti aggiunte)



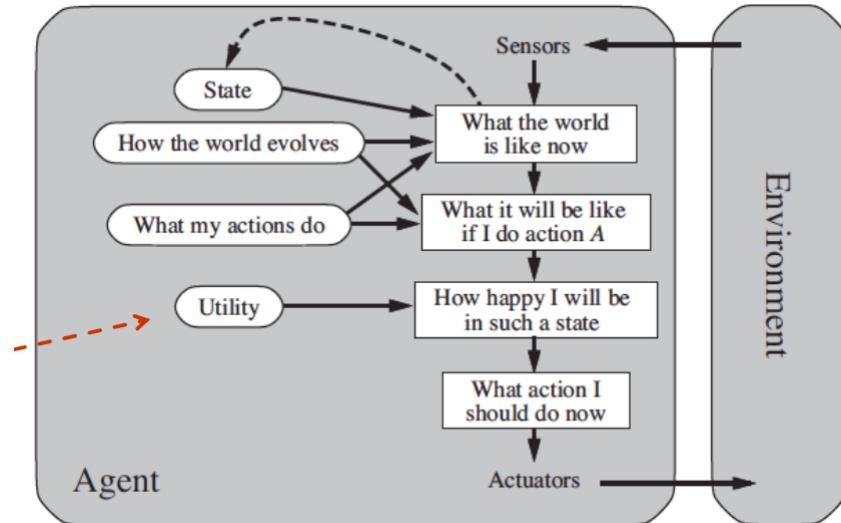
Sono guidati da un obiettivo nella scelta che intraprendono, è stato fornito un **goal esplicito**: per esempio una città da raggiungere.

A volte l'azione migliore dipende dall'obiettivo da raggiungere (*da che parte devo girare?*)

Devo **pianificare una sequenza di azioni** per raggiungere l'obiettivo. Sono meno efficienti ma **più flessibili** rispetto ad un agente reattivo. L'obiettivo può cambiare, non è codificato nelle regole.

Esempio classico: ricerca della sequenza di azioni per raggiungere una data destinazione.

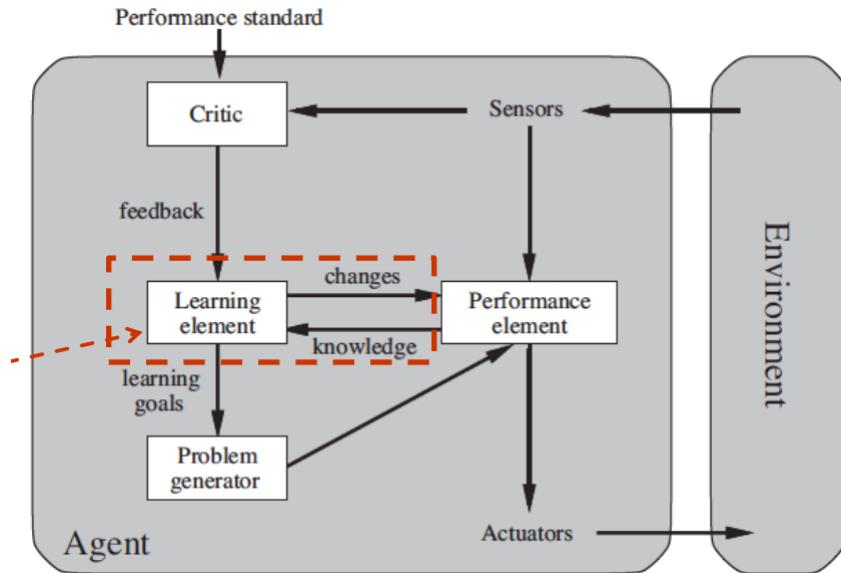
Agenti con valutazione di utilità



Obiettivi alternativi, o più modi per raggiungerlo: l'agente deve decidere verso quali muoversi, quindi è **necessaria una funzione di utilità** che associa ad uno stato obiettivo un numero reale.

Obiettivi più facilmente raggiungibili di altri: la funzione di utilità **tiene conto della probabilità di successo e/o di ciascun risultato (utilità attesa o media)**

Agenti che apprendono



Componente di apprendimento: produce cambiamenti al programma agente. Migliora le prestazioni, adattando i suoi componenti ed apprendendo dall'ambiente

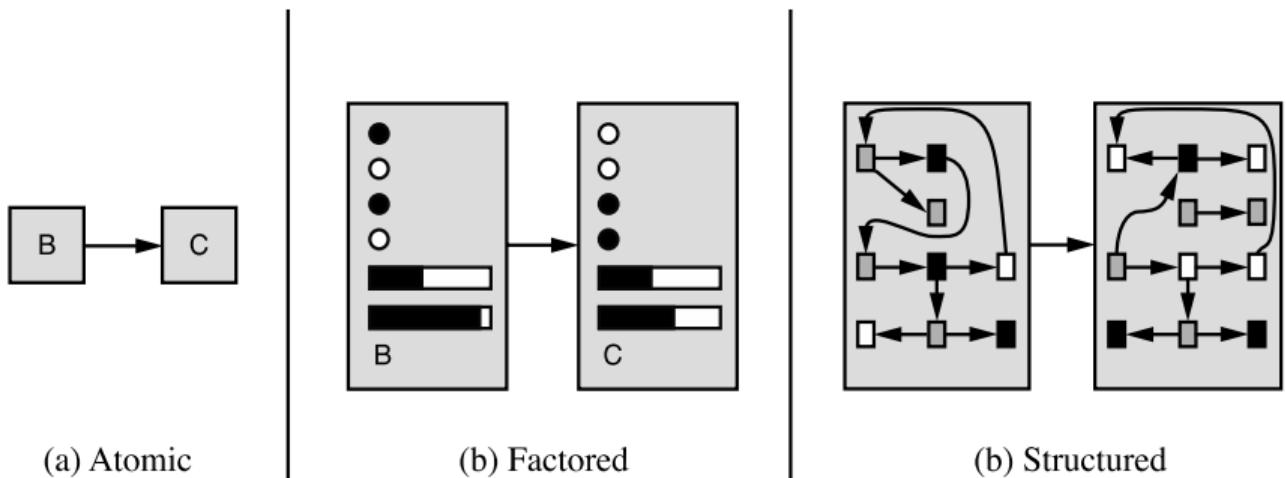
Elemento esecutivo: il programma agente

Elemento critico: osserva e dà feedback sul comportamento

Generatore di problemi: suggerisce nuove situazioni da esplorare

1.4.2 Tipi di rappresentazione

Stati e transizioni



Rappresentazione atomica (stati)

Rappresentazione fattorizzata (+ variabili e attributi)

Rappresentazione strutturata (+ relazioni)

Capitolo 2

Problem Solving

2.1 Agenti Risolutori di Problemi

Problem Solving Questi agenti adottano il paradigma della **risoluzione di problemi** come ricerca in uno spazio di stati (**problem solving**). Sono **agenti con modello** (storia, percezioni) che **adottano una rappresentazione atomica dello stato**. Sono **particolari agenti con obiettivo** che **pianificano l'intera sequenza di azioni** prima di agire.

2.1.1 Processo di risoluzione

Passi da seguire

1. **Determinazioni dell'obiettivo:** un insieme di stati dove l'obiettivo è soddisfatto.
2. **Formulazione del problema:** rappresentazione degli stati e delle azioni.
Fa parte del design "umano".
3. **Determinazione della soluzione** mediante ricerca: un piano d'azione
4. **Esecuzione del piano**
Soluzione algoritmica.

La determinazione dell'obiettivo e la formulazione del problema richiede **tanta intelligenza**, che in fase di design è **spostata sull'umano**. Gli algoritmi sono ancora **"stupidi"**.

Assunzioni sull'ambiente **Statico, osservabile** (so dove sono, es: *viaggio con la mappa*), **discreto** (insieme finito di azioni possibili), **deterministico** (una azione \Rightarrow un risultato. L'agente può eseguire il piano *"ad occhi chiusi"*, niente può andare storto)

Formulazione del problema Un problema può essere **definito formalmente** mediante cinque componenti:

1. **Stato iniziale**
2. **Azioni possibili** nello stato s : **Azioni(s)**
3. **Modello di transizione**
Risultato: stato \times azione \longrightarrow stato
Risultato(s, a): s' , uno stato **successore**
4. **Test obiettivo:** un insieme di stati obiettivo
Goal-Test: stato $\longrightarrow \{\text{true}, \text{false}\}$
5. **Costo del cammino:** somma dei costi delle azioni (costo dei passi).
Costo di un passo: $c(s, a, s')$, mai negativo.

1, 2 e 3 **definiscono implicitamente lo spazio degli stati**. Definirlo esplicitamente può essere molto oneroso, come in quasi tutti i problemi di IA.

2.2 Algoritmi di Ricerca

Il processo che cerca una sequenza di azioni che raggiunge l'obiettivo è detto ricerca.

Algoritmi Gli algoritmi di ricerca prendono in **input** un problema e restituiscono un cammino soluzione, un cammino che porta dallo stato iniziale allo stato goal.

Misura delle prestazioni Trova una soluzione? Quanto costa trovarla? Quanto è efficiente la soluzione?

$$\text{Costo Totale} = \text{Costo della Ricerca} + \text{Costo del Cammino Soluzione}$$

Valuteremo algoritmi sul primo, ottimizzando il secondo.

2.2.1 Ricerca ad Albero

Generazione di un **albero di ricerca sovrapposto allo spazio degli stati**. Ricerca significa **approfondire l'opzione**, mettendo da parte le altre che verranno riprese se non trovo la soluzione.

Quindi l'albero viene generato esplorando i vari nodi partendo dallo stato iniziale. Il nodo è diverso dallo stato: per esempio, in un grafo rappresentante le città, se parto da città A ed esploro l'opzione nodo B, il nodo B avrà come figlio anche città A perché posso tornarci.



Algoritmo Ricerca ad albero, ossia senza controllare se i nodi (**stati**) siano già stati esplorati.

```

function Ricerca-Albero(problema) returns soluzione oppure fallimento
    #Inizializza la frontiera con stato iniziale del problema
    loop do
        if (frontiera vuota)
            return fallimento
        #Scegli* un nodo foglia da espandere e rimuovilo dalla frontiera
        if (nodo contiene uno stato obiettivo)
            return soluzione corrispondente
        #Espandi il nodo e aggiungi i successori alla frontiera

```

* = **strategia**: quale scegliere? I vari algoritmi si differenziano per la strategia di scelta.

Un **nodo** n è una **struttura dati con quattro componenti**

Stato, n.estado

Padre, n.padre

Azione effettuata per generarlo, n.azione

Costo del cammino dal nodo iniziale al nodo, n.costo-cammino

Indicata come $g(b) = \text{padre.costo-cammino} + \text{costo-passo ultimo}$

Frontiera Lista dei **nodi in attesa di essere espansi**, cioè le foglie dell'albero di ricerca. Implementata come una coda con operazioni:

- Vuota(coda)
- Pop(coda) estrae l'ultimo elemento (implementa la strategia)
- Inserisci(elemento, coda)

Diversi tipi di coda hanno differenti funzioni di inserimento e **implementano strategie diverse**.

FIFO → BF

Viene estratto l'elemento più vecchio, cioè in attesa da più tempo. Nuovi nodi aggiunti alla fine

LIFO → DF

Viene estratto l'ultimo elemento inserito. Nuovi nodi aggiunti all'inizio

Con priorità → UC, altri...

Viene estratto l'elemento con priorità più alta in base ad una funzione di ordinamento. All'aggiunta di un nuovo nodo si riordina.

Strategie non informate

- Ricerca in **ampiezza** (BF)
- Ricerca in **profondità** (DF)
- Ricerca in **profondità limitata** (DL)
- Ricerca con **apprendimento iterativo** (ID)
- Ricerca di **costo uniforme** (UC)

Strategie informate Anche dette di **ricerca euristica**: fanno uso di informazioni riguardo la distanza stimata della soluzione.

Valutazione di una strategia

Completezza: se la soluzione esiste viene trovata

Ottimalità (ammissibilità): trova la soluzione migliore, con costo minore

Complessità in tempo: tempo richiesto per trovare la soluzione

Complessità in spazio: memoria richiesta

2.2.2 Breadth-First

Ricerca in ampiezza Esplorare il grafo dello spazio degli stati a livelli progressivi di stessa profondità. Implementata con una coda FIFO. **Algoritmo su albero**:

```
function RicercaAmpiezzaA(problema)      returns soluzione oppure fallimento
    nodo = un nodo con stato = problema.stato-iniziale e costo-di-cammino = 0
    #Stati goal-tested alla generazione: maggior efficienza si ferma appena trova goal
    if (problema.TestObiettivo(nodo.Stato)) return Soluzione(nodo)
    frontiera = una coda FIFO con nodo come unico elemento
    loop do
        if (Vuota(frontiera)) return fallimento
        nodo = Pop(frontiera)
        for each azione in problema.Azioni(nodo.Stato) do #Espansione
            figlio = Nodo-Figlio(problema, nodo, azione) #costruttore: vedi AIMA
            if (Problema.TestObiettivo(figlio.Stato)) return Soluzione(figlio)
            frontiera = Inserisci(figlio, frontiera) #frontiera coda FIFO
```

Algoritmo su grafo evitando di espandere stati già esplorati:

```
function RicercaAmpiezzaG(problema) returns soluzione oppure fallimento
    nodo = un nodo con stato = problema.stato-iniziale e costo-di-cammino = 0
    if (problema.TestObiettivo(nodo.Stato)) return Soluzione(nodo)
    frontiera = una coda FIFO con nodo come unico elemento
    esplorati = insieme vuoto #gestisco stati ripetuti
    loop do
        if (Vuota(frontiera)) return fallimento
        nodo = POP(frontiera) #aggiungi nodo.Stato a esplorati
        for each azione in problema.Azioni(nodo.Stato) do
            figlio = Nodo-Figlio(problema, nodo, azione)
            if (figlio.Stato non in esplorati e non in frontiera)
                if (Problema.TestObiettivo(figlio.Stato)) return Soluzione(figlio)
                frontiera = Inserisci(figlio, frontiera) #in coda
```

Python

```
def breadth_first_search(problem): """Ricerca-grafo in ampiezza"""
    explored = [] # insieme degli stati già visitati (implementato come una lista)
    node = Node(problem.initial_state) #il costo del cammino è inizializzato nel costruttore
    if problem.goal_test(node.state):
        return node.solution(explored_set = explored)
    frontier = FIFOQueue() # la frontiera è una coda FIFO
    frontier.insert(node)
    while not frontier.isempty(): # seleziona il nodo per l'espansione
        node = frontier.pop()
        explored.append(node.state) # inserisce il nodo nell'insieme dei nodi esplorati
        for action in problem.actions(node.state):
            child_node = node.child_node(problem, action)
            if (child_node.state not in explored) and
               (not frontier.contains_state(child_node.state)):
                if problem.goal_test(child_node.state):
                    return child_node.solution(explored_set = explored)
                # se lo stato non è uno stato obiettivo allora inserisci il nodo nella frontiera
                frontier.insert(child_node)
    return None # in questo caso ritorna con fallimento
```

Analisi della complessità spazio-temporiale Assumiamo:

b = fattore di ramificazione (**branching**)

d = profondità del nodo obiettivo più superficiale (**depth**)
Più vicino all'iniziale

m = lunghezza massima dei cammini nello spazio degli stati (**max**)

Analisi:

Strategia **completa**

Strategia **ottimale** se gli operatori hanno tutti lo stesso costo k cioè $g(n) = k \cdot \text{depth}(n)$, dove $g(n)$ è il costo del cammino per arrivare ad n .

Complessità nel tempo (nodi generati)
 $T(b, d) = b + b^2 + \dots + b^d = O(b^d)$, con b figli per ogni nodo.

Complessità nello spazio (nodi in memoria): $O(b^d)$

2.2.3 Depth-First

Ricerca in profondità Implementata da una coda che mette i successori in testa alla lista (LIFO, pila o stack). Algoritmo generale visto all'inizio, su grafo o albero.

Analisi (su albero) Poniamo m lunghezza massima dei cammini nello spazio degli stati e b fattore di diramazione Tempo: $O(b^m)$ che può essere anche $> O(b^d)$

Spazio: $b \cdot m$, frontiera sul cammino perché vengono cancellati i rami completamente esplorati ma mantenuti i fratelli del path corrente.

Non completa (loop) e **non ottimale**, ma drastico risparmio di memoria.

BF, $d = 16 \rightarrow 10$ Esabyte

DF, $d = 16 \rightarrow 156$ Kilobyte

Analisi (su grafo) In caso di DF su grafo si perdono i vantaggi di memoria: torna a tutti i possibili stati (al caso pessimo diventa esponenziale come BF) per mantenere la lista dei visitati, ma così DF diventa **completa** in spazi degli stati finiti (al caso pessimo tutti i nodi vengono espansi).

Rimane non completa in spazi infiniti.

Possibile controllare anche solo i nuovi stati rispetto al cammino radice-nodo corrente senza aggravio di memoria. Si evitano i cicli finiti in spazi finiti ma non i cammini ridondanti.

2.2.4 Depth-First ricorsiva

Ancora più efficiente in occupazione di memoria perché mantiene solo il cammino corrente (m nodi al caso pessimo). Realizzata da un algoritmo ricorsivo "con backtracking" che non necessita di tenere in memoria b nodi per ogni livello, ma salva lo stato su uno stack a cui torna in caso di fallimento per fare altri tentativi. **Algoritmo su albero**:

```
function Ricerca-DF-A (problema) returns soluzione oppure fallimento
    return Ricerca-DF-ricorsiva(CreaNodo(problema.Stato-iniziale), problema)

function Ricerca-DF-ricorsiva(nodo, problema) returns soluzione oppure fallimento
    if problema.TestObiettivo(nodo.Stato) return Soluzione(nodo)
    else
        for each azione in problema.Azioni(nodo.Stato) do
            figlio = Nodo-Figlio(problema, nodo, azione)
            risultato = Ricerca-DF-ricorsiva(figlio, problema)
            if risultato != fallimento then return risultato
    return fallimento
```

Python

```
def recursive_depth_first_search(problem, node):
    """Ricerca in profondità ricorsiva"""
    #controlla se lo stato del nodo è uno stato obiettivo
    if problem.goal_test(node.state):
        return node.solution()
    #in caso contrario continua
    for action in problem.actions(node.state):
        child_node = node.child_node(problem, action)
        result = recursive_depth_first_search(problem, child_node)
        if result is not None: return result
    return None #con fallimento
```

2.2.5 Depth-Limited

Ricerca in profondità limitata Si va in profondità fino ad un certo livello predefinito l .

Completa per problemi di cui si conosce un limite superiore per la profondità della soluzione: ad esempio route-finding limitata dal numero di città - 1

Completo se $d < l$

Non ottimale

Complessità in tempo: $O(b^l)$

Complessità in spazio: $O(b \cdot l)$

2.2.6 Iterative-Deepening



Analisi Miglior compromesso tra BF e DF. Nell'ID, i nodi dell'ultimo livello sono generati una volta, quelli del penultimo 2, del terzultimo 3... quelli del primo d volte.

$$\text{ID: } (d)b + (d-1)b^2 + \dots + 3b^{d-2} + 2b^{d-1} + b^d$$

Complessità in tempo: $O(b^d)$

Complessità in spazio: $O(b \cdot d)$, vs $O(b^d)$ del BF.

2.3 Direzione della Ricerca

Altro aspetto usato per ottimizzare la risoluzione di problemi, la **direzione della ricerca** è un problema ortogonale alla **strategia di ricerca**. La ricerca si può fare

In avanti, guidata dai dati come fatto fin'ora: si esplora lo spazio di ricerca dallo stato iniziale allo stato obiettivo

All'indietro o guidata dall'obiettivo: si esplora lo spazio di ricerca a partire da un goal e riconducendosi a sotto-goal fino a trovare uno stato iniziale.

Conviene procedere nella direzione in cui il fattore di diramazione è minore.

Si preferisce la **ricerca all'indietro** quando

l'**obiettivo è chiaramente definito** (es. theorem proving) o si possono **formulare una serie limitata di ipotesi**

i **dati del problema non sono noti** e la loro **acquisizione può essere guidata dall'obiettivo**

mentre si preferisce la **ricerca in avanti** quando

gli **obiettivi possibili sono molti** (es. design)

abbiamo una **serie di dati da cui partire**

Ricerca bidirezionale Si procede nelle due direzioni fino ad incontrarsi

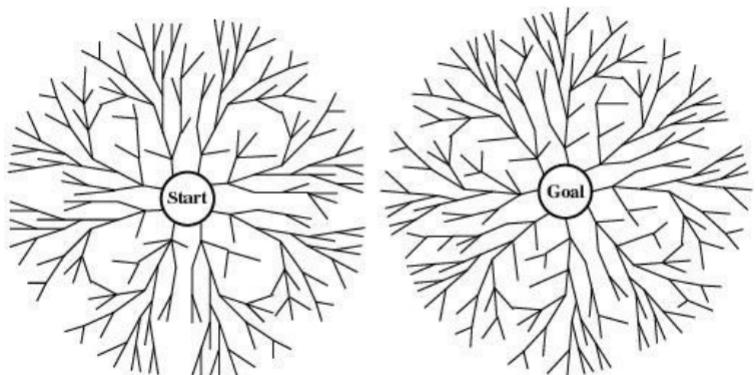
Complessità in tempo: $O(b^{d/2}) = O(\sqrt{b^d})$

Test intersezione in tempo costante, esempio: hash table

Complessità in spazio: $O(b^{d/2}) = O(\sqrt{b^d})$

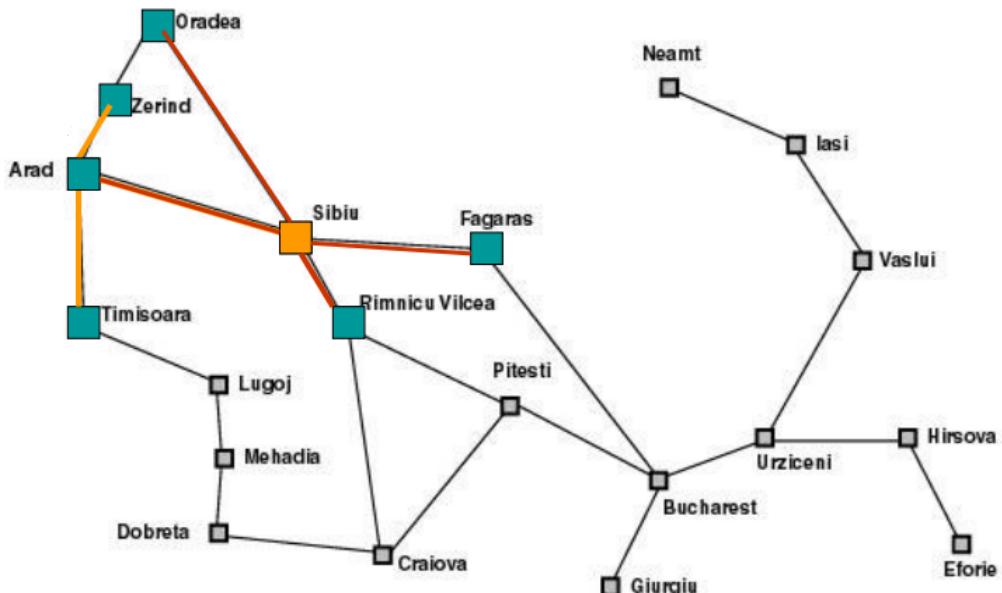
Almeno tutti i nodi in una direzione in memoria, esempio: usando BF

Non è sempre applicabile, ad esempio in casi di predecessori non definiti, troppi stati obiettivo...

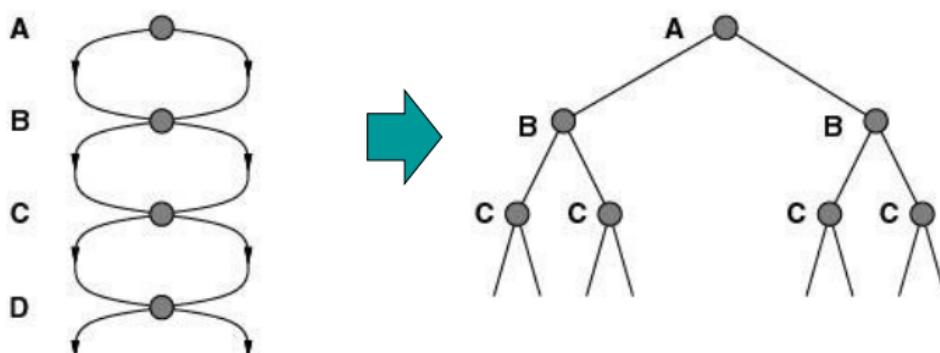


2.4 Problematiche

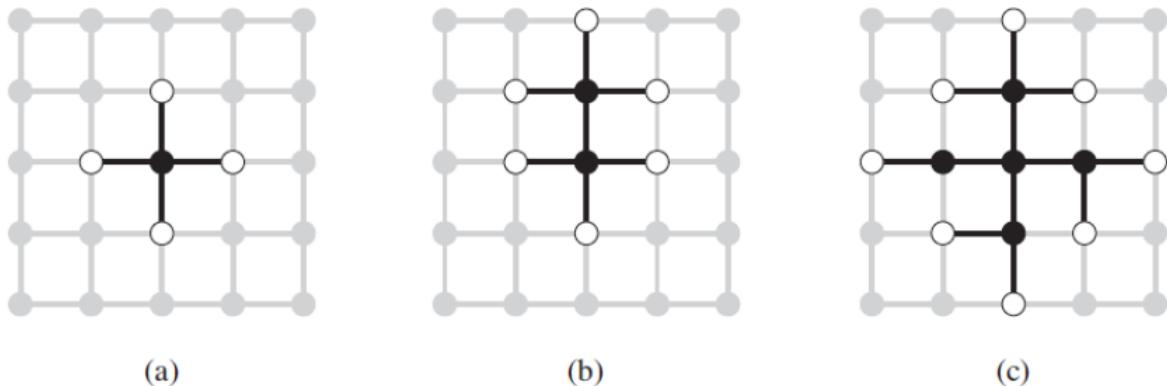
Cammini Ciclici I cammini ciclici potenzialmente rendono gli alberi di ricerca infiniti, anche se con stati finiti.



Ridondanze Su spazi di stati a grafo si generano più volte nodi con lo stesso stato nella ricerca, anche in assenza di cicli.



Un caso è la **ricerca nelle griglie** Visitare stati già visitati fa compiere lavoro inutile. Costo 4^d ma circa $2d^2$ stati distinti.



Come evitarlo?

Compromesso tra spazio e tempo Ricordare gli stati visitati **occupa spazio** ma ci **consente di evitare di visitarli di nuovo**. Gli algoritmi che dimenticano la propria storia sono destinati a ripeterla.

2.4.1 Tre soluzioni

In ordine crescente di costo ed efficacia:

Non tornare nello stato da cui si proviene: si elimina il genitore dai nodi successori.

Non evita i cammini ridondanti.

Non creare cammini con cicli: si controlla che i successori non siano antenati del nodo corrente.

Non generare nodi con stati già visitati/esplorati: ogni nodo visitato deve essere tenuto in memoria per una complessità $O(s)$ dove s è il numero di stati possibili (esempio: hash table per accesso efficiente)

Repetita Il costo può essere alto: in caso di DF la memoria torna da $b \cdot m$ a tutti gli stati, ma diventa una ricerca completa per spazi finiti. Ma **in molti casi gli stati crescono esponenzialmente** (scacchi...)

2.5 Uniform-Cost

Generalizzazione della ricerca in ampiezza (costi diversi tra passi): **si sceglie il nodo di costo $g(n)$ del cammino minore sulla frontiera**, si espande sui contorni di uguale costo (e.g. in km) invece che sui contorni di uguale profondità (BF). Implementata da una **coda ordinata per costo cammino crescente**. **Algoritmo su albero**:

```
function Ricerca-UC-A(problema) returns soluzione oppure fallimento
    nodo = un nodo con stato il problema.stato-iniziale e costo-di-cammino=0
    frontiera = una coda con priorità con nodo come unico elemento
    loop do
        if Vuota?(frontiera) then return fallimento
        nodo = POP(frontiera)
        #Esame post-generaz e vedere costo minore, tipico per coda con priorità
        if problema.TestObiettivo(nodo.Stato) then return Soluzione(nodo)
        for each azione in problema.Azioni(nodo.Stato) do
            figlio = Nodo-Figlio(problema, nodo, azione)
            frontiera = Inserisci(figlio, frontiera) #in coda con priorità
    end
```

Algoritmo su grafo:

```

function Ricerca-UC-G(problema) returns soluzione oppure fallimento
    nodo = un nodo con stato il problema.stato-iniziale e costo-di-cammino=0
    frontiera = una coda con priorita con nodo come unico elemento
    esplorati = insieme vuoto
    loop do
        if Vuota?(frontiera) then return fallimento
        nodo = POP(frontiera);
        if problema.TestObiettivo(nodo.Stato) then return Soluzione(nodo)
        aggiungi nodo.Stato a esplorati
        for each azione in problema.Azioni(nodo.Stato) do
            figlio = Nodo-Figlio(problema, nodo, azione)
            if figlio.Stato non in esplorati e non in frontiera then
                frontiera = Inserisci(figlio, frontiera) #in coda con priorita
            else if figlio.Stato in frontiera con Costo-cammino piu alto then
                sostituisci quel nodo frontiera con figlio

```

Analisi Ottimalità e completezza garantisce purché il costo degli archi sia maggiore di $\epsilon > 0$. Assunto C^* come il costo della soluzione ottima, allora $\lfloor C^*/\epsilon \rfloor$ numero di mosse al caso peggiore, arrotondato per difetto. Tendo ad andare verso tante mosse di costo ϵ prima di una che parta più alta ma poi abbia un path a costo più basso. Complessità: $O(b^{1+\lfloor C^*/\epsilon \rfloor})$.

Quando ogni azione ha lo stesso costo somiglia a BF ma con complessità $O(b^{1+d})$ perché esame e arresto solo dopo aver espanso anche l'ultima frontiera.

2.6 Confronto delle Strategie (albero)

Criteria	BF	UC	DF	DL	ID	Bidirez.
Completa?	Si	Si ¹	No	Si ³	Si	Si
Tempo	$O(b^d)$	$O(b^{1+\lfloor C^*/\epsilon \rfloor})$	$O(b^m)$	$O(b^l)$	$O(b^d)$	$O(b^{d/2})$
Spazio	$O(b^d)$	$O(b^{1+\lfloor C^*/\epsilon \rfloor})$	$O(b \cdot m)$	$O(b \cdot l)$	$O(b \cdot d)$	$O(b^{d/2})$
Ottimale?	Si ²	Si ¹	No	No	Si ²	Si

¹ Per costi degli archi $\geq \epsilon > 0$

² Se gli operatori hanno tutti lo stesso costo

³ Per problemi di cui si conosce un limite alla profondità della soluzione (se $l > d$)

2.7 Conclusioni

Un agente per "problem solving" adotta un paradigma generale di risoluzione dei problemi:

Formula il problema, non automatico

Ricerca la soluzione nello spazio degli stati, automatico

Capitolo 3

Ricerca Euristica

In problemi di complessità esponenziale, come ad es. gli scacchi (10^{120} configurazioni) non è praticabile la ricerca esaustiva. Diventa quindi fondamentale **usare la conoscenza del problema e l'esperienza per riconoscere i cammini più promettenti**, evitando di generare gli altri (**pruning**).

Conoscenza Euristica La **conoscenza euristica** aiuta a fare scelte oculate:

Non evita la ricerca, ma la riduce

Consente, in genere, di trovare una **buona** soluzione in tempi accettabili

Sotto certe condizioni garantisce completezza e ottimalità

3.1 Funzione di Valutazione Euristica

La **conoscenza** del problema è data tramite una **funzione di valutazione** f , che include h detta **funzione di valutazione euristica**:

$$h : n \rightarrow R$$

R = insieme numeri reali. La funzione si applica al nodo, ma dipende solo dallo stato (n.Stato). Per confronto, g dipendeva anche dal cammino fino al nodo. Quindi, la funzione di valutazione

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

dove $g(n)$ è il costo del cammino visto con UC.

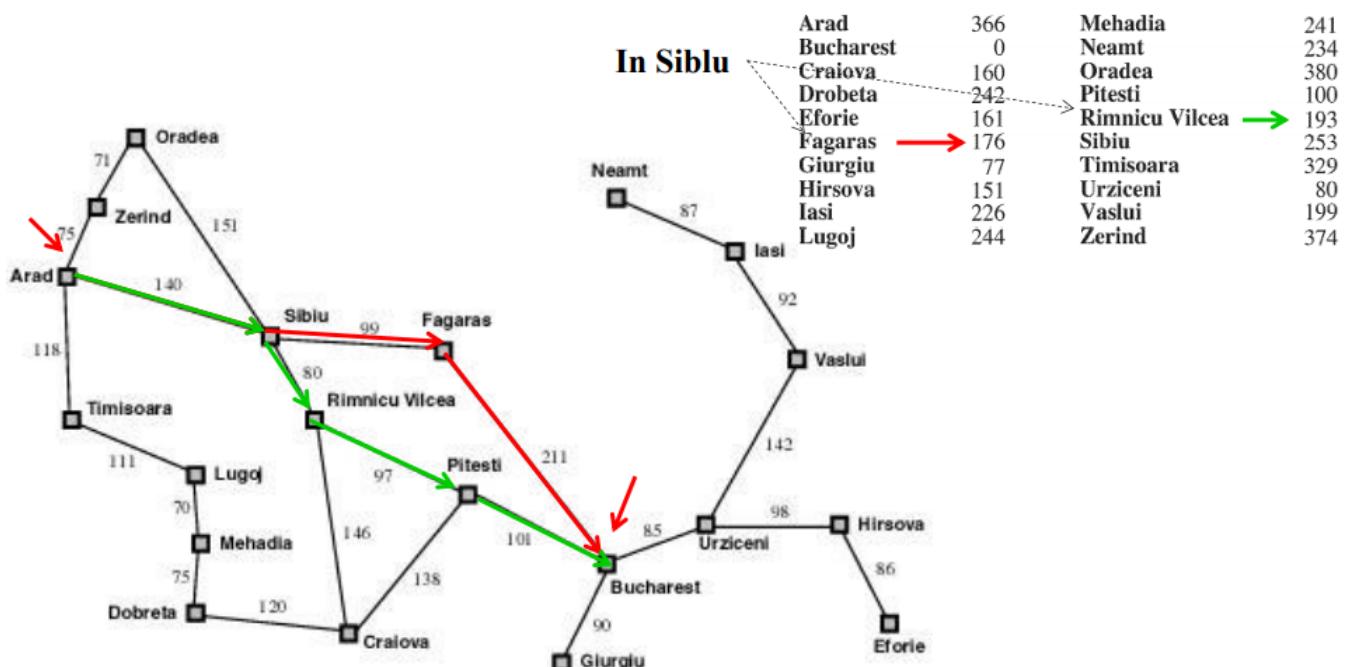
Esempio di euristica h Per procedere preferibilmente verso il percorso migliore, seguendo il "problem-specific information", nel problema del percorso più breve da città a città posso includere nel mio algoritmo le distanze in linea d'aria, oppure il vantaggio in pezzi nella dama o negli scacchi.

3.2 Best-First

Algoritmo di ricerca Best-First Heuristic utilizza lo **stesso algoritmo di Uniform-Cost** ma utilizzando f (stima di costo) per la coda con priorità. La **scelta di f determina la strategia di ricerca**: ad ogni passo si sceglie il nodo sulla frontiera per con valore di f migliore (**nodo più promettente**).

Nota Migliore significa "minore" in caso di un'euristica che stima la distanza della soluzione

Caso Speciale Greedy Best-First, si usa solo h ($f = h$)



Da Arad a Bucarest ...

Greedy best-first: Arad, Sibiu, Fagaras, Bucharest (450)

ma non è l'**Ottimo**: Arad, Sibiu, Rimnicu, Pitesti, Bucarest (418)

3.2.1 Algoritmo A

Si può dire qualcosa di f per avere garanzie di completezza e ottimalità?

Definizione Un algoritmo A è un algoritmo Best-First con una funzione di valutazione dello stato del tipo

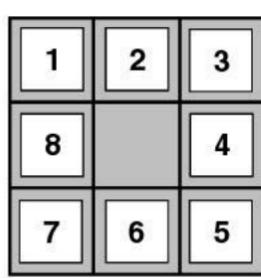
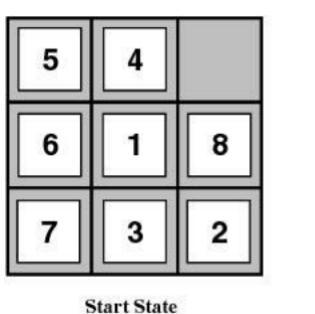
$$f(n) = g(n) + h(n)$$

con $h(n) \geq 0$ e $h(goal) = 0$. $g(n)$ è il **costo del cammino per raggiungere** n e $h(n)$ è **una stima del costo per raggiungere da** n **un nodo** $goal$. Vedremo **casi particolari** dell'algoritmo A:

se $h(n) = 0$, cioè $f(n) = g(n)$, si ha **Ricerca Uniforme** (UC)

se $g(n) = 0$, cioè $f(n) = h(n)$, si ha **Greedy Best First**

Esempio Il gioco dell'otto



$$f(n) = \#\text{mosseFatte} + \#\text{caselleFuoriPosto}$$

$$f(\text{start}) = 0 + 7$$

Dopo $\leftarrow, \downarrow, \uparrow, \rightarrow$ si ha $f = 4 + 7$: stesso stato ma g è cambiata.

$$f(\text{goal}) = ? + 0$$

Algoritmo A è completo

Teorema L'algoritmo A con la condizione $g(n) \geq d(n) \cdot \epsilon$, con $d(n)$ profondità e $\epsilon > 0$ costo minimo dell'arco, è **completo**.

La condizione ci garantisce che non si verifichino condizioni del tipo



e che il costo lungo un cammino non cresca "abbastanza", così da fermarsi per costi alti di g .

Dimostrazione Sia $[n_0 n_1 n_2 \dots n' \dots n_k = goal]$ un cammino soluzione. Sia n' un nodo della frontiera su un cammino soluzione $\rightarrow n'$ prima o poi verrà espanso. Infatti, esistono solo un numero finito di nodi x che possono essere aggiunti alla frontiera con $f(x) \leq f(n')$ (condizione su g).

Quindi, se non si trova una soluzione prima, n' verrà espanso e i suoi successori aggiunti alla frontiera. Tra questi, **anche il suo successore sul cammino soluzione**.

Il ragionamento si può ripetere fino a dimostrare che anche il nostro *goal* sarà selezionato per l'espansione.

3.2.2 Algoritmo A*: La Stima Ideale

Una **funzione di valutazione ideale (oracolo)** $f^*(n) = g^*(n) + h^*(n)$

$g^*(n)$ costo del **cammino minimo** da radice a n

$h^*(n)$ costo del **cammino minimo** da n a *goal*

$f^*(n)$ costo del **cammino minimo** da radice a *goal*, attraverso n

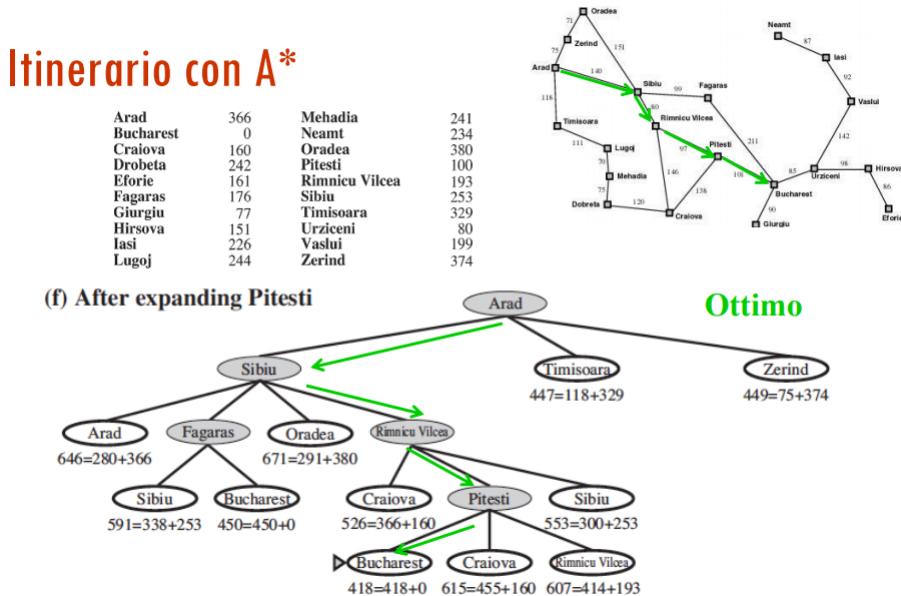
Normalmente $g(n) \geq g^*(n)$ (costo del cammino \geq cammino minimo) e $h(n)$ è una **stima** di $h^*(n)$: si può sotto o sovrastimare la distanza dalla soluzione.

Definizione Euristica ammissibile $\forall n \mid h(n) \leq h^*(n)$, h è una **sottostima**, ad esempio l'euristica della distanza in linea d'aria.

Definizione Algoritmo A*: un algoritmo A in cui h è una funzione euristica ammissibile.

Teorema Gli algoritmi A* sono **ottimali**.

Corollario BF con passi a costo costante e UC sono **ottimali** ($h(n) = 0$)



Osservazioni La componente g fa sì che si abbandonino cammini che vanno troppo in profondità.

h sotto o sovra stima? Una sottostima può farci compiere lavoro inutile, ma **non fa perdere il cammino migliore**: quando trovo il nodo *goal* è il cammino migliore. Invece, una funzione che a volte sovrastima può **farci perdere la soluzione ottimale a causa di tagli per sovrastima**.

Ottimalità Nel caso di ricerca su albero, l'**uso di un'euristica ammissibile è sufficiente a garantire l'ammissibilità \Rightarrow ottimalità di A***.

Nel caso di ricerca su grafo, serve una proprietà più forte: la **consistenza**, anche detta **monotonicità**, perché rischio di scartare candidati ottimi (stato già incontrato) a meno che il primo espanso sia il migliore.

Definizione Euristica consistente se

$$h(goal) = 0$$

Consistenza locale: $\forall n \mid h(n) \leq c(n, a, n') + h(n')$ dove n' è un successore di n e $c(n, a, n')$ è il costo del cammino $n \rightarrow n'$ sull'arco a .

$$\Rightarrow f(n) \leq f(n')$$

Quindi se h è consistente, allora **f non decresce mai lungo i cammini**: da qui il termine monotonia. Esistono euristiche ammissibili che non sono monotone, ma sono rare.

Teorema Un'euristica monotona è ammissibile.

Le euristiche monotone garantiscono che la **soluzione meno costosa venga trovata per prima** e quindi **sono ottimali anche nel caso di ricerca su grafo**.

Non si devono recuperare, tra gli antenati, nodi con costo minore

Lista degli esplorati, stato già esplorato è sul cammino ottimo \Rightarrow posso evitare di inserire il corrente ripetuto senza perdere l'ottimalità

```
if (figlio.Stato non in Esplorati and non in Frontiera)
    Frontiera = Inserisci(figlio, Frontiera)
```

Per la frontiera, volendo evitare stati ripetuti, resta l'**if finale di UC**

```
if (figlio.Stato in Frontiera con costoCammino più alto)
    sostituisci quel nodo frontiera con il figlio
```

Ottimalità di A* Dimostrazione

1. $h(n)$ consistente \Rightarrow i valori di $f(n)$ lungo un cammino sono non decrescenti:

Se $h(n) \leq c(n, a, n') + h(n')$ (def. consistenza)

$g(n) + h(n) \leq g(n) + c(n, a, n') + h(n')$ sommando $g(n)$

ma siccome $g(n) + c(n, a, n') = g(n')$, allora $g(n) + h(n) \leq g(n') + h(n')$

quindi $f(n) \leq f(n')$

2. Ogni volta che A* seleziona un nodo n per l'espansione, il cammino ottimo a tale nodo è stato trovato.

Se così non fosse, ci sarebbe un altro nodo n' della frontiera sul cammino ottimo (a n , ancora da trovare), con $f(n')$ minore (per la monotonia e n successivo di n').

Ma ciò non è possibile perché tale nodo sarebbe già stato espanso.

3. Quando seleziona nodo *goal* è cammino ottimo [$h = 0, f = C^*$]

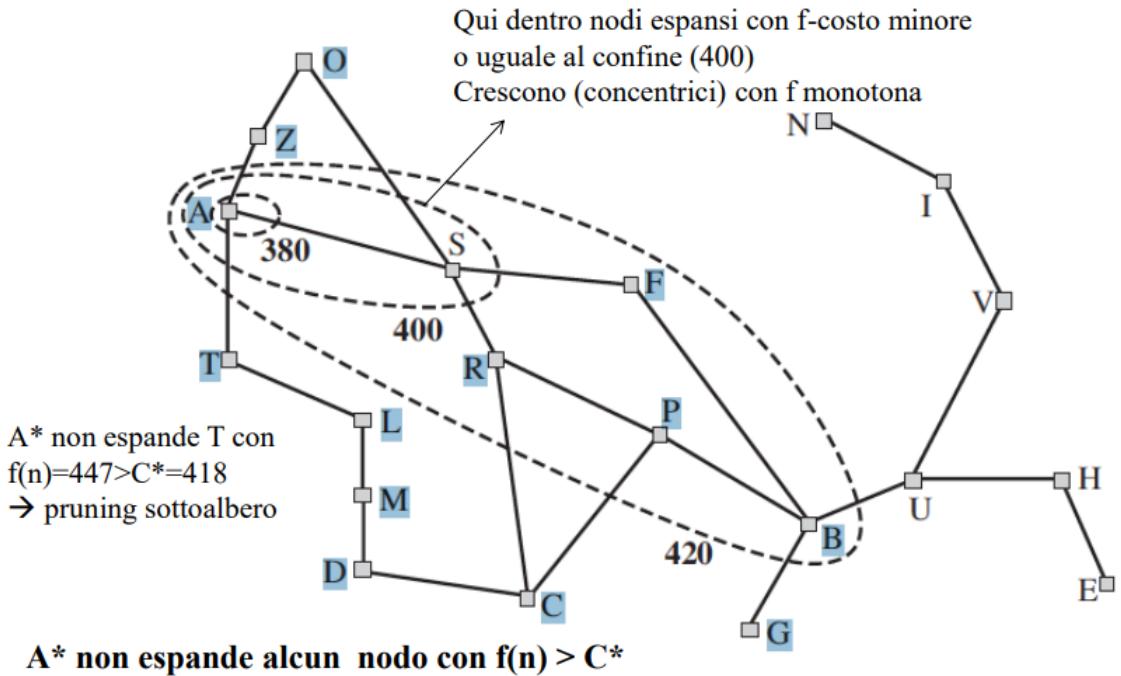
3.2.3 Perché A* è vantaggioso?

A* espande tutti i nodi con $f(n) < C^*$

A* espande *alcuni* nodi con $f(n) = C^*$

A* non espande alcun nodo con $f(b) > C^*$

Quindi alcuni nodi (e suoi sottoalberi) non verranno considerati per l'espansione, ma restiamo ottimali: **pruning**, un' h opportuna, **più alta possibile fra le ammissibili**, fa tagliare molto.



Più f è aderente, più taglio ottenendo ovali più stretti. Cercheremo quindi un' h più alta possibile tra le ammissibili. Se molto bassa, molti (sino a tutti) nodi restano $< C^* \Rightarrow$ espando tutti (a cerchi).

Pruning Il pruning dei sotto-alberi è il punto focale: non li abbiamo già in memoria ed evitiamo di generarli, e ciò è decisivo per i problemi di AI a spazio stati esponenziali.

3.2.4 Conclusioni su A^*

Algoritmo Lo stesso usato per UC

Funzioni Usa $f = g + h$ per la coda di priorità, dove h e g soddisfano le condizioni per algoritmo A e h è una funzione euristica ammissibile per A^* .

Sui grafi necessita di un'euristica monotona.

Completo Discende dalla completezza di A, perché A^* è un A particolare

Ottimale Con euristica monotona

Ottimamente efficiente A parità di euristica nessun altro algoritmo espande meno nodi senza rinunciare ad ottimalità

Problemi Quale euristica?

Occupazione in memoria: $O(b^{d+1})$

3.2.5 Casi speciali di A

$h(n) = 0$ si ha Uniform Cost, cioè $f(n) = g(n)$
 Cioè g non basta

$g(n) = 0$ si ha Greedy Best First, cioè $f(n) = h(n)$
 Ossia h non basta

3.3 Costruire le Euristiche di A*

3.3.1 Valutazione di funzioni euristiche

A parità di ammissibilità, una euristica può essere più efficiente di un'altra nel trovare il cammino soluzione migliore. Questo dipende da quanto **informata** è l'euristica, cioè dal grado di informazione posseduto.

$h(n) = 0$ **minimo** di informazione (BF, o UC)

$h^*(n)$ **massimo** di informazione (oracolo)

In generale, per le euristiche ammissibili,

$$0 \leq h(n) \leq h^*(n)$$

Teorema Se $h_1 \leq h_2$, i nodi espansi da A^* con h_2 sono un **sottoinsieme** di quelli espansi da A^* con h_1 .

Questo perché A^* espande tutti i nodi con $f(n) < C^*$ e sono meno per h maggiore (fa andare più nodi oltre C^*).

\Rightarrow Se $h_1 \leq h_2$, allora A^* con h_2 è **almeno efficiente** quanto A^* con h_1 .

Un'euristica più informata (accurata) riduce lo spazio di ricerca (più efficiente) ma è tipicamente **più costosa da calcolare**.

3.3.2 Confronto di euristiche ammissibili

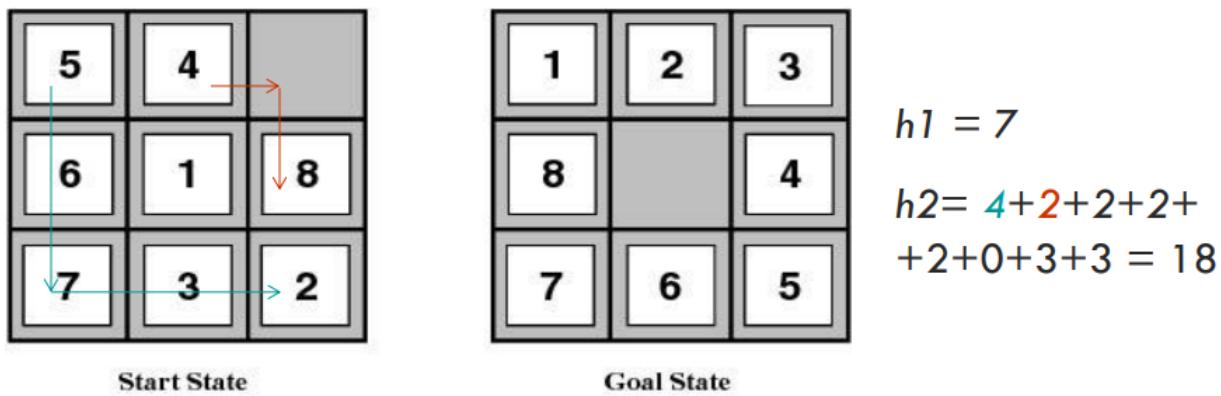
Esempio Due euristiche ammissibili per il gioco dell'otto

h_1 conta il numero di caselle fuori posto

h_2 somma delle distanze Manhattan (orizz/vert) delle caselle fuori posto dalla posizione finale

Manhattan Distance: $h(x, y) = MD((x, y), (x_g, y_g)) = |x - x_g| + |y - y_g|$

$\Rightarrow h_2$ è **più informata** di h_1 , infatti $\forall n \Rightarrow h_1(n) \leq h_2(n)$. Si dice che h_2 **domina** h_1 .



3.3.3 Misura del potere euristico

Come valutare gli algoritmi di ricerca euristica

Fattore di diramazione effettivo Dato N numero di nodi generati, d profondità della soluzione, allora b^* è il fattore di diramazione di un albero uniforme con $N+1$ nodi, soluzione dell'equazione

$$N + 1 = b^* + (b^*)^2 + \dots + (b^*)^d$$

Sperimentalmente, una buona euristica ha un b^* abbastanza vicino ad 1 (< 1.5)

Esempio $d = 5$, $N = 52 \Rightarrow b^* = 1.92$

3.3.4 Capacità di esplorazione

Influenza di b^*

Con $b = 2$

$$d = 6 \quad N = 100$$

$$d = 12 \quad N = 10000$$

Con $b = 1.5$

$$d = 12 \quad N = 100$$

$$d = 24 \quad N = 10000$$

Quindi **migliorando di poco l'euristica si riesce, a parità di nodi espansi, a raggiungere una profondità doppia.**

Quindi

Tutti i problemi dell'IA, o quasi, sono di complessità esponenziale nel generare nodi (ad es. configurazioni possibili), ma c'è esponenziale ed esponenziale. L'euristica può migliorare di molto la capacità di esplorazione dello spazio degli stati rispetto alla ricerca cieca: **migliorando anche di poco l'euristica si riesce ad esplorare uno spazio molto più grande.**

3.4 Come si inventa un'euristica?

Alcune strategie per ottenere euristiche ammissibili, da vedere man mano:

Rilassamento del problema

Massimizzazione di euristiche

Database di pattern disgiunti

Combinazione lineare

Apprendere dall'esperienza

3.4.1 Rilassamento del problema

Spazio degli stati con archi aggiunti.

Gioco dell'otto Nel gioco dell'otto, mossa da A a B possibile se **B adiacente ad A e B libera**.

h_1 e h_2 sono **calcoli della distanza esatta della soluzione** in versioni semplificate del puzzle: uno **spazio degli stati con archi aggiunti**

h_1 nessuna restrizione: sono sempre ammessi scambi tra caselle, si muove ovunque → numero di caselle fuori posto

h_2 solo prima restrizione: sono ammessi spostamenti anche su caselle occupate purché adiacenti → somma delle distanze Manhattan

3.4.2 Massimizzazione di euristiche

Si hanno una serie di euristiche ammissibili h_1, h_2, \dots, h_k , **senza che nessuna domini un'altra**. Allora conviene prendere il **massimo dei loro valori**

$$h(n) = \max\{h_1(n), h_2(n), \dots, h_k(n)\}$$

Se le h_i sono ammissibili, anche la h lo è e **domina tutte le altre**.

Euristiche da sottoproblemi



Il **costo della soluzione ottima del sottoproblema** (sistemare 1, 2, 3, 4) è una **sottostima del costo del problema nel suo complesso** e più accurata della Manhattan.

Database di pattern: si memorizza ogni istanza del sottoproblema con relativo costo della soluzione. Si usa il database per calcolare h_{DB} , estraendo dal DB la configurazione corrispondente allo stato completo corrente.

Sottoproblemi multipli Potremmo poi fare la stessa cosa per altri sottoproblemi: 5-6-7-8, 2-4-6-8... ottenendo altre euristiche ammissibili.

Poi si può prendere il valore massimo: altra euristica ammissibile.

Ma potremmo sommarle ed ottenere un'euristica ancora più accurata?

3.4.3 Pattern Disgiunti

In generale no, perché le **soluzioni ai sottoproblemi interferiscono**: nel caso del gioco dell'otto, condividono alcune mosse perché se sposto 1-2-3-4 sposto anche 4-5-6-7.

La **somma delle euristiche in generale non è ammissibile** perché potremmo sovrastimare, avendo avuto aiuti mutui.

Si deve **eliminare il costo delle mosse che contribuiscono all'altro sottoproblema**: **databse di pattern disgiunti** consentono di sommare i costi (**euristiche additive**).

Sono molto efficaci: gioco del 15 in pochi ms. Ma difficile scomporre per il cubo di Rubik.

3.4.4 Apprendere dall'esperienza

Si fa girare il programma e si raccolgono i dati sottoforma di **coppie** `<stato, h*>`. Si usano i dati per apprendere come predire la h con **algoritmi di apprendimento induttivo**: da istanze note stimiamo h in generale.

Gli algoritmi di apprendimento si concentrano su caratteristiche salienti dello stato (*feature, x_i*). Esempio: numero tasselli fuori posto 5 → costo 14.

Combinazione di euristiche

Quando diverse caratteristiche influenzano la bontà di uno stato, si può usare una combinazione lineare

$$h(n) = c_1 x_1(n) + c_2 x_2(n) + \dots + c_k x_k(n)$$

Gioco dell'otto $h(n) = c_1 \#fuoriPosto + c_2 \#coppieScambiate$

Scacchi $h(n) = c_1 vantaggioPezzi + c_2 pezziAttaccante + c_3 regina + \dots$

Il **peso dei coefficienti può essere aggiustato con l'esperienza**, anche qui **apprendendo automaticamente da esempi di gioco**. $h(goal) = 0$ ma ammissibilità e consistenza **non** automatiche.

Capitolo 4

Algoritmi Evolutivi Basati su A*

4.1 Beam Search

Nel **best first** viene mantenuta tutta la frontiera. Se l'occupazione di memoria è eccessiva, si può ricorrere ad una variante: la **beam search**.

Beam Search La beam search **tiene ad ogni passo solo i k nodi più promettenti**, dove k è detto **ampiezza del raggio** (beam).

Non è completa.

4.2 IDA*

A* con approfondimento iterativo. IDA* combina A* con ID: ad ogni iterazione si ricerca in profondità con un limite (cut-off) dato dal valore della funzione f (e non dalla profondità).

Il limite **f-limit** viene aumentato ad ogni iterazione, fino a trovare la soluzione.

Punto Critico Di quanto viene aumentato f-limit.

Quale incremento? Cruciale la scelta dell'incremento per garantire l'ottimalità. In caso di azioni dal costo fisso, il limite viene incrementato dal costo delle azioni.

Ma in caso di costi variabili? Costo minimo? Si potrebbe, ad ogni passo, fissare il limite successivo al valore minimo delle f scartate (in quanto superavano il limite) all'interazione precedente.

Analisi **Completo e ottimale.** Occupazione in memoria $O(bd)$.

4.3 RBFS

Best-First Ricorsivo: simile a DF ricorsivo ma cerca di usare meno memoria, facendo del lavoro in più.

Tiene traccia del migliore percorso alternativo ad ogni livello. Invece di fare backtracking in caso di fallimento, interrompe l'esplorazione quando trova un nodo meno promettente secondo f . Nel tornare indietro **si ricorda il miglior nodo che ha trovato nel sottoalbero esplorato**, per poterci eventualmente tornare.

Memoria: lineare nella profondità della soluzione ottima.

4.4 A* con memoria limitata

L'idea è quella di utilizzare al meglio la memoria disponibile.

SMA* procede come A* fino ad esaurimento della memoria disponibile. A questo punto "dimentica" il nodo peggiore, dopo avere aggiornato il valore del padre.

A parità di f **si sceglie il nodo migliore più recente e si dimentica il nodo peggiore più vecchio.**

Ottimale se il cammino soluzione sta in memoria.

In algoritmi a memoria limitata, le limitazioni della memoria possono portare a compiere molto lavoro inutile: ad esempio, visitare ripetutamente gli stessi nodi. Diventa quindi **difficile stimare la complessità temporale effettiva**. Le **limitazioni della memoria**, quind, **possono rendere un problema intrattabile** dal punto di vista computazionale.

4.5 Conclusioni

Agenti in ambienti deterministici, osservabili, statici e completamente noti

Ricerca come **scelta della sequenza di azioni**, cioè cammino in uno spazio di stati, **che raggiunge obiettivo**
⇒ il cammino è la soluzione

Attività necessarie

Formulazione del problema

Scelta dell'algoritmo di ricerca adeguato

Identificazione della funzione di valutazione euristica più efficace

Capitolo 5

Oltre la Ricerca Classica

Risolutori "classici" Gli agenti risolutori di problemi *classici* assumono le condizioni viste in precedenza:

Ambienti completamente osservabili

Ambienti deterministici

Si trovano nelle condizioni di produrre un piano d'azione eseguibile "ad occhi chiusi": possono studiare la sequenza d'azioni *offline*, che può essere eseguita senza imprevisti per raggiungere l'obiettivo

5.1 Verso ambienti più realistici

La **ricerca sistematica**, o euristica, nello spazio di stati è **troppo costosa** → Metodi di ricerca locale
Assunzioni sull'ambiente da considerare:

Azioni non deterministiche

Ambiente parzialmente osservabile

→ Piani condizionali, ricerca AND-OR, stati credenza

Ambienti sconosciuti

Problemi di esplorazione (percezioni forniscono nuove informazioni dopo l'azione)

→ Ricerca **online**

5.2 Ricerca Locale

Assunzioni Gli algoritmi visti fin'ora esplorano gli spazi degli stati alla ricerca di un goal e restituiscono un **cammino soluzione**, ma a volte lo stato goal è la soluzione del problema. Gli **algoritmi di ricerca locale** sono adatti per problemi in cui

La sequenza di azioni non è importante: quello che conta è lo stato goal

Tutti gli elementi della soluzione sono nello stato ma alcuni vincoli sono violati

5.2.1 Algoritmi di ricerca locale

Non sono sistematici

Tengono traccia solo del nodo corrente e si spostano su nodi adiacenti

Non tengono traccia dei cammini, poiché non servono a lavoro finito

Efficienti in memoria

Possono trovare soluzioni ragionevoli anche in spazi molto grandi o infiniti, come nel caso di spazi continui

Utili per **risolvere problemi di ottimizzazione**

Stato migliore secondo una funzione obiettivo

Stato di costo minore

Esempio: training di un modello di Machine Learning

5.2.2 Panorama dello spazio degli stati

Esempio di una f euristica di costo della funzione obiettivo (non del cammino)



Uno stato ha una posizione sulla superficie e un'altezza corrispondente al valore della sua funzione di valutazione. Un algoritmo provoca un movimento sulla superficie.

Trovare avvallamento più basso (es: costo minimo) o il picco più alto (es: massimo di un obiettivo)

5.2.3 Algoritmo Hill Climbing

Anche detto **ricerca in salita, steepest ascent/descent**.

Ricerca locale **greedy**.

Vengono generati i successori e valutati. Viene scelto un nodo che migliora la valutazione dello stato attuale (**non si tiene traccia degli altri**, quindi non ho l'albero di ricerca in memoria). La scelta del nodo dipende dall'algoritmo:

Migliore → **Hill Climbing a salita ripida**

Uno a caso tra quelli che migliorano → **Hill Climbing Stocastico**

Il primo → **Hill Climbing con prima scelta**

Se non ci sono stati migliori, l'algoritmo termina con **fallimento**.

```
function Hill-climbing(problema) returns stato-massimo-locale
    nodo-corrente = CreaNodo(problema.Stato-iniziale)
    loop do
        vicino = successore di nodo-corrente di valore piu alto
        if (vicino.Valore <= nodo-corrente.Valore)
            return nodo-corrente.Stato #interrompe la ricerca
        nodo-corrente = vicino #altrimenti, se vicino migliore, continua
```

Si prosegue solo se il vicino più alto è migliore dello stato corrente, se tutti i vicini sono peggiori, si ferma. Non c'è frontiera a cui tornare, si tiene solo uno stato.

Tempo: numero di cicli variabile in base al punto di partenza.

Problema delle 8 Regine Con formulazione a stato completo

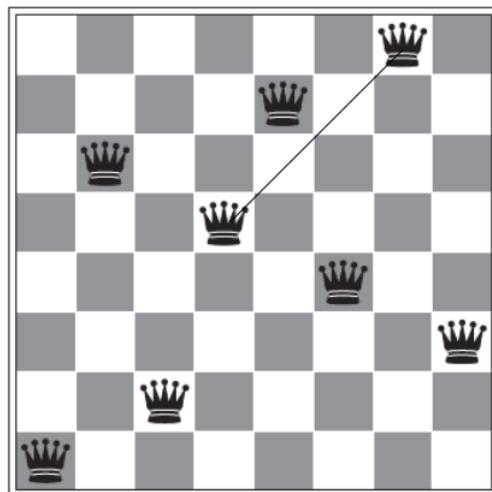
Costo h = numero di coppie di regine che si attaccano a vicenda

I numeri sono i valori dei successori (7×8 , 7 posizioni per ogni regina = su ogni colonna)

Tra i migliori (valore 12) si sceglie a caso

Minimo globale = 0

18	12	14	13	13	12	14	14
14	16	13	15	12	14	12	16
14	12	18	13	15	12	14	14
15	14	14	14	14	14	14	14
14	14	17	15	14	16	16	16
17	14	16	18	15	14	15	14
18	14	14	15	15	14	14	16
14	14	13	17	12	14	12	18



Un minimo locale

$$h = 1$$

Tutti gli stati successori non migliorano la situazione (minimo locale)

Per le 8 regine Hill-Climbing si blocca l'86% delle volte

Ma in media sono 4 passi per la soluzione, e 3 quando si blocca

Su $8^8 = 17$ milioni di stati

Esempio Successo in tre mosse

h qui è l'**euristica di costo della funzione obiettivo** da minimizzare



Problemi con Hill-Climbing

Se la funzione è da ottimizzare, i picchi sono massimi locali o soluzioni ottimali. Nel grafico si possono presentare: **massimi locali, plateau (pianori), spalle e crinali/creste.**

Miglioramenti

1. Consentire un **numero limitato di mosse laterali**
Ossia ci si ferma per $<$ nell'algoritmo, anziché per \leq
 \Rightarrow L'algoritmo sulle 8 regine ha successo nel 94% dei casi, ma impiega in media 21 passi
2. Hill-Climbing Stocastico: si **sceglie a caso tra le mosse in salita**, magari tenendo conto della pendenza.
Converge più lentamente ma a volte trova soluzioni migliori
3. Hill-Climbing con prima scelta: può generare le mosse a caso fino a trovarne una migliore dello stato corrente.
Più efficace quando i successori sono molti (es: migliaia)
4. Hill-Climbing con riavvio casuale (**random restart**): ripartire da un punto scelto a caso.
Se la probabilità di successo è p , saranno necessarie in media $\frac{1}{p}$ ripartenze per trovare la soluzione (Es.: 8 regine, $p = 0.14$, 7 iterazioni cioè 6 fallimenti e un successo).
Tendenzialmente completo: insistendo si generano tutte le possibilità.
Per le regine, 3 milioni di regine in meno di un minuto. Se funziona o no dipende molto dalla forma del panorama degli stati.

5.2.4 Algoritmo di Tempra Simulata

Simulated Annealing L'algoritmo di tempra simulata combina Hill-Climbing con una scelta stocastica ma **non del tutto casuale perché poco efficiente**. L'analogia è col processo di tempra dei metalli: portati a temperature molto elevate e raffreddati gradualmente consentendo di cristallizzare in uno stato a più bassa energia.

Tempra Simulata Ad ogni passo si **sceglie un successore a caso**:

Se migliora lo stato corrente, viene espanso

Se non lo migliora (caso in cui $\Delta E = f(n') - f(n) < 0$), quel nodo viene scelto con probabilità $p = e^{\Delta E/T}$, con p ovviamente $0 \leq p \leq 1$

(Si genera un numero casuale tra 0 e 1, se questo è $< p$ il successore viene scelto, altrimenti no)

p è **inversamente proporzionale al peggioramento**

T (**temperatura**) **decresce al progredire dell'algoritmo**, quindi anche p , secondo un piano definito.
Col progredire, rende improbabili le mosse peggiorative.

Analisi La probabilità di una mossa in discesa diminuisce col tempo, e l'algoritmo si comporta sempre di più come Hill-Climbing.

Se T viene decrementato abbastanza lentamente, con probabilità tendente ad 1 si raggiunge la soluzione ottimale.

Analogia: T corrisponde alla temperatura e ΔE alla variazione di energia

Parametri **Valore iniziale e decremento di T** sono parametri.

I valori per T sono **determinati sperimentalmente**: il **valore iniziale di T** è tale che per valori medi di ΔE , $p = e^{\Delta E/T}$ sia circa 0.5

5.2.5 Algoritmo Local Beam

Versione locale della beam search.

Si tengono in memoria k stati invece che uno solo. Ad ogni passo si generano i successori di tutti i k stati.

Se si trova un goal, ci si ferma

Altrimenti si prosegue con i k migliori tra questi

Note:

Diverso da K restart (che riparte da 0)

Diverso da beam search

5.2.6 Algoritmo Beam Search Stocastico

Si introduce un elemento di casualità, come in un processo di selezione naturale: **diversificare la nuova generazione**. In questa variante stocastica, si scelgono k successori ma **con probabilità maggiore per i migliori**. Terminologia:

Organismo, stato

Progenie, successori

Fitness (valore della f), idoneità

5.3 Algoritmi Genetici

L’Idea Sono varianti della beam search stocastica in cui gli stati successori sono ottenuti *combinando* due stati genitore anziché per evoluzione. Terminologia:

Popolazione di individui (stati)

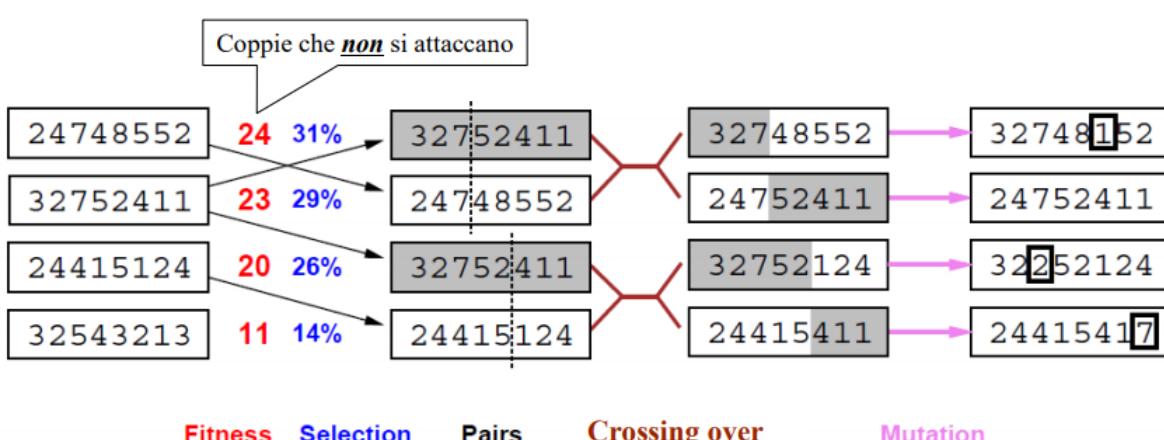
Fitness

Accoppiamenti e mutazione genetica

Generazioni

Funzionamento Popolazione iniziale: k stati/individui generati casualmente. Ogni **individuo è rappresentato come una stringa**: ad esempio 24 bit, o posizione nelle colonne ("24748552") per le 8 regine. Gli individui sono valutati da una funzione di **fitness**: ad esempio numero di coppie di regine che *non* si toccano. Si selezionano gli individui per gli accoppiamenti, con una probabilità proporzionale alla fitness. Le **coppie danno vita alla generazione successiva**: combinando materiale genetico (**crossover**) o con un meccanismo aggiuntivo di mutazione genetica (**casuale**). La popolazione ottenuta dovrebbe essere migliore e la cosa si ripete fino ad ottenere stati abbastanza buoni (**stati obiettivo**).

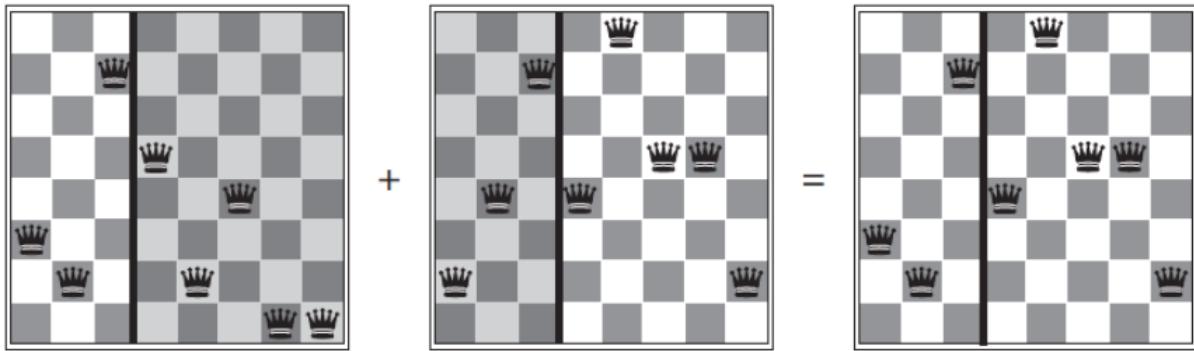
Esempio



Per ogni coppia viene scelto un punto di **crossing over** (la linea tratteggiata) e vengono generati due figli **scambiandosi dei pezzi del DNA**.

Viene infine effettuata una mutazione casuale che dà luogo alla prossima generazione.

Nascita di un figlio



Le parti chiare sono passate al figlio, le parti scure si perdono. Se i genitori sono molto diversi, anche i nuovi stati sono diversi. All'inizio spostamenti maggiori che poi si raffinano.

Algoritmi Genetici

Suggeritivi Area del **Natural Computing**. Usati molto in problemi reali (es.: configurazione di circuiti e scheduling dei lavori).

Combinano la tendenza a salire della beam search stocastica con l'interscambio di informazioni tra thread paralleli di ricerca (blocchi utili che si combinano).

Funziona meglio se il problema (soluzioni) ha componenti significative rappresentate in sottostringhe.

Punto critico: rappresentazione del problema in stringhe.

Spazi Continui Molti casi reali hanno spazi di ricerca continua, fondamentale per il Machine Learning! Lo stato è descritto da variabili continue x_1, \dots, x_n (vettore x), ad esempio la posizione in uno spazio 3D $x = (x_1, x_2, x_3)$.

Apparentemente è complicato: fattori di ramificazione infiniti con gli approcci precedenti. Ma in realtà ci sono molti strumenti matematici per spazi continui, che portano ad approcci anche molto efficienti.

5.3.1 Gradient

Se la f è **continua** e **differenziabile**, ad esempio quadratica rispetto il vettore x , allora il minimo/massimo lo si può cercare utilizzando il **gradiente**, che **restituisce la direzione di massima pendenza del punto**.

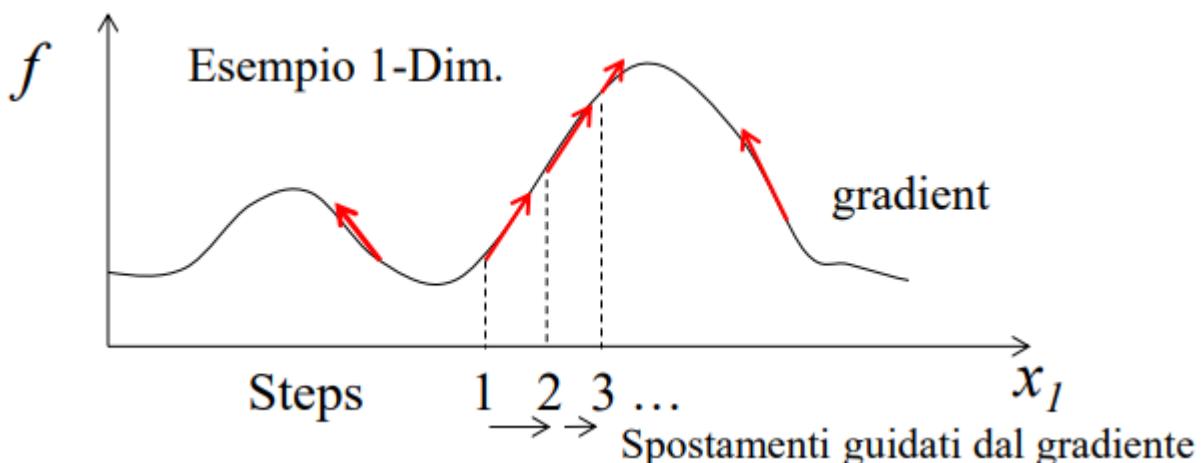
Data f obiettivo

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3} \right)$$

Hill-Climbing iterativo: $x_{new} = x_{old} + \eta \nabla f(x)$ con η dimensione dello step.

Quantifica lo spostamento, senza cercarlo tra gli infiniti possibili successori.

Nota: in generale non è sempre necessario il min/max assoluto. Vedremo nel ML.



5.4 Ambenti più realistici

Problemi classici Gli agenti risolutori di problemi "classici" assumono:

Ambienti **completamente osservabili**

Azioni e ambienti **deterministici**

Piano generato è sequenza di azioni **eseguibile ad occhi chiusi**, generato *offline* ed eseguito senza imprevisti

Le **percezioni non servono**, se non nello stato iniziale

Soluzioni più complesse In un ambiente **parzialmente osservabile e non deterministico** le **percezioni sono importanti: restringono gli stati possibili e informano sull'effetto dell'azione**.

Più che un piano, l'**agente elabora una strategia** che **tiene conto delle diverse eventualità: un piano con contingenza**. Esempio: aspirapolvere con assunzioni diverse.

5.4.1 Azioni non deterministiche

Aspirapolvere imprevedibile Ci sono più stati possibili come risultato dell'azione.

Comportamento: se aspira in una stanza sporca la pulisce... ma **a volte** pulisce anche una stanza adiacente. Se aspira in una stanza pulita, **a volte** rilascia sporco.

Variazioni necessarie al modello Il modello di transizione, quindi, **restituisce un insieme di stati**: l'agente non sa in quale si troverà. Il **piano di contingenza** sarà un piano condizionale e magari con cicli.

Esempio Nell'esempio

Risultati(Aspira, 1) = {5, 7}, cioè aspirando nello stato 1 posso finire nello stato 5 o 7.

Un possibile piano è

```
[ Aspira ,
  if (stato = 5):
    [ Destra , Aspira ]
  else:
    []
]
```



Da sequenza di azioni a piano (albero)

5.4.2 Come si pianifica

Alberi di Ricerca AND-OR

Nodi **OR**: scelte dell'agente (1 sola azione)

Nodi **AND**: le diverse contingenze (scelte dell'ambiente, più stati possibili), **da considerare tutte**

Una **soluzione ad un problema di ricerca AND-OR** è un **albero** che

Ha un nodo obiettivo in ogni foglia

Specifica un'unica azione nei nodi **OR**

Include tutti gli archi uscenti da nodi **AND**

Esempio



Archi in grassetto = soluzione (sottoalbero), la seguente:

Piano: [Aspira: **if** (stato = 5): [Destra , Aspira] **else**: []]

Capitolo 6

I Giochi con Avversario

Premessa Fin'ora abbiamo avuto *Problem Solving* come ricerca. Il paradigma di base era: ambiente osservabile, deterministico, utente singolo e stati atomici.

Da adesso considereremo un **rilassamento delle assunzioni di base**: ambiente multi agente e rappresentazioni degli stati più complesse.

Ci occuperemo di **specializzazioni** del paradigma nell'ambito dei giochi con avversario: quindi i piani d'azione devono tenere conto dell'avversario.

Vedremo **problemi di soddisfacimento di vincoli**, sempre come ricerca in spazio di soluzioni, con stato a struttura fattorizzata.

Questo ci porterà a considerare **sistemi basati su conoscenza**: lo stato è una "base di conoscenza" a cui rivolgere domande, con una rappresentazione in un linguaggio espressivo e **tecniche di "ragionamento" con inferenze**: logica del prim'ordine e calcolo proposizionale.

6.1 Giochi con Avversario

- Regole semplici e formalizzabili
- Ambiente accessibile e deterministico
- Due giocatori, a turni alterni, a **somma zero** (se uno vince, l'altro perde: sommare i punteggi dà 0 o comunque un risultato costante), a informazione perfetta (tutti i giocatori conoscono lo stato attuale del gioco)
- Ambiente multi-agente competitivo: la presenza dell'avversario rende l'**ambiente strategico**, più difficile rispetto a quanto visto fin'ora
- Complessità e vincoli di tempo reale: mossa migliore nel tempo disponibile

⇒ Questa tipologia di giochi sono un po' più simili ai problemi reali

6.1.1 Ciclo *pianifica-agisci-percepisci*

Due agenti a turno Si può pianificare considerando le possibili risposte dell'avversario e le possibili risposte a quelle risposte... e così via.

Una volta decisa la mossa

Si **esegue** la mossa

Si **vede** cosa fa l'avversario

Si **pianifica** la prossima mossa

6.2 Giochi come problemi di ricerca

Stati: configurazioni del gioco

Player(s): a chi tocca muovere nello stato s

Stato iniziale: configurazione iniziale gioco

Actions(s): mosse legali in s

Result(s, a): stato risultante da una mossa

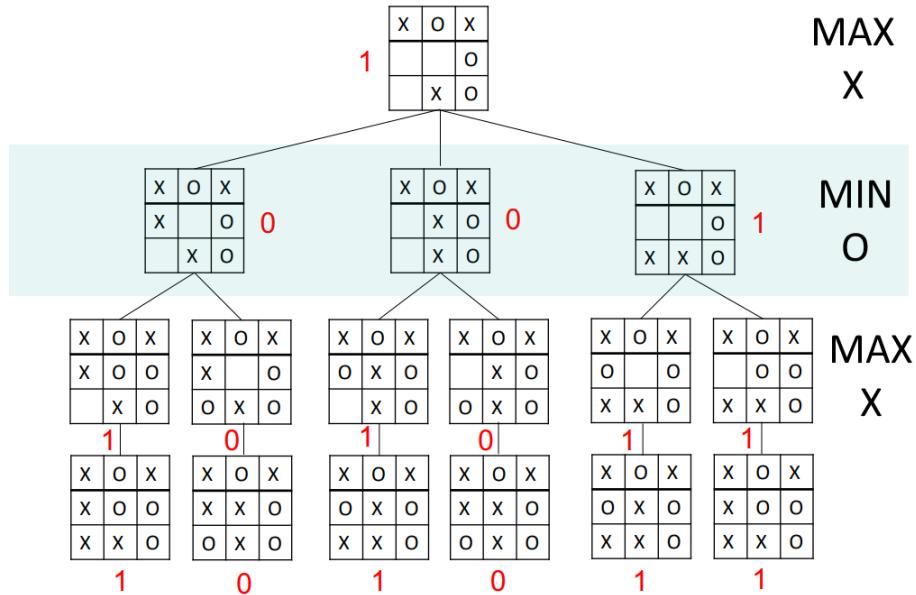
Terminal-Test(s): determina se stato è fine del gioco

Utility(s, p): utilità (o **pay-off**), cioè valore numerico degli stati terminali del gioco per il giocatore p
Es: 1, -1, 0, conteggio punti...

6.2.1 Algoritmo MinMax

Valuto gli stati terminali in base al punteggio/vittoria che ottengo, poi valuto gli stati preterminali in base allo stato terminale in cui mi portano (se mi portano a vittoria o meno a ritroso).

Valuto gli stati intermedi con il valore **minimo** dei risultati se tocca all'avversario (minimo rischio) e col valore **massimo** se tocca a me.



$$\text{Valore MinMax } \text{Minimax}(s) = \begin{cases} \text{Utility}(s, \text{Max}) & \text{se Terminal-Test}(s) \\ \max_{a \in \text{Action}(s)} \text{Minimax}(\text{Result}(s, a)) & \text{se Player}(s) = \text{MAX} \\ \min_{a \in \text{Action}(s)} \text{Minimax}(\text{Result}(s, a)) & \text{se Player}(s) = \text{min} \end{cases}$$

Come conviene esplorare l'albero di gioco? In profondità, perché hanno ampiezza esponenziale.

Algoritmo ricorsivo Di seguito

```

function MINIMAX-DECISION(state) returns an action
    return  $\arg \max_{a \in \text{ACTIONS}(s)} \text{MIN-VALUE}(\text{RESULT}(s, a))$ 

function MAX-VALUE(state) returns a utility value
    if TERMINAL-TEST(state) then return UTILITY(state)
     $v \leftarrow -\infty$ 
    for each a in ACTIONS(state) do
         $v \leftarrow \text{MAX}(v, \text{MIN-VALUE}(\text{RESULT}(s, a)))$ 
    return v

function MIN-VALUE(state) returns a utility value
    if TERMINAL-TEST(state) then return UTILITY(state)
     $v \leftarrow \infty$ 
    for each a in ACTIONS(state) do
         $v \leftarrow \text{MIN}(v, \text{MAX-VALUE}(\text{RESULT}(s, a)))$ 
    return v

```



Costo

Tempo: come DF $\rightarrow O(b^m)$

Spazio: $O(m)$

Per gli scacchi (10^{40} nodi) più di 10^{22} secoli \Rightarrow improponibile un'esplorazione sistematica del grafo, se non per giochi molto semplici

Necessarie **euristiche**

6.2.2 Algoritmo Min-Max Euristico (con orizzonte)

In casi più complessi, dove esplorare stati è troppo costoso (es. scacchi), occorre una **funzione di valutazione euristica**.

Strategia Guardare avanti d livelli

Si espande l'albero di ricerca un certo numero di livelli d , compatibile con il tempo e lo spazio disponibili

Si **valutano gli stati ottenuti** e si propaga indietro il risultato con la regola del Minimax

L'algoritmo è simile a prima ma

```
if Terminal-Test(s) then return Utility(s) → if Cutoff-Test(s, d) then return Eval(s)
```

Posto d limite della profondità consentita

$$H\text{-Minimax}(s, d) = \begin{cases} \text{Eval}(s) & \text{se Cutoff-Test}(s, d) \\ \max_{a \in Action(s)} H\text{-Minimax}(\text{Result}(s, a), d + 1) & \text{se Player}(s) = \text{MAX} \\ \min_{a \in Action(s)} H\text{-Minimax}(\text{Result}(s, a), d + 1) & \text{se Player}(s) = \text{min} \end{cases}$$

|| filetto

$$\begin{aligned} \text{Eval}(s) &= X(s) - O(s) \\ X(s) &\quad \text{righe aperte per X} \\ O(s) &\quad \text{righe aperte per O} \end{aligned}$$

Una configurazione vincente per X viene stimata $+\infty$
Una vincente per O, $-\infty$



Funzione di Valutazione

La funzione di valutazione **Eval** è una **stima dell'utilità attesa** a partire da una certa posizione. La qualità della **funzione** è determinante.

deve essere **consistente** con l'utilità se applicata agli stati terminali del gioco (indurre lo stesso ordinamento)

deve essere **efficiente** da calcolare

deve **riflettere** le probabilità effettive di vittoria (A valutato meglio di B se da A ho più probabilità di vittoria rispetto a B)

combina probabilità con utilità dello stato terminale, può essere appreso con l'esperienza

Esempio Scacchi

Si potrebbe pensare di valutare caratteristiche diverse dello stato: valore del materiale (pedone 1, cavallo/alfiere 3, torre 5, regina 9...), buona disposizione dei pedoni, protezione del re...

Funzione lineare pensata $\text{Eval}(s) = w_1 f_1(s) + w_2 f_2(s) + \dots + w_n f_n(s)$

Ma anche combinazioni non lineari delle caratteristiche: alfiere vale di più a fine partita, due alfieri valgono più del doppio di uno solo...

Problemi Noti

Stati Non Quiescenti: l'esplorazione fino ad un certo livello può mostrare una situazione molto vantaggiosa, ma la mossa successiva risulta estremamente svantaggiosa es. regina mangiata dalla torre.

Soluzione: non fermarsi agli stati non quiescenti ma esplorare un po' di più fino ad arrivare a stati in cui la funzione di valutazione non attende variazioni repentine (**stati quiescenti**)

Effetto Orizzonte: si privilegiano mosse **diverse che hanno solo l'effetto di spingere il problema oltre**, per evitare mosse disastrose che alla fine devono per forza accadere

Ottimizzazione Dobbiamo esplorare ogni cammino? No, **esistono tecniche di potatura** che dimezzano la ricerca pur mantenendo decisione min-max corretta (**potatura alfa-beta**)

Potatura Alfa-Beta ridurre spazi ricerca algoritmi min-max

6.2.3 Potatura Alfa-Beta

Tecnica di *potatura* per ridurre l'esplorazione dello spazio di ricerca in algoritmi minmax

Idea



$$\begin{aligned}
 \minmax(\text{radice}) &= \max(\min(3, 12, 8), \min(2, x, y), \min(14, 5, 2)) = \\
 &= \max(3, \min(2, x, y), 2) = \\
 &= \max(3, z, 2) = 3 \quad \text{con } z \leq 2
 \end{aligned}$$

Algoritmo Consideriamo un nodo v : se c'è una **scelta migliore sopra**, allora quel v non sarà mai raggiunto. Max passerà da α piuttosto che finire a v .

Implementazione Si va avanti in profondità fino al livello desiderato e, propagando indietro i valori, si decide se si può abbandonare l'esplorazione nel sotto-albero.

MaxValue e **MinValue** vengono invocate con **due valori di riferimento iniziali**

$\text{MaxValue} = \alpha$ (inizialmente $-\infty$)

$\text{MinValue} = \beta$ (inizialmente $+\infty$)

Rappresentano **rispettivamente la migliore alternativa per Max e per Min** fino a quel momento

I tagli avvengono quando, nel propagare indietro, si verifica:

$v \geq \beta$ per i nodi Max (taglio β)

$v \leq \alpha$ per i nodi Min (taglio α)

```

function ALPHA-BETA-SEARCH(state) returns an action
  v  $\leftarrow$  MAX-VALUE(state,  $-\infty$ ,  $+\infty$ )
  return the action in ACTIONS(state) with value v

function MAX-VALUE(state,  $\alpha$ ,  $\beta$ ) returns a utility value
  if TERMINAL-TEST(state) then return UTILITY(state)
  v  $\leftarrow -\infty$ 
  for each a in ACTIONS(state) do
    v  $\leftarrow$  MAX(v, MIN-VALUE(RESULT(s,a),  $\alpha$ ,  $\beta$ ))
    if v  $\geq \beta$  then return v  $\leftarrow$  taglio  $\beta$ 
     $\alpha \leftarrow \text{MAX}(\alpha, v)$ 
  return v

function MIN-VALUE(state,  $\alpha$ ,  $\beta$ ) returns a utility value
  if TERMINAL-TEST(state) then return UTILITY(state)
  v  $\leftarrow +\infty$ 
  for each a in ACTIONS(state) do
    v  $\leftarrow$  MIN(v, MAX-VALUE(RESULT(s,a),  $\alpha$ ,  $\beta$ ))
    if v  $\leq \alpha$  then return v  $\leftarrow$  taglio  $\alpha$ 
     $\beta \leftarrow \text{MIN}(\beta, v)$ 
  return v

```

Ordinamento delle mosse La potatura ottimale si ottiene quando ad ogni livello sono generate prima le mosse migliori per chi gioca:

Per i **nodi Max** sono generate prima le mosse con valore più **alto**

Per i **nodi Min** sono generate prima le mosse con valore più **basso**

Complessità $O(b^{m/2})$

Alfa-Beta può arrivare a profondità doppia rispetto a MinMax, ma come avvicinarsi all'ordinamento ottimale?

Ordinamento Dinamico Usando un approfondimento iterativo si possono scoprire ad una iterazione delle informazioni utili per l'ordinamento delle mosse, da usare in una successiva iterazione (**mosse killer**).

Tabella delle trasposizioni: per ogni stato incontrato si memorizza la sua valutazione: situazione tipica è quando $[a_1, b_1, a_2, b_2]$ e $[a_1, b_2, a_2, b_1]$ portano nello stesso stato.

Altri miglioramenti

Potatura in avanti Esplorare solo **alcune mosse ritenute promettenti** e tagliare le altre: beam search, tagli probabilistici basati su esperienza

Database di mosse di apertura e chiusura: nelle prime fasi ci sono poche mosse sensate e ben studiate, mentre per le fasi finali il computer può esplorare in maniera esaustiva e ricordarsi le chiusure migliori

Giochi Multiplayer Invece di aver due valori ho un vettore di valori per nodo. Per propagare all'indietro prendo vettore che migliora la valutazione.

Giochi più complessi

Giochi stocastici, con un lancio di dadi

Giochi parzialmente osservabili: deterministici (mosse deterministiche ma non si conoscono gli effetti delle mosse dell'avversario, es: battaglia navale), **stocastici** (carte distribuite a caso come in molti giochi)

Capitolo 7

Problemi di Soddisfacimento di Vincoli

CSP Sono problemi con una struttura particolare, che si prestano ad algoritmi di ricerca specializzati. Grazie alla struttura si cerca meglio la soluzione: rappresentazione **fattorizzata**, con la quale iniziamo a dire qualcosa sulla struttura dello stato.

Classe ampia Questa classe di problemi è molto ampia, esistono **euristiche generali** da applicare e che consentono la risoluzione di istanze con dimensioni significative.

7.1 Formulazione

Problema descritto da tre componenti:

X, insieme di **variabili**

D, insieme di **domini**

Dove D_i è l'insieme dei valori possibili per X_i

C, insieme di **vincoli**, relazioni tra le variabili

Stato, un **assegnamento parziale/completo** di valori a variabili

$\{X_i = v_i, X_j = v_j, \dots\}$

Stato iniziale: {}

Azioni: assegnamento di un valore ad una variabile, tra quelli leciti

Soluzione (goal test): **assegnamento completo** (tutte le variabili hanno un valore) e **consistente** (i vincoli sono tutti soddisfatti)

Esempio Il problema delle 8 regine

$X = \{Q_1, \dots, Q_8\}$

$D_i = \{1, 2, 3, 4, 5, 6, 7, 8\}$

C_i i vincoli di non attacco

Esempio di vincolo fra Q_1 e Q_2 : $\langle(Q_1, Q_2), \{(1, 3), (1, 4), (1, 5), (1, 6), (1, 7), (1, 8), (2, 5), \dots\}\rangle$

Formulazione incrementale o a stato completo

7.2 Strategie per problemi CSP

Fin'ora potevamo solo ricercare la soluzione nel grafo degli stati, guidati da una metrica definita sullo stato. Adesso possiamo:

Usare delle euristiche specifiche per questa classe di problemi

Fare delle inferenze che ci portano a restringere i domini, quindi a **limitare la ricerca**: **propagazione di vincoli**

Fare **backtracking intelligente**

Tipicamente **un mixto di queste strategie**.

7.2.1 Ricerca in problemi CSP

Ad ogni passo si assegna una variabile: la massima profondità della ricerca è fissata dal numero di variabili n .

Versione ingenua L'ampiezza dello spazio di ricerca è $|D_1| \times |D_2| \times \dots \times |D_n|$, dove $|D_i|$ è la cardinalità del dominio di X_i .

Il fattore di diramazione è pari a $n \cdot d$ al primo passo, $(n - 1) \cdot d$ al secondo... le foglie sarebbero $n! \cdot d^n$

C'è una drastica riduzione dello spazio di ricerca, dovuta al fatto che il **Goal-Test** è **commutativo**: l'ordine con cui si assegnano le variabili non è importante). In realtà, quindi, il fattore di diramazione è pari alla dimensione dei domini d (d^n foglie).

7.2.2 Backtracking

BT Ricerca con backtracking a profondità limitata: **controllo anticipato** della violazione dei vincoli. Intuile andare avanti fino alla fine e poi controllare, si può fare backtracking non appena si scopre che un vincolo è stato violato.

La ricerca è limitata naturalmente in profondità dal numero di variabili, quindi il metodo è completo.

Scelta delle variabili

MRV (minimum remaining values): scegliere la **variabile con meno valori legali residui**, cioè quella più vincolata. Si scoprono prima i fallimenti (**fail first**)

Euristica del grado: scegliere la **variabile coinvolta in più vincoli** con le altre variabili (quella più vincolante, o di grado maggiore).

Da usare a parità di MRV

Scelta dei valori: scegliere il **valore meno vincolante**, quello che esclude meno valori possibili per le altre variabili. Meglio valutare prima assegnamento che ha più probabilità di successo.

Propagazione dei vincoli

Verifica in avanti, Forward Checking: meno costosa, una volta assegnato il valore vado a vedere le variabili direttamente collegate nel grafo dei vincoli (non itero)

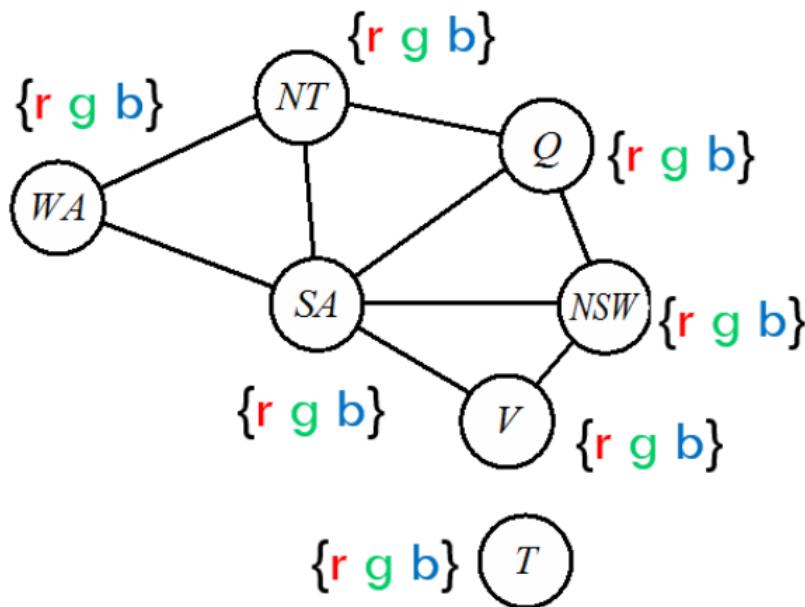
Consistenza di nodo e d'arco: più costosa ma anche più efficace, restringono i valori dei domini delle variabili tenendo conto dei vincoli unari e binari su tutto il grafo (itero finché tutti nodi e archi sono consistenti)

Backtracking ricorsivo per CSP

```

function BACKTRACKING-SEARCH(csp) returns a solution, or failure
  return BACKTRACK({ }, csp)
  
function BACKTRACK(assignment, csp) returns a solution, or failure
  if assignment is complete then return assignment                      trovata soluzione
  var  $\leftarrow$  SELECT-UNASSIGNED-VARIABLE(csp)
  for each value in ORDER-DOMAIN-VALUES(var, assignment, csp) do
    if value is consistent with assignment then                                controllo anticipato
      add {var = value} to assignment
      inferences  $\leftarrow$  INFERENCE(csp, var, value)                         riduce i domini
      if inferences  $\neq$  failure then                                         nessun dominio è vuoto
        add inferences to assignment
        result  $\leftarrow$  BACKTRACK(assignment, csp)
        if result  $\neq$  failure then
          return result
        remove {var = value} and inferences from assignment                  si disfa lo stato
      return failure
  
```

Esempio di FC (Forward Check) Colorare mappa con Rosso, Verde, Blu.



Backtracking cronologico Suppongo di avere $\{Q=R, NSW=G, V=B, T=R\}$ e cerco di assegnare SA. Il fallimento genera un backtracking cronologico: **si provano tutti i valori alternativi per l'ultima variabile**, T, continuando a fallire...

Questo perché **non è la T responsabile del fallimento ma le altre variabili**.

Quindi si può fare un backtracking "intelligente" guidato dalle dipendenze.

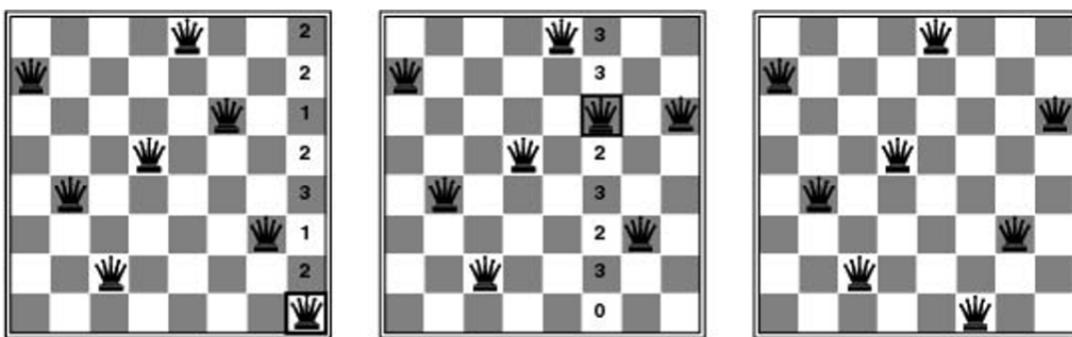
Backtracking guidato dalle dipendenze Considero le alternative solo per le variabili che hanno causato il fallimento $\{Q, NSW, V\}$, l'**insieme dei conflitti**.

Si considerano alternative solo per le variabili che hanno causato il fallimento.

Metodi CSP locali I problemi CSP possono essere affrontati con metodi locali. Esempio: **le regine**.

Euristica dei conflitti minimi (min-conflicts): si sceglie un valore che crea meno conflitti. Solitamente si sceglie a caso una delle regole che violano dei vincoli.

Molto efficace: 1 milione di regine in 50 passi!



Conclusioni Abbiamo visti due domini specifici per paradigma risoluzione problemi ricerca: giochi con avversario (complicazione: ambiente strategico e vincoli di tempo reale, non posso pianificare azioni da eseguire "a occhi chiusi") e CSP (grazie a rappresentazione stato più ricca tecniche problem solving specializzabili e usate per risolvere istanze di problemi di dimensioni maggiori)

Prossimamente: **sistemi basati su conoscenza**. Conoscenza implica capacità inferenziali, con rappresentazione stato ancora più ricco con linguaggio rappresentazione conoscenza. L'inferenza è anch'essa un problema di ricerca in uno spazio di stati.

Capitolo 8

Agenti Basati su Conoscenza

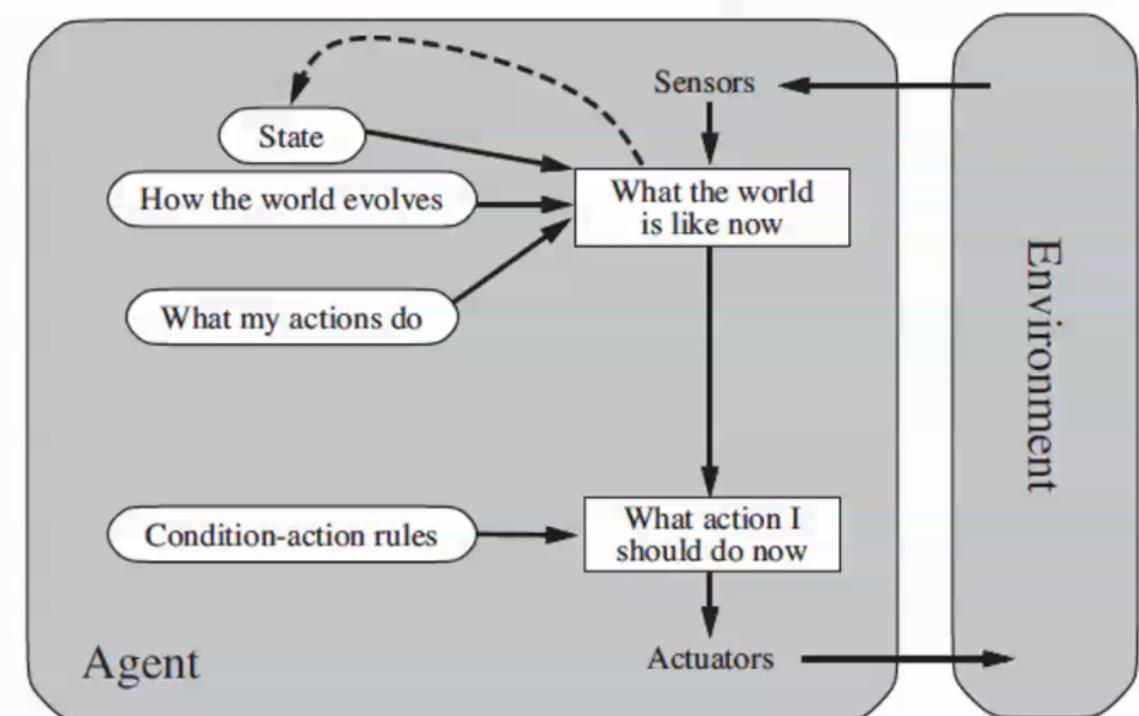
Fin'ora Abbiamo trattato agenti con stato e con obiettivo, in mondi osservabili, con stati atomici e azioni descrivibili in maniera semplice, facendo **enfasi sul processo di ricerca**.

Poiabbiamo trattato descrizioni **fattorizzate**, come nel CSP, che consentono di iniziare a *guardare dentro* lo stato, descritto come un **insieme di caratteristiche rilevanti** (**feature**)

Adesso Vedremo come costruire agenti dotati di capacità inferenziali. Cerchiamo di **migliorare la capacità di ragionamento**, dotando gli agenti di rappresentazioni del mondo più complesse non descrivibili semplicemente.

Agenti basati su conoscenza, con all'interno una **base di conoscenza** (**knowledge base**, o **KB**) con conoscenza espressa in forma esplicita e dichiarativa attraverso un **linguaggio**.

8.1 Agenti Knowledge-Based



La maggior parte dei problemi di I.A. sono *”knowledge-intensive”*. Il mondo è tipicamente **complesso**: serve una rappresentazione parziale e incompleta (**astrazione**) del mondo utile agli scopi dell'agente.

Per ambienti parzialmente osservabili e complessi ci servono linguaggi di rappresentazione della conoscenza più espresivi e **capacità inferenziali**.

La conoscenza può essere codificata a mano, acquisita o estratta da testi ed esperienza.

Approccio dichiarativo vs procedurale Tipicamente la conoscenza in una KB viene espressa in **forma dichiarativa**. L'alternativa è **codificare la conoscenza in maniera procedurale**, attraverso un programma che implementa il processo decisionale. Un agente KB (con conoscenza dichiarativa) è più flessibile: perché è più semplice ricevere conoscenza in maniera incrementale e modificare il comportamento con l'esperienza.

8.1.1 Il mondo del Wumpus

Esempio Agente in 1, 1: esplorare e trovare oro

Misura delle prestazioni

+1000 se trova l'oro, torna in [1, 1] ed esce

-1000 se muore

-1 per ogni azione

-10 se usa la freccia

Percezioni

puzzo nelle caselle adiacenti al Wumpus

brezza nelle caselle adiacenti alle buche

luccichio nelle caselle con l'oro

bump se batte in un muro

urlo se il Wumpus viene ucciso

L'agente non percepisce la sua locazione

Azioni

Avanti

A destra di 90 gradi

A sinistra di 90 gradi

Scaglia la freccia (solo una)

Esce

Ambienti generati a caso, con [1, 1] safe.

4	Stench		Breeze
3	Wumpus	Stench	PIT
2	Stench		Breeze
1	START	Breeze	PIT
	1	2	3
			4

8.1.2 Knowledge-Base

KB Insieme di **enunciati** espressi in un linguaggio di rappresentazione.
L'agente interagisce con una KB con un'interfaccia funzionale **Tell-Ask**:

Tell per aggiungere nuovi enunciati alla KB

Ask per interrogare la KB

Retract per eliminare enunciati

Enunciati in KG rappresentano la conoscenza agente. Le **risposte** α devono essere tali che α è una conseguenza della KB.

Problema Il problema da risolvere: data una base di conoscenza KB che contiene una rappresentazione dei fatti ritenuti veri, vorrei sapere se un certo fatto α è vero di conseguenza, cioè vorrei sapere se

$$\text{KB} \models \alpha \\ (\text{conseguenza logica})$$

In questo caso, i modelli che soddisfano KB sono un sottoinsieme dei modelli che soddisfano α

Programma di un agente KB Di seguito

```
function Agente-KB (percezione) returns azione
    persistent: KB, una base di conoscenza
        t, un contatore, inizialmente a 0, che indica il tempo
        TELL(KB, Costruisci-Formula-Percezione(percezione, t))
        azione = ASK(KB, Costruisci-Query-Azione(t))
        TELL(KB, Costruisci-Formula-Azione(azione, t))
        t = t + 1
    return azione
```

KB vs DB Una **base di conoscenza** è una **rappresentazione esplicita, parziale** e compatta in un linguaggio simbolico: contiene sia fatti esplicativi (es. *Pozzo in [3, 3]*) ma anche fatti generali o regole (es: brezza in caselle adiacenti a pozzi)

Una **base di dati** invece contiene solo fatti specifici e consente solo il recupero.

La differenza è che la **KB ha capacità inferenziale**: si possono **derivare nuovi fatti** da quelli memorizzati esplicitamente.

Trade-Off Fondamentale della Rappresentazione della Conoscenza Trovare il **giusto compromesso tra l'espressività** del linguaggio di rappresentazione e la **complessità** del meccanismo inferenziale.

Più un linguaggio è espressivo meno è efficiente il meccanismo inferenziale. Questi due obiettivi sono in contrasto, bisogna mediare e trovare un **compromesso**.

Formalismi per la Rappresentazione della Conoscenza

Un formalismo per la rappresentazione della conoscenza ha tre componenti:

Sintassi: un linguaggio composto da un vocabolario e da regole per la formazione delle frasi (**enunciati**)

Semantica: stabilisce la corrispondenza tra gli enunciati e fatti del mondo. Se un agente ha un enunciato α nella sua KB, allora crederà che il fatto corrispondente ad α sia vero nel mondo

Meccanismo Inferenziale: codificato o meno tramite regole di inferenza come nella logica, che ci consente di inferire nuovi fatti

Logica come linguaggio Qual è la complessità computazionale del problema $\text{KB} \models \alpha$ nei vari linguaggi logici? Quali algoritmi decisione e strategia ottimizzazione?

Linguaggi logici: calcolo proposizionale (POP) e logica dei predicati (FOL). Sono adatti per la rappresentazione della conoscenza? Da una parte sono anche troppo complessi, in particolare il FOL. Dall'altra, ci sono meccanismi non posseduti da FOL e POP ma che sono utili nelle KB.

Rivistazione di PROP e FOL per rappresentazione conoscenza, con attenzione ad algoritmi e complessità. **Contrazioni:** linguaggi a regole e **programmazione logica**.

8.1.3 Algoritmo TT-entails

$\text{KB} \models \alpha?$

Enumera tutte possibili interpretazioni KB (k simboli, 2^k possibili interpretazioni)

Per ciascuna interpretazione

Se non soddisfa KB , OK

Se soddisfa KB , si controlla che soddisfi anche α

Se si trova anche solo una interpretazione che soddisfa KB e non α , la risposta sarà NO

Algoritmo Di seguito

```

function TT-ENTAILS?(KB,  $\alpha$ ) returns true or false
  inputs: KB, the knowledge base, a sentence in propositional logic
           $\alpha$ , the query, a sentence in propositional logic

  symbols  $\leftarrow$  a list of the proposition symbols in KB and  $\alpha$ 
  return TT-CHECK-ALL(KB,  $\alpha$ , symbols, { })

```

```

function TT-CHECK-ALL(KB,  $\alpha$ , symbols, model) returns true or false
  if EMPTY?(symbols) then
    if PL-TRUE?(KB, model) then return PL-TRUE?( $\alpha$ , model)
    else return true // when KB is false, always return true
  else do
     $P \leftarrow$  FIRST(symbols)
    rest  $\leftarrow$  REST(symbols)
    return (TT-CHECK-ALL(KB,  $\alpha$ , rest, model  $\cup$  { $P = \text{true}$ })
            and
            TT-CHECK-ALL(KB,  $\alpha$ , rest, model  $\cup$  { $P = \text{false}$ }))

```

Esempio $(\neg A \vee B) \wedge (A \vee C) \models (B \vee C)$?

```

TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [A, B, C], {} )

TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [B, C], {A = true} )
TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [C], {A = true, B = true} )
OK TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [], {A = true, B = true, C = true} )
OK TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [], {A = true, B = true, C = false} )
TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [C], {A = true, B = false} )
OK TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [], {A = true, B = false, C = true} )
OK TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [], {A = true, B = false, C = false} )
TT-CHECK-ALL(  $(\neg A \vee B) \wedge (A \vee C)$ ,  $(B \vee C)$ , [B, C], {A = false} ) ...

```

Solo alla fine, **dopo aver provato tutti i possibili assegnamenti**, possiamo rispondere se $(B \vee C)$ è conseguenza logica.

8.2 Algoritmi per la soddisfacibilità (SAT)

Usano una KB come **insieme di clausole**, cioè insiemi di letterali:

{A, B} { \neg B, C, D} { \neg A, F}

La forma a clausole è la **forma normale congiuntiva (CNF)**, cioè una congiunzione di disgiunzioni letterali:

{ $A \vee B$ } \wedge { $\neg B \vee C \vee D$ } \wedge { $\neg A \vee F$ }

Non è restrittiva: è sempre possibile ottenerla con trasformazioni che preservano l'equivalenza logica.

Trasformazione in forma a clausole I passi sono:

- Eliminazione della \Leftrightarrow : $(A \Leftrightarrow B) = (A \Rightarrow B) \wedge (B \Rightarrow A)$
- Eliminazione dell' \Rightarrow : $(A \Rightarrow B) = (\neg A \vee B)$
- Negazioni all'interno (de Morgan):
 - $\neg(A \vee B) = (\neg A \wedge \neg B)$
 - $\neg(A \wedge B) = (\neg A \vee \neg B)$
- Distribuzione di \vee su \wedge : $(A \vee (B \wedge C)) = (A \vee B) \wedge (A \vee C)$

8.2.1 Algoritmo DPLL

DPLL Davis, Putman, Lovemann, Loveland. Parte da una KB in forma a clausole. Enumerazione in profondità di tutte le possibili interpretazioni alla ricerca di un modello. Tre miglioramenti rispetto TTEntails:

Terminazione anticipata: si può decidere sulla verità di una clausola anche con interpretazioni parziali, basta che un letterale sia vero.

Se A è *true*, lo sono anche $\{A, B\}$ e $\{A, C\}$ indipendentemente dai valori di B e di C .

Se anche una sola clausola è falsa, l'interpretazione non può essere un modello dell'insieme delle clausole.

Euristica dei **simboli puri** (o letterali) **puri**: un **simbolo puro** è un simbolo che **appare con lo stesso segno in tutte le clausole**. Ad esempio, nella KB $\{A, \neg B\}, \{\neg B, \neg C\}, \{C, A\}$ i simboli A e B sono puri.

Nel determinare un simbolo se è puro possiamo trascurare occorrenze simbolo in clausole già rese vere.

Se simbolo è puro può essere assegnato a *true* se positivo e *false* se negativo senza escludere modelli utili: se le clausole hanno un modello continueranno ad averlo anche dopo questo assegnamento. L'assegnamento è obbligato.

Euristica delle **clausole unitarie**: una **clausola unitaria** è una **clausola con un solo letterale non assegnato**. Ad esempio, $\{B\}$ è unitaria, ma anche $\{B, \neg C\}$ quando $C = \text{true}$.

Conviene assegnare prima valori ai letterali in clausole unitarie. L'assegnamento è univoco (*true* se positivo, *false* se negativo)

function DPLL-SATISFIABLE?(s) **returns** *true* or *false*

inputs: s , a sentence in propositional logic

$\text{clauses} \leftarrow$ the set of clauses in the CNF representation of s

$\text{symbols} \leftarrow$ a list of the proposition symbols in s

return DPLL($\text{clauses}, \text{symbols}, \{\}$)

function DPLL($\text{clauses}, \text{symbols}, \text{model}$) **returns** *true* or *false*

if every clause in clauses is true in model **then return** *true*

if some clause in clauses is false in model **then return** *false*

$P, \text{value} \leftarrow \text{FIND-PURE-SYMBOL}(\text{symbols}, \text{clauses}, \text{model})$

if P is non-null **then return** DPLL($\text{clauses}, \text{symbols} - P, \text{model} \cup \{P=\text{value}\}$)

$P, \text{value} \leftarrow \text{FIND-UNIT-CLAUSE}(\text{clauses}, \text{model})$

if P is non-null **then return** DPLL($\text{clauses}, \text{symbols} - P, \text{model} \cup \{P=\text{value}\}$)

$P \leftarrow \text{FIRST}(\text{symbols}); \text{rest} \leftarrow \text{REST}(\text{symbols})$

return DPLL($\text{clauses}, \text{rest}, \text{model} \cup \{P=\text{true}\}$) **or**

DPLL($\text{clauses}, \text{rest}, \text{model} \cup \{P=\text{false}\}$)

Miglioramenti di DPLL DPLL è completo e termina sempre. Alcuni miglioramenti:

Analisi di componenti (sottoproblemi indipendenti): se le variabili possono essere suddivise in sottoinsiemi disgiunti

Ordinamento di variabili e valori: scegliere la variabile che compare in più clausole

Backtracking intelligente

Altre ottimizzazioni...

8.2.2 Metodi locali per SAT

Formulazione

Gli stati sono assegnamenti completi

L'obiettivo è un assegnamento che soddisfa tutte le clausole (un **modello**)

Si parte da un assegnamento casuale

Ad ogni passo **si cambia il valore di una proposizione (flip)**

Gli stati sono valutati contando il numero di clausole non soddisfatte, meno sono meglio è (o soddisfatte)

Algoritmi Ci sono molti minimi locali, per sfuggire ai quali bisogna introdurre delle perturbazioni casuali. Ad esempio, hill-climbing con riavvio casuale, simulated annealing...

C'è stata molta sperimentazione per trovare il miglior compromesso tra il grado di avidità e la casualità.

WalkSAT è uno degli algoritmi più semplici ed efficaci.

8.2.3 Algoritmo WalkSAT

Ad ogni passo **sceglie a caso una clausola non ancora soddisfatta**. Poi **sceglie un simbolo da modificare (flip)**, scegliendo con probabilità p (di solito 0.5) tra una delle due seguenti possibilità:

Passo casuale: sceglie un simbolo a caso da flippare

Passo di ottimizzazione: sceglie quello che rende più clausole soddisfatte

Si arrende dopo un numero predefinito di flip.

```

function WALKSAT(clauses, p, max_flips) returns a satisfying model or failure
  inputs: clauses, a set of clauses in propositional logic
            p, the probability of choosing to do a “random walk” move, typically around 0.5
            max_flips, number of flips allowed before giving up

  model  $\leftarrow$  a random assignment of true/false to the symbols in clauses
  for i = 1 to max_flips do
    if model satisfies clauses then return model
    clause  $\leftarrow$  a randomly selected clause from clauses that is false in model
    with probability p flip the value in model of a randomly selected symbol from clause
    else flip whichever symbol in clause maximizes the number of satisfied clauses
  return failure

```

Analisi Se *maxflips* è ∞ e l'insieme clausole è soddisfacibile, prima o poi termina. Ma **non può essere usato per verificare l'insoddisfacibilità** (è **incompleto**). Il problema è decidibile ma l'algoritmo è incompleto.

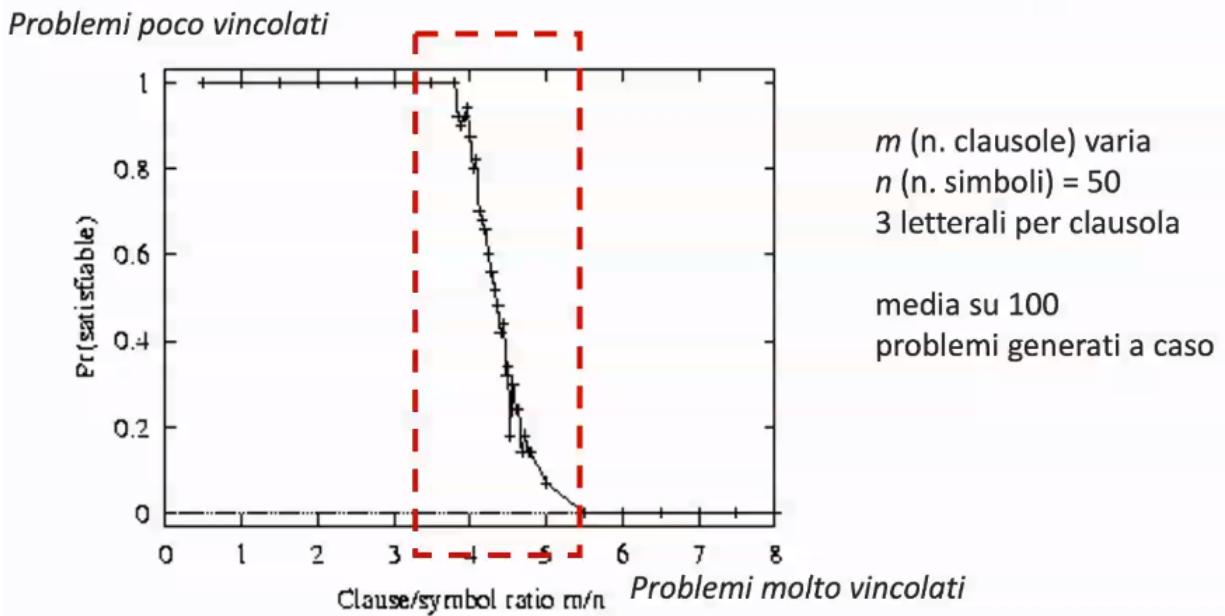
Va bene per cercare un modello, sapendo che c'è. Ma se è insoddisfacibile, non termina.

Lo si usa perché è efficiente in tempo e spazio.

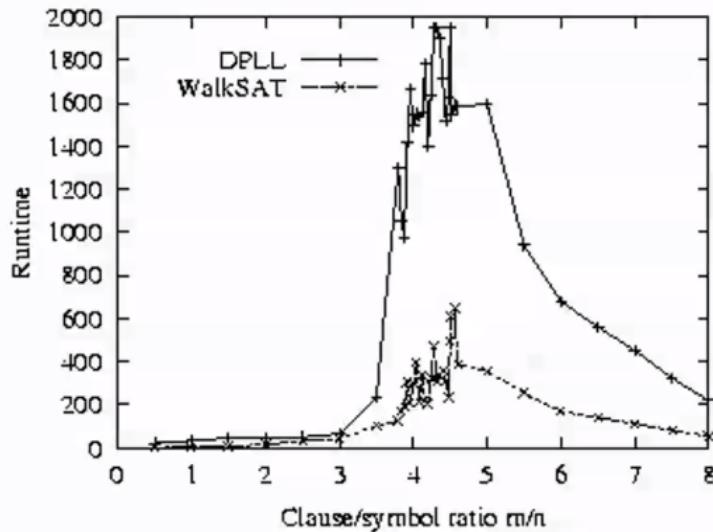
Problemi SAT difficili Se un problema ha molte soluzioni (sotto-vincolato) è più probabile che WalkSAT ne trovi una in tempi brevi. Per esempio, se ha 16 soluzioni su 32, un assegnamento ha il 50% di probabilità di essere giusto, quindi **2 passi in media**.

Più grande è il rapporto più vincolato è il problema.

Probabilità di soddisfabilità in funzione di $\frac{m}{n}$



Confronto tra DPLL e WalkSAT



WalkSAT è molto efficace quando il problema è sottovincolato: cioè quando ci sono molti modelli che soddisfano la KB (molte soluzioni). Ad esempio 16 soluzioni su 32 modelli possibili, quindi un assegnamento ha il 50% di probabilità di essere vero.

8.3 Inferenza come Deduzione

Un altro modo per decidere se $\text{KB} \models A$ è usare un meccanismo di **dimostrazione (deduzione)**.

Si scrive $\text{KB} \vdash A$ (**A è deducibile da KB**) e la deduzione avviene specificando delle regole di inferenza: dovranno derivare solo le formule che sono conseguenza logica e **tutte** le formule che sono conseguenza logica.

Sono le proprietà della **correttezza** e **completezza**.

Correttezza Se $\text{KB} \vdash A$, allora $\text{KB} \models A$

Tutto ciò che è derivabile è conseguenza logica. Le regole preservano la verità

Completezza Se $\text{KB} \models A$, allora $\text{KB} \vdash A$

Tutto ciò che è conseguenza logica è ottenibile tramite il meccanismo di inferenza.

(**Teorema di refutazione**)

Dimostrazione come ricerca Come decidere ad ogni passo quale regola di inferenza applicare? A quali premesse? Come evitare esplosione combinatoria? Problema esplorazione spazio di stati. Ci riguarda perché vogliamo progettare algoritmi di inferenza. Una **procedura di dimostrazione** definisce direzione e strategia di ricerca.

Direzione Nella dimostrazione di teoremi **conviene procedere all'indietro**. Con un'applicazione in avanti delle regole di inferenza non controllata posso ottenere, ad esempio da $A, B: A \wedge B, A \wedge (A \wedge B) \dots$ All'indietro invece, se voglio dimostrare $A \wedge B$, si cerca di dimostrare A e poi B . Se voglio dimostrare $A \Rightarrow B$, assumo A e cerco di dimostrare $B \dots$

Strategia Anche assumendo insieme di regole di inferenza completo se l'algoritmo non è completo potrei non trovare soluzione. La complessità è alta: è un problema **decidibile ma NP-completo**.

8.3.1 Regola di risoluzione per PROP

Meno regole uso meglio è, senza rinunciare alla completezza. L'unica regola di inferenza è la regola di risoluzione, che presuppone la forma a clausole. Con due clausole, una P e l'altra che contiene $\neg P$, possiamo considerare l'OR degli altri elementi togliendo P e $\neg P$. Cioè

$$\frac{\{P, Q\} \quad \{\neg P, R\}}{\{Q, R\}}$$

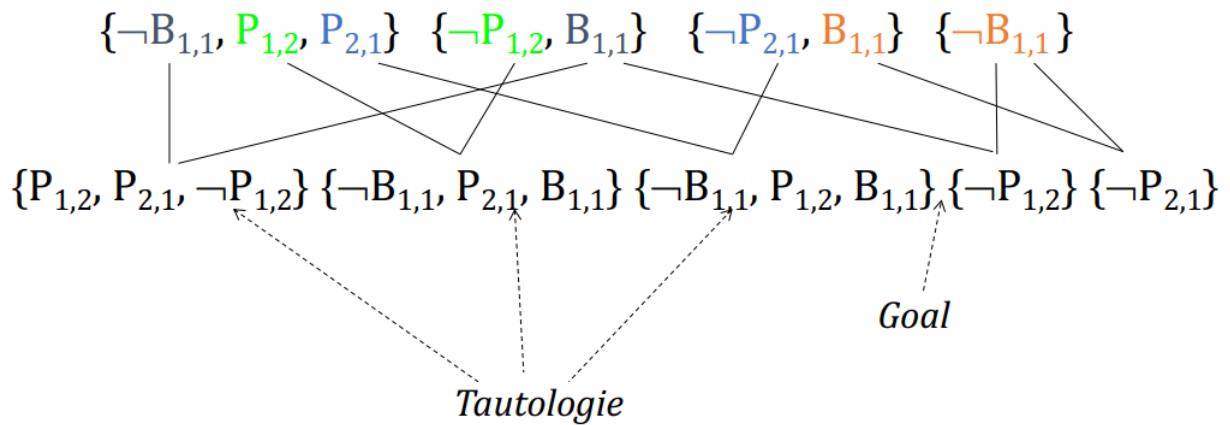
Corretta? Basta pensare ai modelli.

In generale Con l, m **letterali**, simboli di proposizione positivi o negativi. l_i e m_j sono **uguali ma di segno opposto**.

$$\frac{\{l_1, l_2, \dots, l_i, \dots, l_k\} \quad \{m_1, m_2, \dots, m_j, \dots, m_n\}}{\{l_1, l_2, \dots, l_{i-1}, l_{i+1}, \dots, l_k, m_1, m_2, \dots, m_{j-1}, m_{j+1}, \dots, m_n\}}$$

Caso particolare $\frac{\{P\} \quad \{\neg P\}}{\{\}}$: **clausola vuota o contraddizione**.

Grafo di risoluzione



Attenzione! $\{B, N\} \quad \{\neg B, \neg N\} \rightarrow \{\}$ NO!

Diventa o $\{N, \neg N\}$ oppure $\{B, \neg B\}$, un passo alla volta!

Siamo sicuri che basti una regola? Completezza: $\text{KB} \models \alpha \Rightarrow \text{KB} \vdash_{res} \alpha$? Non sempre. Un controesempio è $\text{KB} \models \{A, \neg A\}$ ma non è vero che $\text{KB} \vdash_{res} \{A, \neg A\}$

Ma ho due strumenti:

Teorema di risoluzione KB insoddisfacibile $\Leftrightarrow \text{KB} \vdash_{res} \{\}$ completezza

Teorema di refutazione $\text{KB} \models \alpha \Leftrightarrow \text{KB} \cup \{\neg\alpha\}$ è insoddisfacibile

Nell'esempio $\text{KB} \cup \text{FC}(\neg(A \text{ or } \neg A))$ è insoddisfacibile? Si, perché $\text{KB} \cup \{A\} \cup \{\neg A\} \vdash_{res} \{\}$ in un passo

Conclusioni Abbiamo visto che gli agenti KB che usano il calcolo proposizionale come linguaggio di rappresentazione possono decidere se $\text{KB} \models \alpha$. Il problema è decidibile ma intrattabile (NP al caso peggiore). Esistono algoritmi efficienti e completi che consentono di affrontare problemi di grosse dimensioni, ed esistono algoritmi locali particolarmente efficienti ma non completi. Oppure per via deduttiva.

8.3.2 Logica del Primo Ordine

8.3.3 Inferenza nella Logica del Primo Ordine

Regole d'inferenza per \forall Istanziazione dell'Universale

$$\frac{\forall x \ A[x]}{A[g]}$$

Da $\forall x \ King(x) \wedge \text{Greedy}(x) \Rightarrow \text{Evil}(x)$ si possono ottenere

$$\text{King(John)} \wedge \text{Greedy(John)} \Rightarrow \text{Evil(John)}$$

$$\text{King(Father(John))} \wedge \text{Greedy(Father(John))} \Rightarrow \text{Evil(Father(John))}$$

Regola d'inferenza per \exists Istanziazione dell'Esistenziale

$$\frac{\exists x \ A[x]}{A[k]}$$

Se \exists non compare nell'ambito di \forall , K è una costante nuova (**costante di Skolem**). Altrimenti va introdotta una funzione (**di Skolem**) nelle variabili quantificate universalmente

$$\exists x \ \text{Padre}(x, G) \text{ diventa } \text{Padre}(k, G)$$

$$\forall x \ \exists y \ \text{Padre}(x, y) \text{ diventa } \forall x \ \text{Padre}(x, p(x)) \text{ e non } \forall x \ \text{Padre}(x, k) \text{ altrimenti tutti avrebbero lo stesso padre}$$

Grounding Anche detto **proposizionalizzazione**: creo tante istanze delle formule quantificate universalmente quanti sono gli oggetti menzionati ed elimino i quantificatori esistenziali **skolemizzando**.

La KB diventa proposizionale e possiamo applicare gli algoritmi visti. Sorgono dei problemi: le costanti sono in numero finito, ma **se ci sono delle funzioni il numero delle istanze da creare diventa infinito** ($\text{John}, \text{Padre}(\text{John}), \text{Padre}(\text{Padre}(\text{John})), \dots$)

8.3.4 Teorema di Herbrand

Se $\text{KB} \models A \Rightarrow$ c'è una dimostrazione che coinvolge solo un sottoinsieme finito della KB proposizionalizzata. Si può procedere in modo incrementale

Creo le istanze con le costanti

Creo quelle con un solo livello di annidamento: $\text{Padre}(\text{John}), \text{Madre}(\text{John})$

Creo quelle con due livelli di annidamento: $\text{Padre}(\text{Padre}(\text{John})), \text{Padre}(\text{Madre}(\text{John})), \text{Madre}(\text{Padre}(\text{John})), \text{Madre}(\text{Madre}(\text{John}))$

Se $\text{KB} \not\models A$, il processo non termina \Rightarrow **semidecidibile**.

8.3.5 Regola di risoluzione per il FOL

Abbiamo visto la regola di risoluzione per il PROP, un metodo deduttivo corretto e completo con un'unica regola. Possiamo estenderla al FOL? Si, ma per definirla dobbiamo estendere al FOL la trasformazione in forma a clausole e introdurre il concetto di unificazione.

Forma a clausole Costanti, funzioni e predicati sono come definiti prima, ma **escludiamo la formule atomiche del tipo** ($t_1 = t_2$).

Clausola Una clausola è insieme di letterali (formula atomica eventualmente negata) che rappresenta la loro disgiunzione.

Una KB è un insieme di clausole.

$$\text{Clausola} \Rightarrow \{\text{Letterale}, \dots, \text{Letterale}\}$$

$$\text{Letterale} \Rightarrow \text{FormulaAtomica} \mid \neg\text{FormulaAtomica}$$

Trasformazione in forma a clausole

Teorema \forall formula chiusa α del FOL, è possibile trovare in maniera **effettiva** un insieme di clausole $\text{FC}(\alpha)$ soddisfacibile $\Leftrightarrow \alpha$ era soddisfacibile (viceversa, insoddisfacibile $\Leftrightarrow \alpha$ era insoddisfacibile)

Esempio trasformazione Vediamo un esempio per la frase "tutti coloro che amano tutti gli animali sono amati da qualcuno".

$$\forall x (\forall y \text{ Animale}(y) \Rightarrow \text{Ama}(x, y)) \Rightarrow (\exists y \text{ Ama}(y, x))$$

1. Eliminazione delle implicazioni \Rightarrow e \Leftrightarrow :

$$A \Rightarrow B \text{ diventa } \neg A \vee B, A \Leftrightarrow B \text{ diventa } (\neg A \vee B) \wedge (\neg B \vee A).$$

$$\forall x (\forall y \text{ Animale}(y) \Rightarrow \text{Ama}(x, y)) \underline{\Rightarrow} (\exists y \text{ Ama}(y, x))$$

$$\forall x \neg (\forall y \text{ Animale}(y) \underline{\Rightarrow} \text{Ama}(x, y)) \vee (\exists y \text{ Ama}(y, x))$$

$$\forall x \neg (\forall y \neg \text{Animale}(y) \vee \text{Ama}(x, y)) \vee (\exists y \text{ Ama}(y, x))$$

2. Negazioni all'interno: $\neg\neg A$ diventa A

$$\neg(A \wedge B) \text{ diventa } \neg A \vee \neg B \text{ (De Morgan)}$$

$$\neg(A \vee B) \text{ diventa } \neg A \wedge \neg B \text{ (De Morgan)}$$

$$\neg \forall x A \text{ diventa } \exists x \neg A, \neg \exists x A \text{ diventa } \forall x \neg A$$

$$\forall x \neg (\forall y \neg \text{Animale}(y) \vee \text{Ama}(x, y)) \vee (\exists y \text{ Ama}(y, x))$$

$$\forall x (\exists y \neg (\neg \text{Animale}(y) \vee \text{Ama}(x, y))) \vee (\exists y \text{ Ama}(y, x))$$

$$\forall x (\exists y (\text{Animale}(y) \wedge \neg \text{Ama}(x, y))) \vee (\exists y \text{ Ama}(y, x))$$

3. Standardizzazione delle variabili: ogni quantificatore una variabile diversa

$$\forall x (\exists y (\text{Animale}(y) \wedge \neg \text{Ama}(x, y))) \vee (\exists z \text{ Ama}(z, x))$$

4. Skolemizzazione: eliminazione di quantificatori esistenziali.

Ci sono due quantificatori esistenziali nell'ambito di uno universale, dobbiamo introdurre due funzioni di Skolem

$$\forall x (\text{Animale}(F(x)) \wedge \neg \text{Ama}(x, F(x))) \vee (\text{Ama}(G(x), x))$$

5. Eliminazione di quantificatori universali

Possiamo portarli tutti davanti (forma premessa) ed eliminarli usando la convenzione che le variabili libere sono quantificate universalmente.

Se B non contiene x

$$(\forall x A) \vee B \text{ diventa } \forall x (A \vee B), (\forall x A) \wedge B \text{ diventa } \forall x (A \wedge B)$$

$$\underline{\forall x} (\text{Animale}(F(x)) \wedge \neg \text{Ama}(x, F(x))) \vee (\text{Ama}(G(x), x))$$

$$(\text{Animale}(F(x)) \wedge \neg \text{Ama}(x, F(x))) \vee (\text{Ama}(G(x), x))$$

6. Forma normale congiuntiva: congiunzioni di disgiunzioni letterali

$$A \vee (B \wedge C) \text{ diventa } (A \vee B) \wedge (A \vee C) \text{ Quindi}$$

$$(\text{Animale}(F(x)) \wedge \neg \text{Ama}(x, F(x))) \underline{\vee} (\text{Ama}(G(x), x))$$

$$(\text{Animale}(F(x)) \vee \neg \text{Ama}(G(x), x)) \wedge ((\neg \text{Ama}(x, F(x)) \vee \text{Ama}(G(x), x)))$$

7. Notazione a clausole

$\{\text{Animale}(F(x)), \text{Ama}(G(x), x)\}$
 $\{\neg\text{Ama}(x, F(x)), \text{Ama}(G(x), x)\}$

8. Separazione delle variabili: clausole diverse, variabili diverse

$$\forall x (P(x) \wedge Q(x)) \Leftrightarrow \forall x_1 P(x_1) \wedge \forall x_2 Q(x_2)$$

$\{\text{Animale}(F(x)), \text{Ama}(G(x), x)\} \longrightarrow \{\text{Animale}(F(x_1)), \text{Ama}(G(x_1), x_1)\}$
 $\{\neg\text{Ama}(x, F(x)), \text{Ama}(G(x), x)\} \longrightarrow \{\neg\text{Ama}(x_2, F(x_2)), \text{Ama}(G(x_2), x_2)\}$

La Skolemizzazione non preserva l'equivalenza: $P(a) \models \exists x P(x)$ ma $\exists x P(x) \not\models P(x)$

Unificazione Operazione con la quale si determina se due espressioni possono essere rese identiche mediante una sostituzione di termini a variabili.

Il risultato è la sostituzione che rende le due espressioni identiche, detta **unificatore**, o **FAIL** se le espressioni non sono unificabili.

Sostituzione Insieme finito di associazioni tra variabili e termini in cui ogni variabile compare una sola volta sulla sinistra. Es $\{x_1/A, x_2/f(x_3), x_3/B\}$ significa che A va sostituita a x_1 , $f(x_3)$ sostituita a $x_2\dots$

A sinistra vanno le variabili e a destra costanti e variabili, con la restrizione che la variabile sulla sinistra non può comparire anche sulla destra della stessa "coppia".

Applicazione Sia σ una sostituzione e A espressione, $A\sigma$ è l'istanza generale dalla sostituzione (delle variabili con l'espressione corrispondente)

In AIMA $\text{SUBST}(\sigma, A)$. Le variabili sono **sostituite simultaneamente** e si esegue un solo passo di sostituzione

Espressioni unificabili Se esiste una sostituzione che le rende identiche, cioè se esiste un unificatore.

In genere l'unificatore τ non è unico, quindi vorremmo l'unificatore più generale di tutti (MGU). **Teorema: l'unificatore più generale è unico** a meno dei nomi delle variabili (l'ordine non conta).

Algoritmo di unificazione Input: p, q espressioni.

Output: MGU θ se esiste ($\text{unify}(p, q) = \theta$ tale che $\text{subst}(\theta, p) = \text{subst}(\theta, q)$) altrimenti **FAIL**.

Esplora in parallelo le due espressioni e costruisce l'unificatore strada facendo. Fallisce non appena trova due espressioni non unificabili. Una causa del fallimento sono le sostituzioni del tipo $x = f(x)$: questo controllo si chiama **occurr-check**.

```

function UNIFY(x, y, theta) returns una sostituzione
    # theta , la sostituzione costruita fin'ora (opzionale , vuota di default)
    if theta = FAIL then return FAIL
    else if x = y: return theta
    else if VARIABLE?(x): return UNIFY-VAR(x, y, theta)
    else if VARIABLE?(y): return UNIFY-VAR(y, x, theta)
    else if COMPOUND?(x) and COMPOUND?(y):
        return UNIFY(x.ARGS, y.ARGS, UNIFY(x.OP, y.OP, theta))
    else if LIST?(x) and LIST?(y):
        return UNIFY(x.REST, y.REST, UNIFY(x.FIRST, y.FIRST, theta))
    else return FAIL

function UNIFY-VAR(var, x, theta) returns una sostituzione
    if {var/val} in theta: return UNIFY(val, x, theta) # var ha già un valore
    else if {x/val} in theta: return UNIFY(var, val, theta) # x ha già un valore
    else if OCCUR-CHECK?(var, x): return FAIL # controllo di occorrenza
    else return EXTEND({var/x} , theta)

```

OCCURR-CHECK controlla se var occorre all'interno dell'espressione x. In tal caso, fallisce. Controllo di complessità quadratica.

Attenzione: EXTEND non aggiunge semplicemente ma applica la sostituzione in θ .

8.4 Definizione e Confronto di Euristiche Ammissibili

Mondo dei blocchi Problema classico, parte dei micromondi usati soprattutto per la pianificazione. Serie di blocchi su un tavolo impilati e vogliamo raggiungere certa config. final. Mosse sono spostare blocco su tavolo o su altro blocco, a condizione che blocco sia libero senza blocchi sopra e blocco destinazione anche.

Euristica H1 Numero di blocchi appoggiati su blocco sbagliato (incluso tavolo)

Euristica H2 Numero blocchi con supporto (torre sotto) sbagliato

Sono ammissibili? Sono ammissibili se servono almeno H_1 mosse per giungere in stato goal. H_2 è più accurata perché domina H_1 , valore più alto per ogni stato possibile. Questo perché $H_2 \Rightarrow H_1$, i blocchi contati in H_1 vengono contati anche in H_2 , ma H_2 ha casi non contati da H_1 .

Per trovare euristica bisogna contare quanto dista soluzione e usare questa come euristica.

Euristica = n ammissibile se servono almeno n mosse per arrivare in stato finale, ≥ 0 ovunque ma = 0 solo in stati goal.

Capitolo 9

Strategie di risoluzione

Fin'ora Abbiamo visto metodi di risoluzione per le KB in forma a clausole: l'unificazione e la regola di risoluzione per FOL (estensione rispetto alla regola PROP).

Come rendere più efficiente il meccanismo di risoluzione? Bisogna adottare strategie di risoluzione: tecniche per esplorare in maniera efficiente il grafo di risoluzione, possibilmente senza perdere la completezza.

Un percorso che ci porterà a giustificare i sistemi a regole e le restrizioni del FOL associate.

9.1 Strategia di risoluzione

Distinzione tra

Strategie di cancellazione: *ci sono clausole che possiamo eliminare?*

Strategie di restrizione: *posso usare ad ogni passo solo alcune clausole?*

Strategie di ordinamento: *posso "risolvere" i letterali in un ordine specifico?*

Tutto possibilmente **senza perdere completezza**

9.1.1 Strategie di Cancellazione

Consiste nel **rimuovere** dalla KB, ai fini della dimostrazione, **le clausole che non saranno mai utili nel processo di risoluzione.**

Clausole con Letterali Puri

Letterali che non hanno il loro negato nella KB

$\{\neg P, \neg Q, R\} \{\neg P, \underline{S}\} \{\neg Q, \underline{S}\} \{P\} \{Q\} \{\neg R\}$

Non potranno mai essere risolte con altre clausole per ottenere {}, tanto vale eliminarle. Non si perdono soluzioni

Tautologie

Clausole con due letterali identici e complementari

$\{P(A), \neg P(A), \dots\} \{P(x), Q(y), \neg Q(y)\}$

Nota: non basta che siano unificabili e di segno opposto

$\{\neg P(A), P(x)\} \{P(A)\} \{\neg P(B)\}$ è insoddisfacibile

$\{\overline{P(A)}\} \{\overline{P(B)}\}$ è soddisfacibile

Le tautologie **possono essere generate**, quindi questo controllo deve essere eseguito ad ogni passo

Clausole Sussunte (implicate)

$P(x)$ **sussume** $P(A)$, $P(A)$ **sussume** $\{P(A), P(B)\}$

In generale α **sussume** $\beta \Leftrightarrow \exists \sigma$ con $\alpha\sigma \subset \beta$, cioè se un'istanza di α con la sostituzione σ è un sottoinsieme di β . Ad esempio

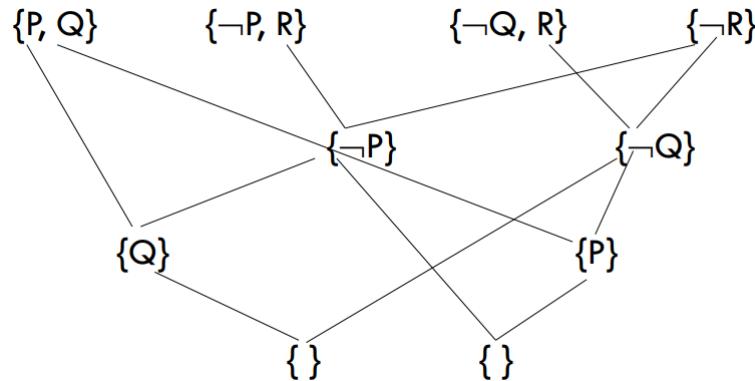
$\{P(x), Q(y)\}$ sussume $\{P(A), Q(v), R(w)\}$ infatti $\{P(x), Q(y)\}\{x/A, y/v\} = \{P(A), Q(v)\} \subset \{P(A), Q(v), R(w)\}$
 β può essere ricavata da α , quindi β può essere eliminata senza perdere soluzioni. Anche le clausole sussunte possono essere generate.

9.1.2 Strategie di Restrizione

Ad ogni passo si sceglie tra un sottoinsieme di possibili clausole

Risoluzione unitaria

Almeno una delle due clausole utilizzate nel passaggio è **unitaria** (un solo letterale)

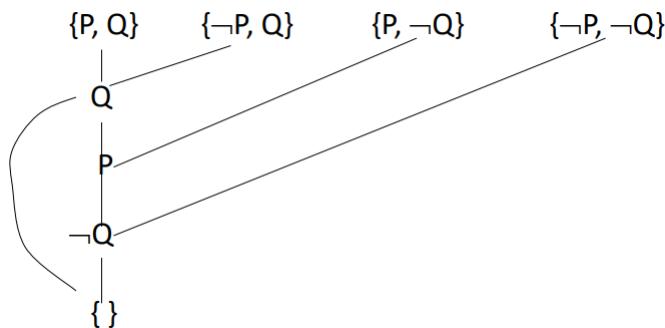


Converge molto rapidamente, perché ad ogni passo si elimina una clausola, ed è facile da implementare. Ma non è completa, esempio: $\{P, Q\} \{\neg P, Q\} \{P, \neg Q\} \{\neg P, \neg Q\} \vdash_{RES} \{\}$ non con risoluzione unitaria.

Completa per **Clausole di Horn**: clausole con **al più** un letterale positivo

Risoluzione lineare

Prendo l'ultima clausola generata con una clausola da input (della KB iniziale), oppure una clausola antenata nella sequenza. Completa per la risoluzione per refutazione.



Risoluzione guidata dal goal/da insieme di supporto (sottoinsieme della KB responsabile dell'insoddisfacibilità)

Almeno una delle due clausole appartiene a questo insieme o a suoi discendenti

Tipicamente, assumendo la KB iniziale consistente, **si sceglie come insieme di supporto iniziale il negato della clausola goal**. Il risultato è che è come procedere all'indietro dal goal. Strategia completa, per la refutazione.



9.1.3 Strategie di Ordinamento

Risoluzione ordinata

Ogni clausola è un **insieme ordinato** di letterali e **si possono unificare solo i primi letterali** delle clausole. L'ordinamento deve essere rispettato nel risolvente.



La risoluzione ordinata è completa per le clausole Horn

9.1.4 Sottoinsieme a regole del FOL

Clausole di Horn definite Clausole con **esattamente** un letterale positivo.
Possano essere riscritte come fatti e regole:

$$\neg P_1 \vee \dots \vee \neg P_k \vee Q$$

$$\neg(P_1 \wedge \dots \wedge P_k) \vee Q$$

Una **KB a regole**

Regola: $P_1 \wedge \dots \wedge P_k \Rightarrow Q$

Fatto: Q

9.1.5 Sistemi a regole logici

Se la KB contiene solo clausole Horn definite, i meccanismi inferenziali sono molto più semplici ed il processo è molto più guidato, senza rinunciare alla completezza: risolutori in tempo lineare per il caso proposizionale. **Nota:** è restrittivo e non coincide né con il PROP né con il FOL.

Uso delle regole in avanti e indietro

Backward Chaining: un'istanza di ragionamento **guidato dall'obiettivo**.

Le regole sono applicate alla rovescia: **programmazione logica** (PROLOG ha le KB a regole e procede applicando le regole in avanti o all'indietro)

Forward Chaining: un'istanza di ragionamento o ricerca **guidato dai dati**

Le regole sono applicate nel senso "antecedente-conseguente". Basi di dati deduttive e sistemi di produzione.

9.1.6 Programmazione Logica

I **programmi logici** sono KB costituite di clausole Horn definite, espressi come fatti e regole con sintassi alternativa:
 $A : -B_1, B_2, \dots, B_n$ (**A testa, B₁, ..., B_n corpo**)

A vero se sono veri B_1, \dots, B_n , in accordo al significato logico di implicazione. Questa è la **interpretazione dichiarativa**.

Interpretazione procedurale: la testa può essere vista come una **chiamata di procedura** e il corpo come una serie di procedure da eseguire in sequenza.

Altre convenzioni: in PL le variabili sono indicate con le lettere maiuscole e le costanti con le lettere minuscole

9.1.7 Risoluzione SLD

Selection Linear Definit-Clauses Strategia lineare nell'input e ordinata, basata su un insieme di supporto (la clausola goal).

SLD è completa per clausole Horn.

Alberi di Risoluzione SLD Dato un programma logico P, l'albero SLD per un goal G è definito come segue:

Ogni nodo dell'albero corrisponde ad un goal (**congiuntivo**)

La radice è $\text{:- } G_1, G_2, \dots, G_k$, il nostro goal

Sia $\text{:- } G_1, G_2, \dots, G_k$ un nodo dell'albero: il **nodo ha tanti discendenti quanti sono i fatti e le regole in P la cui testa è unificabile con G_1**

Se $A := B_1, \dots, B_k \wedge A$ unificabile con $G_1 \wedge \gamma = \text{MGU}(A, G_1) \Rightarrow$ un discendente è il goal $\text{:- } (B_1, \dots, B_k, G_2, \dots, G_k)\gamma$

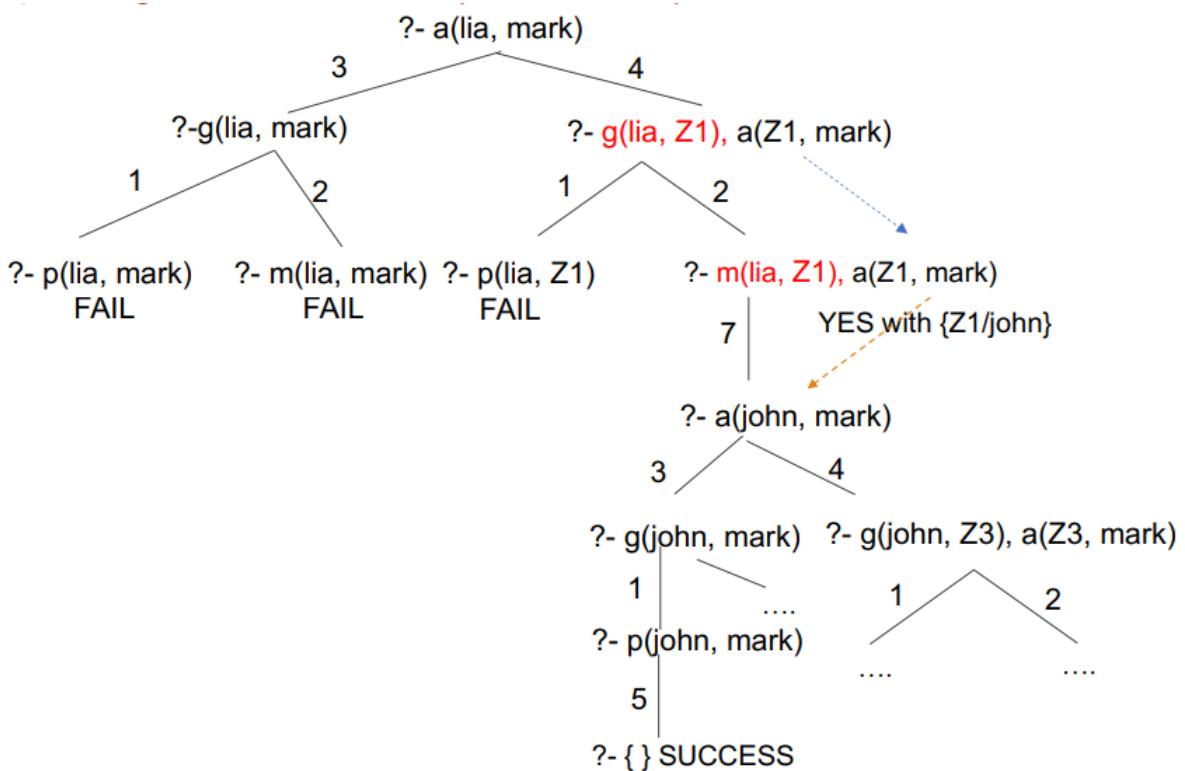
I nodi che sono **clausole vuote** sono **successi**

I nodi che non hanno successori sono **fallimenti**

Esempio

1. Genitore(X, Y) :- Padre(X, Y)
2. Genitore(X, Y) :- Madre(X, Y)
3. Antenato(X, Y) :- Genitore(X, Y)
4. Antenato(X, Y) :- Genitore(X, Z), Antenato(Z, Y)
5. Padre(Gio, Mark)
6. Padre(Gio, Luc)
7. Madre(Lia, Gio)

8. $\text{:- } \text{Antenato}(\text{Lia}, \text{Mark})$ goal negato



Risoluzione SLD Strategia completa per clausole Horn definite.

Se $P \cup \{\neg G\}$ è insoddisfacibile, allora una delle foglie deve essere la clausola vuota (successo).

Non è restrittivo andare in ordine nel risolvere i sottogoal in \wedge . La sostituzione corrispondente è la **risposta calcolata**.

Strategia di visita dell'albero SLD e PROLOG A seconda di come viene visitato l'albero, la clausola vuota potrebbe non essere trovata: la strategia di ricerca può essere la responsabile dell'incompletezza.

In Prolog la visita dell'albero di risoluzione avviene con una ricerca in profondità e backtracking su fallimento, quindi non è una strategia completa. Il programmatore ha la possibilità di controllare l'ordine di generazione perché le regole sono applicate nell'ordine in cui sono scritte nel file.

9.2 Sistemi a Regole in Avanti

Modus Ponens Generalizzato

$$\frac{p'_1 p'_2 \dots p'_n (p_1 \wedge p_2 \wedge \dots \wedge p_n \Rightarrow q)}{(q) \Theta}$$

Dove $\Theta = \text{MGU}(p'_i, p_i) \forall i$

Regola **corretta**:

Si istanziano gli universali

Si istanziano le regole

Si applica il Modus Ponens classico

Più in generale di MP, ma anche più limitata nella forma del conseguente

Esempio Supponiamo di avere nella KB:

King(John)

Greedy(y)

King(x) \wedge Greedy(x) \Rightarrow Evil(x)

Con $\Theta = \{x/John, y/John\}$ si ottiene

King(John), Greedy(John), King(John) \wedge Greedy(John) \Rightarrow Evil(John)

Quindi la conclusione della regola è

Evil(John)

Concatenazione in avanti Un semplice processo inferenziale (FOL FC Ask) **applica ripetutamente il Modus Ponens generalizzato** per ottenere nuovi fatti, fino a che:

Si dimostra quello che si desidera

Nessun fatto nuovo può essere raggiunto

Strategia di **ricerca sistematica in ampiezza**

Esempio "È un crimine per un Americano vendere armi a una nazione ostile. Il paese Nono, un nemico dell'America, ha dei missili, e tutti i missili gli sono stati venduti dal colonnello West, un Americano."

Dimostrare che West è un criminale. **Formalizzazione**:

1. Americano(x) \wedge Arma(y) \wedge Vende(x, y, z) \wedge Ostile(z) \Rightarrow Criminale(x)
2. $\exists x \text{ Possiede}(\text{Nono}, x) \wedge \text{Missile}(x)$
 $\text{Possiede}(\text{Nono}, M_1) \wedge \text{Missile}(M_1)$
3. Missile(x) \wedge Possiede(Nono, x) \Rightarrow Vende(West, x, Nono)
4. Missile(x) \Rightarrow Arma(x)

5. Nemico(x, America) \Rightarrow Ostile(x)
6. Americano(West)
7. Nemico(Nono, America)

In questo caso non ci sono funzioni ed il processo converge: siamo nelle condizioni di un database Datalog (database deduttivi). **Prima iterazione**

1. Possiede(Nono, M₁) \wedge Missile(M₁)
2. Missile(x) \wedge Possiede(Nono, x) \Rightarrow Vende(West, x, Nono)
La regola 3 è soddisfatta con {x/M₁} e viene aggiunto Vende(West, M₁, Nono)
3. Missile(x) \Rightarrow Arma(x)
La regola 4 è soddisfatta con {x/M₁} e viene aggiunto Arma(M₁)
4. Nemico(x, America) \Rightarrow Ostile(x)
5. Nemico(Nono, America)
La regola 5 è soddisfatta con {x/Nono} e viene aggiunto Ostile(Nono)

Seconda iterazione

1. Americano(x) \wedge Arma(y) \wedge Vende(x, y, z) \wedge Ostile(z) \Rightarrow Criminale(x)
La regola 1 è soddisfatta con {x/West, y/M₁, z/Nono} e Criminale(West) viene aggiunto



9.2.1 Analisi di FOL-FC-Ask

Corretta perché il Modus Ponens Generalizzato è corretto.

Completa per KB di clausole Horn definite:

Completa e convergente per calcolo proposizionale e per KB di tipo Datalog (senza funzioni) perché la chiusura deduttiva è un insieme finito

Completa anche con funzioni ma il processo potrebbe non terminare (semidecidibile)

Il metodo descritto è **sistematico ma non troppo efficiente**.

9.2.2 FC Efficiente

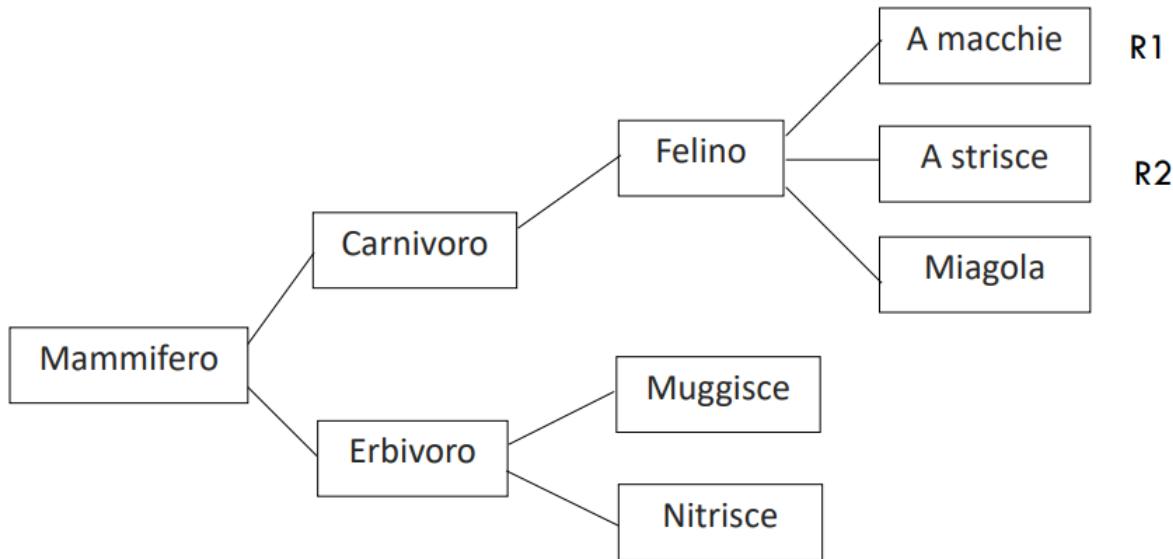
Ordinamento dei congiunti Conviene **soddisfare prima i congiunti con meno istanze nella KB** e che compaiono in più regole: $\text{Missile}(x) \wedge \text{Possiede}(\text{Nono}, x) \Rightarrow \text{Vende}(\text{West}, x, \text{Nono})$ (Tipi di missile << cose possedute) Altre ottimizzazioni sono mutuate dai CSP.

Rete di discriminazione Assunzione: regole diverse possono condividere molte delle precondizioni. Ad esempio:

R1 (Mammifero x) (Felino x) (Carnivoro x)
 (A-Macchie x) \Rightarrow (assert (Leopardo x))

R2 (Mammifero x) (Felino x) (Carnivoro x)
 (A-Strisce x) \Rightarrow (assert (Tigre x))

Idea: codificare gli antecedenti delle regole sotto forma di rete di discriminazione.



Incrementale Ogni nuovo fatto inferito al tempo t deve essere dedotto usando almeno un fatto dedotto al tempo $t - 1$

Si possono guardare solo le regole che hanno premesse unificabili con fatti aggiunti nell'ultima iterazione.

Individuare le regole sui fatti, evitare di ricalcolare le unificazioni...

Altre ottimizzazioni presenti nell'algoritmo RETE...

Ridurre deduzioni irrilevanti Un modo per evitare di ricavare fatti irrilevanti: **lavorando all'indietro dal goal**, non c'è questo problema. Si fa una specie di **pre-processing** per individuare le regole che servono, procedendo all'indietro dal goal e **marcando le regole utili**. Poi si procede in avanti **utilizzando solo le regole marcate**.

Magic Set Dato il goal: Criminal(West), creo una KB \leftarrow KB \cup {Magic(West)}. **Riscrittura delle regole:**

Magic(x) \wedge Americano(x) \wedge Arma(y) \wedge Vende(x, y, z) \wedge Ostile(z) \Rightarrow Criminale(x)

Procedendo poi in avanti saranno utilizzate solo le "regole magiche" in modo mirato.

Combina BC con FC.

Capitolo 10

Machine Learning

10.1 Introduzione al Machine Learning

Apprendimento Principio universale che riguarda gli esseri viventi, al cuore del problema dell'intelligenza sia biologica che artificiale.

L'apprendimento è una sfida strategica per fornire **intelligenza** ai sistemi.

Si tratta di un problema complesso, campo che cresce continuamente e riguarda aspetti teorici e applicativi: apprendimento automatico / **machine learning**.

Nasce con lo scopo di combinare le ambizioni di creare macchine che possono apprendere con sistemi statistici potenti **utilizzabili in vari ambiti** con approfondimento rigoroso nella scienza computazionale.

Macchine che imparano da sé Perché? Lusso o necessità?

- Crescente disponibilità e bisogno di analisi di dati empirici
- Ruolo centrale e metodologico per il **cambio di paradigma della scienza: data-driven**
- Difficile fornire intelligenza attraverso la programmazione (vedi Turing).
Ha molto più senso lasciare che le macchine apprendano da sole, *non si può insegnare loro tutto*
- L'apprendimento è l'unica via...

Dati + HPC + ML moderno → strada verso la nuova era delle IA **Obiettivi**

Costruire sistemi intelligenti adattivi

Data Analysis per costruire sistemi predittivi potenti, strumenti per i data scientist

Modelli come strumenti per risolvere problemi complessi ed interdisciplinari

Apprendimento automatico di un sistema dell'esperienza (serie di esempi) per affrontare un task computazionale.
Partire da esempi

Esempio Classificare lo spam: collezionare 100 mail di spam, 100 non di spam e le uso come esempi che fornisco al sistema: esso apprende dagli esempi per poter essere capace di classificare in futuro spam e non.

Vari utilizzi Robotica, comprensione linguaggio naturale, data mining, sensori, componenti personalizzati... è **pervasivo**.

Quando va applicato? Strumento molto potente, ma va capito quando applicarlo: **ha dei limiti**. Utile l'apprendimento predittivo quando non esiste, o è scarsa o è difficile da formalizzare, la teoria attorno ad un problema. Oppure anche quando i dati sono incerti, rumorosi o incompleti.

Altro caso sono i **componenti personalizzati (ambienti dinamici)**: quei casi in cui il comportamento è specifico per persona (per me è spam, per un'altra persona no). Non si può completamente definire il comportamento a priori. Le **richieste** per il ML sono: fonte di esperienza di apprendimento (dati rappresentativi) e tolleranza alla precisione dei risultati.

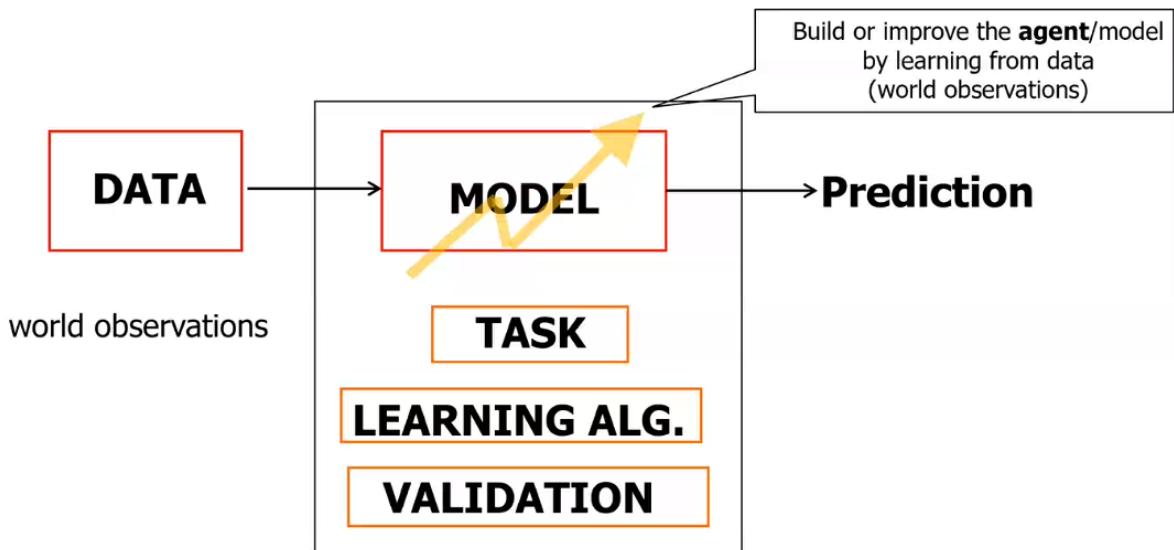
Perché Machine Learning? Si tratta di un'opportunità per imparare nuovi paradigmi computazionali, con un approccio differente rispetto alla programmazione standard ed agli algoritmi classici: trattamento dell'incertezza, tolleranza all'imprecisione...

Tipico dell'area del **soft computing/intelligenza computazionale**.

Per trovare **soluzioni approssimate** di problemi complessi, difficili da formalizzare da algoritmi "fatti a mano".

Per costruire sistemi intelligenti nuovi, robusti e estensivamente applicabili.

Ma non è una metodologia approssimativa! Si tratta di un **approccio rigoroso** atto a **trovare funzioni approssimative** con cui affrontare **problemi complessi** (supportati da risultati empirici e teorici, es. SLT)



Apprendimento L'apprendimento vogliamo vederlo come un'**approssimazione di una funzione non nota attraverso degli esempi**.

Facciamo un esempio intuitivo: il riconoscimento caratteri scritti a mano. L'input è la collezione di immagini (array/matrici di valori rappresentante i colori/livello di grigio) e l'obiettivo è costruire un modello che ne assegna il valore rappresentato.



Ha tutte le **caratteristiche elencate prima**: difficile da formalizzare, presenza di rumore nei dati e di dati ambigui. Ma è molto facile collezionare degli esempi etichettati. ⇒ esempio di applicazione ottima del ML.

Generalizzo il problema. **Apprendimento supervisionato** (classificazione e regressione): supervisore che etichetta gli esempi. Spazio di input: $x \rightarrow f$ categorie o valori reali. Costruisco una funzione attraverso gli esempi.

Supervised Learning

Abbiamo delle coppie date da qualcuno (supervisore) $\langle \text{input } x, \text{output } d \rangle$ per una funzione non nota f

Target Value: valore desiderato d dato dal supervisore in relazione a $f(x)$

Vogliamo trovare una buona approssimazione di f : un'ipotesi h che può essere usata per prevedere un dato nuovo x'

Target: etichetta numerica/categorica

Classificazione: $f(x)$ ritorna la corretta classe di x ipotizzata

Regressione: approssima una funzione obiettivo a valori reali

Sono entrambi **problemi di approssimazione di funzione**, cambia solo il codominio che nel primo caso è a valori discreti.

Inferire funzioni generali da dati noti

Handwriting Recognition

- x : Data from images of the characters.
- $f(x)$: Letter of the alphabet.

Disease diagnosis (from database of past medical records)

- x : Properties of patient (symptoms, lab tests)
- $f(x)$: Disease (or maybe, recommended therapy)
- TR $\langle x, f(x) \rangle$: database of past medical records

Face recognition

- x : Bitmap picture of person's face
- $f(x)$: Name of the person.

Spam Detection

- x : Email message
- $f(x)$: Spam or not spam.

Approccio Non-Supervisionato Non c'è il supervisore: serie di esempi non etichettati, ad esempio per **trovare raggruppamenti naturali dei dati**.

Modello Ha lo scopo di **catturare e descrivere relazioni fra i dati**. Definisce la classe delle funzioni che il sistema di apprendimento può implementare, cioè lo **spazio delle ipotesi**.

Esempi di apprendimento Esempi della forma $(x, f(x))$, con x solitamente vettori di valori, $f(x) = t$ è il **valore target**. La vera funzione f è la funzione target.

Ipotesi La proponiamo noi, h espressione in un qualche **linguaggio** che si crede simile ad f ignota.

Spazio delle ipotesi L'insieme di tutte le ipotesi che possono essere soluzione

Linguaggi per l' h Logica del prim'ordine, equazioni numeriche, ma anche probabilità... **Esempi**

Modelli lineari: la rappresentazione dell' h definisce un spazio parametrico continuo di potenziali ipotesi. Ogni assegnamento w è un'ipotesi differente, ad esempio $h_w(x) = w_1x + w_0 \rightarrow h_w(x) = 0.232x + 246$

Regole simboliche: lo spazio delle ipotesi è basato su rappresentazioni discrete. Sono possibili regole differenti, ad esempio

```
if (x1 = 0 && x2 = 1) then h(x) = 1
else h(x) = 0
```

Modelli probabilistici: stima di $p(x, y)$

Approcci basati su istanze: prevedere il valore medio di y basandosi sui vicini

Algoritmi di apprendimento Ci riporta al concetto dei problemi di ricerca: **ricerca euristica nello spazio delle ipotesi**.

L'apprendimento è quindi la ricerca di una *buona* funzione. Cosa significa buona: **quando ha una buona capacità di generalizzazione**, quindi ha un **basso errore di generalizzazione**. Cioè **alta accuratezza su dati nuovi**.

Capacità di generalizzazione: apprendere dai dati per **applicare in generale**. Altrimenti stiamo costruendo un db.

Learning phase: costruisco il modello da dati noti

Predictive phase (test): applico a nuovi esempi. Valutazione dell'ipotesi predittiva

Performance in ML = accuratezza predittiva stimata dall'errore calcolato sul **test set**.

Teoria Sotto quali condizioni matematiche un modello è in grado di generalizzare? Statistical Learning Theory.

10.2 Concept Learning

Concept Learning Inferire una funzione booleana da esempi d'addestramento positivi e negativi. X spazio delle istanze, C: $X \rightarrow \{\text{tt}, \text{ff}\}$ oppure $\{+, -\}$ oppure $\{0, 1\} \dots$

Si apprende dagli esempi come una ricerca nello spazio delle ipotesi. **Imparare è migliorare, tramite l'esperienza, in qualche attività.**

Migliorare sul task T

rispetto a delle misure di performance P

tramite l'esperienza E

10.2.1 Supervised Learning

Vengono forniti una serie di **esempi** $\langle \text{input}, \text{output} \rangle = \langle x, d \rangle$ per una qualche **funzione sconosciuta f**.

Il **valore obiettivo d** è dato dal **supervisore** in accordo a $f(x)$. L'**obiettivo è trovare una buona approssimazione di f**.

Esempio $\langle x, c(x) \rangle \in D$ (o training set, TR set)

Soddisfa $h : X \rightarrow \{0, 1\}$ **soddisfa** x se $h(x) = 1$

Consistenza Un'ipotesi è consistente con l'intero D se coincide con i valori target forniti per ogni esempio. Cioè se $\forall \langle x, c(x) \rangle \in D \Rightarrow h(x) = c(x)$

L'obiettivo è **ricostruire** quell'*unknown function*. Questo è un **problema inverso (ill posed, o malposto)**, perché può violare l'esistenza, l'unicità o la stabilità della funzione.

Nel caso generale, $|h| = 2^{\text{num istanze}} = 2^{2^n}$ per input binari, con n = dimensione dell'input

10.2.2 Regole congiuntive

Proposizioni fatte con l'AND. Quante regole congiuntive ci sono?

Nel caso generale, ho i **letterali positivi** (esempi: $h_1 = l_1$, $h_2 = (l_1 \wedge l_2)$, $h_3 = \text{true} \dots$). Ognuno lo posso mettere o non mettere all'interno della regola congiuntiva, quindi tutti i modi possibili sono 2^n , se includo anche il \neg diventano $3^n + 1$

Esempio: Enjoy Sport Idea: "giorni in cui il mio amico Aldo preferisce praticare sport".

Task: predire il valore di "enjoy sport" per un giorno arbitrario basandosi sul valore di una serie di attributi

Sky	Temp.	Humid.	Wind	Water	Forecast	Enjoy Sport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Una riga rappresenta un'istanza

10.2.3 Rappresentare le ipotesi

Un'ipotesi h è una **congiunzione di vincoli sugli attributi**. Ogni vincolo può essere:

Un valore specifico, es $Water = Warm$

Un valore ininfluente, es $Water = ?$

Un'ipotesi nulla, nessun valore concesso, es: $Water = \emptyset$, $l_i \wedge \neg l_j$

Esempio di ipotesi h

Sky	Temp.	Humid.	Wind	Water	Forecast
{Sunny}	?	?	Strong	?	Same

Corrispondente alla funzione booleana

$$\text{Sky} = \text{Sunny} \wedge \text{Wind} = \text{Strong} \wedge \text{Forecast} = \text{Same}$$

Ipotesi più specifica $\langle \emptyset \emptyset \emptyset \emptyset \emptyset \emptyset \rangle$

Ipotesi più generale $\langle ? ? ? ? ? ? \rangle$

Learning hypothesis Dati

Istanze X: giorni possibili descritti dagli attributi Sky, Temp, Humid, Wind, Water, Forecast

Funzione Target c: Enjoy Sport $X \rightarrow \{0, 1\}$

Ipotesi H: insieme di ipotesi, che sono congiunzioni di un'insieme finito di letterali.
Es $\langle \text{Sunny} ? ? \text{Strong} ? \text{Same} \rangle$

Esempi d'apprendimento D: esempi positivi e negativi della funzione target.
Es $\langle x_1, c(x_1) \rangle, \dots, \langle x_n, c(x_n) \rangle$

Trovare un'ipotesi $h \in H$ tale che $\forall x \in X, h(x) = c(x)$

Quindi **apprendere = cercare** nello spazio delle ipotesi H .

Assunzione Ogni ipotesi che approssima bene la funzione obiettivo sugli esempi d'apprendimento, approssimerà la funzione obiettivo anche su esempi non osservati.

$$h(x) = c(x) \quad \forall x \in D \quad (\text{cioè } \text{consistente con } D)$$

$$h(x) = c(x) \quad \forall x \in X ?$$

Il problema fondamentale del Machine Learning

10.2.4 Numerare le istanze, concetti, ipotesi

La rappresentazione scelta per H determina lo spazio di ricerca:

Sky: Sunny, Cloudy, Rainy (3 valori)

Temp: Warm, Cold

Humid: Normal, High

Wind: Strong, Weak

Water: Warm, Cold

Forecast: Same, Change

$$\# \text{ istanze distinte} = 3 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 96$$

$$\# \text{ concetti distinti} = 2^{\# \text{ istanze}} = 2^{96}$$

$$\# \text{ ipotesi sintatticamente distinte (congiunzioni)} = 5 \cdot 4 \cdot 4 \cdot 4 \cdot 4 \cdot 4 = 5120 \text{ (es: warm/cold/?/\emptyset)}$$

$$\# \text{ ipotesi semanticamente distinte} = 1 + 4 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 973 \text{ perché tutte le } h \text{ con } \emptyset \text{ equivalgono a } false.$$

In generale la dimensione è molto grande, può addirittura essere infinita. Strutturare bene lo spazio di ricerca H può aiutare drasticamente nel ricercare in maniera efficiente.

Specificare l'ordinamento Considerate due funzioni a valori booleani h_j e h_k definite su X , allora h_j è più generale di o uguale a h_k (scritto $h_j \geq h_k$) $\Leftrightarrow \forall x \in X : [(h_k(x) = 1) \Rightarrow (h_j(x) = 1)]$

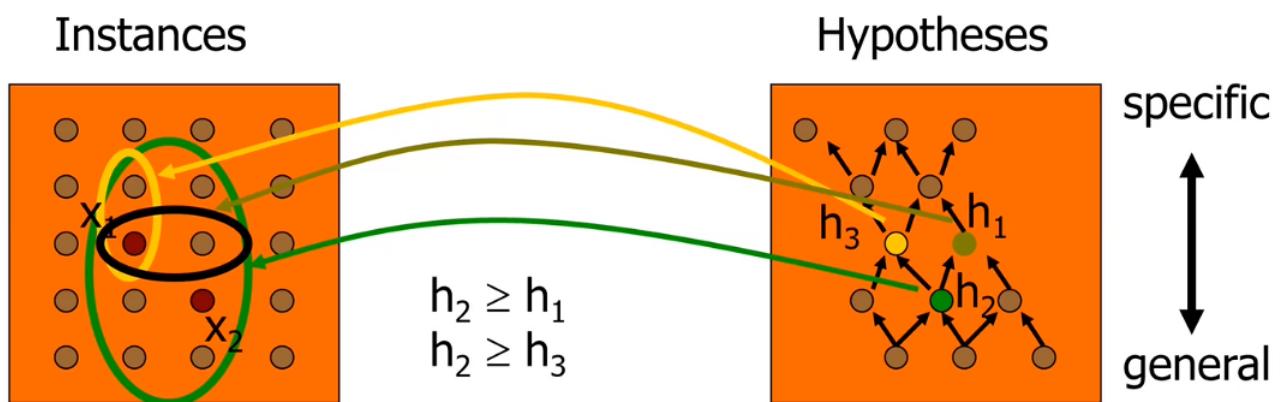
Un esempio:

$$h_1 = \langle \text{Sunny}, ?, ?, \text{Strong}, ?, ? \rangle$$

$$h_2 = \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle$$

h_2 impone meno vincoli rispetto ad h_1 , e quindi classifica un maggior numero di istanze x come positive $h(x) = 1$

Questo impone un **ordinamento parziale** sullo spazio di ipotesi H che viene utilizzato da molti metodi di concept learning. Possiamo usare questo ordinamento parziale a nostro vantaggio per organizzare efficientemente la ricerca in H .



L'idea è di fare una ricerca in un ordine intelligente. Parto dall'ipotesi più specifica e mi sposto verso quella più generale, guardando le istanze che arrivano dalla tabella una ad una. Se l'ipotesi non soddisfa, cerco di generalizzarla (renderla vera) ma col passo minore possibile.

10.2.5 Algoritmo Find-S

Si cerca di usare l'ordinamento parziale per cercare efficientemente le h consistenti, senza enumerare esplicitamente ogni $h \in H$.

1. Inizializziamo h con l'ipotesi più specifica in H
2. Per ciascuna istanza positiva, guardiamo gli attributi dell'ipotesi.
Se gli attributi sono soddisfatti dall'istanza nuova allora ok.
Altrimenti bisogna cambiare gli attributi per generalizzare l'ipotesi e portarla ad 1, ma facendo il minimo possibile. In pseudocodice:

```
for each attributo  $a_i \in h$ 
    if  $a_i$  è soddisfatto da  $x$  then do nothing
    else sostituisci  $a_i$  in  $h$  con il prossimo vincolo più generale che è soddisfatto da  $x$  (cioè rimuovi da  $h$  i letterali che non soddisfano  $x$ )
```

3. Output: ipotesi h

Proprietà

Lo spazio delle ipotesi è descritto da congiunzioni di attributi (limitante!)

Find-S produce l'ipotesi più specifica in H consistente con gli esempi d'allenamento **positivi**

L'ipotesi in output h sarà anche consistente con gli esempi negativi, se il concetto obiettivo è contenuto in H , perché $c \geq h$.

Limitazioni del Find-S

Non c'è tolleranza al rumore: ignorando gli esempi negativi non c'è modo di sapere se gli esempi d'apprendimento sono inconsistenti.

Non si può dire se chi apprende converge verso il concept obiettivo, nel senso che non può determinare se ha trovato l'*unica* ipotesi consistente con gli esempi.

Perché preferire l'ipotesi più specifica?

Cosa succede se ci sono più ipotesi massimalmente specifiche

Version Spaces

Idea Come output una descrizione dell'insieme di *tutte* le h consistenti con D .

Possiamo farlo senza enumerarle tutte

$$\text{Consistent}(h, D) := \forall \langle x, c(x) \rangle \in D \ h(x) = c(x)$$

Definizione Il **version space** $VS_{H,D}$, rispetto allo spazio delle ipotesi H e al training set D , è il **sottoinsieme delle ipotesi di H consistenti con tutti gli esempi d'apprendimento**

$$VS_{H,D} = \{h \in H \mid \text{Consistent}(h, D)\}$$

10.2.6 Algoritmo List-Then-Eliminate

Irrealistico perché richiede enumerazione esaustiva di tutte le ipotesi di h in H

1. *VersionSpace*: una lista contenente tutte le ipotesi di H
2. \forall esempio d'apprendimento $\langle x, c(x) \rangle$

rimuoviamo da *VersionSpace* tutte le ipotesi inconsistenti con l'esempio d'apprendimento, cioè con $h(x) \neq c(x)$

3. Output: la lista di ipotesi in *VersionSpace*

Ma vogliamo sfruttare ordinamento parziale

10.2.7 Rappresentare i Version Spaces

Definizioni

General boundary G di $VS_{H,D}$, è l'insieme dei membri (di H consistenti con D) **massimamente generici** (*maximally general*)

Specific boundary S di $VS_{H,D}$, è l'insieme dei membri (di H consistenti con D) **massimamente specifici** (*maximally specific*)

Teorema: ogni membro del VS sta tra questi due confini

$$VS_{H,D} = \{h \in H \mid \exists s \in S, \exists g \in G \Rightarrow g \geq h \geq s\}$$

Dove con $x \geq y$ si intende che x è più generale o uguale a y .

10.2.8 Algoritmo Candidate Elimination

Pseudocodice Dati G (insieme delle ipotesi massimamente generiche) e S (insieme delle ipotesi massimamente specifiche).

Per ogni esempio d'addestramento $d = \langle x, c(x) \rangle$

Se d è un esempio **positivo**

Rimuovi da G ogni ipotesi inconsistente con d (def. di VS)

Per ogni ipotesi $s \in S$ inconsistente con $d \rightarrow$ **Generalizza S**

Rimuovi s da S

Aggiungi a S tutte le generalizzazioni minime h di s tali che

- h è consistente con d
- $\exists g \in G$ più generale di h

Rimuovi da S tutte le ipotesi che sono più generali di un'altra ipotesi in S

Se d è un esempio **negativo**

Rimuovi da S tutte le ipotesi inconsistenti con d ($h = 1$)

Per ogni ipotesi $g \in G$ inconsistente con $d \rightarrow$ **Specializza G**

Rimuovi g da G

Aggiungi a G tutte le specializzazioni minime h di g tali che

- h è consistente con d
- $\exists s \in S$ più specifico di h

Rimuovi da G tutte le ipotesi che sono meno generali di un'altra ipotesi in G

Grazie all'ordinamento parziale posso controllare di non avere cose più generali/specifiche e mantenere la consistenza.

10.3 Bias Induttivo ed il suo ruolo

Per ora abbiamo usato uno spazio delle ipotesi **biased**, vincolato. Abbiamo assunto che lo spazio delle ipotesi contenesse C , cioè che l'obiettivo C fosse espresso in \wedge e non in \vee . Quindi il nostro H non è in grado di rappresentare un concetto target disgiuntivo.

Supponiamo sistema di apprendimento unbiased, con H che esprime ogni concetto insegnabile: H è l'insieme di tutti i possibili sottoinsiemi di X : $\mathcal{P}(X)$.

Se $|X| = 96$, $|\mathcal{P}(X)| = 2^{96}$ cioè circa 10^{28} .

Cosa succede nella generalizzazione?

Definizione Un **unbiased learner** è impossibilitato a generalizzare: ogni istanza non osservata, viene classificata come positiva da precisamente metà delle ipotesi nel version space e come negativa dall'altra metà. $\forall h$ consistente con x_i test $\Rightarrow \exists h'$ identica ad h eccetto che $h'(x_i) \neq h(x_i)$
 $h \in VS \Rightarrow h' \in VS$ (identiche su D).

Inutilità dell'apprendimento senza bias Un learner che non fa assunzioni a priori riguardo l'identità del concetto target non ha basi razionali per classificare qualsiasi istanza non osservata.

Il bias non è solo assunto per efficienza ma anche **necessario per la generalizzazione**.

Un semplice esempio:

$$\langle x, c(x) \rangle, H = \{x, \neg x, 0, 1\}$$

$$TR \langle 0, 0 \rangle, VS = \{x, 0\}$$

TS $\langle 1, ? \rangle \rightarrow$ può essere 1 o 0... a meno che non si siano usate tutte le x come TR

Bias Induttivo Considero:

Un algoritmo di concept learning L

Istanze X, target concept C

Esempi d'apprendimento $D_C = \{\langle x, c(x) \rangle\}$

$L(x_i, D_C)$ denota la classificazione assegnata all'istanza x_i da L dopo l'apprendimento su D_C

Il **bias induttivo** di L è un set minimo di asserzioni B tale che per ogni concetto obiettivo C e dati di addestramento corrispondenti D_C ottengo

$$\forall x_i \in X [B \wedge D_C \wedge x_i] \vdash L(x_i, D_C)$$

Dove con $A \vdash B$ indico che A è conseguenza logica di B

10.4 Sistemi Induttivi e Sistemi Deduttivi Equivalenti



Tre learner con bias differenti

Lookup table: mantiene gli esempi, classifica x se e solo se corrisponde ad un esempio precedentemente osservato.

Non ha bias induttivo \Rightarrow non ha generalizzazione

Version space candidate elimination algorithm, il cui bias è: lo spazio di ipotesi contiene il target concept (cioè è una congiunzione di attributi)

$$|H| = 973 \text{ vs } 10^{128}$$

Find-S, il cui bias: lo spazio delle ipotesi contiene il target concept e tutte le istanze sono istanze negative a meno che l'opposto non sia dedotto da altra conoscenza (vista come esempio positivo)

10.5 Modelli Lineari

Ingredienti

Dati di allenamento

Spazio delle ipotesi H

- Costituisce l'**insieme delle funzioni che possono essere realizzate dal sistema di apprendimento**
- Si assume che la funzione da apprendere f possa essere rappresentata da una ipotesi $h \in H$
- Si seleziona h attraverso i dati di apprendimento
- Oppure si assume che almeno una delle ipotesi $h \in H$ sia *simile* ad f (**approssimazione**)

Algoritmo di apprendimento, cioè un algoritmo di **ricerca nello spazio delle ipotesi**

Ad esempio: adattamento dei parametri liberi del modello al task

Nota: H non può coincidere con l'insieme di tutte le funzioni possibili e la ricerca potrebbe essere esaustiva (**Bias Induttivo**)

Compito supervisionato Un task

Dati

Esempi d'apprendimento come $\langle x, d \rangle = \langle \text{input}, \text{output} \rangle$ per una qualche funzione sconosciuta f
Target value: il valore desiderato d è dato dal supervisore a seconda di $f(x)$

Trovare

Una *buona* approssimazione di f

Cioè un'ipotesi h che può essere usata per previsioni su dati non osservati x'

Target d : un'etichetta numerica/categorica

Classificazione: $f(x)$ ritorna la corretta classe (ipotizzata) per $x \Rightarrow f(x)$ è una funzione a valori discreti

Regressione: approssimare una funzione obiettivo a valori reali (in R o R^n)

Entrambe sono task di approssimazione di funzioni

10.5.1 Problemi di Regressione

Processo di stima di una funzione a valori reali basandosi su un'insieme finito di **esempi rumorosi**: coppie $\langle x, f(x) + \text{rumore casuale} \rangle$.

Un esempio è una tabella $x \mapsto$ valore come la seguente

x	target
1	2.1
2	3.9
3	6.1
4	8.4
5	9.8

Si può ipotizzare che $f(x) = 2x$, a mente: abbiamo applicato un algoritmo di machine learning assumendo che la funzione sia lineare del tipo $y = w_1x + w_0$

Definizione Assumiamo un modello $h_w(x)$ espresso come $y = w_1x + w_0$ Trovare, a partire dall'esempio, un'ipotesi h che esegue il **fitting** dei dati dai dati osservati x e y al meglio possibile.

Assumendo che y sia legata linearmente ad un'altra variabile x o altre variabili $y = w_1x + w_0 + \text{rumore}$, dove le w sono libere e *rumore* è l'errore misurato negli obiettivi.

Costruiamo il modello cercando di trovare i valori di w per predire e stimare y per altri valori di x **non osservati**.

Costruire il modello

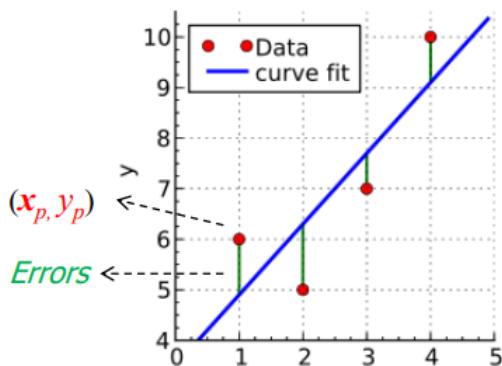
Apprendimento tramite LMS Dato l'insieme di l esempi di training $\langle x_p, y_p \rangle$, trovare $h_w(x)$ nella forma $h_w(x) = w_1x + w_0$ che **minimizza l'errore atteso sui dati d'apprendimento**.

Per l'errore utilizziamo la **least mean square (LMS)**:

$$\text{loss}(h_w) = E(w) = \sum_{p=1}^l (y_p - h_w(x_p))^2 = \sum_{p=1}^l (y_p - (w_1x_p + w_0))^2$$

con valori al quadrato per non avere negativi.

Perché? Perché LMS per il migliore h ?



Minimizzo la distanza (linee verdi) dalla media (linea blu). Approccio standard per approssimare la soluzione di sistemi **sovra-determinati**, cioè con più equazioni che incognite.

Diverse medie avranno diversi errori. Minimizzare gli errori è un buon modo per trovare la migliore approssimazione/fitting dei dati (cioè il nostro $h_w(x)$ o la linea blu). $E(w)$ quantifica le linee verdi.

Come risolvere Un **punto di minimo** è un **punto stazionario**, cioè dove il gradiente è nullo. Bisogna cercare fra quelli a gradiente nullo

$$\frac{\partial E(w)}{\partial w_i} = 0$$

quindi per la regressione lineare (a due parametri)

$$\frac{\partial E(w)}{\partial w_0} = 0 \quad \frac{\partial E(w)}{\partial w_1} = 0$$

Come impostare il gradiente (per ogni pattern p)

Hence we omit p for x

Basic rules:

$$\begin{aligned} \frac{\partial}{\partial w} k &= 0, \quad \frac{\partial}{\partial w} w = 1, \quad \frac{\partial}{\partial w} w^2 = 2w \\ \frac{\partial(f(w))^2}{\partial w} &= 2f(w) \frac{\partial(f(w))}{\partial w} \end{aligned}$$

Der. sum = sum of der.

$$\begin{aligned} \frac{\partial E(\mathbf{w})}{\partial w_i} &= \frac{\partial(y - h_{\mathbf{w}}(x))^2}{\partial w_i} = \\ &= 2(y - h_{\mathbf{w}}(x)) \frac{\partial(y - h_{\mathbf{w}}(x))}{\partial w_i} = 2(y - h_{\mathbf{w}}(x)) \frac{\partial(y - (w_1x + w_0))}{\partial w_i} \end{aligned}$$

$$\frac{\partial E(\mathbf{w})}{\partial w_0} = -2(y - h_{\mathbf{w}}(x))$$

$$\frac{\partial E(\mathbf{w})}{\partial w_1} = -2(y - h_{\mathbf{w}}(x)) \cdot x$$

Derivata della somma = somma delle derivate.
Per $\frac{\partial E(w)}{\partial w_1} = \frac{\partial y}{\partial w_1} - \left(\frac{\partial w_1 x}{\partial w_1} + \frac{\partial w_0}{\partial w_1} \right) = 0 - (x + 0)$

Riassumendo Dato un insieme di l esempi d'apprendimento come coppie $\langle x_p, y_p \rangle$, trovare i valori di w costruendo $h_w(x)$ espressa come $w_1x + w_0$.

Trovata $h_w(x)$, si può usare per esprimere nuovi valori. La linea diretta è meno interessante, preferiamo il gradiente per l'approccio della ricerca locale.

Il gradiente ci dà la direzione di **ascesa**, quindi per il minimo basta prendere l'opposto.

$$w_{new} = w_{old} + \eta \cdot \Delta w$$

con $\Delta w = -\text{gradiente di } E(w)$

Δw è il "passo" che si fa in direzione del gradiente e η è la **costante di apprendimento**.

$$\Delta w_0 = -\frac{\partial E(w)}{\partial w_0} = 2(y - h_w(x)) \quad \Delta w_1 = -\frac{\partial E(w)}{\partial w_1} = 2(y - h_w(x)) \cdot x$$

Error Correction (Delta Rule) Regola che cambia la w proporzionalmente all'errore:

Target y – output $h = \text{Err} = 0 \rightarrow$ nessuna correzione

Output $>$ target $\Rightarrow y - h < 0$ **output troppo grande**

$\Rightarrow \Delta w_0$ negativa \Rightarrow **ridurre** w_0 e

se (input $x > 0$) Δw_1 negativa \Rightarrow **ridurre** w_1 altrimenti aumenta w_1

Output $<$ target $\Rightarrow y - h > 0$ **output troppo piccolo...**

Consente la **ricerca in uno spazio d'ipotesi infinito**. Può essere facilmente applicata sempre per H continuo a perdita differenziabile, **non solo a modelli lineari**.

Efficiente? Ci sono molte migliorie, come il metodo di Newton, gradiente coniugato...

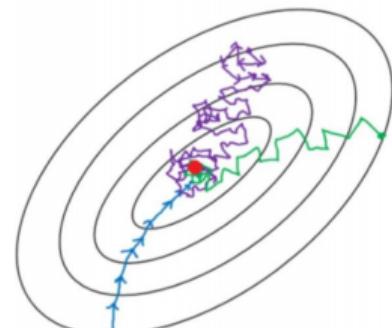
Per pattern l

$$\Delta w_0 = -\frac{\partial E(w)}{\partial w_0} = 2 \sum_{p \rightarrow l} (y_p - h_w(x_p)) \quad \Delta w_1 = -\frac{\partial E(w)}{\partial w_1} = 2 \sum_{p \rightarrow l} (y_p - h_w(x_p)) \cdot x_p$$

Possiamo aggiornare w dopo aver ripetuto un'epoca di dati d'apprendimento l

\rightarrow Algoritmo Batch (blu)

Altrimenti possiamo aggiornare w dopo ogni pattern p
 \rightarrow Algoritmo On-Line (stochastic gradient descent, viola e verde)



Input multidimensionale Nel caso standard ho input multidimensionale, da 2 a migliaia di input/variabili: il pattern di input è un **vettore**.

DATA notation X è una matrice $l \times n$ (l righe, n colonne (variabili)) $p = 1..l$, $j = 1..n$

Pattern	x_1	x_2	x_j	x_n
Pat 1	x_{11}	x_{12}	...	x_{1n}
:				
Pat p	x_{p1}	x_{p2}	...	x_{pn}

omettendo p

x_p o x_i vettori: p -esima o i -esima riga nella tabella, cioè pattern p o i

x_{pj} scalare, anche scritto come $(x_p)_j$: componente j del pattern p

Per il target y tipicamente si usa y_p con $p = 1..l$
Nota: con una variabile (cioè fin'ora) $x_p = x_{p1}$

Ogni riga corrisponde ad un generico x vettore:
esempio, pattern, istanza, sample...

x_i , x_j scalari: componenti i o j di un dato pattern,

Notazione Si assumono x e w come vettori colonna.

Il numero di dati è l , la dimensione del vettore input è n , i target sono y_p (precedentemente d_i o t_i) con $p = 1..l$
 w_0 è il threshold, bias, offset...

Spesso conviene includere la costante $x_0 = 1$ così possiamo scrivere la prima equazione come $w^T x = x^T w$ con
 $x^T = [1, x_1, x_2, \dots, x_n]$, $w^T = [w_0, w_1, \dots, w_n]$

Riassunto

Dato un set di l esempi d'apprendimento $\langle x_p, y_p \rangle$

Trovare il vettore di pesi w che minimizza l'errore atteso sui dati d'apprendimento

$$E(w) = \sum_{p=1}^l (y_p - x_p^T w)^2 = \|y - Xw\|^2$$

$$\Delta w_i = -\frac{\partial E(w)}{\partial w_i} = 2 \sum_{p=1}^l (y_p - h_w(x_p)) \cdot x_{pi} = 2 \sum_{p=1}^l (y_p - x_p^T w) \cdot x_{pi}$$

10.5.2 Gradient Descent Algorithm

Semplice algoritmo sul quale si basano molti di quelli attuali

1. Inizio con un vettore di pesi $w_{initial}$ inizializzato con valori piccoli, fisso η ($0 < \eta < 1$)
2. Si calcola $\Delta w = -$ "gradiente di $E(w)$ " = $-\frac{\partial E(w)}{\partial w}$ ($\forall w_i$)
3. Si fa il passo, calcolando $w_{new} = w_{old} + \eta \cdot \Delta w$ ($\forall w_i$)
 Con η dimensione del passo, o **coefficiente di apprendimento**
4. Ripeto da 2 finché converge o $E(w)$ è *sufficientemente piccolo*

$\frac{\Delta w}{l}$ LMN (least mean squares)

Batch version (Δw dopo ogni epoca di dati l)

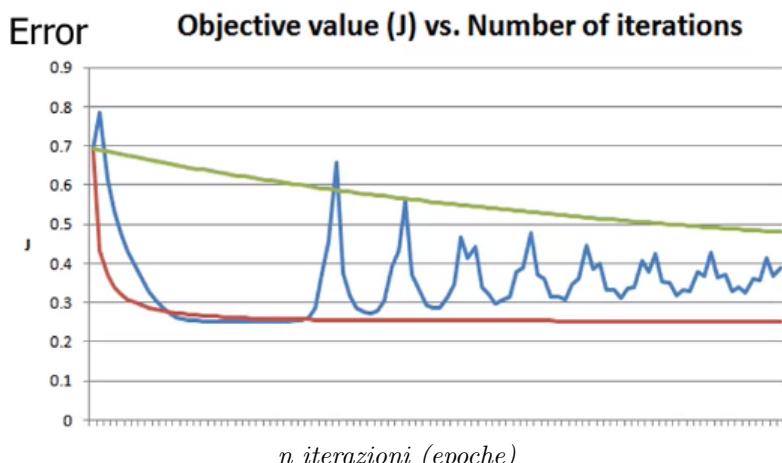
On-Line version: migliorare w dopo ogni pattern p (tramite $\Delta_p w = x_p(y_p - x_p^T w)$) senza aspettare la somma totale su l .

Può essere il più veloce, ma necessita di un η più piccole: progredisce con ogni esempio che osserva.

eta: tasso di apprendimento, trade-off su velocità/stabilità.

Può essere gradualmente ridotto a 0: garantisce la convergenza, evita oscillazione intorno al minore.

Esempi di curva di apprendimento Curva di apprendimento: elettrocardiogramma del modello. Se prendo modello, applico i dati e l'algoritmo, la curva di apprendimento mi mostra l'andatura del mio modello.



Vantaggi dei modelli lineari Se funzionano bene sono modelli fantastici

Semplici

Tutti i dati in w

Facili da interpretare: praticati in medicina, biologia, chimica...

Dati rumorosi **ammessi**

Statistici contenti (possiedono proprietà interessanti)

Fenomeni lineari: un sogno per la scienza, ideali per creare una "legge naturale"

Baseline per l'apprendimento, prima cosa a cui pensare: è un problema lineare?

Usato/incluso in modelli più complessi

Ma hanno molte limitazioni evidenti: soprattutto su funzioni non lineari.

Possiamo introdurre, al posto delle x originali delle trasformazioni non lineari: x^2, x^3, \dots
 $w_0 + w_1x + w_2x^2 + w_3x^3 \dots$ che corrisponde ad una **regressione polinomiale** =

$$y(x, w) = \sum_{j=0}^M w_j x^j$$

si può sempre risolvere con Least Square Solution. Questo perché ad esempio posso chiamare $x^2 = z_2$ e niente cambia nella tecnica risolutiva.

Linear Basis Expansion (LBE)

$$h_w(x) = \sum_{k=0}^K w_k \phi_k(x)$$

Aumento il vettore di input con variabili aggiuntive, ciascuna è trasformazione di x tramite la funzione $\phi_k : R^n \rightarrow R$
 ϕ può essere una qualsiasi trasformazione: la norma, il quadrato, logaritmo...

Sto espandendo l'insieme delle variabili considerate, quindi $K > n$, lineare nei parametri e in ϕ ma non in x : **possiamo usare i soliti algoritmi lineari** di prima.

Esempi LBE è anche detto **approccio a dizionario**: le ϕ sono le parole a disposizione per comporre le "frasi": diventa più espressivo, modella le relazioni non lineari.

Svantaggio: se introduco troppe trasformazioni possiamo avere un modello *tropppo* complesso rispetto alle necessità (**overfitting**).

Trade-Off sulla complessità del modello:

Modello troppo semplice non fa buon fitting, cioè **underfitting**, soluzione biased

Modello troppo complesso: sensibile a perturbazione dei dati, molta varianza ed **overfitting**

Metodi di **regolarizzazione del modello** per **trovare bilanciamento sul controllo della complessità del modello**. Penalizzando le funzioni troppo complesse (riducendo il valore dei pesi) mantenendo la flessibilità spazio ipotesi.

Rasoio di Occam: *la spiegazione più semplice è spesso la più corretta.*

Preferire le ipotesi più semplici che corrispondono ai dati.

Questo è un **concetto fondamentale in ML**, che troveremo di nuovo quando "razionalizzeremo" alla fine. Ora introduciamo un approccio per i modelli lineari: regolarizzazione con **ridge regression**.

10.6 Ridge Regression

Tikhonov Regularization Bassa varianza, meno overfitting. Modello smoothed.

Introduce vincoli che controllano la somma dei valori nostri parametri $|w_j|$ favorendo modelli sparsi, abbassando il valore di w . Se alcuni li **portiamo bassi**, in particolare a **0**, anche se abbiamo introdotto molte funzioni alcune **spariscono**.

$$\text{loss}(h_w) = \sum_p (y_p - h_w(x_p))^2 + \lambda \|w\|^2$$

con nuova aggiunta alla nostra funzione di loss: $\lambda \|w\|^2$, con λ **coefficiente di regolarizzazione** (parametro costante) e $\|w\|^2 = \sum w_i^2$

Quando minimizzo l'errore cerco anche di tenere i w_i bassi, perché la loss deve riuscire ad **abbassare entrambi i termini**.

Effetto: decadimento del peso (in pratica aggiungo $2\lambda w$ al gradiente)

$$w_{\text{new}} = w_{\text{old}} + \eta \cdot \Delta w - 2\lambda w_{\text{old}}$$

Se λ è alto allora modello fortemente normalizzato, se λ basso allora do molta importanza ai dati quindi rischio overfitting.

Limitazione delle funzioni base expansion

Avere una basis function lungo ogni dimensione di uno spazio di input a dimensione D richiede un numero combinatorio di funzioni.

Ad esempio, un polinomio generale di ordine 3 usa tutte le combinazioni di input a causa dei prodotti $x_1x_2, x_2x_3, \dots, x_1x_2x_3 \dots \Rightarrow D^3$

ϕ sono fissate *prima* di osservare i dati.

In altri modelli, vedremo come cavarsela con un numero minore di basis function, scegliendole tra i dati di apprendimento: ϕ dipenderà da w e il modello non sarà lineare (es: **reti neurali**)

In altri modelli ancora, la computazione del nuovo spazio è fatta implicitamente attraverso funzioni kernel e di controllo della complessità del modello (es: **SVM**)

10.7 Classificazione

Classificazione di modelli lineari Riusiamo il modello lineare.

Attribuiamo una categoria ad es 0/1, -1/+1...

Usiamo l'iperpiano (wx) assumendo valori positivi o negativi.

Decidiamo se x appartiene a zona positiva o negativa. Vogliamo trovare w tramite l'apprendimento.

Decision boundary $w^T x = w_1x_1 + w_2x_2 + w_0 = 0$ punto d'**intersezione** fra i due piani.

Metto uno scalino: tutte le volte che ero nella zona positiva attribuisco il valore d'uscita 1, dall'altra parte ad esempio 0 o -1...

$$h(x) = \begin{cases} 1 & \text{se } wx + w_0 \geq 0 \\ 0 & \text{altrimenti} \end{cases} \quad \text{oppure } h(x) = \text{sign}(wx + w_0)$$

Usando x_p e includendo w_0 in w posso avere $h(x_p) = \text{sign}(x_p^T w) = \text{sign}\left(\sum_{i=0}^n x_{pi} w_i\right)$

Iperpiano separatore Insieme dei punti per cui il nostro modello è 0. Quando ≥ 0 possiamo dire 1 altrimenti 0.

LTU: linear threshold unit.

10.7.1 Problema d'apprendimento per classificatori lineari

Dato un set di l esempi d'apprendimento, trovare w per cui **minimizzare** la somma di quadrati

$$E(w) = \sum_{p=1}^l (y_p - \mathbf{x}_p^T w)^2 = \|\mathbf{y} - \mathbf{X}w\|^2$$

Errore minimo: se $y_p = 1$ allora $\mathbf{x}_p^T w \rightarrow 1$. Se $y_p = 0/-1$ allora $\mathbf{x}_p^T w \rightarrow 0/-1$

Non usiamo $h(x)$ perché in classificazione ci starei mettendo $h(x) = \text{sign}(wx)$ che **non è differenziabile**.

Abbiamo ancora l'algoritmo iterative gradient descent

$$\Delta w_i = -\frac{\partial E(w)}{\partial w_i} = \sum_{p=1}^l (y_p - \mathbf{x}_p^T w) \cdot x_{pi}$$

Learning Rule Δw come correzione d'errore.

Se misclassified (perché la funzione target è diversa) allora posso aumentare i w_i corrispondenti agli errori proporzionalmente al Δ attraverso η .

Riassumendo

Il modello viene **addestrato** (su un TR set) con LS (LMS) su wx , tramite il simple gradient descent algorithm usato per la regressione lineare.

Il modello viene **usato** per classificare applicando la funzione threshold $h(x) = \text{sign}(wx)$

L'errore può essere computato come errore di classificazione o numero di pattern mal classificati (non sono dal mean square error)

$$L(h(\mathbf{x}_p), d_p) = \begin{cases} 0 & \text{se } h(\mathbf{x}_p) = d_p \\ 1 & \text{altrimenti} \end{cases} \quad \text{mean_err} = \frac{1}{l \cdot \sum_{i=1}^l L(h(x_i), d_i)}$$

$$\text{Con } \text{num_err} = \sum_{i=1}^l L(h(x_i), d_i)$$

$$\text{Accuratezza} \quad \text{Media dei classificati correttamente} = \frac{l - \text{num_err}}{l}$$

10.7.2 Algoritmo di Apprendimento

Come per la regressione: vedi gradient descent algorithm. Anche il linear basis expansion e Tikhonov possono essere usati.

Notare che: l'algoritmo di apprendimento

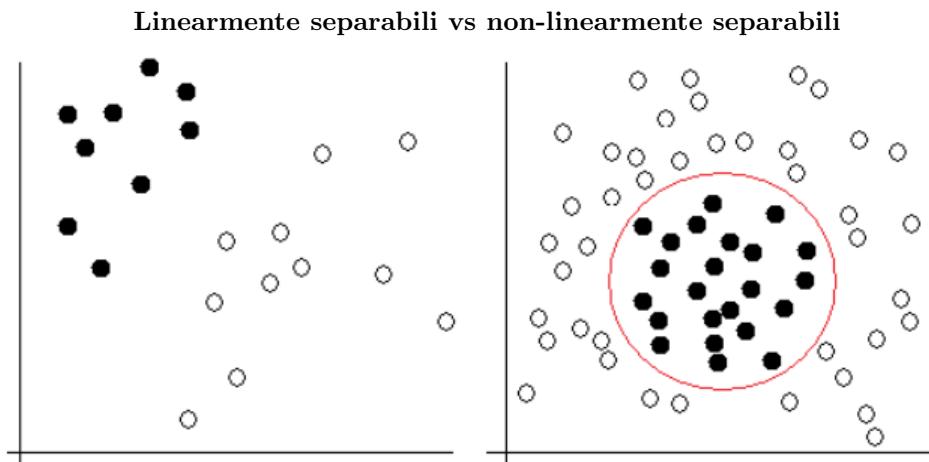
Converge asintoticamente al MinLS

Non necessariamente al numero minimo di errori di classificazione

Limitazioni In geometria, due insiemi di punti in uno spazio a due dimensioni sono **linearmente separabili** quando i due set di punti possono essere completamente separati da una singola linea.

In generale, **due gruppi sono linearmente separabili in uno spazio ad n -dimensioni se possono essere separati da un iperpiano a $(n - 1)$ -dimensioni**.

Il linear decision boundary può fornire soluzioni esatte solo per insiemi linearmente separabili.



10.8 Conclusione sui modelli lineari

Approccio base e ben fondato sia per regressione che per classificazione: la conoscenza è rappresentata in maniera molto compatta, ma con forti assunzioni sulle relazioni fra i dati.

Un **algoritmo di correzione iterativa dell'errore** (LMS) che cerca continuativamente nello spazio delle ipotesi (base di molti altri approcci ML).

Una visione delle limitazioni degli approcci lineari e delle necessità di modelli di ML più flessibili ed i loro problemi: un'estensione dei modelli lineari per task non lineari, un'introduzione al controllo della complessità (regolarizzazione).

10.9 Alberi di Decisione

Cambia sostanzialmente lo spazio delle ipotesi perché cambia la forma delle ipotesi a disposizione. Si rappresenta graficamente come un albero di decisione: albero dove ogni nodo ha un test di attributo per le foglie corrispondono al classificatore finale (Si/No).

L'albero di decisione rappresenta la disgiunzione di congiunzioni di vincoli sul valore degli attributi

outlook = sunny and humid = normal or

outlook = overcast or

outlook = rain and wind = weak

Oppure, **if-then-else**. H di decision tree è capace di esprimere qualsiasi funzione finita a valori discreti (propositionale).

10.9.1 Algoritmo Top-Down Induction (ID3)

Dato il TR set, l'algoritmo costruisce un Decision Tree ricercando nello spazio dei DT. La costruzione è top-down e la ricerca è greedy.

La domanda fondamentale è *quale attributo testare ora?*, **quale informazione ci fornisce il maggior information gain?: selezionare il miglior attributo**.

Per ogni possibile valore dell'attributo scelto viene generato un nodo discendente e gli esempi vengono partizionati secondo questo valore.

Il processo è ripetuto per ogni nodo successore finché tutti gli esempi sono correttamente classificati o sono finiti gli attributi.

Best Attribute In base all'**entropia**

Entropia: quantità di informazione portata da un segnale, più è stabile meno entropia porta quindi meno informazione, più oscilla più entropia più informazione. Nel nostro caso, il concetto di entropia misura impurità di una collezione d'esempi.

Data S collezione di esempi training, p_+ proporzione degli esempi positivi in S e p_- degli esempi negativi

$$\text{Entropy}(S) = -(p_+) \cdot \log_2(p_+) - (p_-) \cdot \log_2(p_-) \text{ (assumendo } 0 \cdot \log_2(0) = 0\text{)}$$

Se ho tutti positivi o tutti negativi, l'entropia è 0. Nei casi misti entropia alta.

Bassa omogeneità non va bene (alta impurità, alta entropia)

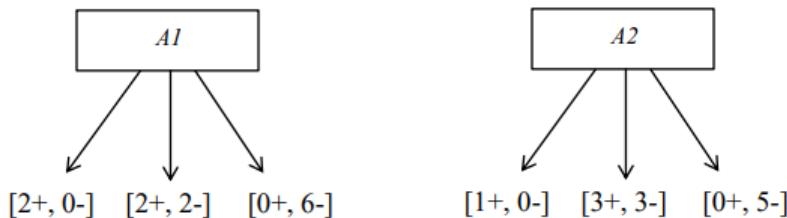
Notare che $0 \leq p \leq 1$, quindi $0 \leq \text{entropia} \leq 1$

Nei casi peggiori, esempio $7+7-$, si entra nella zona scomoda di indecisione. Più nodi devono portare più discriminazione.

Vogliamo discriminare, quindi seleziono l'attributo migliore: quello che dà miglior riduzione dell'entropia conoscendo valore dell'attributo

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \left(\frac{|S_v|}{|S|} \cdot \text{Entropy}(S_v) \right)$$

Un esempio



$$\begin{aligned} G(S, A1) &= E(S) - \left[\frac{2}{12} \cdot \left(-\frac{2}{2} \log_2(\frac{2}{2}) - \frac{0}{2} \log_2(\frac{0}{2}) \right) + \frac{4}{12} \left(-\frac{2}{4} \log_2(\frac{2}{4}) - \frac{2}{4} \log_2(\frac{2}{4}) \right) + \frac{6}{12} \left(-\frac{0}{12} \log_2(\frac{0}{12}) - \frac{6}{12} \log_2(\frac{6}{12}) \right) \right] \\ &= E(S) - \frac{2}{12}(0) - \frac{4}{12}(1) + \frac{6}{12}(0) = E(S) - \frac{1}{3} \\ G(S, A2) &= E(S) - \frac{1}{12}(0) - \frac{6}{12}(1) - \frac{5}{12}(0) = E(S) - \frac{1}{2} \end{aligned}$$

In questo esempio si sceglierrebbe $A1$

Più alto è l'information gain, più efficace è nel classificare i dati. Se ottengo tutti sottoinsiemi con positivi e 0 negativi, e 0 positivi e negativi ottengo una classificazione pulita.

Entropia misura l'omogeneità (impurità) della classe del sottounsieme di esempi, seleziono A che massimizza.

Dopo aver separato, mi aspetto tanti sottoinsiemi omogenei con valori più bassi quindi più alto guadagno.

Quindi l'entropia guida la funzione di apprendimento.

Perché funziona? I casi con entropia molto bassa pesano molto, sono molto omogenei (es $6+1-$)

Problemi con l'information gain Così com'è favorisce situazioni con molti valori possibili. Ma ci sono casi dove non sono significativi. Elemento correttivo può essere

$$\text{GainRatio}(S, A) = \frac{\text{Gain}(S, A)}{\text{SplitInformation}(S, A)}$$

$$\text{SplitInformation}(S, A) = - \sum_{i=1}^c \left(\frac{|S_i|}{|S|} \cdot \log_2 \frac{|S_i|}{|S|} \right)$$

Splitinf misura l'entropia di S rispetto ai valori di A, più uniformemente dispersi sono i dati meglio è.

GainRatio penalizza

Ha un denominatore che può succedere vada a 0 quando $|S_i|$ è circa uguale a $|S|$ per qualche valore i ($\log(1) = 0$ per esempio)

Un caso estremo $|S_1| = 0$ e $|S_2| = n$, $\text{SplitInformation} = -0 - 0 = 0$ che farebbe esplodere GainRatio e farebbe scegliere questo attributo che in realtà è costante e non darebbe informazione

Per mitigare questo effetto, calcolo il Gain per ogni attributo e applico GainRatio solo ad attributi con $\text{Gain} > \text{media}$.

10.9.2 Ricerca nello spazio delle ipotesi dei Decision Trees

Ricerca hill-climbing attraverso lo spazio dei possibili DT dal più semplice al più complesso (**ricerca locale senza backtracking**). Quindi si rischia ottimo locale e non ottimalità. Maniene una singola ipotesi corrente in memoria e lo spazio ipotesi è completo (rappresenta tutte le funzioni a valori discreti). Usa tutti gli esempi disponibili (non è incrementale).

Può terminare prima, accettando classi rumorose senza accuratezza del 100% sul TR.

Due tipi di bias

Qual è il bias induttivo dei DT?

Alberi più corti preferiti rispetto alberi più lunghi
Legato all'aspetto greedy, con cui sceglio il miglior attributo

Preferiamo seguire attributi con maggiore information gain più vicini alla radice, per primi

Nota che il DT non è limitato nel rappresentare tutte le possibili funzioni: la restrizione non è sullo spazio delle ipotesi, è sulla strategia di ricerca

Preferenza o search bias è legato alla strategia con cui ci muoviamo nello spazio delle ipotesi
Es IDT cerca in uno spazio ipotesi completo ma la strategia è incompleta

Restrizione o language bias è sulle ipotesi esprimibili o considerabili
Es Candidate Elimination su spazio ipotesi incompleto (solo and) ma strategia completa

Perché bias di ricerca preferibile a quello di linguaggio? Per evitare di limitare a priori linguaggi di espressione perché così non ho più possibilità di trovare determinate soluzioni. Se mantengo tutte le possibilità me la gioco con la strategia di ricerca

In ML tipicamente si usano **approcci flessibili** o con **capacità di approssimazione universale** senza escludere a priori una possibile funzione target.

Ovviamente c'è il rischio di overfitting.

Perché preferire ipotesi più breve L'ipotesi più semplice che dà un buon fitting è quella che dà una buona risposta con maggiore probabilità.

Ricordare il controllo della complessità del modello tramite la regolarizzazione dei modelli lineari.

10.9.3 Problemi nell'apprendimento con DT

La risoluzione costruirà il C4.5

Overfitting h overfitta i dati di apprendimento se c'è un'ipotesi alternativa $h' \in H$ tale che

$$E_D(h) < E_D(h') \wedge E_x(h') < E_x(h)$$

cioè h' si comporta peggio sul TR e meglio sui dati non osservati, altrimenti detto **peggio su errore empirico e meglio su errore atteso**.

Approcci flessibili possono facilmente incontrare l'overfitting.

"Evitare" l'overfitting "Avoiding" non va bene, non si può evitare l'overfitting che è un **problema inherente** e gli strumenti lo controllano ma non ci sollevano dalla responsabilità di gestire la questione.

Una tecnica comune è quella di fermare l'apprendimento presto prima di avere una classificazione perfetta, oppure consentire la crescita e l'overfitting ma fare un **pruning a posteriori** tornando ad un albero più piccolo.

Come agire?

Training e **validation set**

Si divide in due parti e si usa validation per decidere quando fermarsi o quando effettuare pruning

Minimum description lenght, uso una misura di complessità e mi fermo quando cresce troppo oltre una certa misura (similitudine con Tikhonov)

Ogni nodo è candidato per il pruning: consiste nel rimuovere un sottoalbero radicato in un certo nodo, che diventa una foglia assegnata, solo se risulta un albero non peggiore sul validation set.

Dopo aver creato l'albero con il TR set, converto ogni cammino in un **set di regole equivalenti**: ogni cammino corrisponde ad una regola, ogni nodo intermedio ad una precondizione e ogni foglia classificata ad una postcondizione. Il **pruning generalizza ogni regola rimuovendo le precondizioni che migliorano l'accuratezza** sul VL set e/o sul training con un'euristica pessimistica.

Valori continui Si trovano le soglie per cui cambia la classificazione, ad esempio prendendo la media tra due valori consecutivi classificati diversamente.

Dati incompleti Si usano gli esempi per indovinare l'attributo mancante: si prende il valore più comune tra gli esempi nella solita classe oppure si assegna ad ogni valore una probabilità e si assegna all'attributo mancante il valore seguendo la probabilità assegnata.

10.10 Validazione

Ogni volta che si aumenta la precisione c'è il rischio di andare in overfitting. Bisogna quindi **valutare la capacità di generalizzazione dell'ipotesi**: ruolo essenziale della validazione. Con cenni di fondamenti teorici. Aspetto teorico e pratico per un **uso consapevole del ML**.

Quando un modello di ML è un buon modello? Usare ML vs usare *bene* ML.

Ricordiamo apprendimento: ricerca di una buona funzione in uno spazio.

Inductive learning hypothesis: qualunque h che approssima *bene* f sui TR la approssimerà bene anche su istanze non viste x . **Misure:**

Classificazione: Mean Square Error per la perdita, **accuratezza** e **Mean Error Rate** per il risultato
Ma anche precisione, specificità, sensitività (per falsi positivi e negativi)

Regressione: **Mean Square Error**, Root MSE, **Mean Absolute Error**, Max Absolute Error
Ma anche statistica come R (correlazione coefficiente/indice)...

Ovviamente errore alto accuratezza bassa. sia per TR che per test...

basso fitting alto errore training, bassa generalizz alto errore di test

Due obiettivi della validazione:

Model Selection Stima performance/prestazioni su modelli differenti di apprendimento per scegliere il migliore da generalizzare
Include la ricerca dei parametri migliori per il modello. **Ritorna un modello**

Model Assessment Scelto il modello finale, stimare/valutare il rischio/errore di predizione su dati di test (nuovi). Misuro la qualità del modello finale. **Ritorna una stima** "predice all'85% di accuratezza"

Obiettivo importante è separare il data set così da perseguire gli obiettivi: fitting, model selection e assessment. Tre parti: **training, validation e test**. Sono **insiemi disgiunti**. Ad esempio 50% dati in TR, 25% in VL e 25% in TS.

TR, per il **training**

VL, per la **validation**

TS, per il **test**

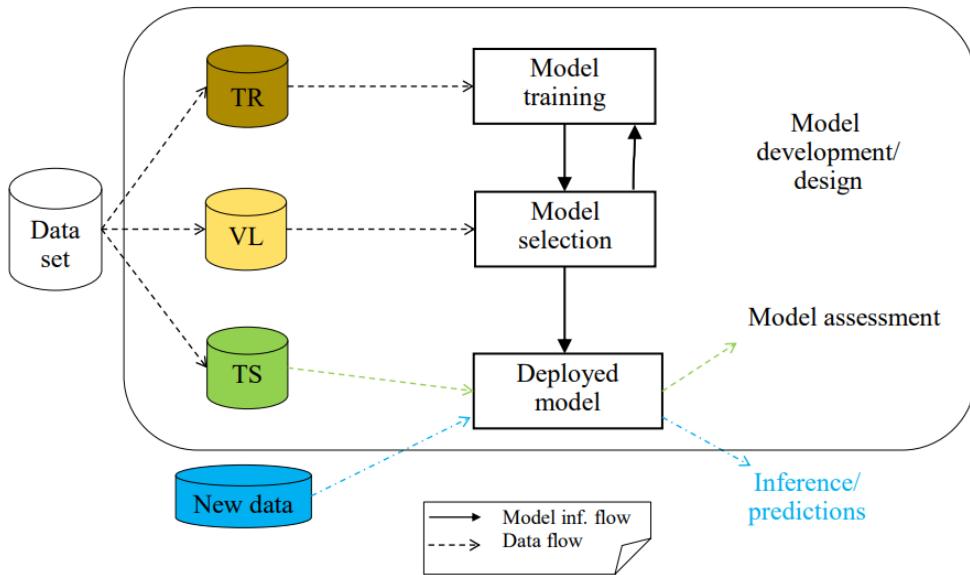
TR+VL solitamente chiamato development/design set, usato per costruire.

La stima in model selection è solo per scopi di model selection, **non è buona stima di rischio**, la quale va fatta sul TS. Il **TS non può essere usato per la model selection**.

Posso usare il TS più volte? **No**, perché altrimenti starei facendo una model selection e non assessment affidabile. Non possiamo riusarlo su esempi futuri. **Blind Test Set**: se vedi la soluzione non è un test!

Si avrebbe una stima overottimistica

Regola d'oro: mantenere una separazione fra gli obiettivi e usa set separati: TR per training, VL per model selection e TS per risk estimation.



Meta-Algoritmo

1. Separare TR, VL e TS
2. Selezionare il miglior $h_{w,\lambda}()$ cambiando λ
3. Per ciascun λ , allenare il modello cambiando w che minimizza errore (**fitting TR**). Miglior significa minimo errore su TR (**fitting**).
Una volta finito, cambiare λ e riprovare i vari w .
4. Selezionare il miglior modello su λ che ha un minor errore su VL
5. Opzionalmente, si può fare fitting su TR+VL (unico training set) col il miglior λ trovato
6. Una volta finito, valutare $h_{w,\lambda}(x)$ sul TS

Iper-Parametro Un **iper-parametro** è un parametro non appreso, ad esempio λ . Ricercare l'iper-parametro migliore può essere un ciclo su una griglia di possibilità. Può essere automatizzato e parallelizzato (prove indipendenti). Si sceglie il modello migliore su VL, NON si usa TS per model selection.

Esercizio

Controesempio

K-Fold Cross-Validation Si divide il dataset D in k subset mutualmente esclusivi D_1, D_2, \dots, D_k . Si addestra l'algoritmo su D senza D_i e si testa su D_i . Si fa la media dei risultati sui D_i . Questa tecnica usa tutti i dati per training e validation o testing e può essere usata sia per il VL che il TS.

Esempio di model selection e assesment Dividere i dati in TR e TS. Usare K-Fold CV sul TR per trovare gli iperparametri migliori del modello (es: ordine del polinomio, λ del ridge regression...): grid-search con tanti valori possibili. Es: un K-Fold CV per $\lambda = 0.1$, uno per $\lambda = 0.01\dots$ e scelgo il miglior λ tramite la media sul VL per ogni fold.

Si addestra su tutto il TR il modello finale e si valuta sul TS esterno.

Verso SLT

Mettendo tutto insieme La capacità di generalizzazione di un modello, rispetto all'errore di training ed eliminando over e underfitting

Mettendo tutto insieme Il ruolo della complessità di un modello

Mettendo tutto insieme Il ruolo del numero di dati

10.11 SLT

Statistical Learning Theory Approssimare una funzione $f(x)$ sconosciuta con valori target d . Minimizzare la funzione rischio

$$R = \int L(d, h(x)) dP(x, d)$$

con x vettore. R indica il **vero errore su tutti i dati**.

Dato

il valore dal supervisore

la distribuzione di probabilità $P(x, d)$

una funzione di **loss** (o costo) $L(h(x), d) = (d - h(x))^2$

allora bisogna cercare h in H tale che minimizzi R . Abbiamo però solo il dataset finito $TR = (x_i, d_i)$ ($i = 1..l$)
Per cercare h bisogna minimizzare l'errore in fitting cioè il **rischio empirico** (errore di training)

$$R_{emp} = \frac{1}{l} \sum_{i=1}^l (d_i - h(x_i))^2$$

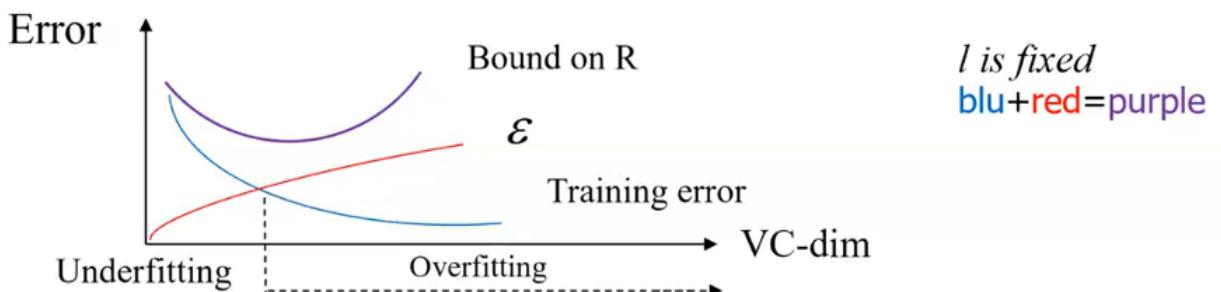
ERM: Empirical Risk Minimization come principio induttivo. Possiamo usare R_{emp} per approssimare R ?
La teoria esiste: **Statistical Learning Theory**. Ingrediente fondamentale: VC-Dim (Vapnik-Chervonenkis) misura della complessità di H (flessibilità per fitting dati)
Si ha, con probabilità $1 - \delta$, che (R rischio garantito)

$$R \leq R_{EMP} + \epsilon \left(\frac{1}{l}, VC, \frac{1}{\delta} \right)$$

Con ϵ detta **VC-Confidence**, funzione direttamente proporzionale a VC (VC-dim) ed inversamente proporzionale a l (numero dei dati) e δ (**confidenza**, guida la probabilità che valga il bound: δ piccolo 0.01 \Rightarrow probabilità che valga il bound dello 0.99).

Intuizione: l alto $\rightarrow R$ basso. Fissato l , alto VC-dim $\rightarrow R_{emp}$ basso ma R potrebbe aumentare (overfitting)

Structural Risk Minimization



Intuizione

l alto, R basso

VC-dim alto, R_{EMP} riduce ma R potrebbe aumentare (overfitting)

Bisogna mettersi dove R è minore (linea viola)

SLT Permette di inquadrare formalmente il problema della generalizzazione e underfitting/overfitting, fornendo limitazioni superiori analitiche e quantitative al rischio R di predizione.

ML è ben fondato: rischio di learning (errore di generalizzazione) può essere analiticamente limitato e solo pochi concetti sono fondamentali

Si può trovare una buona approssimazione dell' f da esempi, avendo un buon numero di dati e un'adeguata complessità (misurabile con la VC-dim)

Porta a nuovi modelli (SVM) e altri modelli che considerano direttamente il controllo della complessità nella costruzione.

Fonda uno dei principi

esistono altri principi?

come misurare complessità?

Modelli lineari: è intuitivamente legato al numero di parametri liberi (dim input)

Decision trees: numero di nodi.

Conclusioni Flessibilità modelli ML. Potenza ML senza controllo produce risultati illusori. Controllare tradeoff tra fitting e complessità. Fondamentale il ruolo dei processi di validazione

ML ben fondato teoricamente Domande fondamentali nella SLT.

10.12 Support Vector Machine

SVM Sostanzialmente è un modello lineare, ma è nato dalla Statistical Learning Theory. Diventata famosa, dopo anni di sviluppo teorico, quando usando immagini come input ha avuto un'accuratezza comparabile a quella di reti neurali sviluppate ad hoc.

SVM è attualmente usato in tutti i campi di applicazione dell'apprendimento supervisionato, e anche usato per la regressione.

Argomento tipico introdotto alla fine di un corso di ML, perché discuterlo nei dettagli richiede una visione approfondita del ML. Saremmo comunque tentati di usarlo, attraverso uno dei tanti tool: ci dà una prima **visione critica**. Anche opportunità per vedere almeno un modello allo stato dell'arte.

Ripartiremo dai modelli lineari per la classificazione, reintegrandoli per problemi non lineari.

Obiettivi

1. Introdurre il **concetto di controllo di complessità** del modello **attraverso un approccio di ottimizzazione**

Questo per **approssimare direttamente** la minimizzazione del rischio strutturale (Structural Risk Minimization)

Max Margin Classifier

2. **Usare efficientemente la linear basis expansion** attraverso il concetto di **kernel**

Così otteniamo un altro approccio flessibile per apprendimento non lineare e supervisionato
Kernel

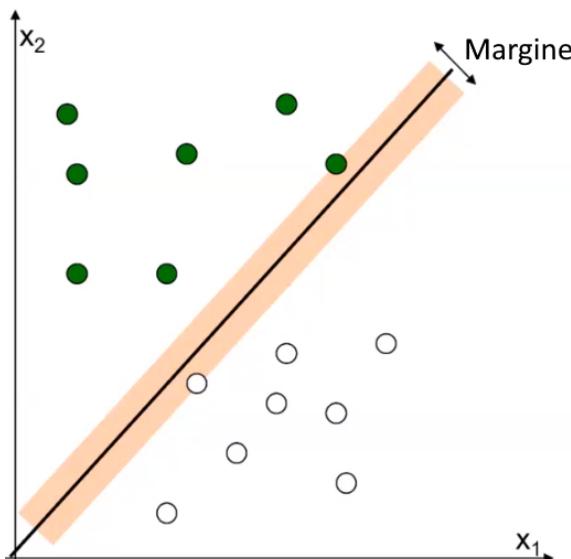
3. Evitare misinterpretazioni nell'uso dell'SVM e del ML

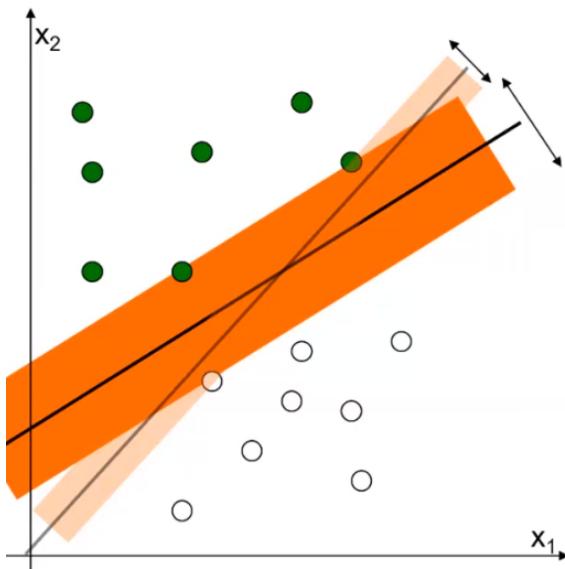
10.12.1 Maximum Margin Classifier

Soluzioni In generale per un problema ci possono essere più soluzioni ottime, ma vanno davvero tutte bene? Per noi all'interno del ML **no**, e vediamo come deciderlo.

Classificazione a massimo margine Su un problema di classificazione binaria, iniziando con l'assumere di avere un problema linearmente separabile (**hard margin SVM**).

Non tutti gli iperpiani che risolvono un task sono uguali.





Variando l'iperpiano separatore il **margine** cambia.

Margine Il margine è il **doppio della distanza tra l'iperpiano e i punti più vicini**.

Intuitivamente Il max margin classifier intuitivamente è la **massima distanza dai datapoint, la massima "safe zone"**.

Rappresentazione canonica nell'SVM I pallini nella figura rappresentano i vettori, quelli verdi sono i positivi (x^+) e i bianchi i negativi (x^-).

I vettori di supporto sono quei vettori per cui l'iperpiano vale **esattamente 1** per i positivi, o -1 per i negativi. **Quelli su bordo del margine**.

Cioè $|w^T x_i + b| = 1$. Tutti i punti sono correttamente classificati se $(w^T x_i + b)y_i \geq 1$ per ogni i che sostanzialmente indica che tutti i punti sono lontani dal margine.

Considereremo il problema di apprendere un modello lineare per classificazione binaria, cioè una funzione $h : \mathbb{R}^n \rightarrow \{-1, 1\}$, $h(x) = \text{segno}(w^T x + b)$, basandosi su esempi (x_i, y_i)

Training Problem: trovare (w, b) tale che tutti i punti sono correttamente classificati e il **margine è massimizzato**.

Margine Il margine è dato dalla formula $\frac{2}{|w|}$, dove $|w|^2 = (w^T w)$

Quindi massimizzo margine \Leftrightarrow minimizzo $|w| \Leftrightarrow$ minimizzo $\frac{|w|^2}{2}$

Th La VC-dim dell'SVM è inversa al margine, quindi **controllo della complessità tramite il margine** vedi SLT

L'iperpiano ottimo è quello che massimizza il margine e risolve il problema.

Possiamo ora impostare il tutto come un problema di ricerca operativa.

Training Problem Trovare (w, b) tale che tutti i punti sono classificati correttamente e il margine è massimizzato.

Forma Primale hard margin **Training Problem** diventa minimizzare $\frac{|w|^2}{2}$ (cioè minimizzare $w^T w$) in maniera che $(w x_i + b)y_i \geq 1 \quad \forall i$

Diventa un problema di ottimizzazione quadratica, risolvibile con software appropriati.

Diretta ottimizzazione di complessità del modello. La funzione obiettivo, che è l'ottimizzazione stessa (e fornisce la soluzione con errore 0 perché si assume che il problema sia linearmente separabile), è convessa in w .

Forma Duale Nella formula duale si calcolano gli α , moltiplicatori di Lagrange, che sono legati ai vettori di supporto e ad una forma speciale della soluzione, più elegante.

Con gli α si può calcolare (w, b) , $w = \sum \alpha_i y_i x_i$ con $i = 1 \dots l$ e $b = y_k - w^T x_k$ per qualsiasi $\alpha_k > 0$, con l numero dei dati.

$$h(x) = \text{segno}(w^T x + b) = \text{segno} \left(\sum_{i=1}^l \alpha_i y_i x_i^T x + b \right) = \text{segno} \left(\sum_{i \in \text{SV}} \alpha_i y_i x_i^T x + b \right)$$

Si ottiene $\alpha_i \neq 0 \Leftrightarrow x_i$ è un **vettore di supporto**. La soluzione è **sparsa** e formulata solamente in termini dell'**SV**: l'iperpiano dipende solo dai vettori di supporto.

In questa formulazione della soluzione non è neanche necessario calcolare esplicitamente (w, b)

Soft Margin SVM Nell'hard margin, visto fin'ora, si potrebbe avere troppa restrittività. Possiamo concedere un po' di errore, per tollerare dati rumorosi e avere un margine più ampio.

Soluzione: consentire errori introducendo le variabili **slack** ξ_i , non nulle per qualche i dove consente l'errore.

Forma Primale soft margin Minimizzare $\frac{|w|^2}{2} + C \cdot \sum_i \xi_i$ in maniera che $(wx_i + b)y_i \geq 1 - \xi_i$ $\xi_i \geq 0 \quad \forall i$

C è un **iperparametro** definito dall'utente che **guida il numero di errori concessi**:

C basso \rightarrow vengono consentiti molti errori di training \rightarrow possibile underfitting

C alto \rightarrow non vengono consentiti errori TR \rightarrow margine piccolo e possibile overfitting

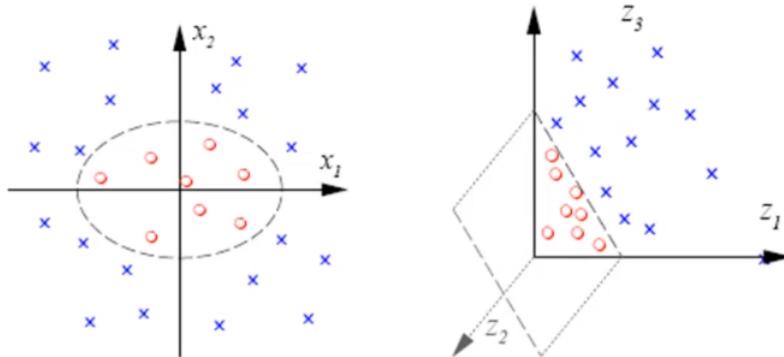
Essendo C un iperparametro, v'è trovato tramite la validation (**model selection**), quindi rientra in gioco il trade-off. L'hard margin **non** si ottiene con $C = 0$, poiché bisogna considerare gli ξ_i presenti nei vincoli: si darebbe totale libertà alle variabili slack. $C = 0$ significa "puoi anche fare il 100% di errori nel TR". L'hard margin, quindi, si ottiene con un C **molto alto**.

10.12.2 Kernel

Fin'ora abbiamo parlato solo di problemi linearmente separabili, per quanto riguarda i casi non lineari?

Kernel Modo per usare efficientemente la linear basis expansion, così da ottenere un nuovo approccio flessibile per modelli supervisionati non lineari.

Sappiamo che si può trasformare lo spazio originale in maniera da risolvere i problemi non lineari, tramite la ϕ . I separatori lineari nel nuovo spazio corrispondono a separatori non lineari nello spazio originale.



Nell'esempio $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ a cui $(x_1, x_2)^T \mapsto (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$

Bisogna quindi mappare i datapoint dello spazio in input verso un **feature-space** a dimensioni maggiori dove sono linearmente separabili.

Quindi al posto di x possiamo usare $\phi(x)$. Sappiamo però che usare un high dimensional feature space può essere sia computazionalmente oneroso sia portare all'overfitting se non si opera un controllo della dimensione del feature space e della complessità del classificatore (in termini di dimensioni dell'input, cioè se uso troppe ϕ).

$$h_w(x) = \text{segno} \left(\sum_k w_k \phi_k(x) \right)$$

Kernel Approccio per mappare implicitamente il feature space **nel contesto di un modello regolarizzato**, perché la complessità è legata al margine.

Grazie alla regolarizzazione automatica del SVM, la complessità del classificatore può essere tenuta bassa qualunque sia la dimensione del nuovo feature space.

Se proietto in uno spazio anche molto ampio, ma riesco a tenere un margine ampio, la complessità è comunque regolare.

Nell'SVM non è necessario calcolare w e i dati tramite prodotti scalari $h(x) = \text{segno}(\sum_{i \in SV} \alpha_i y_i x_i^T x + b)$

$$h_w(x) = \text{segno}\left(\sum_k w_k \phi_k(x)\right) \Rightarrow h(x) = \text{segno}\left(\sum_{i \in SV} \alpha_i y_i \phi^T(x_i) \phi(x)\right)$$

Ma non è nemmeno necessario calcolare direttamente ϕ , ma si può **gestire implicitamente** tramite una funzione Kernel

$$h(x) = \text{segno}\left(\sum_{i \in SV} \alpha_i y_i K(x_i, x)\right)$$

Con $K(x, y) = \phi(x)^T \phi(y)$

10.12.3 Kernel notevoli

Lineare $K(x_i, x_j) = x_i^T x_j$

Mappa $\phi : x \mapsto \phi(x)$ dove $\phi(x) = x$

Polinomiale di potenza p $K(x_i, x_j) = (1 + x_i^T x_j)^p$

Mappa $\phi : x \mapsto \phi(x)$ dove $\phi(x)$ ha dimensione esponenziale in p

RBF (Radial-Basis Function) Gaussiana $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$

Mappa $\phi : x \mapsto \phi(x)$ dove $\phi(x)$ ha dimensione **infinita**.

Scelta molto popolare, notare l'iperparametro σ . Molto flessibile e utilizzabile per costruire decision boundary attorno ad ogni punto. σ piccolo significa Kernel a punta stretta, cioè pattern simili solo se molto vicini e il classificatore risponde con la classe del punto più vicino.

Molto potente ma portato all'overfitting.

Il design di nuovi kernel per tipi di dato speciali è un argomento di ricerca attuale.

10.12.4 Sintesi su SVM

Trade-off sul parametro C e la funzione kernel K (e i suoi parametri)

Risolvere il problema di ottimizzazione per trovare α

Il costo computazionale scala con l (numero dei dati) invece di n (dimensione dello spazio).

Modularità: basta cambiare il kernel

Modello finale

$$h(x) = \text{segno}\left(\sum_{i \in SV} \alpha_i y_i K(x_i, x)\right)$$

10.12.5 Uso pratico

Evitare la misinterpretazione

La tipica misinterpretazione nell'uso dell'SVM è pensare che non faccia overfitting. Esso può verificarsi se non si opera un'attenta selezione dei parametri C , kernel, parametri del kernel...

La trattazione隐式 di un spazio a grandi dimensioni è **nel feature space**, non nell'input space (assumendo di proiettarci gli input significativi)

Le tecniche di **validazione** viste precedentemente per il model selection (C , kernel e parametri) e model evaluation dovrebbero essere usate rigorosamente anche qua.

Una piccola guida all'uso (da LIBSVM):

1. Trasformare i dati nel formato di un software SVM
2. Scalare i dati
3. Considerare il kernel RBF $K(x, y) = e^{-\gamma||x-y||^2}$ con $\gamma = \frac{1}{2\sigma^2}$
4. Usare il cross-validation per trovare i parametri C e γ (**VL set** \neq **TR set**)
5. Testare (**su un set separato ed esterno**)

Data processing Alcuni esempi:

Colori {rosso, verde, blu} $\rightarrow (0,0,1), (0,1,0), (1,0,0)$

Per scalare valori continui in un range con la formula $\frac{x - \min}{\max - \min}$

Model selection grid-search Per esempio per trovare C e γ nell'RBF: con una tabella con tutte le combinazioni di valori crescenti possibili per trovare intervalli buoni, ad esempio

$$C = 2^{-5}, 2^{-3}, \dots, 2^{15}$$

$$\gamma = 2^{-15}, 2^{-13}, \dots, 2^3$$

Per poi usare una grid-search più fine, una volta trovato il range interessato (es. $2^2, 2^3$).

10.13 K-Nearest Neighbors

10.13.1 1-Nearest Neighbor

Supervised Learning Alcune definizioni

Memorizza i dati di training $\langle x_j, y_j \rangle$

Training finito, non c'è da fare altro.

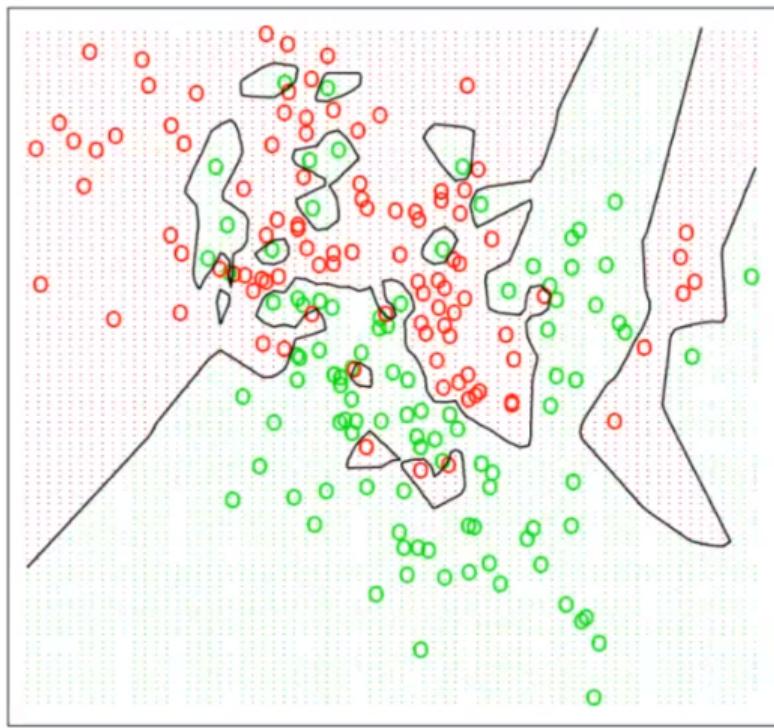
Viene fornito un input x di dimensione n

Per rispondere, troviamo il training example più vicino x_i

Trovare l'indice i con la minima distanza da x $i(x) = \arg \min_j d(x, x_j)$

La distanza è la classica distanza euclidea, per esempio, cioè: $d(x, x_j) = ||x - x_j|| = \sqrt{\sum_{t=1}^n (x_t - x_{jt})^2}$

Come output y_i , elemento x_i del TR più vicino a x input

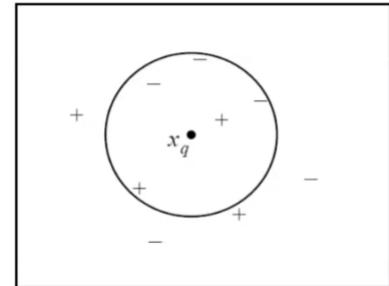


Può creare superficie di separazione molto flessibile, anche troppo. Non c'è errore sui dati di TR, mentre i decision boundaries non sono lineari, ma anzi molto irregolari.

Smoothing Un possibile smoothing della 1-NN è la 5-NN:

1-NN ritorna + per x_q

5-NN ritorna - per x_q



Smooting su un set di vicini per dati rumorosi.

10.13.2 K-Nearest Neighbors

Un modo naturale per classificare un nuovo punto è **guardare i suoi vicini, e prenderne la media**:

$$\text{avg}_k(x) = \frac{1}{k} \cdot \sum_{x_i \in N_k(x)} y_i$$

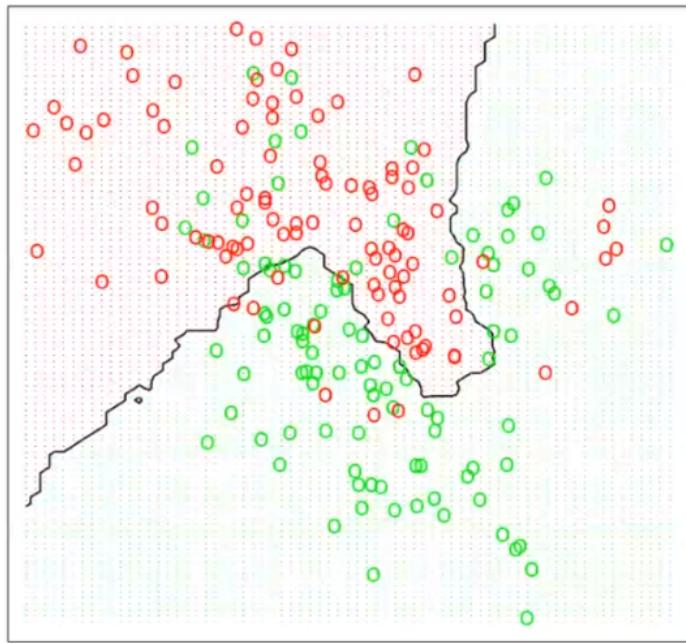
dove $N_k(x)$ è il **vicinato di x che contiene esattamente k vicini** in accordo a d

Se c'è una classe dominante nell'intorno di x osservato, allora è probabile che prenderemo quella classe anche come osservazione. Di conseguenza, la **regola di classificazione è la regola di maggioranza tra i membri di $N_k(x)$.**

Come prima, $h(x) = \begin{cases} 1 & \text{se } \text{avg}_k(x) > 0.5 \\ 0 & \text{altrimenti} \end{cases}$ per $y_i = \{0, 1\}$ target

Per i task di **regressione** si usa direttamente l'**avg**, la media sui K-nn.

15-NN



Ancora molto flessibile

Accetta errori di classificazione nei dati TR

Decision boundary non lineare, ma ancora (seppur meno) irregolare

Decision boundary si adatta alle densità locali delle classi

10.13.3 Complessità

Trade-off Come sempre c'è un trade-off tra underfitting e overfitting sui possibili valori di K

Abbiamo la solita curva ad U sul grafico con K , muovendosi da un caso estremamente flessibile ($K = 1$, overfitting) fino ad un modello molto rigido ($K = l$, underfitting), con una media per tutti i dati.

Il K migliore, come sempre, è il giusto bilanciamento tra la complessità del modello rispetto al problema e, come sempre, si trova tramite il validation set e il model selection.

Con più classi? Per ciascuna classe vado a contare nell'intorno quante occorrenze ho e rispondo con la classe più frequente.

$$h(\mathbf{x}) = \arg \max_v \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} l_{v,y_i}$$

$$l_{v,y_i} = \begin{cases} 1 & \text{se } v = y_i \\ 0 & \text{altrimenti} \end{cases}$$

10.13.4 Un estremo

K-NN non è un vero a proprio modello, ma è un estremo. Non fa un'ipotesi globale, quindi non ha un modello per cui fare il fitting. Memorizziamo tutti i dati d'esempio, senza un modello parametrico. Per contro, basti pensare al modello parametrico lineare con un insieme di parametri w .

Ha altre caratteristiche, fa una **stima locale** con delle funzioni costanti locali, in contrasto ai modelli globali di stima/approssimazione della funzione target.

Metodi basati sulla distanza, pigri, basati sulla memoria, basati sugli esempi. La **distanza** è la misura critica, definita quella ho definito quali sono i più vicini.

10.13.5 Qualche limite di K-NN

Costo computazionale Notare che il K-NN fa un'approssimazione locale della funzione target per ogni nuovo esempio da predire: il **costo computazionale è spostato verso la fase di predizione**.

Inoltre, alto costo di retrieval: computazionalmente intenso per ogni nuovo input, calcolo della distanza dall'esempio di test verso *tutti* i vettori in memoria

Tempo proporzionale al numero di pattern memorizzati

Possibilità di ottimizzazione ad-hoc, ad esempio con algoritmi di indicizzazione dei pattern

Costoso anche in spazio, tutti i dati sono memorizzati.

Curse of dimension Quando abbiamo molte variabili in input (n alto), i metodi K-NN spesso falliscono per la *maledizione della dimensionalità* (curse of dimension):

Quando la dimensionalità d aumenta, il volume (lato d) dello spazio aumenta così velocemente che i dati disponibili diventano sparsi

vale a dire che la quantità di dati necessaria a supportare il risultato, cioè avere un intorno relativamente buono, cresce **esponenzialmente** con la dimensionalità

Curse of noisy Feature irrilevanti: se il target dipende solo sul alcune delle molte feature in x (es 2 su 20), possiamo fare il retrieval di un pattern "simile" con la similarità dominata dal grande numero di feature irrilevanti.

Cresce con la dimensionalità.

10.13.6 Considerazioni

I metodi basati sulla distanza o sulla metrica (linea comune tra, ad esempio, tra gli approcci K-means, K-NN e kernel-based) sono **basati sulla similitudine**.

Cosa significa simile? Misura quando una coppia di pattern sono *simili*, data una metrica sto ponendo un forte bias. La metrica tipicamente, dipende dal dominio: stringhe in un linguaggio, in biologia.... Basata magari sulla conoscenza pregressa di un esperto...

Ci chiediamo: è **possibile imparare la metrica?** Si.

Capitolo 11

Vari Modelli

11.1 Overview sul supervised learning

Spazi discreti, ad esempio: congiunzione di letterali, decision tree (regole proposizionali), grammatiche induttive, programmazione logica induttiva...

Spazi continui, ad esempio: Multiple Linear Regression, Linear Threshold Unit (LTU), regressione polinomiale, k-means, SVM, reti neurali...

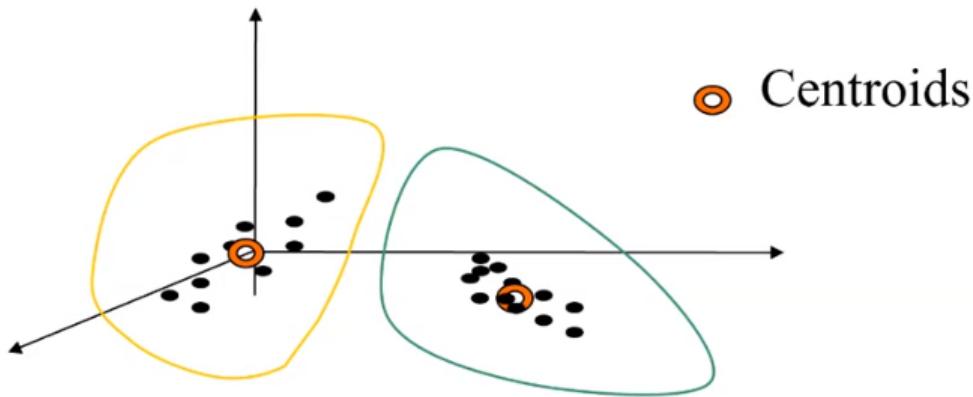
Probabilistici/generativi: modelli parametrici tradizionali (stima di densità, analisi del discriminante, regressione polinomia), modelli grafici (reti Bayesiane, Naive Bayes, modelli di Markov, modelli di Markov nascondenti...)

Instance-Based: nearest neighbors

11.2 Unsupervised learning

No teacher Nessun supervisore, TR è un insieme di dati senza etichetta $\langle x \rangle$

Si affrontano compiti diversi, come il raggruppamento naturale di insiemi di dati, ad esempio **clustering**: ripartire dati in gruppi, sottoinsieme di dati "simili". **Centroide**: un centro che rappresenta gli elementi del cluster.



Utilità del clustering In natura:

Tassonomia, per organizzare le forme di vita: cosa sono i mammiferi?

Apprendimento umano: i bambini imparano a riconoscere le facce familiari prima di comprendere il linguaggio umano

In **data analysis**:

Explorative data analysis, per scoprire la struttura latente e trovare dei pattern tipici

Preprocessing per altri approcci ML, es. riduzione della dimensionalità

Collezionare dati non etichettati è più facile che collezionare dati etichettati: ad esempio 10 milioni di immagini, da frame di YouTube, raggiungendo un riconoscimento degli oggetti allo stato dell'arte (22'000 categorie)

Clustering Ripartire dati in gruppi di dati simili. I pattern all'interno di un cluster valido sono più simili tra loro rispetto a qualsiasi altro pattern appartenente ad un altro cluster. (Con pattern si intende l'elemento del TR, quindi l'esempio di training)

Il centroide, o **prototipo**, è il "baricentro" del cluster: un po' come il rappresentante del cluster, alla minima distanza da tutti gli altri.

L'**obiettivo** è quindi ottenere una ripartizione dello spazio in x in regioni/cluster approssimate dal cluster center o **prototipo**.

H: insieme dei quantizzatori $x \rightarrow c(x)$ che rappresenta il cluster, da spazio continuo a spazio discreto. Esempio di funzione di loss: misurare l'ottimalità del vettore quantizzatore. Una funzione di loss comune potrebbe essere la **squared error distortion**

$$L(h(x_i)) = \|x_i - c(x_i)\|^2$$

Il valore medio sulla distribuzione degli input è la distorsione media, o ricostruzione, o **errore di quantizzazione**.

11.3 K-Means

Il più semplice e più usato algoritmo che implementa la squared error. Popolare perché facile da implementare e efficiente in generale, ma ha diverse limitazioni che vedremo.

L'**algoritmo** base è il seguente:

1. Si scelgono k cluster, che coincidono con k pattern scelti a caso o k punti all'interno del volume contenente i pattern
2. Assegniamo ogni pattern al cluster più vicino, cioè al cluster il cui centro è più vicino al pattern (il winner)
- ↑
3. Si ricalcolano i centri dei cluster (centroide geometrico, vale a dire la media) usando l'attuale appartenenza ai cluster
4. Se il criterio di convergenza non viene raggiunto, ritorno allo step due, cioè:

Non c'è, o è minimo, il riassetto dei nuovi pattern

Diminuzione minima nell'errore quadratico

Alcuni difetti

Come scegliere il winner: dati i centri dei cluster c_1, \dots, c_k (vettori di dimensione n , come x)

Per ogni x il winner si determina similmente al winner del k-NN: $i^*(x) = \arg \min_i \|x - c_i\|^2$, trovo i tale che la distanza minima (solo l'indice)

$$\|x - C_i\|^2 = \sum_j^n (x_j - c_{ij})^2$$

Ora x appartiene al cluster i^*

Per ogni cluster i , la nuova media (centroide) è:

$$c_i = \frac{1}{|\text{cluster}_i|} \cdot \sum_{j \mid x_j \in \text{cluster}_i} x_j$$

Limitazioni

k è scelto a priori: trial-and-error per trovare il k più adatto

I minimi locali della loss $L(x)$ rendono il metodo dipendente dall'inizializzazione.

Si può provare più volte con inizializzazioni differenti, magari tramite euristica

Può funzionare bene per cluster compatti e ipersferici, non altre forme

Non ha proprietà di visualizzazione, non consente di proiettare dati su un spazio di dimensione minore

Valutazione Come valutare un metodo di clustering? Più sottile rispetto al solito, perché non sappiamo in generale a priori come si dovrebbe comportare. Cosa caratterizza un "buon" clustering da un clustering poco interessante?

Soggettivo: non esistono "gold-standard", a parte sottodomini particolari

Le misure obiettive esistono, come l'errore di quantizzazione (ma dipende dal numero di centroidi, se metto un centroide per dato ho errore di quantizzazione pari a zero)

11.4 Dimensionality Reduction

Approccio al preprocessing Tra la moltitudine di approcci, si merita una menzione in questo corso il **dimensionality reduction** (unsupervised learning): $\langle x_1, x_2, \dots, x_n \rangle \rightarrow \langle x'_1, x'_2, \dots, x'_m \rangle$ con $n > m$

Si può ridurre la dimensionalità degli elementi mantenendone le informazioni principali. Questo si può fare selezionando alcune feature o **creando una combinazione**, ad esempio tramite il **Principal Component Analysis** (PCA)

PCA In nuovi assi sono computati tramite gli assi dove i dati variano di più.

Selezione delle feature Scegliere un sottoinsieme delle feature (feature engineering, selezionandole automaticamente in base a ridondanza o rilevanza nel task cioè le più informative). Tanti approcci, problema difficile tanto quanto il problema dell'apprendimento.

Outlier detection Individuare valori inusuali inconsistenti con la maggior parte delle osservazioni (es. a causa di errori di misurazione)

11.5 Altri task nell'ML

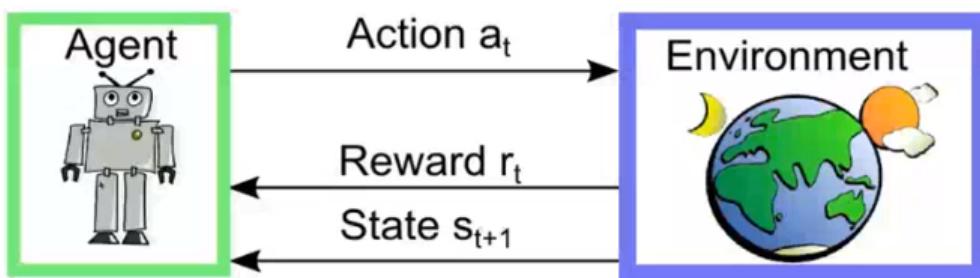
11.5.1 Reinforcement Learning

Apprendimento con criterio di giusto o sbagliato.

Si tratta nell'adattamento di sistemi autonomi (specialmente in robotica) rispetto al mondo. L'algoritmo apprende una politica che indica come agire in accordo ad una data osservazione del mondo. Ogni **azione** ha un impatto nell'ambiente, e l'**ambiente fornisce feedback** che guida l'algoritmo di apprendimento.

Invece di una supervisione in ogni passo, abbiamo informazione riguardo vittoria/perdita (**reward** o **punishment**) nello stato finale.

L'azione deve massimizzare la quantità di reward, e l'apprendimento decide quali azioni sono responsabili per le vittorie e le perdite



11.5.2 Semi-supervised learning

Combina esempi etichettati e non (tipicamente in numero maggiore) per generare una funzione/classificazione appropriata.

Learn to rank Esempio per i motori di ricerca, l'input è una serie di oggetti e l'output è la loro classifica

11.5.3 On-Line Learning

Nuovi esempi appresi e usati per fare il learning nel tempo

11.5.4 Apprendimento di domini strutturati e apprendimento relazionale

I domini di input/output possono essere strutturati sottoforma di sequenze (segnali, serie temporali...) o anche più complessi: alberi, grafi, reti.

Pattern di vettori con relazioni con altri componenti. Input: molecole, pagine web come grafi di reti... Output: nodi di una rete sociale, alberi di parsing da una frase....

11.5.5 Reti Neurali

Neural Networks Per learning supervisionato e non supervisionato.

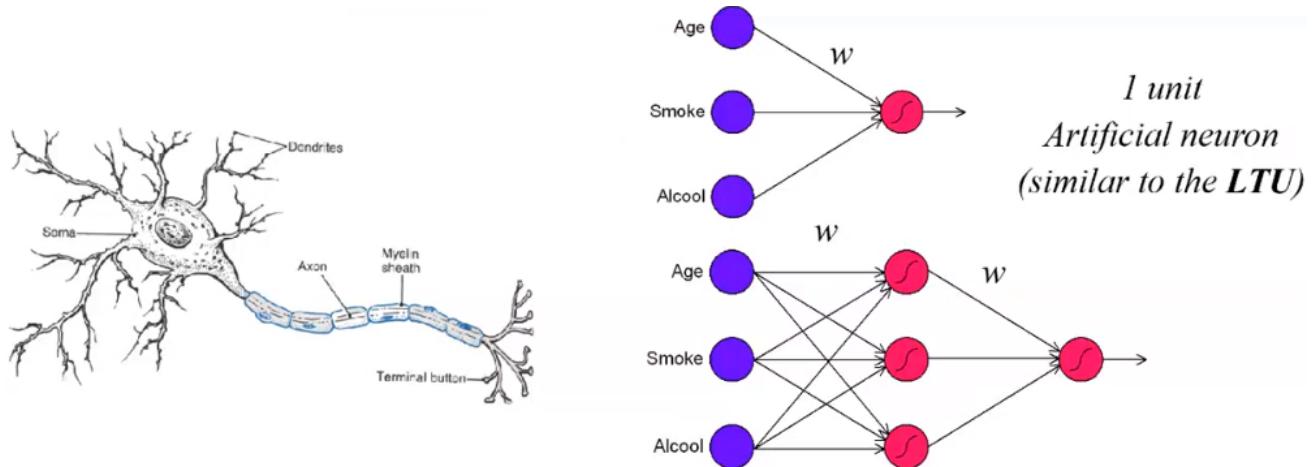
Lo abbiamo accennato: è vicino alla LTU (Linear Threshold Unit). La LTU è l'unità base di una **Artificial Neural Network: il perceptron**.

Una NN (neural network) è una rete di tali unità non lineari, con una capacità di approssimazione universale (qualsiasi tipo di funzione).



Si compone così un approccio molto flessibile per qualsiasi obiettivo del ML. Approccio più comune: gradient descent.

Ispirazione biologica Vengono studiate come paradigmi di calcolo sin dagli anni '40. Adesso rappresentano un insieme di potenti modelli di calcolo per approssimazione di funzioni con capacità predittive, supportate da una rigorosa fondazione teorica (learning theory)



Perché i layer interni? Le 3 f impilate nell'immagine precedente sono i cosiddetti **layer nascosti**.

Consentono al modello di estrarre, attraverso l'apprendimento, una nuova rappresentazione dei dati con l'apprendimento delle features.

La nuova rappresentazione semplifica il task di classificazione al layer finale (la f in fondo).

Possiamo considerarle come una linear basis expansion $\phi(x)$, dove le ϕ sono apprese e quindi dipendono anche da w $\phi(x,w)$

Questo comporta un problema di ottimizzazione non lineare, quindi i metodi diretti non funzionano più (ma la discesa di gradiente sì).

11.5.6 Deep Learning

DL Intuitivamente: architetture NN multilayer profonde



Aspetti simili tra approcci differenti

Livelli multipli di unità di processo non lineari

Apprendimento supervisionato o non supervisionato delle **feature representation** in ogni layer, con i livelli che formano una gerarchia da low-level a high-level features/representations (livelli differenti di astrazione)

Apprendere la rappresentazione dei dati Per esempio, è difficile dire quali sono le feature utili per riconoscere un volto.

Si scoprono in maniera automatica le feature necessarie. Si aumenta il livello di astrazione attraverso diversi layer. Ad esempio, un'immagine può essere una matrice di pixel, o in maniera più astratta come un insieme di bordi/aree, regioni dalla forma particolare...

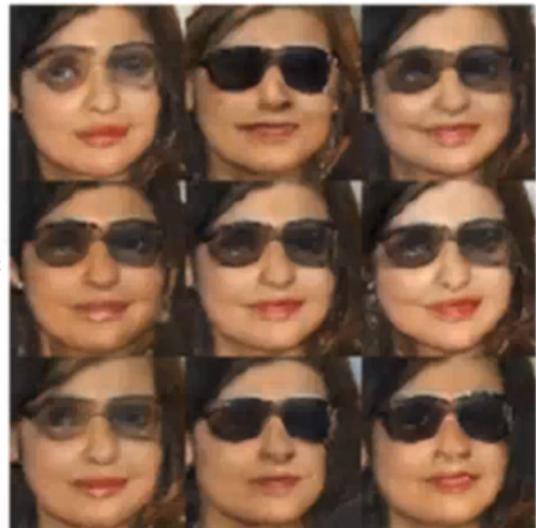
L'organizzazione viene scoperta **in maniera del tutto autonoma**.

Si comporta bene quando l'input ha qualche struttura (spaziale, temporale... come immagini, linguaggio, suoni...)

Riutilizzo Le feature di basso livello introdotte possono essere riusate per essere ricomposte in vari modi, per generalizzare esempi non visti. Ad esempio, combinando l'immagine di un uomo con occhiali con quella di una donna si può ottenere una donna con occhiali. Questo perché nella rappresentazione si sono **separati i concetti** di volto e occhiali. Combinazione geometrica.

Internal representations can be combined

$$\begin{array}{c} \text{[Man with glasses]} \\ - \\ \text{[Man without glasses]} \\ + \\ \text{[Woman's face]} \\ = \end{array}$$



Quindi non ha bisogno di avere tutte le combinazioni da vedere, ma può generarle da sé.

Vantaggi Sfruttano la composizionalità della rappresentazione interna, ottenendo un guadagno esponenziale nella potenza di rappresentazione. Questo grazie al fatto che i concetti semplici sono rappresentati nei layer della rete e possono essere usati come primitive dal layer successivo per rappresentare concetti più complessi, evitando rappresentazione combinatorie esplicite e l'apprendimento delle features.

Quindi **meno esempi necessari per avere una buona capacità di generalizzazione**.

Risultati Le reti neurali profonde sono sempre esistite, ma erano difficili da istruire in passato. Combinando le nuove tecniche di apprendimento per modelli così grandi, high-performance computing (es. GPU) e collezioni di dati molto grandi sia dall'industria che da applicazioni reali (es. milioni di immagini).

Oggi funzionano molto bene.

Capitolo 12

Applicazioni e oltre II A

Quando applicare ML Per risolvere problemi reali difficili da trattare tradizionalmente: complementari a programmazione classica, analisi e algoritmi.

Esempi: face recognition, voice automation...

Utile anche in casi dove non c'è sufficiente conoscenza umana (es. predire la forza dei legami delle molecole delle proteine) o comportamenti personalizzati (es. classificare mail o pagine web in accordo alle preferenze utente).

Esempi famosi Riconoscere lo ZIP code, OCR (riconoscimento caratteri scritti a mano) → MNIST, dataset benchmark per riconoscimento OCR.

GO (DeepMind, Google): enorme spazio di ricerca, difficile valutare posizione sui bordi. Value Network per valutare le posizioni e politica per scegliere le mosse. Queste reti neurali deep allenate giocando contro umani o contro sé stesse. Si introduce un nuovo algoritmo di ricerca che combina Monte Carlo con value e policy network, algoritmo collegato a minimax e alpha-beta

Veicoli a guida autonoma.

Sfide generali Costruire macchine autonome intelligenti/che apprendono (robotica, HRI, motori di ricerca)
Costruire strumenti potenti per le nuove sfide in data analysis

Aprire nuove aree in CS: problemi interdisciplinari innovativi (il limite è la fantasia, ML nell'era di un cambiamento nel paradigma della scienza dove le innovazioni sono sempre più data-driven)

Sfide Bio-chemioinformatics, medicina personalizzata, tossicologia. Problemi aperti enormi che richiedono strumenti di analisi molto flessibili.

Dati web, fino a social e big data.

Crescita dei dati aprono le porte ad aree di ricerca molto grandi.

12.0.1 Pattern Recognition

Assegnare etichetta ad un input/segnale tramite machine learning. Difficile da formalizzare, dati ambigui...

12.0.2 Data Mining/Knowledge Discovery

Da intersezione fra ML e database.

KDD: data management e preprocessing, come estrarre, interpretare e valutare le informazioni, oltre aspetti etici e di privacy.

DM: aspetto descrittivo, clustering e ricerca di pattern.

Nel ML focus sul learning e metodi, in DM ho focus su dati/pattern.

12.0.3 Computational Intelligence

Area in relazione alla IA, sottoarea specificamente collegata a metodi legati al trattamento di problemi complessi con tolleranza ad imprecisione, incertezza, approssimazione... il cosiddetto **soft computing**.

12.0.4 Deep Learning

Estrarre la conoscenza direttamente dai dati: representation learning. Gerarchia...