

# QUESTI APPUNTI NON SONO ANCORA FINITI



PER ALTRI APPUNTI CONSULTARE IL SITO:

[https://luigi-v.github.io/Appunti\\_Universita/](https://luigi-v.github.io/Appunti_Universita/)

## 10. CONOSCENZA INCERTA E RAGIONAMENTO

Dagli agenti logici passiamo a scenari nei quali è presente molta incertezza così da non riuscire a gestirli con un approccio formale come quello logico. Con la logica riusciamo a fare ragionamenti e riusciamo a dedurre nuova conoscenza andando a modellare la KB. Negli scenari più reali, risulta difficile applicare approcci logici, siccome la realtà racchiude incertezze. In questo caso, l'incertezza viene modellata attraverso la probabilità, quindi avremo non più un modello logico ma un modello probabilistico.

### Esempio:

Sia  $A_t = \text{avviamoci all'aeroporto } t \text{ minuti prima del volo}$ . Mi permetterà  $A_t$  di giungere in tempo per il volo?

Problemi:

- Osservabilità parziale (stato della strada, piani di altri autisti, etc);
- Sensori rumorosi (sensori che percepiscono informazioni errate o disturbate);
- Incertezza nei risultati dell'azione (foratura gomma);
- Elevata complessità nella modellazione e nella predizione del traffico.

Pertanto, un approccio puramente logico non è fattibile:

- Comporta rischio di falsità: “ $A_{25}$  mi porterà in tempo a destinazione”, oppure
- Porta a conclusioni troppo deboli per il decision making:
  - $A_{25}$  mi porterà a destinazione in tempo a patto che non ci siano incidenti sul ponte, non piove, gomme integre etc.
  - Ragionevolmente si può dire che  $A_{1440}$  mi porterebbe a destinazione in tempo, ma dovrei pernottare in aeroporto.

In tal caso abbiamo tante possibilità corrette, ma l'agente deve fornire un'unica risposta.

### **DECISIONE RAZIONALE:**

Un agente logico non è in grado di agire poiché non conosce con quali azioni raggiungere l'obiettivo. L'informazione in possesso dell'agente non può garantire i possibili esiti di  $A_{90}$ , ma può fornire un grado di credenza sul loro raggiungimento.

La cosa giusta da fare dipende:

1. Dall'importanza relativa ai vari obiettivi;
2. Dalla probabilità e dalla misura del loro raggiungimento.

### **INADEGUATEZZA DELL'APPROCCIO LOGICO:**

Consideriamo un esempio di diagnosi medica:

$$\forall p \text{ Sintomo}(p, \text{MalDiDenti}) \Rightarrow \text{Malattia}(p, \text{Carie})$$

Sbagliato! Non tutti i pazienti che accusano mal di denti hanno carie:

$$\forall p \text{ Sintomo}(p, \text{MalDiDenti}) \Rightarrow \text{Malattia}(p, \text{Carie}) \vee \text{Malattia}(p, \text{Gengivite}) \vee \text{Malattia}(p, \text{Ascesso})....$$

Elenco lunghissimo di cause. Regola causale:

$$\forall p \text{ Malattia}(p, \text{Carie}) \Rightarrow \text{Sintomo}(p, \text{MalDiDenti})$$

Non tutte le carie causano dolore. Bisogna elencare tutte le cause del mal di denti sul lato sinistro.

### **PROBABILITÀ:**

Il mondo cambia da proposizioni che sono vere o false a proposizioni che hanno un certo **grado di credenza** (rappresentato da un valore di probabilità) del modello di agente. Date le evidenze disponibili so che  $A_{25}$  mi porterà in tempo a destinazione con probabilità 0.04. Le asserzioni probabilistiche sintetizzano gli effetti di:

- **Pigrizia:** mancata enumerazione di eccezioni, condizioni, etc., sia perché richiede troppo lavoro, sia perché le regole risulterebbero difficili da usare;
- **Ignoranza teorica:** assenza di fatti rilevanti. Esempio la scienza medica non ha una teoria completa per il suo dominio;
- **Ignoranza pratica:** anche se conosciamo tutte le regole potremmo essere incerti perché non sono state fatte tutte le misurazioni.

Nello scenario che andremo ad analizzare ora, l'agente non ha più certezze in quanto fa uso delle probabilità.

### **DIFFERENZE CON ONTOLOGIE SPECIFICHE:**

#### **Probabilità soggettiva:**

- Le probabilità mettono in relazione proposizioni e stato della conoscenza dell'agente (probabilità condizionate), cioè:

$$P(A_{25} | \text{nessun incidente}) = 0.06$$

Queste non sono asserzioni sul mondo reale.

- Le probabilità di proposizioni cambiano con nuove evidenze, cioè:

$$P(A_{25} | \text{nessun incidente, alle 5 a.m.}) = 0.15$$

Il grado di credenza è **diverso** dal grado di verità. Le formule sono sempre vere o false. Una probabilità di 0.8 indica un grado di credenza nella verità dell'80%.

## PRENDERE DECISIONI INCERTE:

Supponiamo che io abbia le seguenti convinzioni:

$$P(A_{25} \text{ mi porta in tempo a destinazione} | \dots) = 0.04$$

$$P(A_{90} \text{ mi porta in tempo a destinazione} | \dots) = 0.70$$

$$P(A_{120} \text{ mi porta in tempo a destinazione} | \dots) = 0.95$$

$$P(A_{1440} \text{ mi porta in tempo a destinazione} | \dots) = 0.9999$$

**NOTA:** i numeri come pedice di A rappresenta t ( $A_t$  = avviamoci all'aeroporto t minuti prima del volo).

Il tempo non è l'unico fattore, dipende anche dalle mie preferenze tra perdere il volo rispetto a passare del tempo ad attenderlo, etc.

La Teoria dell'Utilità viene usata per rappresentare ed inferire preferenze. Assieme alla Teoria della Probabilità forma:

**Teoria delle Decisioni = Teoria della Probabilità + Teoria dell'Utilità**

$$\text{Maximize expected utility : } a^* = \underset{a}{\operatorname{argmax}} \sum_s P(s | a) U(s)$$

## DECISION-THEORETIC AGENT CHE SELEZIONA AZIONI RAZIONALI:

Un agente che fa uso della teoria della decisione funziona nel modo seguente:

L'agente prendendo in input una percezione, calcola le probabilità per le azioni, dopo sceglie quella che ci dà la **highest expected utility**.

```
function DT-AGENT(percept) returns an action
  persistent: belief_state, probabilistic beliefs about the current state of the world
              action, the agent's action

  update belief_state based on action and percept
  calculate outcome probabilities for actions,
    given action descriptions and current belief_state
  select action with highest expected utility
    given probabilities of outcomes and utility information
  return action
```

## SINTASSI DI PROPOSIZIONI:

Una **variabile casuale** può assumere diversi valori, ed ogni valore ha una certa probabilità. Le variabili casuali si dividono in:

- **Booleans**, possono assumere o TRUE o FALSE;
- **Discrete**, assumono più di due valori possibili. Esempio il *tempo atmosferico* che può assumere <soleggiato, piovoso, nuvoloso, neve>;
- **Continuous**, esprimono una distribuzione come una funzione parametrizzata di valori (che non vengono trattate);

I **valori dei domini** sono **esaurivi**, cioè sono presenti tutti i possibili valori, e sono **mutuamente esclusivi**.

Le **proposizioni elementari**, costruite assegnando un valore ad una variabile casuale:

$$\text{Tempo} = \text{soleggiato}, \text{Carie} = \text{Falso} (\neg \text{carie}).$$

Le **proposizioni complesse**, formate da proposizioni elementari e connettivi logici standard:

$$\text{Tempo} = \text{soleggiato} \vee \text{Carie} = \text{falso}.$$

Un **evento atomico** rappresenta una specifica completa del mondo dell'agente che è incerto.

Se il mondo consiste solo di 2 variabili booleane Carie e MalDiDenti, allora ci sono 4 eventi atomici distinti:

$$\begin{aligned} \text{Carie} = \text{falso} \wedge \text{MalDiDenti} = \text{falso} \\ \text{Carie} = \text{falso} \wedge \text{MalDiDenti} = \text{vero} \\ \text{Carie} = \text{vero} \wedge \text{MalDiDenti} = \text{falso} \\ \text{Carie} = \text{vero} \wedge \text{MalDiDenti} = \text{vero} \end{aligned}$$

Gli eventi atomici sono mutuamente esclusivi (massimo 1 si verifica) ed esaurivi (almeno 1).

## PROBABILITÀ A PRIORI:

Sono probabilità secondo le quali una certa variabile casuale assuma un certo valore senza avere nessun'altra conoscenza.

- $P(\text{Carie}=\text{ver})=0$  o  $P(\text{Tempo}=\text{soleggiato})=0.72$  corrisponde alla confidenza prima dell'arrivo di qualunque (nuova) evidenza.

Ogni possibile mondo  $w$  è associato con un valore di probabilità tale che:

$$0 \leq P(w) \leq 1$$

$$\sum_{w \in \Omega} P(w) = 1$$

Esempio: se lanciamo due dadi (distinguibili) ci sono 36 possibili mondi da considerare: (1,1), (1,2), ..., (1,6)  $P(w)=1/36$

Una **distribuzione di probabilità** fornisce valori per tutti i possibili assegnamenti:

- Es  $P(\text{tempo}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$  (normalizzata, i.e., somma a 1)

$$P(\text{Dadi}) = \langle 1/36, \dots, 1/36 \rangle$$

Una **distribuzione di probabilità congiunta** per un insieme di variabili casuali fornisce le probabilità di ogni evento atomico di tali variabili.

### Esempio:

$P(\text{Tempo}, \text{Carie})$  = una matrice  $4 \times 2$  con valori:

<b>Tempo</b> =	<b>soleggiato</b>	<b>piovoso</b>	<b>nuvoloso</b>	<b>neve</b>
<i>Carie</i> = vero	0.144	0.02	0.016	0.02
<i>Carie</i> = falso	0.576	0.08	0.064	0.08

Ogni domanda su un dominio può essere risposta con una distribuzione congiunta.

### **PROBABILITÀ CONDIZIONATA:**

La **probabilità condizionata** è la probabilità che una variabile (es *a*) assume un certo valore sapendo già che un dato evento è accaduto (*b*):

$$P(a | b) = P(a \wedge b) / P(b) \quad \text{if } P(b) > 0 \quad \left| P(\text{doubles} | \text{Die}_1=5) = \frac{P(\text{doubles} \wedge \text{Die}_1=5)}{P(\text{Die}_1=5)} \right.$$

Questa probabilità condizionata è importante per un agente, siccome esso parte da una data conoscenza.

La **regola del prodotto** offre una formulazione alternativa:

$$P(a \wedge b) = P(a | b) P(b) = P(b | a) P(a)$$

Questa regola ci permette di definire la probabilità che due eventi sono accaduti.

### Esempio:

$$P(\text{Tempo}, \text{Carie}) = P(\text{Tempo} | \text{Carie}) P(\text{Carie})$$

(Vista come un insieme di  $4 \times 2$  equazioni, non più come una matrice multipla)

La regola del prodotto anziché avere solo due variabili possiamo generalizzarla ad *n*, questa è chiamata **Chain rule**, che è derivata attraverso successive applicazioni della regola del prodotto:

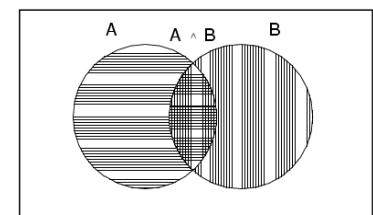
$$\begin{aligned} P(X_1, \dots, X_n) &= P(X_1, \dots, X_{n-1}) P(X_n | X_1, \dots, X_{n-1}) \\ &= P(X_1, \dots, X_{n-2}) P(X_{n-1} | X_1, \dots, X_{n-2}) P(X_n | X_1, \dots, X_{n-1}) \\ &= \dots \\ &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

### **ASSIOMI DI PROBABILITÀ:**

Per ogni coppia di proposizioni A, B:

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$  and  $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Questi 3 assiomi sono la base su cui si può dimostrare tutto che è stato visto.



### **INFERENZA TRAMITE ENUMERAZIONE:**

Adesso ciò che interessa è implementare un agente che risponde a delle domande definendo un processo inferenziale, facendo un'inferenza per enumerazione, cioè enumerando tutti i possibili valori.

L'approccio si basa sulla **distribuzione congiunta di probabilità**:

	<b>malidenti</b>		$\neg$ <b>malidenti</b>	
	$\text{prende}$	$\neg \text{prende}$	$\text{prende}$	$\neg \text{prende}$
<b>carie</b>	.108	.012	.072	.008
$\neg$ <b>carie</b>	.016	.064	.144	.576

La probabilità di una proposizione  $\phi$  è data dalla somma degli **eventi atomici**  $\omega$  su cui  $\phi$  diventa vera:

$$P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$$

### Esempio:

Proposizione: *malidenti* =vera, qual è la probabilità che l'evento *malidenti* sia vero?

Bisogna sommare tutti gli eventi atomici in cui il *malidenti* è vero:

$$P(\text{malidenti}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2 \quad (\text{probabilità marginale})$$

Oppure, qual è la probabilità che l'evento *caries* sia vero?

$$P(\text{caries}) = 0.108 + 0.012 + 0.072 + 0.008$$

Spesso siamo interessati a calcolare le **probabilità condizionali** di alcune variabili, date le evidenze su altre.

### Esempio:

Proposizione: Quale è la probabilità di avere *carie* sapendo di avere *maldimenti*?

$$P(\text{carie} | \text{maldimenti}) = P(\text{carie} \wedge \text{maldimenti}) / P(\text{maldimenti}) = (0.108 + 0.012) / (0.108 + 0.012 + 0.016 + 0.064) = 0.6$$

Il suo negato è il complemento:

$$P(\neg \text{carie} | \text{maldimenti}) = P(\neg \text{carie} \wedge \text{maldimenti}) / P(\text{maldimenti}) = 0.4$$

### **NORMALIZZAZIONE:**

Nei calcoli, il denominatore è lo stesso e può essere visto come una **costante di normalizzazione**  $\alpha$ . Serve che il risultato è compreso tra 0 e 1 e che la somma sia 1.

$$\begin{aligned} P(\text{Carie} | \text{maldimenti}) &= \alpha P(\text{Carie}, \text{maldimenti}) \\ &= \alpha [P(\text{Carie}, \text{maldimenti}, \text{prende}) + P(\text{Carie}, \text{maldimenti}, \neg \text{prende})] \\ &= \alpha [0.108, 0.016] + [0.012, 0.064] \\ &= \alpha [0.12, 0.08] = [0.6, 0.4] \quad \text{Non ci serve conoscere } P(\text{maldimenti})! \end{aligned}$$

L'idea generale è calcolare la distribuzione sulla variabile di query fissando variabili di evidenze (maldimenti) e sommando variabili nascoste (prende).

### **INFERENZA PER ENUMERAZIONE:**

In genere, siamo interessati a:

- La distribuzione congiunta a posteriori delle variabili di query X (Carie nell'esempio);
- Dati valori specifici e per le variabili di evidenza E (MalDiDenti nell'esempio).

Siano le **variabili nascoste H** = Y – X – E, il risultato richiesto è ottenuto sommando le variabili nascoste:

$$P(X | E=e) = \alpha P(X, E=e) \alpha \sum_h P(X, E=e, H=h)$$

I termini nella sommatoria rappresentano entry congiunte, perché X, E ed H insieme esauriscono l'insieme di variabili casuali.

### **INFERENZA PROBABILISTICA:**

X rappresenta le query, e rappresenta le evidenze e P la distribuzione congiunta.

Il problema di questo algoritmo è la distribuzione congiunta di variabili perché la dimensione è molto grande.

Quindi **complessità di tempo** nel caso peggiore  $O(d^n)$  dove d è la più grande aritma.

**Complessità di spazio**  $O(d^n)$  per memorizzare la distribuzione congiunta.

Come trovare i numeri per  $O(d^n)$  entry?

```
function ENUMERA-CONGIUNTA-ASK(X, e, P) returns una distribuzione su X
    inputs: X, la variabile della query
            e, i valori osservati per le variabili E
            P, una distribuzione congiunta sulle variabili {X} ∪ E ∪ Y
            /* Y = variabili nascoste */
    Q(X) ← una distribuzione su X, inizialmente vuota
    for each valore  $x_i$  di X do
         $Q(x_i) \leftarrow \text{ENUMERA-CONGIUNTA}(x_i, e, Y, [], P)$ 
    return NORMALIZZA(Q(X))

function ENUMERA-CONGIUNTA(x, e, variabili, valori, P) returns un numero reale
    if VUOTA(variabili) then return P(x, e, valori)
    Y ← PRIMO(variabili)
    return  $\sum_y \text{ENUMERA-CONGIUNTA}(x, e, RESTO(variabili), [y|valori], P)$ 
```

### **INDIPENDENZA:**

Quando gli eventi sono indipendenti possono essere trattati parallelamente. A e B sono **indipendenti** se e solo se:

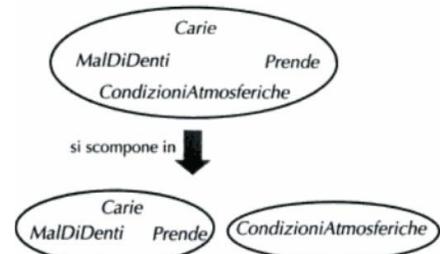
$$P(A|B)=P(A) \text{ or } P(B|A)=P(B) \text{ or } P(A,B)=P(A)P(B).$$

### Esempio:

$$P(\text{maldimenti}, \text{prende}, \text{carie}, \text{condizioni}) = P(\text{maldimenti}, \text{prende}, \text{carie})P(\text{condizioni})$$

Da 32 entry si riducono a 12; per n lanci di moneta indipendenti,  $O(2^n) \rightarrow O(n)$ .

L'indipendenza assoluta è molto potente ma rara nella realtà.



## INDIPENDENZA CONDIZIONALE:

Una cosa che succede più frequente è l'indipendenza condizionale che suppone che gli eventi non siano del tutto indipendenti tra loro.

Esempio:

$$P(MalDiDenti, Prende, Carie) \text{ ha } 2^3 - 1 = 7 \text{ entry indipendenti}$$

Se ho una carie, la probabilità che lo strumento appuntito si blocca non dipende dal fatto che io abbia mal di denti:

$$(1) P(\text{prende} | \text{maldidenti, carie}) = P(\text{prende} | \text{carie})$$

La stessa indipendenza vale se io non ho una carie:

$$(2) P(\text{prende} | \text{maldidenti, } \neg \text{carie}) = P(\text{prende} | \neg \text{carie})$$

Prende è **condizionalmente indipendente** da MalDiDenti data Carie:

$$P(\text{Prende} | \text{MalDiDenti, Carie}) = P(\text{Prende} | \text{Carie})$$

Affermazioni equivalenti:

$$P(\text{MalDiDenti} | \text{Prende, Carie}) = P(\text{MalDiDenti} | \text{Carie})$$

$$P(\text{MalDiDenti}, \text{Prende} | \text{Carie}) = P(\text{MalDiDenti} | \text{Carie}) P(\text{Prende} | \text{Carie})$$

Decomposizione della distribuzione congiunta completa tramite **chain rule**:

$$\begin{aligned} & P(\text{MalDiDenti}, \text{Prende}, \text{Carie}) \\ &= P(\text{MalDiDenti} | \text{Prende, Carie}) P(\text{Prende}, \text{Carie}) \\ &= P(\text{MalDiDenti} | \text{Prende, Carie}) P(\text{Prende} | \text{Carie}) P(\text{Carie}) \\ &= P(\text{MalDiDenti} | \text{Carie}) P(\text{Prende} | \text{Carie}) P(\text{Carie}) \end{aligned}$$

cioè  $2 + 2 + 1 = 5$  entry indipendenti

Nella maggior parte dei casi, l'uso dell'indipendenza condizionale riduce la dimensione della rappresentazione della distribuzione congiunta da esponenziale in n a lineare in n.

L'indipendenza condizionale è la nostra forma più semplice e solida di conoscenza di ambienti incerti.

## REGOLA DI BAYES:

La **regola di Bayes** può essere usata nell'inferenza, viene semplicemente riscritta delle regole condizionali precedenti:

$$\text{Regola del prodotto: } P(a \wedge b) = P(a | b) P(b) = P(b | a) P(a) \quad \rightarrow \quad \text{Bayes' rule: } P(a | b) = P(b | a) P(a) / P(b)$$

Oppure, in forma distribuita:

$$P(Y | X) = P(X | Y) P(Y) / P(X) = \alpha P(X | Y) P(Y)$$

Perché è utile:

- Costruiamo un condizionale dal suo inverso;
- Spesso un condizionale è complicato ma l'altro è semplice;
- Descrive un passaggio di "aggiornamento" dal precedente  $P(a)$  al successivo  $P(a | b)$ .

Utile per valutare la probabilità diagnostica dalla **probabilità causale**:

$$P(\text{Cause} | \text{Effect}) = P(\text{Effect} | \text{Cause}) P(\text{Cause}) / P(\text{Effect})$$

- $P(\text{Effect} | \text{Cause})$  descrive la direzione causale;
- $P(\text{Cause} | \text{Effect})$  descrive la relazione diagnostica.

La regola di Bayes è molto utilizzata perché quando si hanno le probabilità condizionate si hanno delle variabili che sono *cause* e altre che sono *effetto* di quelle cause, il poter invertire l'ordine aiuta siccome si conosce più la probabilità di avere un effetto data una cerca causa.

Esempio diagnosi medica (regola di Bayes):

Dai casi passati sappiamo che:  $P(\text{symptoms} | \text{disease})$ ,  $P(\text{disease})$ ,  $P(\text{symptoms})$ .

Per un nuovo paziente conosciamo i sintomi e cerchiamo una diagnosi quindi  $P(\text{disease} | \text{symptoms})$ :

- La meningite provoca un torcicollo il 70% delle volte;
- La probabilità a priori di meningite è 1/50000;
- La probabilità a priori di torcicollo è 1%.

Qual è la probabilità che un paziente con un torcicollo abbia la meningite?

$$P(m | s) = P(s | m) * P(m) / P(s) = 0.7 * 1/50000 / 0.01 = 0.0014$$

Perché la probabilità condizionale per la direzione diagnostica non viene memorizzata direttamente?

- La conoscenza diagnostica è spesso più fragile ai cambiamenti di valori alle variabili rispetto alla conoscenza causale;

Per esempio, se c'è un'improvvisa epidemia di meningite, la probabilità incondizionata di meningite  $P(m)$  salirà; quindi, anche  $P(m | s)$  dovrebbe salire mentre la relazione casuale  $P(s | m)$  non è influenzata dall'epidemia, poiché riflette come funziona la meningite.

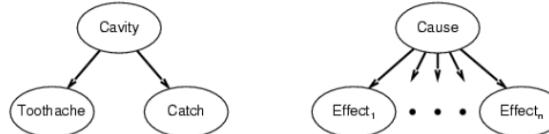
## INDIPENDENZA CONDIZIONALE (REGOLA DI BAYES):

Assume che tutti gli effetti siano indipendenti tra di loro, e lo rende molto semplicistico.

$$\begin{aligned} \mathbf{P}(\text{Carie} \mid \text{malidenti} \wedge \text{prende}) \\ = \alpha \mathbf{P}(\text{malidenti} \wedge \text{prende} \mid \text{Carie}) \mathbf{P}(\text{Carie}) \\ = \alpha \mathbf{P}(\text{malidenti} \mid \text{Carie}) \mathbf{P}(\text{prende} \mid \text{Carie}) \mathbf{P}(\text{Carie}) \end{aligned}$$

Questo è un esempio di modello **naïve Bayes**:

$$\mathbf{P}(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = \mathbf{P}(\text{Cause}) \prod \mathbf{P}(\text{Effect}_i \mid \text{Cause})$$



Il numero totale di parametri è **lineare** in n.

### Esempio nel mondo del Wumpus:

Abbiamo un labirinto con pozzi che vengono rilevati nei quadrati vicini attraverso il segnale brezza. Ogni cella contiene un pozzo con probabilità 0.2 (eccetto (1,1)). Dove dovrebbe andare l'agente, se c'è brezza a (1,2) e (2,1)? La pura inferenza logica non può concludere nulla su quale quadrato sia più probabile che sia sicuro!

Assumiamo di **sapere** che:

$$\begin{aligned} b &= \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1} \\ \text{known} &= \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1} \end{aligned}$$

Siamo interessati a rispondere a **query** come:

$$\mathbf{P}(P_{1,3} \mid \text{known}, b)$$

La **risposta** può essere calcolata elencando l'intera distribuzione di probabilità congiunta.

Siano Unknown le variabili  $P_{i,j}$  eccetto  $P_{1,3}$  e Known:

$$\mathbf{P}(P_{1,3} \mid \text{known}, b) = \sum_{\text{Unknown}} \mathbf{P}(P_{1,3}, \text{unknown}, \text{known}, b)$$

Ma significa esplorare tutti i possibili valori delle variabili sconosciute e ci sono  $2^{12} = 4096$  termini (crescita esponenziale nel numero di stanze).

Possiamo farlo meglio (più velocemente)?

### Variabili casuali booleane:

$$\begin{aligned} P_{ij} &= \text{pozzo nella casella } (i,j) \\ B_{ij} &= \text{brezza nella casella } (i,j) \\ (\text{solamente per le caselle osservate } B_{1,1}, B_{1,2} \text{ e } B_{2,1}) \end{aligned}$$

1.4	2.4	3.4	4.4
1.3	2.3	3.3	4.3
1.2 B	2.2	3.2	4.2
1.1	2.1 B	3.1	4.1

### Distribuzione completa delle probabilità congiunte

$$\mathbf{P}(P_{1,2}, \dots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1}) = \mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,2}, \dots, P_{4,4}) * \mathbf{P}(P_{1,2}, \dots, P_{4,4})$$

$$\mathbf{P}(P_{1,2}, \dots, P_{4,4}) = \prod_{i,j} \mathbf{P}(P_{ij})$$

Regola del prodotto

$$\mathbf{P}(P_{1,2}, \dots, P_{4,4}) = 0.2^n * 0.8^{16-n}$$

I pozzi sono distribuiti indipendentemente

la probabilità di pozzi è 0,2 e ci sono n pozzi

### Osservazione (Indipendenza condizionale):

Le brezze osservate sono condizionatamente indipendenti dalle altre variabili date le variabili note (bianco), di frontiera (giallo) e di query.

Dividiamo l'insieme delle variabili nascoste in frontiera e altre variabili:

$$\text{Unknown} = \text{Fringe} \cup \text{Other}$$

Dall'indipendenza condizionale abbiamo:

$$\mathbf{P}(b \mid P_{1,3}, \text{known}, \text{unknown}) = \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe})$$

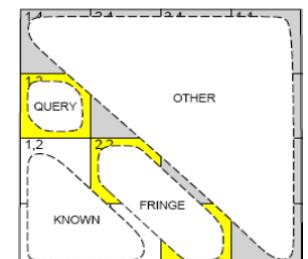
Ora, sfruttiamo questa formula.

Query iniziale ->

$$\mathbf{P}(P_{1,3} \mid \text{known}, b)$$

$$\begin{aligned} &= \alpha \sum_{\text{Unknown}} \mathbf{P}(P_{1,3}, \text{known}, \text{unknown}, b) \\ &= \alpha \sum_{\text{Unknown}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{unknown}) * \mathbf{P}(P_{1,3}, \text{known}, \text{unknown}) \\ &= \alpha \sum_{\text{Fringe}} \sum_{\text{Other}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe}, \text{other}) * \mathbf{P}(P_{1,3}, \text{known}, \text{fringe}, \text{other}) \\ &= \alpha \sum_{\text{Fringe}} \sum_{\text{Other}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe}) * \mathbf{P}(P_{1,3}, \text{known}, \text{fringe}, \text{other}) \\ &= \alpha \sum_{\text{Fringe}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe}) * \sum_{\text{Other}} \mathbf{P}(P_{1,3}, \text{known}, \text{fringe}, \text{other}) \\ &= \alpha \sum_{\text{Fringe}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe}) * \sum_{\text{Other}} \mathbf{P}(P_{1,3}) \mathbf{P}(\text{known}) \mathbf{P}(\text{fringe}) \mathbf{P}(\text{other}) \\ &= \alpha \mathbf{P}(\text{known}) \mathbf{P}(P_{1,3}) \sum_{\text{Fringe}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe}) \mathbf{P}(\text{fringe}) \sum_{\text{Other}} \mathbf{P}(\text{other}) \\ &= \alpha' \mathbf{P}(P_{1,3}) \sum_{\text{Fringe}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe}) \mathbf{P}(\text{fringe}) \end{aligned}$$

$$\alpha' = \alpha \cdot \mathbf{P}(\text{known}) \sum_{\text{Other}} \mathbf{P}(\text{other}) = 1$$



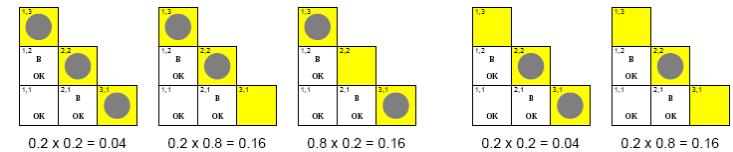
### Soluzione:

$$\mathbf{P}(P_{1,3} \mid \text{known}, b) = \alpha' \mathbf{P}(P_{1,3}) \sum_{\text{Fringe}} \mathbf{P}(b \mid P_{1,3}, \text{known}, \text{fringe}) \mathbf{P}(\text{fringe})$$

Esploriamo i possibili modelli (valori) di frontiera compatibili con l'osservazione b.

$$\begin{aligned} \mathbf{P}(P_{1,3} \mid \text{known}, b) \\ = \alpha' \langle 0.2(0.04 + 0.16 + 0.16), 0.8(0.04 + 0.16) \rangle \\ = \langle 0.31, 0.69 \rangle \end{aligned}$$

$$\mathbf{P}(P_{2,2} \mid \text{known}, b) = \langle 0.86, 0.14 \rangle$$



## 11. RAGIONAMENTO PROBABILISTICO

Riassumendo, il problema è gestire l'incertezza, possiamo gestirla utilizzando la probabilità e fare inferenza, ma questo è molto oneroso siccome è esponenziale rispetto al numero di variabili e quindi tutto ciò non è fattibile nella pratica.

Possiamo ridurre questo tempo eccessivo con le proprietà di indipendenza delle distribuzioni di probabilità, in particolare, l'indipendenza delle variabili è un evento molto raro però l'indipendenza condizionata è più frequente e questo ci permette di ottenere vantaggi nella computazione.

Le **reti bayesiane** mettono in pratica questi vantaggi sottoforma di un formalismo grafico che permette di andare a rappresentare questi problemi in termini di variabili che dipendono una dall'altra.

I grafici al lato riassumono quello visto nei precedenti capitoli, tra cui distribuzioni di probabilità congiunta dove bisogna fare la sommatoria su tutti i valori delle altre variabili di istanza, ma anche la regola di Bayes sull'indipendenza condizionata che ci permetteva di invertire la formula.

**Modelli Naive Bayes:** modelli nei quali si assume indipendenza condizionale tra le varie variabili, modello semplicistico con una singola variabile che dipende da tutte le altre.

Basic laws:  $0 \leq P(\omega) \leq 1$ ,  $\sum_{\omega \in \Omega} P(\omega) = 1$ ,  $P(A) = \sum_{\omega \in A} P(\omega)$

Random variable  $X(\omega)$  has a value in each  $\omega$

- ▶ Distribution  $P(X)$  gives probability for each possible value  $x$
- ▶ Joint distribution  $P(X,Y)$  gives total probability for each combination  $x,y$

Summing out/marginalization:  $P(X=x) = \sum_y P(X=x, Y=y)$

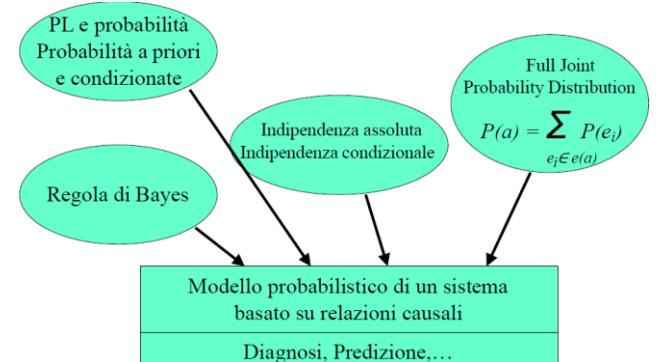
Conditional probability:  $P(Y|X) = P(X,Y)/P(X)$

Chain rule:  $P(X_1, \dots, X_n) = \prod_i P(X_i | X_1, \dots, X_{i-1})$

Bayes Rule:  $P(X|Y) = P(Y|X)P(X)/P(Y)$

Independence:  $P(X,Y) = P(X)P(Y)$  or  $P(X|Y) = P(X)$  or  $P(Y|X) = P(Y)$

Conditional Independence:  $P(X|Y,Z) = P(X|Z)$  or  $P(X,Y|Z) = P(X|Z)P(Y|Z)$



$$P(MalDiDenti, Prende, Carie) =$$

$$P(MalDiDenti | Carie) P(Prende | Carie) P(Carie)$$

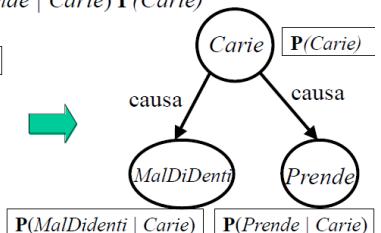
$$P(MalDiDenti, Prende, Carie)$$

	toothache	¬ toothache		
catch	.108	.012	.072	.008
cavity	.016	.064	.144	.576
¬ cavity				

Bayes

$$P(Cause|Effect) =$$

$$\frac{P(Effect|Cause) P(Cause)}{P(Effect)}$$



$$P(MalDiDenti | Prende, Carie) = P(MalDiDenti | Carie)$$

relazione di indipendenza condizionata

### RETI BAYESIANE:

Ci sposteremo quindi da una **Distribuzione Congiunta di Probabilità** (inefficienti siccome esponenziali alle variabili) ad una **rete bayesiana** che non è altro che una rappresentazione grafica di un insieme di variabili casuali e relazioni di indipendenza tra esse.

Queste reti sono un **grafo orientato aciclico** annotato con distribuzioni di probabilità condizionate:

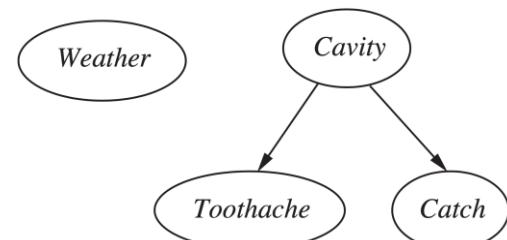
- Insieme di nodi, uno per variabile casuale;
- Insieme di archi orientati che connettono coppie di nodi. Se c'è un arco dal nodo  $X$  al nodo  $Y$  si dice che  $X$  è **parent** di  $Y$  ( $X$  ha un'influenza diretta su  $Y$ )
- Ad ogni nodo  $X_i$  è associata una distribuzione di probabilità condizionale che quantifica l'effetto dei parents sul nodo:

$$P(X_i | Parents(X_i))$$

- Per variabili discrete, la distribuzione condizionale è rappresentata come una tabella (**CPT Conditional Probability Table**) che fornisce la distribuzione su  $X_i$  per ogni combinazione dei valori dei nodi parents (in direzione causale).

La topologia della rete codifica le asserzioni di indipendenza condizionale:

- Weather è indipendente dalle altre variabili;
- Maldidenti e prende sono condizionalmente indipendenti data Carie (ognuna dipende da carie ma non c'è relazione causale tra le due).



### Esempio:

Abbiamo un allarme installato a casa abbastanza affidabile nell'andare a identificare i furti, ma qualche volta si attiva l'allarme in occasione di terremoti (nella zona sono frequenti, a Los Angeles). Ci sono due vicini, Jhon e Mary, che ci chiamano (siamo i proprietari dell'allarme) quando sentono l'allarme. Jhon chiama sempre quando sente l'allarme suonare ma alcune volte lo confonde con il nostro telefono che squilla. Mary, invece, gli piace ascoltare musica ad alto volume, ed alcune volte non ci telefona quando l'allarme suona. Vogliamo stimare la probabilità di un furto.

Le variabili sono 5: Allarme – Telefonate da parte dei due vicini – Furto – Terremoto.

Tramite reti bayesiane, possiamo andare a rappresentare un nodo per ogni variabile casuale e poniamo un arco quando il valore di una variabile è condizionato dal valore di un'altra variabile.

Variabili:

Burglary, Earthquake (Cause), Alarm, JohnCalls, MaryCalls (Effect). Non è un fatto di sintassi, ma definisce quanto il modello è un buon modello della realtà (le relazioni di indipendenza condizionale codificate nel modello sono una sufficiente, per gli scopi dati, approssimazione della realtà).

La topologia della rete riflette la conoscenza causale:

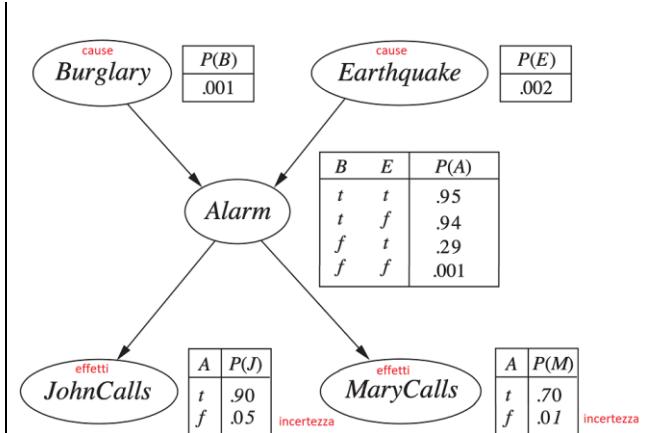
- Il furto con scasso ed il terremoto possono attivare l'allarme;
- Il fatto che Mary o John possano chiamare dipende dall'allarme;
- Mary e John non sono attivati dal furto o dal terremoto e non interagiscono tra di loro.

Permette una rappresentazione della conoscenza incerta e la rete non modella il fatto che Mary ascolta musica ad alto volume e che John confonde il suono del telefono con quello dell'allarme. Tutti questi fattori (e altri potenzialmente infiniti) sono riassunti dall'incertezza associata ai link tra alarm, Mary e John.

Ogni nodo ha una CPT Conditional Probability Table (per variabili discrete).

Ogni riga contiene, per ogni valore del nodo, la probabilità condizionale per un *conditioning case* (una possibile combinazione di valori dei nodi *parents*).

Un nodo senza parents ha una sola riga che rappresenta la probabilità a priori di ogni possibile valore della variabile, nell'esempio Burglary).



### Esempio 1 (Gestione del traffico):

Un comune può decidere se bloccare o no le auto per una giornata nel caso in cui si verifichi uno dei seguenti casi:

- Viene raggiunto il livello massimo di inquinamento;
- Si verifica una congestione delle strade.

A seconda della situazione i cittadini dovranno decidere se spostarsi con i mezzi pubblici o prende la macchina.

Avremo **MP** (Mezzo Pubblico) e **M**(macchina) le quali dipendono dal **Blocco** (*B*) e questo dipende dall'*inquinamento* (*I*) e *congestione* (*C*).

### Esempio 2 (l'albero di Jack):

Un giorno Jack si accorge che il suo albero di mele perde le foglie. Jack sa che se l'albero è secco allora è normale che perda le foglie ma Jack sa anche che la perdita delle foglie può essere sintomo di malattia per il suo albero.

La rete consiste di 3 nodi:

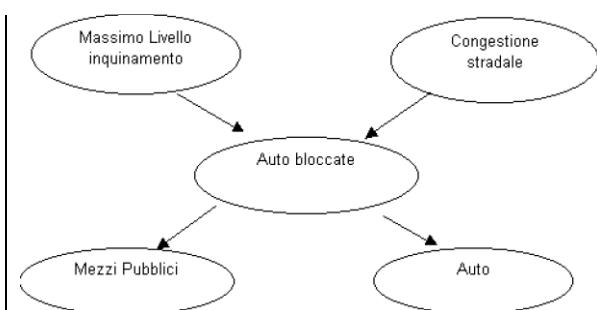
- Malato, Secco e Perde le foglie
- Malato può essere malato o no
- Secco può essere secco o no
- Perde le foglie può essere sì o no

La dipendenza casuale è tra **Malato** e **Perde le foglie** e **Secco** e **Perde le foglie**. Ad ogni nodo è associata una tabella di probabilità, che possono essere a priori o condizionate. Ad esempio:

- $P(\text{Malato} = \text{"malato"}) = 0.1$
- $P(\text{Malato} = \text{"no"}) = 0.9$
- $P(\text{Secco} = \text{"secco"}) = 0.1$
- $P(\text{secco} = \text{"no"}) = 0.9$

<i>B</i>	<i>E</i>	<i>P(A B,E)</i>	<i>P(¬A B,E)</i>
<i>T</i>	<i>T</i>	.95	.05

conditioning case



	Secco="secco"		Secco="No"	
	Malato="Malato"	Malato="No"	Malato="Malato"	Malato="No"
Perde="si"	0.95	0.85	0.90	0.02
Perde="no"	0.05	0.15	0.10	0.98

$P(\text{Perde le foglie} | \text{Malato, Secco})$

## SEMANTICA RETI BAYESIANE:

**Semantica:** una rete è una rappresentazione di una distribuzione di congiunta probabilità. Uno specifico elemento di una distribuzione di probabilità congiunta (evento atomico) è definito come:

$$P(X_1=x_1, \dots, X_n=x_n) \text{ abbreviato in } P(x_1, \dots, x_n)$$

Nelle reti bayesiane, sfruttando l'indipendenza condizionale, abbiamo che ogni variabile dipende solo dai genitori, quindi viene semplificato nel modo seguente. Il valore dell'elemento è:

$$P(x_1, \dots, x_n) = \prod_{i=1, n} P(x_i | \text{Parents}(X_i))$$

- Una distribuzione di probabilità congiunta è definita come il prodotto delle distribuzioni condizionali locali (CPT), date le asserzioni di indipendenza condizionale codificate dalla topologia della rete;
- Le CPT sono la **decomposizione** di una distribuzione di probabilità congiunta.

## PROBABILITÀ DELL'EVENTO ATOMICO:

$\text{Parents}(X_i)$  denota i valori specifici delle variabili in  $X_i$  definiti in  $x_1, \dots, x_n$  (l'elemento corrispondente della CPT di  $x_i$ )

$$P(x_1, \dots, x_n) = \prod_{i=1, n} P(x_i | \text{Parents}(X_i))$$

↑  
La probabilità di  $x_i$  condizionata dai nodi genitori di  $x_i$

### Esempio:

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$

$$P(x_i | \text{Parents}(X_i))$$

La probabilità di  $a$  condizionata dai nodi genitori di  $a$  è:

$$P(a | \neg b, \neg e)$$

Applico iterativamente la product rule  $P(a \wedge b) = P(a | b) P(b)$ :

...

$$= P(j | m \wedge a \wedge \neg b \wedge \neg e) P(m | a \wedge \neg b \wedge \neg e) P(a | \neg b \wedge \neg e) P(\neg b | \neg e) P(\neg e)$$

Utilizzo le relazioni di indipendenza condizionale codificate nel grafo:

$$P(j | m \wedge a \wedge \neg b \wedge \neg e) = P(j | a) \quad j \text{ dipende solo da } a$$

...

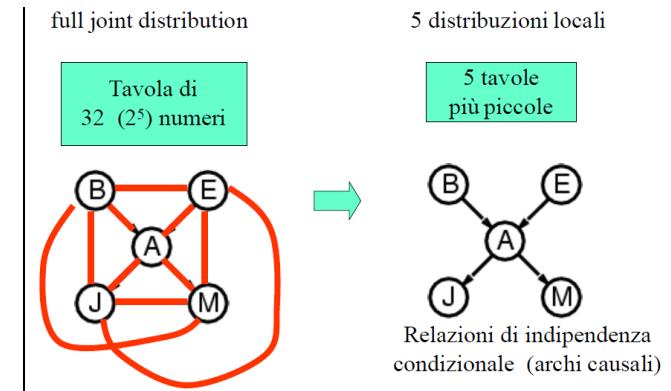
ed ottengo:

$$= P(j | a) P(m | a) P(a | \neg b, \neg e) P(\neg b) P(\neg e)$$

Il risultato finale è dato quindi da  $0.90 * 0.70 * 0.001 * 0.999 * 0.998 = 0.00062$ .

Alla fine, si riesce a rispondere ad una query su un evento atomico facendo un numero di calcoli molto inferiore rispetto a quello che avevamo visto in precedenza.

In termini numerici, in questo caso abbiamo solo 5 variabili casuali, mentre con la tabella di distribuzione congiunta (full joint distribution) i valori nella tavola di verità erano 32, nella rete bayesiana abbiamo 5 tabelle più piccole



Il vantaggio è la **compattezza**, la rete è una rappresentazione più compatta della distribuzione di probabilità condizionale.

Topologia + CPTs = rappresentazione compatta di una distribuzione di probabilità condizionale.

Se ogni variabile ha non più di  $k$  parents, la rete completa di  $n$  variabili (booleane) richiede  $O(n * 2^k)$  numeri a differenza della distribuzione di probabilità condizionale che cresce esponenzialmente:  $O(2^n)$ .

Per la rete di allarme precedente:  $1+1+4+2+2=10$  mentre l'equivalente distribuzione di probabilità condizionale  $25-1=31$ .

Se  $n=30$  e  $k=5$  abbiamo 960 numeri nella RB contro 1 miliardo.

## COSTRUZIONE DI UNA RETE:

Data la semantica della rete, come la si costruisce? Come si vede dall'esempio precedente:

- Applicando iterativamente la product rule e si ottiene (**chain rule**):

$$P(x_1, \dots, x_n) = \prod_{i=1, n} P(x_i | X_{i-1}, \dots, X_1)$$

- Utilizzando l'equazione:

$$P(x_1, \dots, x_n) = \prod_{i=1, n} P(X_i | \text{Parents}(X_i))$$

- Si ottiene

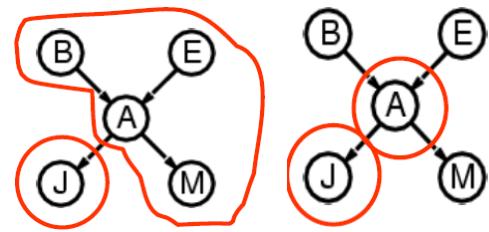
$$P(X_i | X_{i-1}, \dots, X_1) = P(x_i | \text{Parents}(X_i))$$

- La probabilità di  $X_i$  condizionata da tutti gli altri nodi è la probabilità di  $X_i$  condizionata dai nodi genitori.

In questo caso j dipende solo da a:

$$P(X_i | X_{i-1}, \dots, X_1) = P(x_i | Parents(X_i))$$

Dato l'ordinamento dei nodi  $X_n, \dots, X_1$ , per ogni nodo  $X_i$ , la probabilità condizionata dal nodo rispetto a tutti gli altri nodi è la probabilità condizionata rispetto ai nodi genitori. Ogni nodo deve essere condizionalmente indipendente dai suoi predecessori nell'ordinamento dei nodi, dati i suoi nodi genitori. I genitori di ogni nodo devono essere tutti e soli i nodi che influenzano direttamente il nodo.



L'ordine di selezione dei nodi da aggiungere alla rete, sceglieremo prima le cause e poi gli effetti, questo porta a costruire reti ottime. Infatti, l'ordine di scelta dei nodi nella costruzione porta a reti diverse.

**Primo modo:** costruzione di una rete a partire dal significato numerico

- Si sceglie un ordinamento dei nodi;
- Si scrive un nodo alla volta;
- Si verifica la indipendenza condizionale del nuovo nodo dai precedenti;
- Conseguentemente si scrivono gli archi.

$$P(X_i | X_{i-1}, \dots, X_1) = P(x_i | Parents(X_i))$$

La probabilità di  $X_i$  condizionata da tutti gli altri nodi è la probabilità di  $X_i$  condizionata dai nodi genitori.

Esempio:

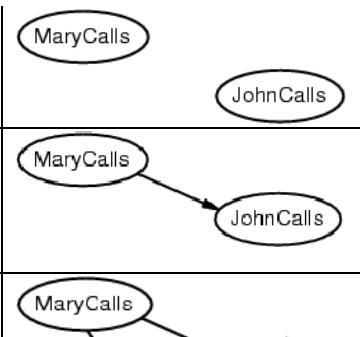
Ordine (diagnostico) dei nodi (si parte prima dagli effetti e poi le cause):  $M, J, A, B, E$

- Inserisco il nodo  $M \rightarrow$  non ha parents
- Inserisco il nodo  $J$

Jhon dipende da Mary? -  $P(J | M) = P(J)?$

- **No**, J non è condizionalmente indipendente da M.

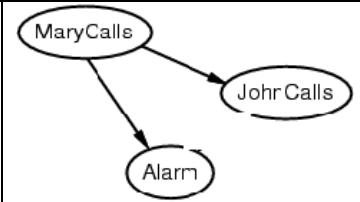
Se Mary chiama significa che probabilmente c'è stato un allarme e quindi è più probabile che anche John chiami.



Inserisco il nodo A:

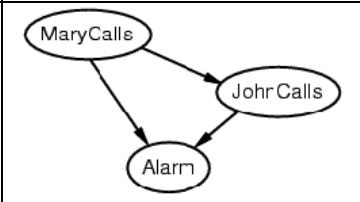
- $P(A | J, M) = P(A | J)?$  **No**

La probabilità che ci sia un allarme dipende dal fatto che Mary abbia chiamato.



- $P(A | J, M) = P(A)?$  **No**

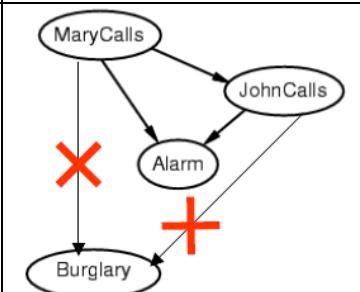
Se sia Mary che John chiamano, la probabilità che ci sia un allarme cambia.



Inserisco il nodo B:

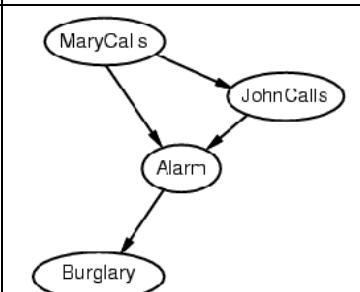
- $P(B | A, J, M) = P(B | A)?$  **Yes**

La probabilità di *Burglary* avendo come evidenza lo stato di *Alarm* non cambia se si aggiungono come evidenze lo stato di *MaryCalls* e *JohnCalls*.



- $P(B | A, J, M) = P(B)?$  **No**

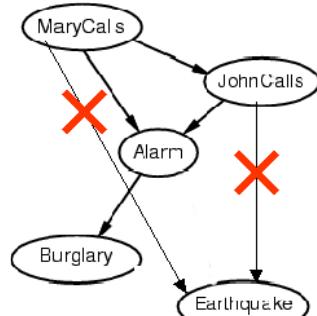
La probabilità di *Burglary* cambia se si ha l'evidenza dello stato di *Alarm*. Siccome il furto dipende dall'allarme.



Inserisco il nodo E:

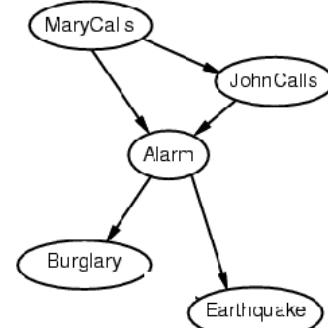
- $P(E | B, A, J, M) = P(E | A, B)$ ? Yes

*Earthquake* non cambia conoscendo lo stato delle chiamate di Mary e John.



- $P(E | B, A, J, M) = P(E | B)$ ? No

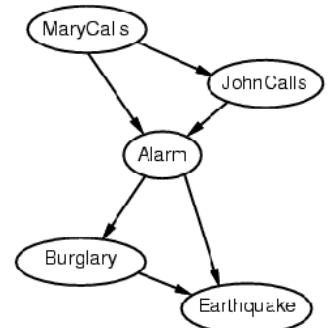
Se ho evidenza ci sia un *Alarm* la probabilità che ci sia un *Earthquake* cambia.



Inserisco il nodo E:

- $P(E | B, A, J, M) = P(E | A)$ ? No

Se so che c'è stato un *Burglary* in presenza di allarme, *Burglary* può spiegare l'allarme e quindi influenzare la probabilità di *Earthquake*.



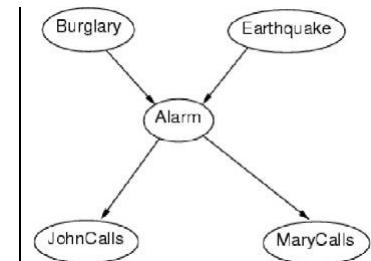
Risultato finale con l'ordinamento:

M, J, A, B, E

sintomo, sintomo, causa intermedia, causa iniziale, causa iniziale

Un numero maggiore di link rispetto a:

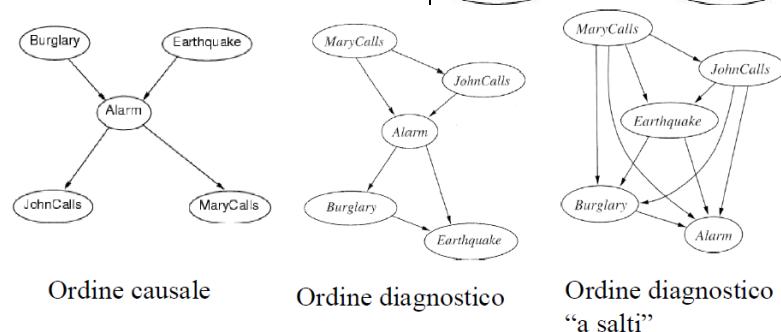
Alcuni link sono innaturali e di difficile stima (Probabilità di *Earthquake*, dato *Alarm* e *Burglary*)



**Secondo modo:** costruzione di una rete causale:

- Si definiscono le cause prime;
- Si connettono le cause prime agli effetti diretti;
- Si procede allo stesso modo interpretando gli effetti come nuove cause.

L'ordine di scelta causale porta a costruire reti ottime.

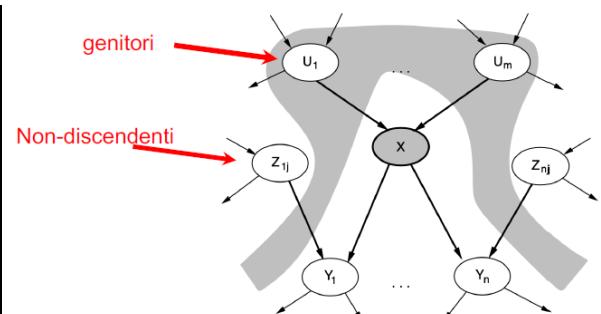


### SEMANTICA LOCALE:

L'indipendenza tra nodi può essere codificata tramite una **semantica locale**. In questo modo definendo una semantica locale per ogni nodo possiamo definire la semantica per l'intera rete (global semantics).

**Semantica locale (topologica):** ogni nodo è condizionatamente indipendente dai suoi non-descendenti dati i suoi genitori. È condizionato solo dai valori che assumono i suoi genitori.

In altre parole, ogni nodo è condizionatamente indipendente da tutti gli altri nodi data la sua **Coperta di Markov**, ovvero genitori + figli + genitori di figli. Tutti questi nodi non influenzano x.



## INFERNZA ESATTA SU RETI BAYESIANE:

Vogliamo usare una rete bayesiana per fare inferenza, facendola usare ad un agente e quando questo fa inferenza non è interessato solo a sapere se un evento può accadere oppure no, ma ha anche delle conoscenze, quindi la risposta che vuole ottenere è in funzione di quello che lui già sa.

Con l'inferenza esatta si vuole calcolare la probabilità CONDIZIONATA (a posteriori) di un insieme di **query variables**, dato un evento (un assegnamento di valori ad un insieme di **evidence variables**).

Utilizzando la product rule:

$X$  variabile query,  $E$  variabili evidenza,  $e$  valori osservati per  $E$ ,  $Y$  variabili Hidden i cui valori sono  $y$ :

Vogliamo rispondere alla query:

$$P(X | e)$$

Distribuzione di probabilità di  $X$  condizionata da  $e$

La formula diventa:

$$P(X | e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y)$$

Somma su tutti i possibili valori  $y$  delle variabili non osservate

In una rete bayesiana i termini  $P(X, e, y)$  (probabilità dell'EVENTO ATOMICO) possono essere scritti come prodotti di probabilità condizionate prese dalla rete (CPT):

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | Parents(X_i))$$

Esempio:

Query:  $P(Burglary | JohnCalls=true, MaryCalls=true)$

Inferenza in verso diagnostico utilizzando CPT scritte in verso causale (Teorema di Bayes):

$$P(B | j, m) = \alpha P(B, j, m) = \alpha \sum_e \sum_a P(B, j, m, e, a)$$

Somma su tutti i possibili valori  $y$  delle variabili hidden  $e, a$

Calcolo per  $Burglary = true$  utilizzando:  $P(x_i | Parents(X_i))$

$$P(b | j, m) = \alpha \sum_e \sum_a P(b) P(j|a) P(m|a) P(e) P(a|b, e)$$

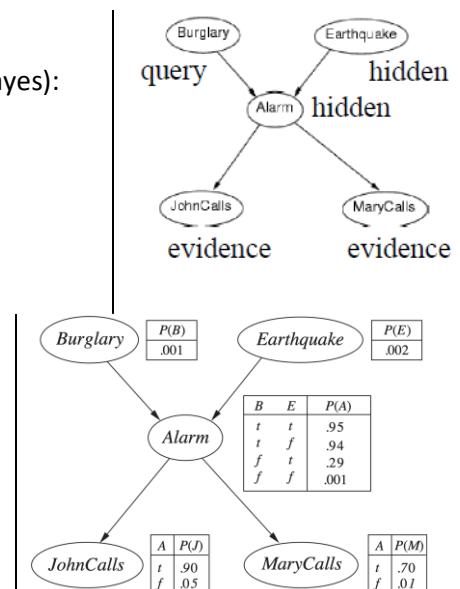
Devo sommare 4 termini (2\*2 combinazioni di valori di  $e$  ed  $a$ ).

Ogni termine è un prodotto di 5 numeri. Con  $n$  variabili booleane  $O(n2^n)$ .

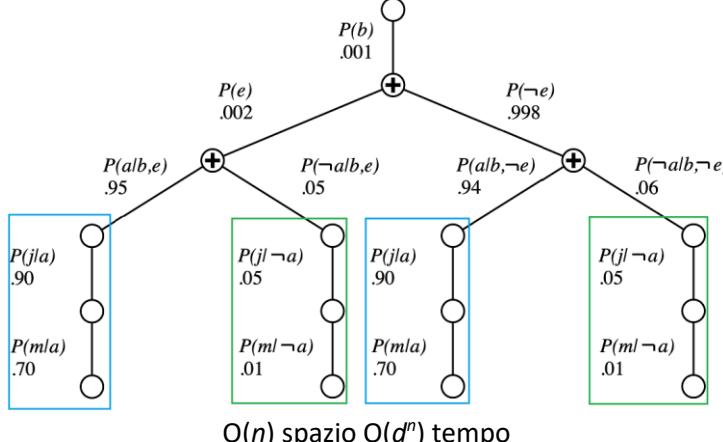
$$P(b | j, m) = \alpha \sum_e P(e) \sum_a P(j|a) P(m|a) P(a|b, e)$$

$$P(B | j, m) = \alpha <0,00059224, 0,0014919>$$

Normalizzando a 1 con  $\alpha$ :  $<0,284, 0,716>$



Graficamente viene una visita in profondità:



L'algoritmo effettua dei calcoli ripetitivi, come si può notare con le "foglie". Si possono effettuare delle ottimizzazioni.

## INFERNZA ATTRAVERSO L'ELIMINAZIONE DI VARIABILI:

eseguire somme da destra a sinistra, memorizzando risultati intermedi (**fattori**) per evitare la ricomputazione:

$$\begin{aligned} P(B | j, m) &= \alpha \underbrace{P(B)}_{\overline{B}} \sum_e \underbrace{P(e)}_{\overline{E}} \sum_a \underbrace{P(a | B, e)}_{\overline{A}} \underbrace{P(j | a)}_{\overline{J}} \underbrace{P(m | a)}_{\overline{M}} \\ &= \alpha P(B) \sum_e P(e) \sum_a P(a | B, e) P(j | a) f_M(a) \\ &= \alpha P(B) \sum_e P(e) \sum_a P(a | B, e) f_J(a) f_M(a) \\ &= \alpha P(B) \sum_e P(e) \sum_a f_A(a, b, e) f_J(a) f_M(a) \\ &= \alpha P(B) \sum_e P(e) f_{\bar{A}JM}(b, e) \quad (\text{sum out } A) \\ &= \alpha P(B) f_{\bar{E}\bar{A}JM}(b) \quad (\text{sum out } E) \\ &= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b) \end{aligned}$$

Vettore di due elementi  
 $P(m|a)$  e  $P(m| \neg a)$

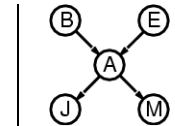
## VARIABILI IRRILEVANTI:

Per velocizzare ancora di più possiamo verificare se ci sono variabili irrilevanti, cioè i loro valori di probabilità non influenzano il risultato finale.

Identifichiamo tale variabile analizzando la rete bayesiana e quelle irrilevanti sono quelle che non fanno parte degli antenati né della query né delle evidenze. Ad esempio, se ho una query su J, tutti i dati su M non mi interessano.

Consideriamo la query  $P(JohnCalls | Burglary=true)$

$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b, e) P(J|a) \sum_m P(m|a)$$



La somma su  $m$  è 1;  $M$  è irrilevante per la query

**Teorema 1:** Y è irrilevante a meno che:

$$Y \in Ancestors(\{X\} \cup \mathbf{E})$$

Here,  $X = JohnCalls$ ,  $\mathbf{E} = \{Burglary\}$ , and  
 $Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$   
so  $MaryCalls$  is irrelevant

Un algoritmo di eliminazione variabili può quindi eliminare tutte queste variabili prima di valutare la query.

## COMPLESSITÀ DELL'INFERENZA ESATTA:

**Reti singolarmente connesse** (o *polialberi*):

- Due nodi qualsiasi sono collegati al massimo da un percorso non orientato
- Il costo in tempo e spazio dell'eliminazione variabili è  $O(d^k n)$

**Reti a connessioni multiple:**

- Può ridursi a 3SAT  $\rightarrow$  NP-Hard
- Equivalenti a contare modelli 3SAT  $\rightarrow$  NP-Hard

Fare **clustering** significa trasformare una rete bayesiana classica in un polialbero, accorpando più nodi ed ovviamente la tabella di CPT diventa complicata.

## INFERENZA MEDIANTE SIMULAZIONE STOCASTICA:

Dobbiamo usare **metodi approssimati di inferenza** perché l'inferenza esatta su reti Bayesiane è **NP-Hard**.

Per l'inferenza approssimata dobbiamo stimare i valori di probabilità attraverso delle simulazioni (o campionamento):

- Generare  $N$  campioni da una distribuzione di campionamento  $S$ ;
- Calcolare una probabilità a posteriori approssimativa  $\hat{P}$  (**P segnato**);
- Mostrare che converge alla probabilità reale  $P$ , così da poter rispondere alla query con  $\hat{P}$ .

Tecniche che vedremo (di complessità crescente) per ricavarci  $\hat{P}$ :

- Campionamento da una rete vuota**;
- Campionamento di rigetto**: respingere i campioni in disaccordo con le evidenze;
- Pesatura di verosimiglianza**: utilizzare le evidenze per ponderare i campioni;
- Catena di Markov Monte Carlo (MCMC)**: campione da un processo stocastico la cui distribuzione stazionaria è la vera probabilità a posteriori.

## CAMPIONAMENTO DA UNA RETE VUOTA:

Partendo da una rete bayesiana, vogliamo ricavarci la probabilità di un determinato evento atomico. Vogliamo poter calcolare la distribuzione di probabilità congiunta. L'algoritmo è il seguente:

Prende in input la rete bayesiana, per un dato evento effettua un campionamento per tutti i valori di questo evento. Per far sì che il valore sia corretto, andiamo ad associare dei valori che rispettano ciò che è stato definito nella rete bayesiana; quindi, tenendo in considerazione la probabilità di  $X_i$  dati i suoi parenti (il valore associato dipende dai valori dei genitori). Il campione viene creato partendo dalle variabili causali (quelle che non hanno genitori, assegnando un valore random).



```

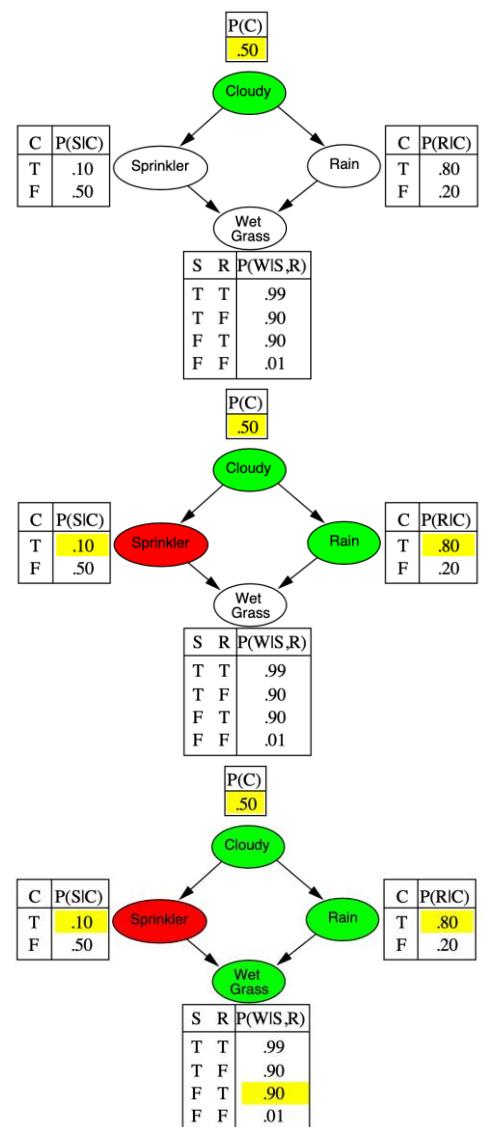
function PRIOR-SAMPLE(bn) returns an event sampled from bn
  inputs: bn, a belief network specifying joint distribution  $\mathbf{P}(X_1, \dots, X_n)$ 
  x  $\leftarrow$  an event with n elements
  for i = 1 to n do
    xi  $\leftarrow$  a random sample from  $\mathbf{P}(X_i | parents(X_i))$ 
      given the values of  $Parents(X_i)$  in x
  return x

```

## Esempio:

Consideriamo la rete bayesiana a lato, abbiamo che l'erba bagnata (Wet Grass) è l'effetto finale che può essere causato da due possibili eventi, ovvero ha piovuto (Rain) o l'irrigatore è stato acceso (Sprinkler), mentre l'ultimo evento è nuvoloso (Cloudy) che influenza i due eventi precedenti siccome se il tempo è nuvoloso, l'irrigatore non viene acceso siccome dovrebbe piovere (ad esempio 80% pioverà).

Una volta definita la rete facciamo partire l'algoritmo, iniziando per ordine topologico, ovvero da Cloudy. A questa variabile viene associato un valore di probabilità che dipenderà da  $P(C)$  (siccome non ha genitori). Il valore lo si sceglie in base ad una distribuzione di probabilità, in questo caso i parents sono equiprobabili perché abbiamo il 50% e sceglierà a random se è vero o falso. Supponiamo che viene associato a Cloudy il valore TRUE.



A questo punto si passa ai figli, andando a campionare il valore della variabile Sprinkler. In questo caso, il valore associato ad esso è probabilmente FALSE perché, poiché la variabile C è TRUE, abbiamo il 10% di volte che S è TRUE. Mentre il valore associato alla variabile R è probabilmente TRUE perché nell'80% dei casi lo è quando Cloudy è TRUE.

Infine, per generare il valore da associare alla variabile W bisogna tenere in considerazione i suoi genitori ( $P(X_i | Parents(X_i))$ ), in tal caso sarà TRUE, andando a vedere in tabella abbiamo che lo è il 90% delle volte.

**NOTA:** L'algoritmo produce solo un evento, pertanto verrà eseguito più volte.

L'algoritmo verrà lanciato un certo numero di volte, una volta avuti i campioni andremo a contare quante volte un evento è capitato.

Facendo crescere il numero di campioni, il valore probabilistico che andremo ad ottenere con il rapporto (numero campioni in cui è vero/numero totale di campioni) coinciderà con la probabilità che quell'evento accade (abbiamo detto essere P).

Fondamentalmente, l'algoritmo produce delle probabilità consistenti.

Probabilità che PriorSample generi un evento particolare

$$SPS(x_1 \dots x_n) = \prod_{i=1}^n P(x_i | parents(X_i)) = P(x_1 \dots x_n)$$

cioè la vera probabilità a priori (distr. congiunta rete Bayesiana)

$$\text{Ad esempio } SPS(t, f, t, t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t, f, t, t)$$

Sia  $N_{PS}(x_1 \dots x_n)$  numeri di eventi in cui  $x_1, \dots, x_n$  è generato

- Esempio: se ho 1000 campioni per la rete WetGrass, e in 511 Rain=true, allora  $P(\text{Rain}=true) = 0,511$ .

$$\begin{aligned} \text{Allora abbiamo } \hat{P}(x_1, \dots, x_n) &= \lim_{N \rightarrow \infty} N_{PS}(x_1, \dots, x_n)/N \\ &= SPS(x_1, \dots, x_n) \\ &= P(x_1 \dots x_n) \end{aligned}$$

In altre parole, le stime derivate da PriorSample sono CONSISTENTI

$$\hat{P}(x_1, \dots, x_n) \approx P(x_1 \dots x_n)$$

## CAMPIONAMENTO DI RIGETTO:

Questo algoritmo è semplice ma inefficiente, rispetto al precedente, teniamo in considerazione delle **evidenze**.

In questo caso vogliamo stimare  $p$  di  $x$  dato  $e$ , quindi va ad applicare l'algoritmo di campionamento a rete vuota  $N$  volte. Successivamente, tra tutti i campioni considera solo quelli **utili** ovvero che soddisfano le evidenze mentre quelli che non le soddisfano non li consideriamo perché non sono utili per la probabilità che sto cercando di calcolare.

Infine, va a normalizzare sul numero di campioni utili.

$\hat{P}(X|e)$  stimato da campioni concordanti con l'evidenza  $e$

```
function REJECTION-SAMPLING( $X, e, bn, N$ ) returns an estimate of  $P(X|e)$ 
  local variables:  $N$ , a vector of counts over  $X$ , initially zero
  for  $j = 1$  to  $N$  do
     $x \leftarrow \text{PRIOR-SAMPLE}(bn)$ 
    if  $x$  is consistent with  $e$  then
       $N[x] \leftarrow N[x]+1$  where  $x$  is the value of  $X$  in  $x$ 
  return NORMALIZE( $N[X]$ )
```

### Esempio:

Ad esempio, stima  $P(Rain|Sprinkler = \text{true})$  usando 100 campioni

- ▶ 27 campioni hanno  $Sprinkler = \text{true}$
- ▶ di questi 8 hanno  $Rain = \text{true}$  e 19 hanno  $Rain = \text{false}$

$$\hat{P}(Rain|Sprinkler = \text{true}) = \text{NORMALIZE}(\langle 8, 19 \rangle) = \langle 0.296, 0.704 \rangle \quad (\text{la normalizzazione è } 8/27 \text{ e } 19/27)$$

(simile a una procedura di stima empirica nel mondo reale)

Quindi il campionamento di rigetto restituisce stime a posteriori **consistenti**. Il problema è che il tutto è irrimediabilmente costoso se  $P(e)$  è piccolo (rifiuterà moltissimi campioni). Infatti,  $P(e)$  diminuisce esponenzialmente al crescere del numero di variabili evidenza. Più evidenze ci sono più l'algoritmo è inefficiente.

$$\begin{aligned}\hat{P}(X|e) &= \alpha N_{PS}(X, e) && (\text{algorithm defn.}) \\ &= N_{PS}(X, e)/N_{PS}(e) && (\text{normalized by } N_{PS}(e)) \\ &\approx P(X, e)/P(e) && (\text{property of PRIORSAMPLE}) \\ &= P(X|e) && (\text{defn. of conditional probability})\end{aligned}$$

### **PESATURA DI VERO SIMIGLIANZA:**

L'idea è di fissare le variabili evidenza, campionare solo le variabili non di evidenza e pesare ogni campione in base alla probabilità che si accordi con le evidenze (gli eventi in cui è improbabile che le prove siano verificate devono pesare di meno nel conteggio). L'algoritmo produce il campione insieme ad un peso e va a considerare tutte le variabili della rete bayesiana e se  $X_i$  rispetta quella condizione (ha valore in  $e$ ), vuol dire che è una variabile di evidenza e quindi devo vedere quanto il valore che ho in  $X$ , si accorda con le variabili che ho assegnato nella rete bayesiana quindi vado ad aggiornare il peso.

Nel caso in cui non lo è, faccio il campionamento secondo la probabilità condizionata rispetto ai suoi genitori.

function LIKELIHOOD-WEIGHTING( $X, e, bn, N$ ) returns an estimate of  $P(X|e)$   
local variables:  $\mathbf{W}$ , a vector of weighted counts over  $X$ , initially zero

```
for  $j = 1$  to  $N$  do
     $\mathbf{x}, w \leftarrow \text{WEIGHTED-SAMPLE}(bn)$ 
     $\mathbf{W}[x] \leftarrow \mathbf{W}[x] + w$  where  $x$  is the value of  $X$  in  $\mathbf{x}$ 
return NORMALIZE( $\mathbf{W}[X]$ )
```

function WEIGHTED-SAMPLE( $bn, e$ ) returns an event and a weight

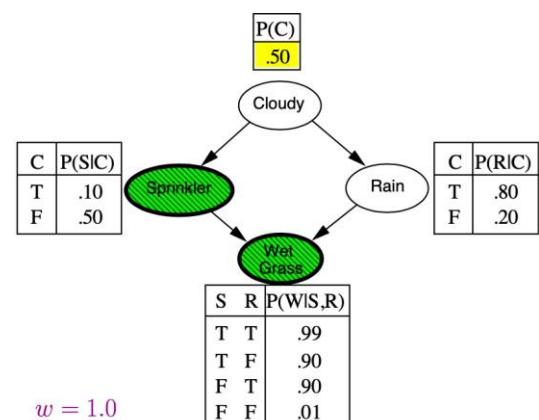
```
 $\mathbf{x} \leftarrow$  an event with  $n$  elements;  $w \leftarrow 1$ 
for  $i = 1$  to  $n$  do
    if  $X_i$  has a value  $x_i$  in  $e$ 
        then  $w \leftarrow w \times P(X_i = x_i | \text{parents}(X_i))$ 
        else  $x_i \leftarrow$  a random sample from  $P(X_i | \text{parents}(X_i))$ 
return  $\mathbf{x}, w$ 
```

### Esempio:

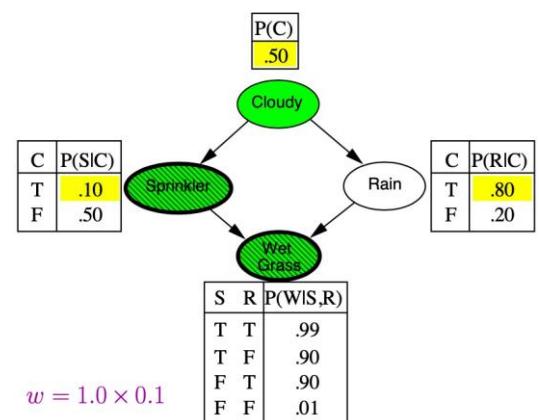
Supponiamo che diamo all'algoritmo la rete a lato, con la rispettiva Query=  $P(\text{Rain} | \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$ .

Alcuni valori di verità sono già fissati e l'algoritmo non li può cambiare, pertanto, l'algoritmo dovrà restituire un vettore che darà un valore a C e R, ed un altro valore che rappresenta il **peso w** che mi dice quanto è importante il campione dei quattro valori di verità.

L'algoritmo parte sempre dalla radice, prende  $X_1$  che sarebbe Cloudy ma essa non è una variabile di evidenza, quindi, va a campionare, trovandoci nello stesso stato del precedente esempio, ovvero ha il 50% di avere TRUE o FALSE.

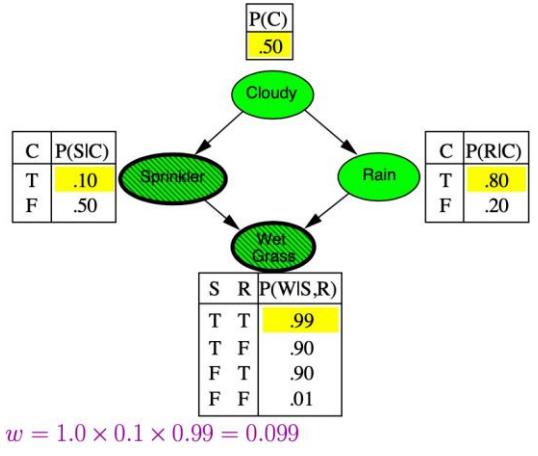


Supponiamo che Cloudy è TRUE e prendiamo la seconda variabile (Sprinkler), andiamo nella funzione dove chiede se X è una variabile di evidenza e in questo caso è sì; pertanto, bisogna aggiornare il peso che sarà w per la probabilità della variabile S data i suoi genitori C. Pertanto, il calcolo sarà  $w=1.0 \times 0.1$ .



Adesso si prende Rain, esso non è in evidenza quindi usiamo sempre la probabilità condizionata data dai genitori, quindi ci troviamo che 80% è TRUE e pertanto glie lo assegniamo.

L'ultima variabile W è in evidenza, pertanto si prende il valore di probabilità condizionato da due TRUE, questo poi sarà moltiplicato per il peso ed il peso totale sarà  $w=1.0 \times 0.1 \times 0.99=0.099$ .



#### ANALISI PESATURA DI VERO SIMIGLIANZA:

L'algoritmo calcola due cose, ovvero il campionamento delle variabili non di evidenza e il calcolo del peso.

La **probabilità di campionamento** per WeightedSample è:

$$S_{ws}(z,e) = \prod_{i=1}^l P(z_i | parents(Z_i))$$

Dove **z** sono tutte le variabili tranne quelle di evidenza, mentre **parents(Z<sub>i</sub>)** può includere sia variabili nascoste che variabili di evidenza.

Bisogna prestare attenzione alle evidenze solo negli **antenati** → ignorando le evidenze che non sono presenti tra gli antenati di Z<sub>i</sub>.

Il **peso** per un dato campione **z** ed evidenze **e** è:

$$w(z,e) = \prod_{i=1}^m P(e_i | parents(E_i))$$

quindi la probabilità pesata di un campione è:

$$SWS(z,e)w(z,e) = \prod_{i=1}^l P(z_i | parents(Z_i)) \prod_{i=1}^m P(e_i | parents(E_i)) = P(z,e)$$

La quale coincide con la probabilità dell'evento atomico, quindi la probabilità pesata restituisce stime **consistenti** ma le prestazioni peggiorano ancora con molte variabili di evidenza perché alcuni campioni hanno quasi tutto il peso totale.

## 12. APPRENDIMENTO TRAMITE OSSERVAZIONI

Un aspetto importante degli agenti è dato dalla capacità di apprendere. Fino ad ora abbiamo visto che l'agente percepisce informazioni tramite sensori e la decisione presa dipende da ciò che c'è dietro. Quello che vedremo ora, sarà la capacità dell'agente di sfruttare le informazioni che riceve dall'ambiente, per capire se in futuro potrà cambiare le sue decisioni.

L'apprendimento è **essenziale** in ambienti sconosciuti perché rende l'agente intelligente, in maniera tale che sia in grado di adattarsi all'ambiente che lo circonda.

L'apprendimento è utile come metodo di costruzione del sistema, esporre l'agente alla realtà piuttosto che cercare di descriverla, siccome la conoscenza da dover inserire nel sistema è troppo ampia e complessa da specificare.

L'apprendimento modifica i meccanismi di decisione dell'agente al fine di migliorarne le prestazioni.

Lo schema riassume l'architettura di un agente capace di apprendere:

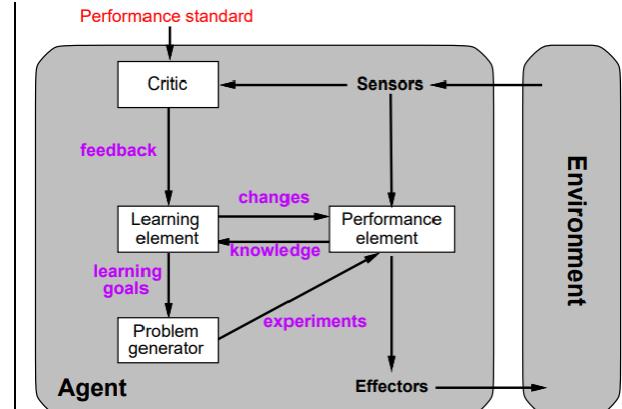
I **sensori** e **attuatori** permettono di interagire con l'ambiente.

Il **performance element** è l'agente che riceve le percezioni dai sensori e decide la migliore azione.

Vogliamo progettare quindi il **Learning element**, che rappresenta la componente chiave e quella che definiremo, perché è il modulo che va a cambiare il comportamento dell'agente; affinché l'agente si adatti in base alle sue performance.

Il **feedback** indica come si sta comportando l'agente, ci dicono se l'azione effettuata ha avuto esito positivo o meno.

Il **problem generator** permette all'agente di esplorare scenari che non erano stati già visitati.



Il **learning element** dipende:

- Dal tipo di **Performance element** usato, cioè da che tipo di agente usiamo (es. agenti basati su algoritmi di ricerca, basati su logica, basati su probabilità);
- Da quali componenti del **Performance element** devono essere appresi (dipende dal tipo di Performance element usato);
- Da come è **rappresentato** quel particolare componente funzionale;
- Dal tipo di **feedback** disponibile per l'apprendimento.

Scenari d'esempio:

Performance element	Componente	Rappresentazione	Feedback
Alpha–beta pruning search	Eval. Function	Weighted linear function	Win/loss
Logical agent	Transition model	Successor–state axioms	Outcome
Utility-based agent	Transition model	Rete Bayesiana	Outcome
Simple reflex agent	Percept–action function	Rete Neurale	Correct action

Supponiamo che l'agente usi l'algoritmo di **Alpha-Beta pruning** (ricerca con avversari). La componente usata è la **funzione di valutazione** e cambiandola si ottiene un comportamento differente dall'algoritmo, ad esempio, cambiando i pesi della funzione così da fare scelte differenti. La funzione di valutazione è **rappresentata** come funzione lineare pesata. Infine, il tipo di **feedback** che si ha è vittoria o sconfitta.

Due **tipologie di apprendimento**:

- **Supervisionato**: risposte corrette per ogni istanza;
- **Rinforzato**: ricompense occasionali.

### TIPI DI APPRENDIMENTO (FEEDBACK):

#### Apprendimento supervisionato:

- Apprendere una funzione da esempi di input/output, siccome abbiamo a disposizione questi;
- In ambienti completamente osservabili può vedere i risultati delle azioni ed imparare a predirli;
- In ambienti parzialmente osservabili gli effetti possono essere invisibili.

#### Apprendimento non supervisionato:

- Imparare a riconoscere pattern nell'input senza alcuna indicazione dell'output.

#### Apprendimento per rinforzo:

- L'agente apprende basandosi sul rinforzo, tramite ricompense.

## 12.1. APPRENDIMENTO INDUTTIVO

L'agente vuole capire se a partire da un insieme di dati (input e output), è possibile capire qual è il nesso tra input e output così da poter decidere l'azione da fare in base al prossimo input.

L'**apprendimento induttivo** rappresenta la forma più semplice di apprendimento, apprendere una funzione dagli esempi; **f** rappresenta la funzione obiettivo, l'input è il vettore **x** mentre l'output **f(x)** è il risultato di una funzione f.

Un esempio è la coppia  $x, f(x)$ , es.  $\begin{pmatrix} \begin{array}{c|c|c} \text{input} & \text{output} \\ \hline O & O & X \\ \hline X & & X \end{array} & , & \begin{array}{c} \text{configurazione} \\ \hline +1 \end{array} \end{pmatrix}$

Ciò che vogliamo fare è ricavare il legame tra input e output; quindi, vogliamo cercare un'ipotesi **h** tale che approssimi f ( $h \approx f$ ) dato un training set di esempi.

Il problema è che noi abbiamo solo una porzione di f, non abbiamo tutti i dati possibili ma solo un sottoinsieme quindi potremmo anche trovare una funzione  $h(x) = f(x)$  su tutte le x che abbiamo come input ma in ogni caso è una funzione differente perché non conosciamo il comportamento di f su altri campioni.

Questo è un modello altamente semplificato di apprendimento reale, infatti:

- Ignora la conoscenza precedente
- Assume la presenza di un ambiente deterministico ed osservabile
- Assume che siano forniti degli esempi
- Assume che l'agente voglia apprendere f perché?

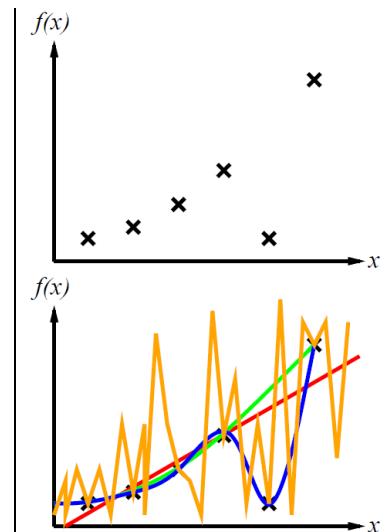
### CURVA FITTING:

Possiamo considerare **ogni coppia (x-f(x))** come punti nel piano cartesiano.

L'obiettivo è trovare la funzione **h** che concorda con **f**, sui dati del training set.

Questo problema si chiama **CURVE FITTING**, cioè trovare la migliore **h** che mi vada a coprire tutte le **x** che sono sul piano.

Ogni retta/curva nel piano rappresenta una possibile approssimazione di f, dove alcuni punti del piano sono più vicini o più lontani rispetto ad una determinata curva.



Quale **h** scegliere? Si deve preferire l'ipotesi più semplice consistente con i dati (**Rasoio di Occam**).

Un altro concetto che influenza questo problema è il **dominio delle funzioni** (es. polinomi, funzioni trigonometriche).

Dal punto di vista dell'apprendimento si dice che un problema di apprendimento è **realizzabile** se lo spazio delle ipotesi contiene la funzione reale. Questo perché la funzione h si prende dallo spazio delle ipotesi, ma se all'interno dello spazio non c'è f si è certi che non si può definire una funzione consistente e si dice **non realizzabile**. Non si sa in anticipo se è non realizzabile siccome la f non la conosciamo, quello che si fa è trovare un compromesso tra complessità del problema e la dimensione dello spazio di ipotesi, ad esempio si considera un h non complesso altrimenti il problema non viene risolto.

### Esempio (Training set – problema di classificazione):

Considerando il seguente training set con 12 campioni ed ognuno ha un vettore di 10 attributi (discreti) con un valore. Tali dati rappresentano situazioni dove converrebbe aspettare/non aspettare per un tavolo.

Example	Attributes										Target WillWait
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	
X <sub>1</sub>	T	F	F	T	Some	\$\$\$	F	T	French	0–10	T
X <sub>2</sub>	T	F	F	T	Full	\$	F	F	Thai	30–60	F
X <sub>3</sub>	F	T	F	F	Some	\$	F	F	Burger	0–10	T
X <sub>4</sub>	T	F	T	T	Full	\$	F	F	Thai	10–30	T
X <sub>5</sub>	T	F	T	F	Full	\$\$\$	F	T	French	>60	F
X <sub>6</sub>	F	T	F	T	Some	\$\$	T	T	Italian	0–10	T
X <sub>7</sub>	F	T	F	F	None	\$	T	F	Burger	0–10	F
X <sub>8</sub>	F	F	F	T	Some	\$\$	T	T	Thai	0–10	T
X <sub>9</sub>	F	T	T	F	Full	\$	T	F	Burger	>60	F
X <sub>10</sub>	T	T	T	T	Full	\$\$\$	F	T	Italian	10–30	F
X <sub>11</sub>	F	F	F	F	None	\$	F	F	Thai	0–10	F
X <sub>12</sub>	T	T	T	T	Full	\$	F	F	Burger	30–60	T

### Caratteristiche ristorante:

1. **Alternate**: whether there is a suitable alternative restaurant nearby.
2. **Bar**: whether the restaurant has a comfortable bar area to wait in.
3. **Fri/Sat**: true on Fridays and Saturdays.
4. **Hungry**: whether we are hungry.
5. **Patrons**: how many people are in the restaurant (values are *None*, *Some*, and *Full*).
6. **Price**: the restaurant's price range (\$, \$\$, \$\$\$).
7. **Raining**: whether it is raining outside.
8. **Reservation**: whether we made a reservation.
9. **Type**: the kind of restaurant (French, Italian, Thai, or burger).
10. **WaitEstimate**: the wait estimated by the host (0–10 minutes, 10–30, 30–60, or >60).

Sulla base di queste premesse vogliamo scoprire il criterio (la funzione) che modella il comportamento della persona.

Possiamo andare a riorganizzare i campioni come segue:

cioè ogni campione è costituito da un vettore  $x_i$  a 10 dimensioni di valori discreti delle caratteristiche e un'etichetta di destinazione  $y_i$  che è booleana.

Possiamo rappresentare questo training set tramite gli alberi di decisione.

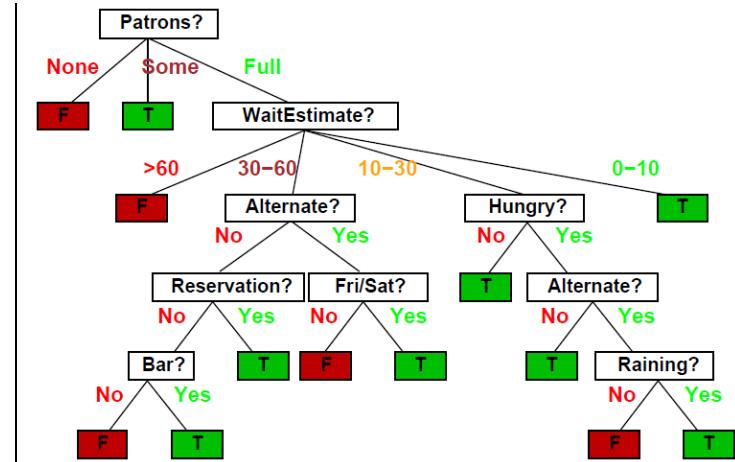
$$x_1 = \left\{ \begin{array}{l} T \\ F \\ F \\ T \\ Some \\ $$$ \\ F \\ T \\ French \\ 0-10 \end{array} \right\}, y_1 = T \quad \text{sample 1}$$

$$x_2 = \left\{ \begin{array}{l} T \\ F \\ F \\ T \\ Full \\ \$ \\ F \\ F \\ Thai \\ 30-60 \end{array} \right\}, y_2 = F \quad \text{sample 2}$$

### ALBERO DI DECISIONE:

Un **albero di decisione** rappresenta una funzione che prende in input un vettore di valori e restituisce una **decisione**, un singolo valore. È una possibile rappresentazione per le ipotesi dove i **nodi** sono gli **attributi del dataset** mentre le **foglie** sono le **classificazioni** (vero/falso nell'esempio). Il nodo che rappresenta un attributo ha come **archi uscenti** i possibili **valori** di quell'attributo. Prendendo un campione e seguendo il percorso dalla radice alla foglia, riesco a classificarlo.

L'albero di decisione rappresenta la tabella dell'esempio precedente:



Gli alberi di decisione possono esprimere qualsiasi funzione degli attributi in input.

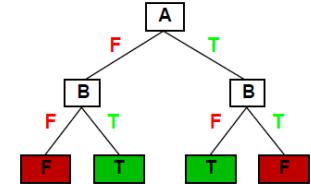
Un albero definisce un **predicato obiettivo** (*WillWait*):

$$\forall s \text{ WillWait}(s) \Leftrightarrow (P_1(s) \vee P_2(s) \vee \dots \vee P_n(s)) \text{ con } P_i \text{ path per arrivare alla foglia T.}$$

(Il predicato attendere è soddisfatto se e solo se uno dei percorsi da radice a foglia è soddisfatto)

In sostanza, esiste un albero decisionale coerente per qualsiasi training set con un path verso la foglia per ogni esempio (a meno che f sia non deterministico in x), ma probabilmente non si generalizzerà a nuovi esempi. È preferibile ricercare alberi di decisione più **compatti**.

A	B	A xor B
F	F	F
F	T	T
T	F	T
T	T	F



### SPAZIO DELLE IPOTESI:

Quanti alberi di decisione si possono creare con **n** attributi booleani (n colonne e 2 valori)?

- = numero di funzioni booleane;
- = numero di tabelle di verità distinte con  $2^n$  righe e tutte le possibili tabelle sono  $2^{2^n}$ .

Ad esempio, con 6 attributi booleani, ci sono 18,446,744,073,709,551,616 alberi di decisione.

#### Esempio:

Quante ipotesi puramente congiuntive si possono formulare (*Hungry  $\wedge$   $\neg Rain$* )?

Ogni attributo può essere vero, falso o ignorato quindi abbiamo  $3^n$  ipotesi congiuntive distinte.

Uno spazio delle ipotesi più espressivo aumenta la possibilità che la funzione obiettivo possa essere espressa, aumenta il numero di ipotesi consistenti con il training set e può portare a delle predizioni peggiori.

## 12.2. APPRENDIMENTO DEGLI ALBERI DI DECISIONE

Soluzione banale per la creazione di alberi decisionali è costruire un percorso fino ad una foglia per ogni campione.

- Memorizza semplicemente le osservazioni (non estrae nessun modello dai dati);
- Produce un'ipotesi consistente ma eccessivamente complessa (ricorda il rasoio di Occam);
- È improbabile che si generalizzi bene a nuove osservazioni.

Dovremmo mirare a costruire l'albero più piccolo (*più è piccolo più il senso dei dati viene evidenziato*) che sia coerente con le osservazioni. Un modo per farlo è testare in modo ricorsivo prima gli attributi più importanti.

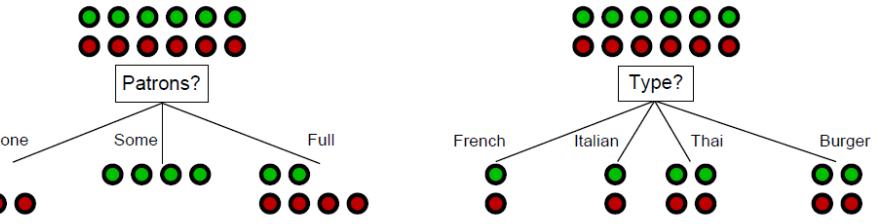
Bisogna capire, quindi, quali sono gli attributi utili per la classificazione.

## Esempio:

Meglio testare prima Patrons oppure Type?

**Patrons** rappresenta una scelta migliore poiché fornisce maggiori informazioni riguardo la classificazione.

L'idea è che un buon attributo divide gli esempi in sottoinsiemi che sono idealmente tutti positivi o tutti negativi.



## ALGORITMO DTL (DECISION TREE LEARNING):

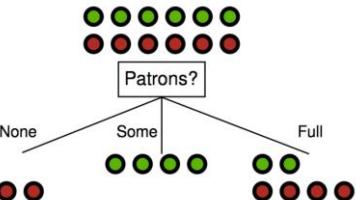
L'**obiettivo** è quello di trovare un albero piccolo e consistente con esempi di training e l'**idea** è di scegliere ricorsivamente l'attributo "più significativo" come radice del (sotto)albero. Se l'algoritmo si trova con tutti campioni della stessa classe, l'algoritmo può terminare e associare a quella foglia la classificazione degli esempi.

La parte fondamentale è scegliere **best** e costruire un albero incentrato su best; quindi, per ogni valore di best va a definire i diversi sottoinsiemi di esempi con valore di best= $v_i$ , e chiama ricorsivamente la funzione, dividendo i campioni in base al valore dell'attributo.

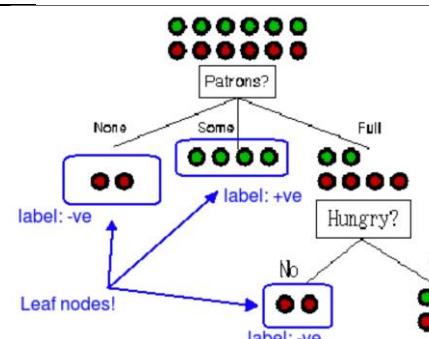
```
function DTL(examples, attributes, default) returns a decision tree
  if examples is empty then return default
  else if all examples have the same classification then return the classification
  else if attributes is empty then return MODE(examples)
  else
    best ← CHOOSE-ATTRIBUTE(attributes, examples)
    tree ← a new decision tree with root test best
    for each value  $v_i$  of best do
      examples $_i$  ← {elements of examples with best =  $v_i$ }
      subtree ← DTL(examples $_i$ , attributes - best, MODE(examples))
      add a branch to tree with label  $v_i$  and subtree subtree
  return tree
```

Quattro idee alla base dell'algoritmo

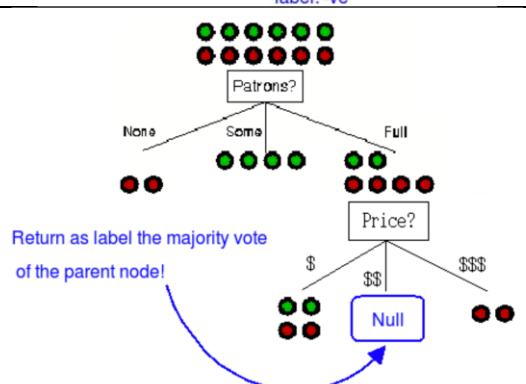
- Se sono presenti campioni positivi e negativi, scegliere l'attributo migliore per dividerli, ad esempio, prova **Patrons** nella radice.



- Se tutti i campioni rimanenti sono tutti positivi o tutti negativi, abbiamo raggiunto un nodo foglia. Assegna l'etichetta come positiva (o negativo), ad esempio →



- Se non sono rimasti campioni, significa che non è stato osservato quel campione. Restituisce un valore predefinito calcolato da classificazione maggioritaria al genitore del nodo, ad esempio, in quel caso verrà associato false, perché è quello maggiore.



- Se non ci sono attributi rimasti, ma ci sono sia campioni positivi che negativi, significa che questi campioni hanno esattamente lo stesso valore delle caratteristiche ma classificazioni diverse. Può succedere perché:
  - Alcuni dati potrebbero essere errati;
  - Gli attributi non forniscono informazioni sufficienti per descrivere completamente la situazione (cioè ci mancano altri attributi utili);
  - Il problema è veramente **non-deterministico**, cioè dati due campioni che descrivono esattamente le stesse condizioni, possiamo prendere diverse decisioni.

Soluzione: chiamarlo nodo foglia e assegnergli il voto di maggioranza come etichetta.

## INFORMAZIONE:

L'aspetto chiave dell'algoritmo DTL è il come fa a giudicare un attributo come buono o cattivo. Un modo possibile è **calcolare il contenuto dell'informazione (entropia)** rappresentata da un certo attributo. L'incertezza rappresenta un valore di entropia alto di questa quantità, mentre più il risultato è certo allora più ci vuole una quantità inferiore di informazione per rappresentarlo. Ad esempio, per il nodo R abbiamo:

$$I(R) = \sum_{i=1}^L -P(c_i) \log_2 P(c_i)$$

dove  $\{c_1, \dots, c_L\}$  sono le L classi presenti nel nodo e  $P(c_i)$  è la probabilità di ottenere la classe  $c_i$  al nodo.

### Esempio:

Al nodo radice del problema del Ristorante  $c_1=True$  e  $c_2=False$ , ci sono 6 campioni veri e 6 falsi. Pertanto, abbiamo:

$$P(c_1) = \frac{\text{no. of samples} = c_1}{\text{total no. of samples}} = \frac{6}{6+6} = 0.5$$

$$P(c_2) = \frac{\text{no. of samples} = c_2}{\text{total no. of samples}} = \frac{6}{6+6} = 0.5$$

$$I(\text{Root}) = -0.5 \times \log_2 0.5 - 0.5 \times \log_2 0.5 = 1 \text{ bit}$$

La totale incertezza corrisponde ad 1 mentre la totale certezza corrisponderà a 0.

In generale, la quantità di informazioni sarà massima quando tutte le classi sono ugualmente probabili e sono minime quando il nodo è omogeneo (tutti i campioni hanno le stesse etichette).

L'informazione in una risposta quando è a priori  $\langle P_1, \dots, P_n \rangle$  è (chiamata anche **entropia** a priori):

$$I(\langle P_1, \dots, P_n \rangle) = \sum_{i=1}^n -P_i \log_2 P_i$$

## SCEGLIERE IL MIGLIOR ATTRIBUTO:

Quindi possiamo sfruttare l'entropia per decidere quali attributi conviene selezionare per prima, in altre parole, quali attributi dividono i campioni in classi che richiedono meno informazioni per andare avanti, quindi più certezze verso la classificazione. Più formalmente, un attributo A dividerà i campioni ad un nodo in diversi sottoinsiemi (o nodi figlio)  $E_{A=v_1}, \dots, E_{A=v_M}$  dove A ha M valori distinti  $\{v_1, \dots, v_m\}$ . Generalmente ogni sottoinsieme  $E_{A=v_i}$  avrà campioni con etichette diverse; quindi, se andiamo lungo quel ramo avremo bisogno di ulteriori  $I(E_{A=v_i})$  bit di informazione.

### Esempio:

Nel problema del ristorante, al ramo  $\text{Patrons} = \text{Full}$  del nodo radice, abbiamo  $c_1 = \text{True}$  con 2 campioni e  $c_2 = \text{False}$  con 4 campioni, quindi:

$$I(E_{\text{Patrons}=\text{Full}}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9183 \text{ bits}$$

Indichiamo quindi con  $P(A=v_i)$  come la probabilità che un campione segua il ramo  $A=v_i$ . Nel problema del ristorante, al nodo radice abbiamo:

$$\begin{aligned} P(\text{Patrons} = \text{Full}) &= \frac{\text{No. of samples where Patrons} = \text{Full}}{\text{No. of samples at the root node}} \\ &= \frac{6}{12} = 0.5 \end{aligned}$$

Dopo aver testato l'attributo A abbiamo bisogno di un resto di bit per classificare i campioni:

$$\text{Remainder}(A) = \sum_{i=1}^M P(A = v_i) I(E_{A=v_i})$$

Più sono omogenei i campioni, più  $I(E_{A=v_i})$  diventa basso, moltiplicandolo per la probabilità di avere quel valore per l'attributo A. Ottengo i bit per continuare il processo di classificazione.

Il **guadagno di informazioni** nel test dell'attributo A è la differenza tra il contenuto dell'informazione originale ed il contenuto delle nuove informazioni, cioè:

$$\begin{aligned} \text{Gain}(A) &= I(R) - \text{Remainder}(A) \\ &= I(R) - \sum_{i=1}^M P(A = v_i) I(E_{A=v_i}) \end{aligned}$$

dove  $\{E_{A=v_1}, \dots, E_{A=v_M}\}$  sono i nodi figlio di R dopo il test dell'attributo A.

L'idea chiave alla base della funzione SCEGLI-ATTRIBUTO dell'algoritmo consiste nello scegliere l'attributo che dà il massimo guadagno (GAIN) di informazioni.

## Esempio:

Nel problema del ristorante, dobbiamo decidere sulla radice se scegliere *Patrons* o *Type*:

$$\begin{aligned} Gain(Patrons) &= 1 - \frac{2}{12}(-\log_2 1) - \frac{4}{12}(-\log_2 1) - \frac{6}{12}(-\frac{2}{6}\log_2 \frac{2}{6} - \frac{4}{6}\log_2 \frac{4}{6}) \\ &\approx 0.5409 \text{ bits.} \end{aligned}$$

Con calcoli simili:

$$Gain(Type)=0$$

Confermando che *Patrons* è un attributo migliore di *Type*. Infatti, alla radice *Patrons* fornisce il maggior guadagno di informazioni.

## **RIASSUMENDO:**

Supponiamo di avere  $p$  esempi positivi e  $n$  esempi negativi alla radice:

→  $I(p/(p+n), n/(p+n))$  bit richiesti per classificare un nuovo esempio.

Esempio, per 12 ristoranti d'esempio,  $p = n = 6$ , quindi abbiamo bisogno di 1 bit.

Un attributo divide gli esempi  $E$  nei sottoinsiemi  $E_i$ , ognuno dei quali (si spera) richiedano meno informazione per completare la classificazione. Sia  $E_i$  caratterizzato da esempi positivi  $p_i$  e negativi  $n_i$ .

→  $I(p_i/(p_i+n_i), n_i/(p_i+n_i))$  bit richiesti per classificare un nuovo esempio.

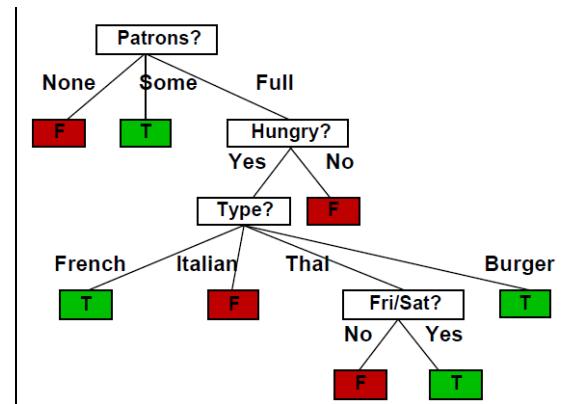
→ il numero di bit atteso per classificare un esempio su tutti i branch è:

$$\sum_i \frac{p_i+n_i}{p+n} I(p_i/(p_i+n_i), n_i/(p_i+n_i))$$

Per *Patrons?* è 0.459 bit, per *Type* è (ancora) di un 1 bit.

→ scegliere l'attributo che minimizza la quantità d'informazione necessaria rimasta.

Albero di decisione appreso da 12 esempi:



Sostanzialmente più semplice dell'albero visto prima. Uno scarso quantitativo di dati non giustifica la formulazione di un'ipotesi più complessa.

## **IMPURITÀ:**

Un punto di vista diverso è considerare **I(nodo)** come una misura di **impurità**. Più sono "misti" i campioni in un nodo (cioè proporzioni uguali di tutte le label di classe), maggiore è il valore di impurità. D'altra parte, un nodo **omogeneo** (cioè ha campioni di una sola classe) avrà impurità zero.

Il valore  $Gain(A)$  può quindi essere visto come la **quantità di riduzione dell'impurità** se dividiamo secondo A.

Ciò offre l'idea intuitiva che possiamo costruire alberi di decisione in modo ricorsivo cercando di ottenere nodi foglia che siano più puri possibile.

## **MISURAZIONE DELLE PRESTAZIONI:**

Come facciamo a sapere se l'albero di decisione è riuscito ad approssimare bene il campione in input?

Come facciamo a sapere che  $h \approx f$  ( $h$  approssima  $f$ )?

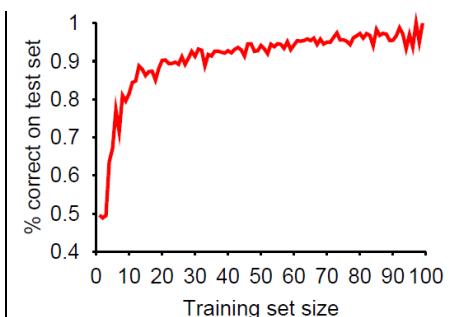
1. Usare teoremi sulla teoria dell'apprendimento computazionale/statistico;

2. Provare  $h$  su un nuovo insieme di esempi di test (usando la stessa distribuzione sullo spazio dell'esempio come insieme di training).

**Curva d'apprendimento** = % corretti sull'insieme di test in funzione delle dimensioni del training stesso.

**Overfitting**: rischio di utilizzare predici osservabili irrilevanti per generare un'ipotesi che sia d'accordo con tutti gli esempi nel training set. Il modello sceglie attributi che sono in realtà non rilevanti e potrei ignorarlo, si cerca di risolvere il problema con il pruning.

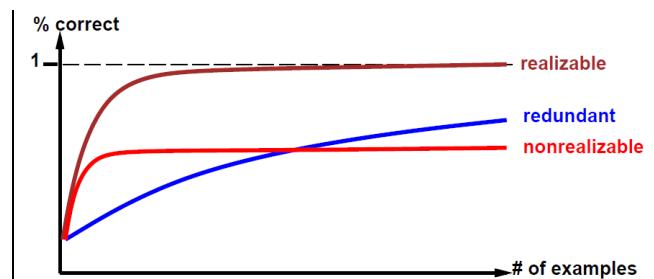
**Tree pruning**: termina la ricorsione quando # errori/gain è piccolo.



La curva di apprendimento dipende da:

- **Realizzabilità** (può esprimere la funzione target) vs **non realizzabilità**: La non realizzabilità può essere dovuta ad attributi mancanti o ad una classe d'ipotesi limitata;
- Espressività ridondante (ad esempio, molti attributi irrilevanti).

La realizzabilità con l'aumentare dei campioni ci porta quasi all'1 mentre la non realizzabilità si assesta su un valore ed anche con l'aumentare dei campioni restiamo su quel valore perché mancano informazioni cruciali per arrivare a percentuali alte.



### GENERALIZZAZIONE E OVERFITTING:

Quello che capita molto spesso è che la funzione  $h$  si va a adattare troppo ai dati di training, cioè va a catturare degli aspetti che sono irrilevanti per la funzione  $f$ . In tal caso, accade che quando inviamo nuovi dati la funzione non generalizza bene, invece di comportarsi come  $f$  va a considerare questi nuovi dati che sono irrilevanti per la classificazione e da delle risposte differenti. L'**overfitting** è quando la funzione  $h$  va a modellare troppo i dati di training, al contrario si ha l'**underfitting** quando la funzione  $h$  non riesce a modellare  $f$ , probabilmente lo spazio delle ipotesi o i dati sono pochi.

Qualitativamente, l'overfitting aumenta con la dimensione dello spazio delle ipotesi e del numero di attributi, ma diminuisce con il numero di esempi.

### DECISION TREE PRUNING:

L'Idea è combattere l'overfitting "generalizzando" gli alberi decisionali calcolati da DTL, sfoltendo i nodi irrilevanti.

Per la potatura dell'albero decisionale, ripetere quanto segue su un albero decisionale appreso:

- Trova un nodo di test **terminale**  $n$  (ha solo foglie come figli);
- Se il test è irrilevante, cioè ha un basso **information gain**, eliminalo sostituendo n con un nodo foglia.

**NOTA:** i nodi dove abbiamo poco guadagno di informazioni sono proprio le foglie.

Per vedere se eliminare o meno un nodo, nella pratica, si utilizzare un **test di significatività statistica**.

Un risultato ha **rilevanza statistica**, se la probabilità che possano derivare dall'ipotesi nulla (cioè l'assunzione che non vi sia un pattern sottostante) è molto bassa (di solito 5%).

### 12.3. VALUTARE E SCEGLIERE LA MIGLIORE IPOTESI

Il problema è che vogliamo apprendere un'ipotesi che si adatta meglio ai dati futuri. L'intuizione è che funziona solo se il training-set è "rappresentativo" per il processo sottostante.

L'idea è suppone che il training set sia indipendente e identicamente distribuito, vuol dire che gli elementi sono rappresentativi così che la funzione  $h$  può generalizzare.

Più formale, una sequenza di  $E_1, \dots, E_n$  delle variabili casuali è **indipendente e distribuita in modo identico** (IID), se è:

- **Indipendente**, la probabilità di avere un certo esempio non è condizionata dagli altri eventi, cioè  $P(E_j | E_{(j-1)}, E_{(j-2)}, \dots) = P(E_j)$ ;
- **Identicamente distribuito**, tutti gli esempi sono equamente probabili, cioè  $P(E_i) = P(E_j)$  per tutti gli  $i$  e  $j$ .

Esempio, una sequenza di lanci di dadi è IID.

**Ipotesi di stazionarietà:** assumiamo che l'insieme  $E$  di esempi sia IID in futuro.

### TASSI DI ERRORE E CROSS-VALIDATION:

Per andare a misurare quanto la funzione che abbiamo appreso si avvicina alla funzione  $f$ , andiamo a definire il **tasso di errore**. Il problema è che, avendo il training set, per sapere se il modello si sta comportando bene bisogna valutarlo su dati che non sono stati visti durante il training set e quindi viene valutato su un insieme nuovo di dati, chiamato **test set**.

Dato un problema di apprendimento induttivo  $\langle H, f \rangle$ , definiamo il **tasso di errore** di un'ipotesi  $h \in H$  come la frazione di errori, ovvero quante volte sbaglia rispetto al numero totale di valori del dominio:

$$\frac{\#\{x \in \text{dom}(f) \mid h(x) \neq f(x)\}}{\#\text{dom}(f)}$$

Un basso tasso di errore sul training set non significa che un'ipotesi si generalizzi bene.

La pratica di suddividere i dati disponibili per l'apprendimento in:

- un **training set**, da cui l'algoritmo di apprendimento produce un'ipotesi  $h$ ;
- un **test set**, che viene utilizzato per valutare  $h$ ,

è chiamato **holdout cross validation** (non è consentito sbirciare nel set di test).

Il problema ora è quanto grande deve essere il training set rispetto al test set:

- piccolo training set → scarse ipotesi, la funzione  $h$  non riesce a capire come si comporta  $f$ ;
- piccolo test set → scarsa stima dell'accuratezza.

Quello che si fa è una ***k-fold cross validation***, cioè eseguiamo  $k$  cicli di apprendimento, ciascuno con  $1/k$  dei dati come test set e una media sui  $k$  tassi di errore.

#### Esempio:

$k = 5$  e  $k = 10$  sono popolari → buona accuratezza con  $k$  volte il tempo di calcolo.

Se  $k=5$  con un campione di 1000, questo lo si divide in 5 gruppi (ognuno di 200 campioni). I primi 200 sono usati come test set e i restanti 800 come training set, questo si ripete per 5 volte, dopodiché, si effettua la media.

Esiste una versione estrema chiamata ***leave-one-out cross validation (LOOCV)*** che è il caso in cui si usa un solo campione come test e il resto come training e questo viene ripetuto  $n$  volte.

#### **SELEZIONE DEL MODELLO:**

Il problema della ***selezione del modello*** consiste nel determinare, a partire dai dati, un buon spazio di ipotesi in cui a cercare  $h$  (evitando l'overfitting).

Possiamo risolvere il problema di "apprendere  $f$  dalle osservazioni" in un processo in due parti:

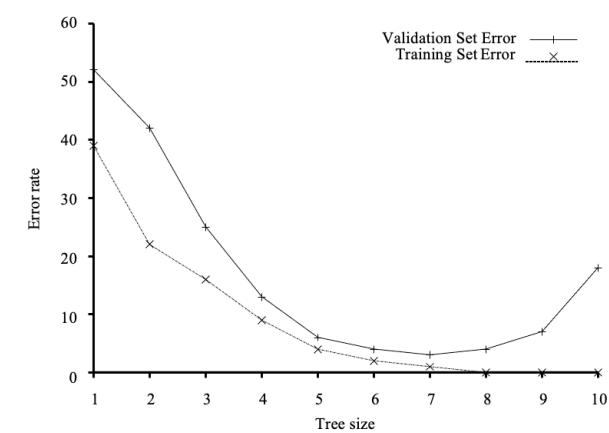
- la ***selezione del modello*** determina uno spazio di ipotesi  $H$ ;
- l'***ottimizzazione*** risolve il problema di apprendimento induttivo indotto ( $H, f$ ).

L'idea è risolvere le due parti insieme, iterando su "dimensione", ma serve una nozione di "dimensione" (ad es. numero di nodi in un albero decisionale).

Un problema concreto è trovare la "dimensione" che meglio bilancia overfitting e underfitting per ottimizzare la precisione del test set.

Si vanno a costruire tanti alberi di decisione con dimensione crescente e si calcola il tasso di errore, sia per il training che per il test set.

Nello schema si sceglie la dimensione 7 siccome successivamente la curva (tasso di errore) ricomincia a salire.



#### **DA ERROR-RATE A LOSS FUNCTION:**

Quello appena misurato si chiama ***error-rate***, ovvero quante volte  $h$  risponde in maniera diversa da  $f$ . In realtà, quando si vanno a costruire questi modelli non interessa misurare l'errore ma misurare la ***loss function***, cioè andare a vedere il tipo di errore ricevuto.

Ad esempio, supponiamo di creare un classificatore di mail, è molto peggio classificare ham (mail legittime) come spam che viceversa (perdita del messaggio). In questo caso, con l'error-rate sono classificati allo stesso modo (conta 1) mentre con la loss-function si può dare un peso differente rispetto al tipo di errore.

Così facendo, anziché parlare di errore si parla di utilità attesa, e quello che si vuole è ***massimizzare l'utilità attesa*** (MEU).

Quindi, l'apprendimento automatico dovrebbe massimizzare l'utilità (non solo ridurre al minimo i tassi di errore).

L'apprendimento automatico si occupa tradizionalmente di utilità sotto forma di "funzioni di loss".

La ***loss function***  $L$  è definita impostando  $L(x, y, \hat{y})$  come la quantità di utilità persa dalla predizione  $h(x) = y$  invece di  $f(x) = y$ .

Se  $L$  è indipendente da  $x$ , usiamo spesso  $L(y, \hat{y})$ .

#### Esempio:

$$L(\text{spam}, \text{ham}) = 1, \text{ mentre } L(\text{ham}, \text{spam}) = 10.$$

**Nota:**  $L(y, \hat{y}) = 0$  (nessuna perdita se hai esattamente ragione)

Le funzioni di loss popolari sono:

absolute value loss	$L_1(y, \hat{y}) :=  (y - \hat{y}) $
squared error loss	$L_2(y, \hat{y}) := (y - \hat{y})^2$
0/1 loss	$L_{0/1}(y, \hat{y}) := 0, \text{ if } y = \hat{y}, \text{ else } 1$

L'idea è massimizzare l'***utilità attesa*** scegliendo l'ipotesi  $h$  che riduce al minimo la loss attesa su tutte le coppie  $(x, y) \in f$ .

Sia  $\mathcal{E}$  l'insieme di tutti i possibili esempi e  $\mathbf{P}(X, Y)$  la distribuzione di probabilità a priori sulle sue componenti, allora la **generalization loss** per un'ipotesi  $h$  rispetto a una loss function  $L$  è:

$$\text{GenLoss}_L(h) := \sum_{(x,y) \in \mathcal{E}} L(y, h(x)) \cdot \mathbf{P}(x, y)$$

(se si ha il training set e la distribuzione di probabilità di  $X$  e  $Y$  possiamo definire la perdita generalizzata per una certa ipotesi rispetto ad una funzione di loss = si somma il prodotto per ogni campione di training moltiplichiamo la loss per la probabilità che quel campione appaia, quindi è pesata rispetto alla distribuzione di probabilità)

Quindi bisogna scegliere  $h$  che minimizza questa loss generalizzata, e l'ipotesi migliore è:

$$h^* := \underset{h \in \mathcal{H}}{\operatorname{argmin}} \text{GenLoss}_L(h).$$

Sia  $L$  una funzione di perdita ed  $E$  un insieme di esempi con  $\#(E) = N$ , allora chiamiamo:

$$\text{EmpLoss}_{L,E}(h) := \frac{1}{N} \left( \sum_{(x,y) \in E} L(y, h(x)) \right)$$

$$\hat{h}^* := \underset{h \in \mathcal{H}}{\operatorname{argmin}} \text{EmpLoss}_{L,E}(h)$$

Infine, la **loss empirica** e  $\hat{h}^*$  è la migliore ipotesi stimata.

Ci sono quattro ragioni per cui  $\hat{h}^*$  può differire da  $f$ :

- **Realizzabilità**: allora dobbiamo accontentarci di un'approssimazione  $\hat{h}^*$  di  $f$ ;
- **Varianza**: diversi sottoinsiemi di  $f$  danno diverse  $\hat{h}^*$  → abbiamo bisogno di più esempi;
- **Rumore**: se  $f$  non è deterministico, allora non possiamo aspettarci risultati perfetti;
- **Complessità computazionale**: se  $H$  è troppo grande per essere esplorato sistematicamente, usiamo un sottoinsieme e otteniamo un'approssimazione.

## REGOLARIZZAZIONE:

Per bilanciare complessità e loss è la **regolarizzazione**, ovvero si vuole selezionare un modello e quello che vogliamo scegliere è quello con la loss migliore, ma minimizzando la loss potremmo creare un modello troppo complesso e si va in overfitting.

Quello che si fa è andare a definire il **costo** come la somma pesata tra loss e complessità.

Sia  $\lambda \in \mathbb{R}$  (*lambda*),  $h \in \mathcal{H}$  ed  $E$  un insieme di esempi, chiamiamo:

$$\text{Cost}_{L,E}(h) := \text{EmpLoss}_{L,E}(h) + \lambda \text{Complexity}(h)$$

il costo totale di  $h$  su  $E$ .

Il processo per trovare un'ipotesi di minimizzazione dei costi totali

$$\hat{h}^* := \underset{h \in \mathcal{H}}{\operatorname{argmin}} \text{Cost}_{L,E}(h)$$

si chiama **regolarizzazione**. La complessità è chiamata **funzione di regolarizzazione** o complessità di ipotesi.

## 12.4. REGRESSIONE E CLASSIFICAZIONE CON MODELLI LINEARI

Nell'esempio precedente ci siamo concentrati sulla classificazione perché quello che volevamo classificare (il codominio) era booleano, più nel preciso, quando un valore è discreto parliamo di classificazione, mentre quando è continuo parliamo di **regressione**.

Chiamiamo un problema di apprendimento induttivo  $\langle H, f \rangle$  un **problema di classificazione** se il *co-dominio*( $f$ ) è discreto, e un **problema di regressione** se *co-dominio*( $f$ ) è continuo, cioè non discreto (di solito a valori reali).

Il caso più semplice di regressione è quando abbiamo una **funzione univariata**, è una funzione con un solo argomento.

**NOTA:** un mapping tra spazi vettoriali è detto lineare se conserva l'addizione vettoriale e la moltiplicazione scalare.

Una funzione univariata, lineare  $f: \mathbb{R} \rightarrow \mathbb{R}$  è della forma:

$$f(x) = w_1 x + w_0 \text{ per qualche } w_i \in \mathbb{R}.$$

Dato un vettore  $w := (w_0, w_1)$ , definiamo  $h_w(x) := w_1 x + w_0$

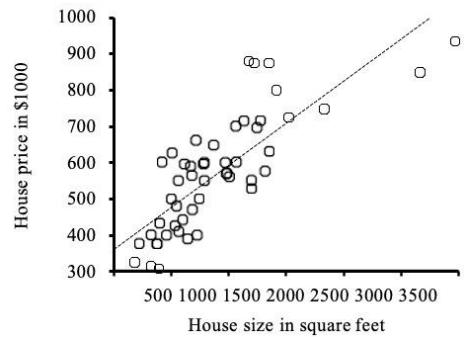
Dato un insieme di esempi  $E \subseteq \mathbb{R} \times \mathbb{R}$ , il compito di trovare l' $h_w$  che meglio si adatta a  $E$  è chiamato **regressione lineare**.

### Esempio:

Supponiamo di avere prezzi delle case rispetto ai feet<sup>2</sup> (metri quadri) nelle case vendute a Berkeley.

Vogliamo trovare i valori  $w_1$  e  $w_0$  che rappresentano la retta che minimizza l'errore quadratico, così che se viene fornito un prezzo, la soluzione dice quanto è grande la casa in metri quadri.

Ipotesi di funzione lineare che minimizza la perdita (quadrato dell'errore  $y = 0,232x + 246$ ).



L'idea è ridurre al minimo la loss come quadrato dell'errore  $L_2$  su  $\{(x_i, y_i) | i \leq N\}$  (utilizzato già da Gauss).

$$\text{Loss}(\mathbf{h}_w) = \sum_{j=1}^N L_2(y_j, h_w(x_j)) = \sum_{j=1}^N (y_j - h_w(x_j))^2 = \sum_{j=1}^N (y_j - w_1 x_j + w_0)^2$$

Il compito è trovare:

$$\mathbf{w}^* := \underset{\mathbf{w}}{\operatorname{argmin}} \text{Loss}(\mathbf{h}_w)$$

Ricordando che  $\sum_{j=1}^N (y_j - w_1 x_j + w_0)^2$  è minimizzato, quando le derivate parziali rispetto ai  $w_i$  sono zero, cioè quando:

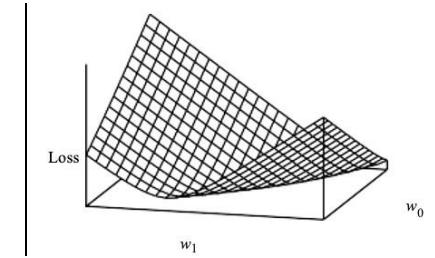
$$\frac{\partial}{\partial w_0} \left( \sum_{j=1}^N (y_j - w_1 x_j + w_0)^2 \right) = 0 \quad \text{and} \quad \frac{\partial}{\partial w_1} \left( \sum_{j=1}^N (y_j - w_1 x_j + w_0)^2 \right) = 0$$

Queste equazioni hanno una soluzione unica, andando a risolvere l'equazione iniziale:

$$w_1 = \frac{N(\sum_j x_j y_j) - (\sum_j x_j)(\sum_j y_j)}{N(\sum_j x_j^2) - (\sum_j x_j)^2} \quad w_0 = \frac{(\sum_j y_j) - w_1(\sum_j x_j)}{N}$$

Andando a disegnare i valori di  $w_1$  e  $w_0$  si ottiene il grafico:

Possiamo variare i valori  $w_i$  affinché la loss vada verso un minimo (il punto più basso è il minimo di loss, ovvero ciò che si vuole).



NOTA: molte forme di apprendimento implicano la regolazione dei pesi per ridurre al minimo le perdite.

Lo **spazio dei pesi** è lo spazio di tutte le possibili combinazioni di pesi. La minimizzazione della perdita in uno spazio di peso è chiamata **adattamento del peso**.

Lo spazio dei pesi della regressione lineare univariata è  $R^2$  → possiamo tracciare graficamente la funzione di perdita su  $R^2$ .

### DISCESA DEL GRADIENTE:

Il problema è che non usiamo sempre  $R^2$ , ci vuole un approccio più generale che funziona per qualsiasi loss. Possiamo andare a cercare il valore di minimo usando un algoritmo di ricerca locale, in particolare l'algoritmo (hill-climbing) **discesa del gradiente**.

In pratica, se si prende il grafico precedente, il gradiente in un punto ci dà la direzione di crescita, si prende la direzione opposta (ovvero la discesa) e si individua il minimo. L'algoritmo aggiorna i singoli pesi andando nella direzione opposta del gradiente calcolato con la derivata rispetto ad ogni singolo peso.

Il parametro  $\alpha$  è chiamato **learning rate**, che è la grandezza dei passi per andare verso il minimo, più è piccolo più tempo è richiesto ma la probabilità aumenta di arrivarci al valore minimo. Può essere una costante fissa o può decadere man mano che l'apprendimento procede.

```
function gradient_descent(f, w, alpha)
    inputs: a differentiable function f and initial weights w = (w0, w1).
    loop until w converges do
        for each wi do
            wi ← wi - alpha ∂/∂wi(f(w))
        end for
    end loop
```

La **discesa del gradiente** aggiorna i pesi con la seguente formula:

$$\frac{\partial \text{Loss}(w)}{\partial w_i} = \frac{\partial (y - h_w(x))^2}{\partial w_i} = 2(y - h_w(x)) \frac{\partial (y - w_1 x + w_0)}{\partial w_i}$$

e così abbiamo:

$$\frac{\partial \text{Loss}(w)}{\partial w_0} = -2(y - h_w(x)) \quad \frac{\partial \text{Loss}(w)}{\partial w_1} = -2(y - h_w(x))x$$

Collegandolo negli aggiornamenti della discesa del gradiente:

$$w_0 \leftarrow w_0 - \alpha - 2(y - h_w(x)) \quad w_1 \leftarrow w_1 - \alpha - 2(y - h_w(x))x$$

Analogamente per  $n$  esempi di addestramento  $(x_j, y_j)$ :

$$w_0 \leftarrow w_0 - \alpha \left( \sum_j -2(y_j - h_w(x_j)) \right) \quad w_1 \leftarrow w_1 - \alpha \left( \sum_j -2(y_j - h_w(x_n)) x_n \right)$$

Questi aggiornamenti costituiscono la **regola di apprendimento** della discesa del gradiente batch per la regressione lineare univariata. La convergenza all'unico minimo di perdita globale è garantita (a patto che scegliamo  $\alpha$  sufficientemente piccolo), ma potrebbe essere molto lenta.

### REGRESSIONE LINEARE MULTIVARIATA:

Una **funzione multivariata** o n-aria è una funzione con uno o più argomenti. Possiamo usarla per la regressione lineare multivariata. Le formule usate prima possono essere usate ma bisogna passare da coppie di valori a vettori.

L'idea è che ogni esempio  $\vec{x}_j$  è un vettore di  $n$  elementi e lo spazio delle ipotesi è l'insieme di funzioni:

$$h_{sw}(\vec{x}_j) = w_0 + w_1 x_{j,1} + \dots + w_n x_{j,n} = w_0 + \sum_i w_i x_{j,i}$$

Il trucco è introduciamo  $x_{j,0} := 1$  e usiamo la notazione matriciale:

$$h_{sw}(\vec{x}_j) = \vec{w} \cdot \vec{x}_j = \vec{w}^t \vec{x}_j = \sum_i w_i x_{j,i}$$

Il miglior vettore di pesi,  $w^*$ , **minimizza** la perdita di errore quadratico sugli esempi:

$$w^* := \underset{w}{\operatorname{argmin}} \left( \sum_j L_2(y_j)(w \cdot \vec{x}_j) \right)$$

La discesa del gradiente raggiungerà il minimo (unico) della funzione di perdita; l'equazione di aggiornamento per ogni peso

$$w_i \leftarrow w_i - \alpha \left( \sum_j x_{j,i} (y_j - h_w(\vec{x}_j)) \right)$$

Possiamo anche risolvere analiticamente il  $w^*$  che **minimizza** la perdita.

Sia  $\vec{y}$  il vettore di output per gli esempi di addestramento e  $X$  sia la matrice di dati, ovvero la matrice di input con un esempio  $n$  dimensionale per riga.

Allora la soluzione è  $w^* = (X^t X)^{-1} X^t \vec{y}$  che **minimizza l'errore quadratico**.

### REGRESSIONE LINEARE MULTIVARIATA (REGOLARIZZAZIONE):

**NOTA:** la regressione lineare univariata non soffre di overfit, ma nel caso multivariato potrebbero esserci "dimensioni ridondanti" che si traducono in un overfitting.

L'idea è utilizzare la **regolarizzazione** con una funzione di complessità basata sui pesi.

$$\text{Complexity}(h_w) = L_q(w) = \sum_i |w_i|^q$$

**Avvertenza:** non confonderlo con le loss function  $L_1$  e  $L_2$ .

Il problema è quale  $q$  dovrebbe essere scelto ( $L_1$  e  $L_2$  minimizzano la somma dei valori assoluti / dei quadrati), ma questo dipende dall'applicazione.

**NOTA:** la regolarizzazione  $L_1$  tende a produrre un **modello sparso**, ovvero imposta molti pesi a 0, dichiarando di fatto gli attributi irrilevanti. Le ipotesi che scartano gli attributi possono essere più facili da comprendere per un essere umano e potrebbero avere meno probabilità di overfit.

### CLASSIFICAZIONI LINEARI CON SOGLIA RIGIDA:

L'idea è che il risultato della regressione lineare può essere utilizzato per la classificazione. Fino ad ora avevamo un insieme di dati di training (i punti nello spazio), determinato i pesi della funzione (lineare) quindi una retta, si può usare questa retta per fare classificazione dicendo che tutto ciò che è presente al di sopra della retta appartiene ad una specifica classe e tutto ciò che sta sotto la retta appartiene ad un'altra classe.

Esempio (Verifica del divieto di test nucleari):

Grafici dei parametri dei dati sismici: magnitudo dell'onda di volume  $x_1$  vs. magnitudo dell'onda superficiale  $x_2$ .

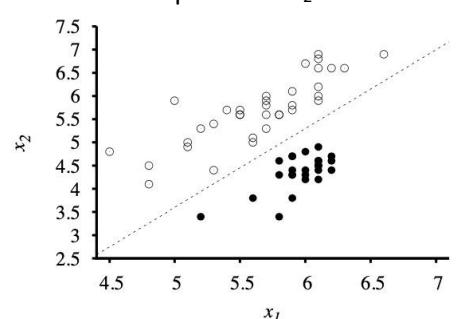
Bianco: terremoti, nero: esplosioni sotterranee

Inoltre:  $h_{w^*}$  come confine di decisione  $x_2 = 17x_1 - 4.9$ .

Un confine di decisione è una linea (o una superficie, in dimensioni superiori) che separa due classi di punti. Un confine di decisione lineare è chiamato separatore lineare e i dati che ne ammettono uno sono chiamati linearmente separabili.

Per l'esempio, il separatore lineare è definito da  $-4.9 + 1.7x_1 - x_2 = 0$ , le esplosioni sono caratterizzate da  $-4.9 + 1.7x_1 - x_2 > 0$ , terremoti da  $-4.9 + 1.7x_1 - x_2 < 0$ .

Trucco: se introduciamo la coordinata fittizia  $x_0 = 1$ , allora possiamo scrivere l'ipotesi di classificazione come  $h_w(x) = 1$  se  $w^*x > 0$  e 0 altrimenti.



## CLASSIFICAZIONI LINEARI CON SOGLIA RIGIDA (REGOLA DEL PERCETTRONE):

Quindi  $h_w(x) = 1$  se  $w^*x > 0$  e 0 altrimenti è ben definito, come scegliere  $w$ ?

Anziché scrivere  $h_w(x) = T(w^*x)$ , possiamo scrivere  $T(z) = 1$ , se  $z > 0$  e  $T(z) = 0$  altrimenti, chiamiamo  $T$  **funzione di soglia**.

Il problema è che  $T$  non è derivabile e  $\frac{\partial T}{\partial z} = 0$  è definito, non si può usare:

- Nessuna soluzione in forma chiusa impostando;
- Nemmeno i metodi di discesa del gradiente nello spazio dei pesi funzionano.

C'è un'altra tecnica dove, possiamo apprendere i pesi ripetendo la seguente regola:

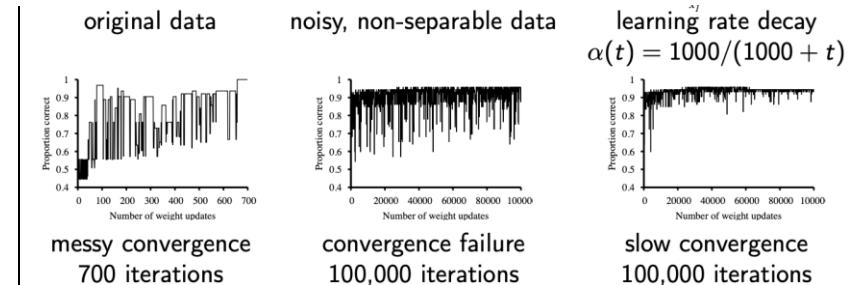
Dato un esempio  $(x, y)$ , la regola di **apprendimento del perceptron** è

$$w_i \leftarrow w_i + \alpha \cdot (y - h_w(x)) \cdot x_i$$

Poiché stiamo considerando la classificazione 0/1, ci sono tre possibilità:

1. Se  $y = h_w(x)$ , allora  $w_i$  rimane invariato;
2. Se  $y = 1$  e  $h_w(x) = 0$ , allora  $w_i$  è incrementato/diminuito sé  $x_i$  è positivo/negativo (vogliamo rendere  $w^*x$  più grande in modo che  $T(w^*x) = 1$ );
3. Se  $y = 0$  e  $h_w(x) = 1$ , allora  $w_i$  è decrescente/aumentato se  $x_i$  è positivo/negativo (vogliamo per ridurre  $w^*x$  in modo che  $T(w^*x) = 0$ ).

Curve di apprendimento (plots dell'accuratezza totale del **training set** rispetto al numero di iterazioni) per la regola del percettrone sui dati di terremoti/esplosioni:



Trovare l'ipotesi dell'errore minimo è NP hard, ma è possibile con il decadimento del **learning rate**.

## CLASSIFICAZIONI LINEARI CON REGRESSIONE LOGISTICA:

Finora abbiamo visto che passando l'output di una funzione lineare attraverso una funzione di soglia  $T$  si ottiene un classificatore lineare. Ma il problema è che la natura difficile di  $T$  porta problemi:

- $T$  non è differenziabile ne continuo, l'apprendimento tramite la regola del percettrone diventa imprevedibile;
- $T$  è "eccessivamente preciso" vicino al confine, necessita di giudizi più graduati.

L'idea è ammorbidente la soglia, approssimarla con una funzione differenziabile.

Usiamo la funzione **logistica standard**:

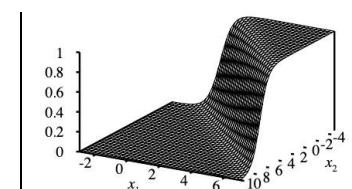
$$l(x) = \frac{1}{1+e^{-x}}$$

Quindi abbiamo:

$$h_w(x) = l(w \cdot x) = \frac{1}{1+e^{-(w \cdot x)}}$$

**(Ipotesi di regressione logistica nello spazio peso)** Grafico di un'**ipotesi di regressione logistica** per i dati di terremoto/esplosione. Il valore in  $(w_0, w_1)$  è la probabilità di appartenere alla classe etichettata 1.

In tal caso, parliamo intuitivamente della **scogliera** nel classificatore.



## REGRESSIONE LOGISTICA:

Il processo di adattamento del peso in  $h_w(x) = \frac{1}{1+e^{-(w \cdot x)}}$  viene chiamato **regressione logistica**.

Non esiste una soluzione semplice in forma chiusa, ma la discesa del gradiente è una regressione logistica semplice.

Poiché le nostre ipotesi hanno un output continuo, utilizziamo la funzione di perdita di errore quadratica  $L_2$ .

Per un esempio  $(x, y)$  calcoliamo le derivate parziali (tramite **chain rule**):

$$\begin{aligned} \frac{\partial}{\partial w_i} (L_2(w)) &= \frac{\partial}{\partial w_i} (y - h_w(x))^2 \\ &= 2 \cdot h_w(x) \cdot \frac{\partial}{\partial w_i} (y - h_w(x)) \\ &= -2 \cdot h_w(x) \cdot l'(w \cdot x) \cdot \frac{\partial}{\partial w_i} (w \cdot x) \\ &= -2 \cdot h_w(x) \cdot l'(w \cdot x) \cdot x_i \end{aligned}$$

La derivata della funzione logistica soddisfa  $l'(z) = l(z)(1-l(z))$ , quindi:

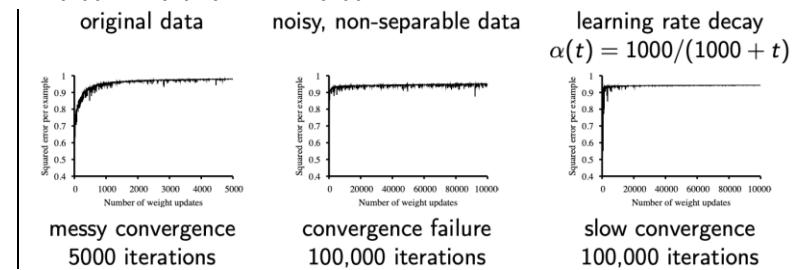
$$l'(\mathbf{w} \cdot \mathbf{x}) = l(\mathbf{w} \cdot \mathbf{x})(1 - l(\mathbf{w} \cdot \mathbf{x})) = h_{\mathbf{w}}(\mathbf{x})(1 - h_{\mathbf{w}}(\mathbf{x}))$$

La regola per l'aggiornamento logistico (aggiornamento del peso per minimizzare la perdita) è:

$$\mathbf{w}_i \leftarrow \mathbf{w}_i + \alpha \cdot (y - h_{\mathbf{w}}(\mathbf{x})) \cdot h_{\mathbf{w}}(\mathbf{x}) \cdot (1 - h_{\mathbf{w}}(\mathbf{x})) \cdot \mathbf{x}_i$$

Rifacendo le curve di training:

Risulta che l'**aggiornamento logistico** sembra funzionare meglio dell'**aggiornamento del percettrone**.



## 12.5. MODELLI NON PARAMETRICI

Ricapitolando, si vuole fare apprendimento tramite esempi (campioni). Dato un training set, insieme di valori/vettori  $\mathbf{x}$  con associata una label di classificazione, l'obiettivo è cercare di apprendere una funzione  $h$  di ipotesi che va a modellare la funzione  $f$  effettiva che corrisponde al mapping tra  $\mathbf{x}$  e  $y$ . Abbiamo visto alcuni casi di classificazione dove la  $y$  è discreta o continua (in tal caso diventa un problema di regressione) e poi con gli alberi di decisione. Si è poi stimata una funzione lineare per poi passare a funzioni multivariate.

Quindi, a partire dalla coppia  $\mathbf{x}$  e  $y$  si è definito i valori dei pesi  $w$  dove  $\mathbf{x}^*w$  approssima  $y$ , che è il risultato dell'apprendimento.

I modelli precedenti sono parametrici perché hanno come parametro quanti pesi dobbiamo stimare (es. nel classificatore lineare bisogna determinare  $w_0$  e  $w_1$ ). I metodi di apprendimento parametrico sono spesso semplici ed efficaci, ma semplificano eccessivamente ciò che realmente succede.

Esistono altri tipi di modelli, chiamati **modelli non parametrici** dove si vuole che la complessità del modello sia in funzione dei dati. In un classificatore lineare (il più semplice) abbiamo molti dati, accadrà che il modello non si adatta al funzionamento dei dati siccome non ha la capacità espressiva dell'ipotesi di partenza, cioè lo spazio delle ipotesi troppo piccolo; pertanto, non si riesce a modellare la funzione presente dietro ai dati. Con i modelli non parametrici, il parametro esterno non è presente, quindi il modello che si va a costruire dipende solamente dai dati; pertanto, più dati si hanno più il modello riuscirà a catturare tutti gli aspetti del dataset, in altre parole, l'apprendimento non parametrico permette alla complessità delle ipotesi di crescere con i dati; quindi, lo spazio delle ipotesi è in funzione del dataset.

L'apprendimento basato sulle istanze è non parametrico siccome costruisce delle ipotesi direttamente dai dati di training.

### MODELLI NEAREST-NEIGHBOR:

Il primo modello non parametrico si chiama **modello K-NN** che è basato sulla vicinanza.

L'idea chiave è che i vicini sono simili:

- Sia  $X$  un insieme di esempi etichettati (il training set);
- Dato un punto  $x$  da classificare (un nuovo campione non presente nel training set), si calcola l'insieme  $U$  dei  $k$  punti dell'insieme  $X$  più vicini a  $x$ , secondo una determinata metrica;
- Si calcola la classe  $C$  più frequente all'interno dell'insieme  $U$ ;
- Si assegna  $x$  a  $C$ ;

Un aspetto importante è come definire e quanto deve essere grande il **vicinato  $N$** :

- Se è troppo piccolo, non ci sono data point;
- Se è troppo grande, la densità è la stessa ovunque;
- Una soluzione è quella di definire  $N$  per contenere  $k$  point, dove  $k$  è grande abbastanza da assicurare una stima significativa (di solito tra 5 e 10).

Un altro aspetto è come determinare i punti vicini (Quale data point è più vicino a  $x$ ?). Abbiamo quindi bisogno di una distanza metrica  $D(x_1, x_2)$  e la distanza euclidea  $D_E$  è la più popolare. Quando però ogni dimensione misura qualcosa di diverso è inappropriato usare  $D_E$ , è importante standardizzare la scala di ogni dimensione, in tal caso, la distanza di Mahalanobis è una soluzione. Invece, le caratteristiche discrete dovrebbero essere trattate diversamente, magari con la distanza di Hamming.

La difficoltà di determinare i punti vicini è che più aumenta la dimensione, più è complesso determinare i vicini.

## SUPPORT-VECTOR MACHINES:

L'approccio utilizzato fino a qualche anno fa è il support-vector machines (SVM oppure support-vector networks) che sono modelli di apprendimento supervisionato (non parametrico) per la classificazione e la regressione.

Le caratteristiche del modello SVM sono:

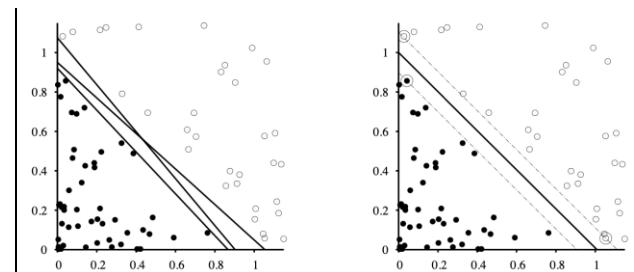
- I dati di training vengono separati tramite un separatore così da separare gli elementi di una classe da un'altra classe costruendo un separatore di **margine massimo**, ovvero un confine di decisione con la maggiore distanza possibile dai punti di esempio. Questo li aiuta a **generalizzare** bene.
- Si può utilizzare anche quando lo spazio non è separabile usando una funzione kernel, può incorporare i dati in uno spazio a dimensione superiore, dove è linearmente separabile dal **trucco del kernel**.
- L'iperpiano di separazione è un'ipersuperficie nei dati originali, danno priorità agli esempi critici (**vettori di supporto**), ovvero tutti i punti vicino al separatore sono quelli più interessanti, tutti gli altri non vengono memorizzati siccome non servono all'algoritmo per fare classificazione (**migliore generalizzazione**).

Uno degli approcci più diffusi per l'apprendimento supervisionato è "off-the-shelf".

Quindi, dato un dataset E linearmente separabile, il **separatore di margine massimo** è il **separatore lineare** s che **massimizza** il **margine**, ovvero la distanza di E da s.

### Esempio:

Tutte le righe a sinistra sono validi separatori lineari che separano la classe dei punti neri dalla classe dei punti bianchi:



Ci aspettiamo che il separatore del margine massimo a destra si generalizzi meglio.

L'idea è ridurre al minimo la generalized loss prevista anziché la loss empirica.

## TROVARE IL SEPARATORE DI MARGINE MASSIMO:

Rispetto agli algoritmi precedenti, le label non sono 1/0 ma 1/-1 per indicare le due classi, pertanto abbiamo un training set  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  dove:

- $y_i \in \{-1, 1\}$  (invece di  $\{1, 0\}$ );
- $x_i \in \mathbb{R}^p$  (classificazione multilineare).

L'obiettivo è trovare l'iperpiano che va a separare al massimo i punti con  $y_i = -1$  da quelli con  $y_i = 1$ . Questo iperpiano possiamo rappresentarlo con un insieme di punti tale che si ha  $\{x | (w^*x) + b = 0\}$ , dove:

- w è il vettore normale (non necessariamente normalizzato) dell'iperpiano;
- Il parametro b determina l'offset dell'iperpiano dall'origine lungo il vettore normale w.

L'idea è utilizzare la discesa del gradiente per cercare lo spazio di tutti w e b per massimizzare le combinazioni.

## CASO SEPARABILE:

Innanzitutto definiamo come si può definire il margine, ovvero il margine è delimitato dai due iperpiani descritti da  $(w \cdot x) + b = -1$  (**limite inferiore**) e  $(w \cdot x) + b = 1$  (**limite superiore**), mentre la distanza tra loro è  $\frac{2}{\|w\|_2}$  per massimizzare il margine, minimizzare  $\|w\|_2$  mantenendo  $x_i$  fuori dal margine.

I vincoli sono che:  $(w^*x_i) + b \geq 1$  per  $y_i = 1$  e  $(w^*x_i) + b \leq -1$  per  $y_i = -1$  o semplicemente  $y_i(w^*x_i - b) \geq 1$  per  $1 \leq i \leq n$ .

Problema di ottimizzazione: minimizzare  $\|w\|_2$  mentre  $y_i(w^*x_i - b) \geq 1$  per  $1 \leq i \leq n$ .

Dopo alcuni passaggi arriviamo ad una **rappresentazione alternativa** (diventando un problema di ottimizzazione):

$$\underset{\alpha}{\operatorname{argmax}} \left( \sum_j \alpha_j - \frac{1}{2} \left( \sum_{j,k} \alpha_j \alpha_k y_j y_k (x_j \cdot x_k) \right) \right) \quad \text{sotto i vincoli } \alpha_j \geq 0 \text{ e } \sum_j \alpha_j y_j = 0.$$

Questa equazione ha tre proprietà importanti:

- L'espressione è convessa, quindi il massimo può essere trovato in modo efficiente ed è unico;
- I dati entrano nell'espressione solo sotto forma di prodotti scalari di coppie di punti, quindi una volta che sono stati calcolati gli  $\alpha_i$  ottimali, abbiamo:

$$h(x) = \operatorname{sign} \left( \sum_j \alpha_j y_j (x \cdot x_j) - b \right)$$

- I pesi  $\alpha_j$  associati a ciascun punto sono zero tranne che in corrispondenza dei vettori di supporto, i punti più vicini al separatore, non c'è bisogno di tenere in memoria tutti i punti.

Una volta trovato un vettore ottimale  $\alpha$ , utilizziamo  $w = \sum_j \alpha_j x_j$

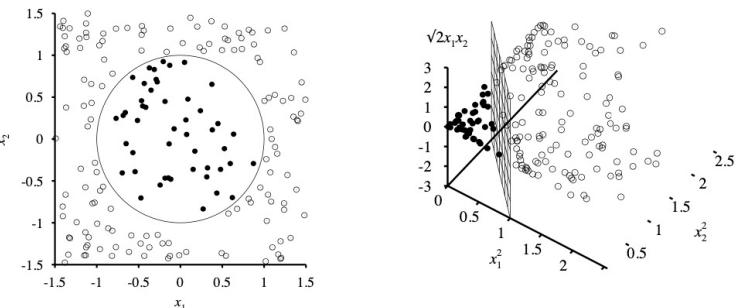
## SUPPORT-VECTOR MACHINES (KERNEL TRICK):

Quando i ***dati non sono separabili***, l'idea è di passare da una determinata dimensione ad una superiore. È dimostrabile che se abbiamo un certo numero di campioni si può passare da uno spazio non separabile ad uno separabile aumentando la dimensione.

### Esempio:

- A sinistra: il vero limite di decisione è  $x_1^2 + x_2^2 \leq 1$ ;
- A destra: mappare uno spazio di input tridimensionale  $\langle x_1^2, x_2^2, \sqrt{2}x_1x_2 \rangle$  è separabile da un iperpiano.

Risultato: Mappiamo ogni vettore di input  $x$  su una  $F(x)$  con  $f_1 = x_1$ ,  $f_2 = x_2$  e  $f_3 = \sqrt{2}x_1x_2$ .



Esistono diverse funzioni  $F(x)$  per l'aumento della dimensione ed hanno diverse proprietà con relativi vantaggi.

Notare che, visto che va fatto il prodotto tra  $x_i \cdot x_j$  nella formula precedente, in questo caso sarà  $F(x_i) \cdot F(x_j)$  nell'equazione SVM. In realtà, se definiamo  $F()$  secondo determinati criteri, come ad esempio fare il quadrato delle  $x$ , cioè  $x_1^2$  e  $x_2^2$  e sulla terza dimensione è proprio  $\sqrt{2}x_1x_2$ , allora per calcolare il prodotto equivale a fare il prodotto delle  $x^2$ , riassumendo:

$$\text{Se } F(x) = \langle x_1^2, x_2^2, \sqrt{2}x_1x_2 \rangle \text{ allora } F(x_i) \cdot F(x_j) = (x_i \cdot x_j)^2$$

Chiamiamo la funzione  $(x_i \cdot x_j)^2$  una ***funzione kernel*** (ce ne sono altre).

Sia  $X$  un insieme non vuoto, a volte indicato come insieme di indici. Una funzione simmetrica  $K: X \times X \rightarrow \mathbb{R}$  è chiamata funzione kernel su  $X$  se e solo se:

$$\sum_{i,j=1}^n c_i c_j K(x_i, x_j) \geq 0 \text{ for any } x_i \in X, n \in \mathbb{N}, \text{ and } c_i \in \mathbb{R}$$

Quindi la formula precedente per spazi separabili, nel caso di spazi non separabili, diventa:

$$\underset{\alpha}{\operatorname{argmax}} \left( \sum_j \alpha_j - \frac{1}{2} \sum_{j,k} \alpha_j \alpha_k y_j y_k K(x_j, x_k) \right) \quad \text{dove } K \text{ è una } \textcolor{red}{\text{funzione kernel}}.$$

La funzione  $K(x_j, x_k) = (1 + x_j \cdot x_k)^d$  è una funzione kernel corrispondente a uno spazio delle caratteristiche la cui dimensione ed esponenziale in  $d$ . Si chiama ***kernel polinomiale***.

## 12.6. ENSEMBLE LEARNING

Abbiamo visto che a partire dai dati possiamo costruire modelli che però possono avere problemi (over/underfitting) e potrebbero essere non performanti. L'idea dell'***ensemble learning*** è cercare di costruire più modelli (non uguali) così da catturare caratteristiche differenti del problema, cosicché mettendoli assieme è più probabile che la classificazione sia fatta in maniera corretta e performano meglio.

L'idea dell'ensemble learning è quella di selezionare una raccolta, o insieme, di ipotesi,  $h_1, h_2, \dots, h_n$  (anziché una sola ipotesi) e combinare le loro previsioni con un criterio, ad esempio facendo la media, votazione o attraverso un altro livello di apprendimento automatico. Chiamiamo le singole ipotesi ***modelli base*** e loro combinazione un ***modello ensemble***.

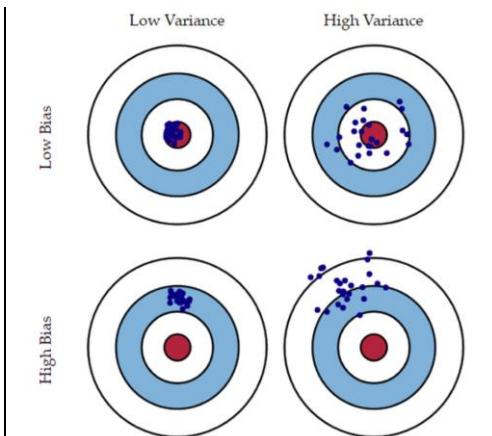
Dal punto di vista teorico, un modello di ensemble learning va ad affrontare due problemi di cui soffrono i classificatori:

1. Ridurre il ***bias*** (distorsione);
2. Ridurre la ***varianza***.

Tramite dei passaggi si può separare l'errore ottenuto in 3 errori:

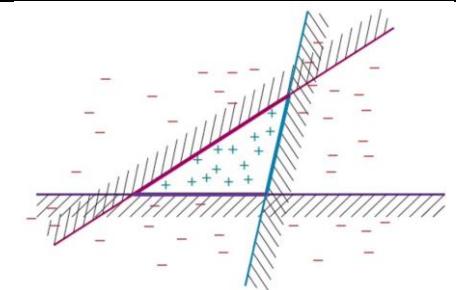
$$\mathbb{E}[(y - t)^2] = \underbrace{(y_* - \mathbb{E}[y])^2}_{\text{bias}} + \underbrace{\text{Var}(y)}_{\text{variance}} + \underbrace{\text{Var}(t)}_{\text{Bayes error}}$$

↗ Perdita attesa



Un ensemble di ***tre classificatori*** lineari può rappresentare una regione triangolare che non potrebbe essere rappresentata da un ***singolo classificatore*** lineare.

Un ensemble di ***n classificatori*** lineari consente di realizzare più funzioni.



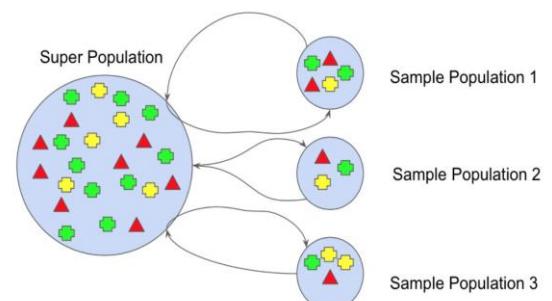
## BAGGING (TECNICA DI ENSEMBLE LEARNING):

L'idea è di costruire tanti training set (non solo uno come in precedenza) in maniera casuale facendo campionamento, può capitare che lo stesso campione rientri in più training set. A partire dai K training set, si costruiscono i modelli di base, dopodiché si fa classificazione ed il risultato lo si può unire secondo delle funzioni (tipo per maggioranza).

Più formalmente, nel **bagging**, generiamo K distinti training set campionando con sostituzione dal training set originale.

Il bagging tende a ridurre la **varianza** ed è un approccio standard quando i dati sono limitati o quando si ritiene che il modello di base sia in overfitting.

Lo si usa per lo più su alberi di decisione.



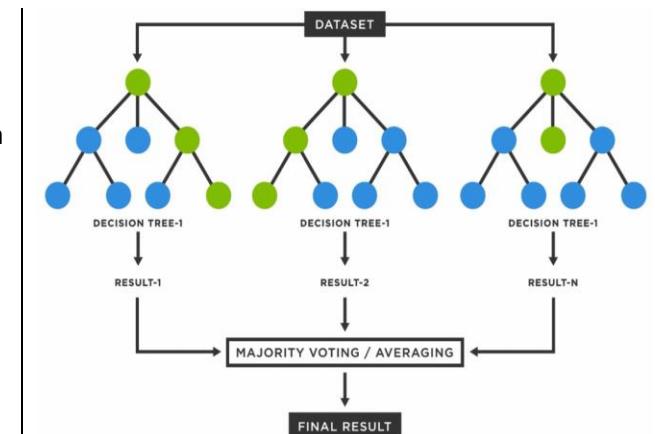
## RANDOM FORESTS (TECNICA DI ENSEMBLE LEARNING):

Un'altra tecnica usata con gli alberi di decisione sono i **random forest**. L'idea è costruire tanti alberi di decisione differenti tra loro in modo che gli attributi non si presentano sempre nello stesso ordine ma vengono mischiati casualmente, siccome in base all'ordine degli attributi abbiamo un comportamento differente. I dati di training sono gli stessi ma la differenza sta come vengono costruiti gli alberi.

Il modello della random forest è una forma di **bagging di alberi decisionali** in cui adottiamo ulteriori passaggi per rendere l'insieme di alberi K più diversificato, per ridurre la **varianza**.

Le foreste casuali possono essere utilizzate per la classificazione o la regressione.

L'idea chiave è di variare casualmente le scelte degli attributi (piuttosto che gli esempi di addestramento).

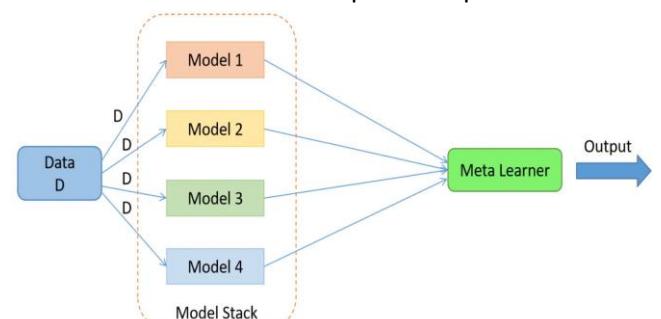


## STACKING (TECNICA DI ENSEMBLE LEARNING):

Un'altra tecnica è lo **stacking** (o generalizzazione impilata), si costruisce uno stack dove abbiamo più livelli per classificare.

In questo modello, i dati sono sempre gli stessi (al contrario della tecnica precedente) e si va a costruire usando modelli differenti (es. albero di decisione, modello logistico...), dopodiché viene costruito e allenato un altro modello che prende l'output di ogni modello usato.

Questo modello di solito riduce il **bias**, ovvero riducendo l'errore di classificazione.

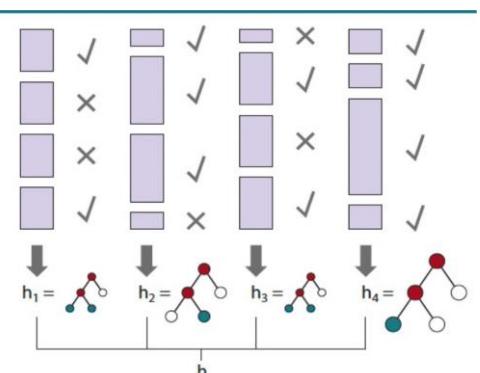


## BOOSTING (TECNICA DI ENSEMBLE LEARNING):

Il modello più utilizzato è il **boosting**, l'idea è introdurre dei pesi. Si costruisce un modello  $h_1$ , lo si testa e si evidenziano i campioni che ha sbagliato a classificare, così va a costruire un altro modello  $h_2$  in cui i campioni mal classificati in  $h_1$  hanno un peso maggiore. In sostanza si costruisce man mano un nuovo modello che riesca a classificare correttamente i campioni classificati male al passo precedente. Il numero di classificatori (ipotesi  $h_i$ ) è un parametro.

Più formalmente, abbiamo un training set pesato, in cui ogni esempio ha un peso associato  $W_j \geq 0$  che descrive quanto l'esempio dovrebbe contare durante il training.

Le ipotesi che si sono comportate meglio nel training hanno un peso maggiore nella votazione.



## 13. APPRENDIMENTO STATISTICO

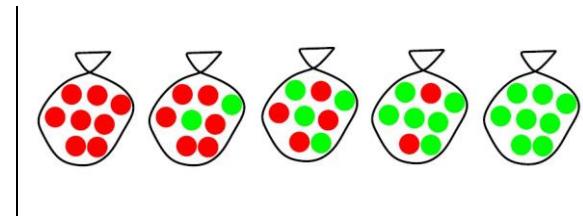
Vogliamo fare apprendimento ma abbiamo bisogno di:

- **Dati:** istanziazioni di alcune o tutte le variabili casuali che descrivono il dominio. Rappresentato dalle prove;
- **Ipotesi:** teorie probabilistiche di come funziona un dominio.

Esempio:

Supponiamo ci siano cinque tipi di sacchetti di caramelle:

- $h_1$ : 100% caramelle alla ciliegia;
- $h_2$ : 75% caramelle alla ciliegia + 25% caramelle al lime;
- $h_3$ : 50% caramelle alla ciliegia + 50% caramelle al lime;
- $h_4$ : 25% caramelle alla ciliegia + 75% caramelle al lime;
- $h_5$ : 100% caramelle al lime.



Il problema è che, dati un insieme di caramelle, supponiamo di aver preso un sacchetto a caso ed estratto una serie di caramelle, qual è la probabilità di estrarre una caramella a ciliegia o lime?

Questo è un esempio di apprendimento probabilistico, ovvero si apprende dai campioni presi dal sacchetto.

### FORMULAZIONE DEL PROBLEMA:

Dato un nuovo sacchetto:

- Una variabile d'ipotesi  $H$  con valori  $h_1, h_2, \dots, h_5$  denota il tipo di sacchetto;
- $D_i$  è una variabile casuale (ciliegia o lime);
- Dopo aver visto  $D_1, D_2, \dots, D_N$  vogliamo predire il sapore (ossia il valore) di  $D_{n+1}$ .

### APPENDIMENTO BAYESIANO COMPLETO:

Considera l'**apprendimento bayesiano** di una distribuzione di probabilità nello **spazio delle ipotesi** ( $P(h_1), P(h_2), \dots$ ). Calcola la probabilità di ogni ipotesi  $h_i$  dai dati forniti e su questa base formula delle predizioni.

Cioè, le previsioni sono fatte usando tutte le ipotesi, pesate dalle loro probabilità, piuttosto che usando solo una singola ipotesi "migliore". In questo modo, l'apprendimento è ridotto a un'inferenza probabilistica.

$H$  variabile d'ipotesi, con valori  $h_1, h_2, \dots, P(H)$  distribuzione a priori.

La **j-esima osservazione  $d_j$**  fornisce il risultato della variabile casuale  $D_j$  (ciliegia o lime) partendo dai dati di training  $d=d_1, \dots, d_n$  precedentemente in possesso.

Partendo dai dati disponibili fino a questo momento, la probabilità condizionata di ogni ipotesi si calcola:

$$P(h_i | d) = \alpha P(d | h_i) P(h_i)$$

(Applicando la regola di Bayes, la probabilità di avere quella sequenza di estrazione data una certa ipotesi, moltiplicata per la probabilità che si presenti quell'ipotesi)

Dove  $P(d | h_i)$  viene chiamato **verosimiglianza** e  $d$  sono valori osservati da  $D$ , la quale ci dice, data una sequenza quanto si accorda con l'ipotesi. Ad esempio, se su 10 caramelle si prendono 8 lime e 2 ciliegie, allora si accorda sugli ultimi sacchetti. In generale, possiamo generalizzare la predizione a qualsiasi  $X$  sconosciuta:

$$\begin{aligned} P(X | d) &= \sum_i P(X | d, h_i) P(h_i | d) \\ &= \sum_i P(X | h_i) P(h_i | d) \\ &= \sum_i P(X | h_i) P(d | h_i) P(h_i) / P(d) \end{aligned}$$

Data una qualsiasi variabile  $X$  e un dataset  $d$ , possiamo dire qual è la probabilità che  $X$  assuma un certo valore, effettuando la sommatoria su tutte le possibili ipotesi (le quali sono variabili nascoste).

Assumendo che  $h_i$  determini una distribuzione di probabilità su  $X$ . Le predizioni sfruttano una probabilità media ponderata sulla verosimiglianza rispetto alle ipotesi.

Esempio:

Una distribuzione per  $P(h_i)$  è  $P(h_1, h_2, h_3, h_4, h_5) = <0.1, 0.2, 0.4, 0.2, 0.1>$ . Se supponiamo l'estrazione di caramelle:



Di che tipo di sacchetto si tratta? Quale sarà il sapore della prossima caramella?

La verosimiglianza dei dati è calcolata partendo dal presupposto che le osservazioni siano **indipendentemente e identicamente distribuite** così che:

$$P(d | h_i) = \prod_j P(d_j | h_i)$$

Supponiamo l'estrazione di caramelle di alcuni sacchetti, lime come sopra  $P(d | h_3) = 0.5^{10}$  perché  $h_3$  la metà delle caramelle sono lime. Mentre se si calcola  $P(d | h_4) = 0.75^{10}$ .

## PROBABILITÀ CONDIZIONATA DELLE IPOTESI:

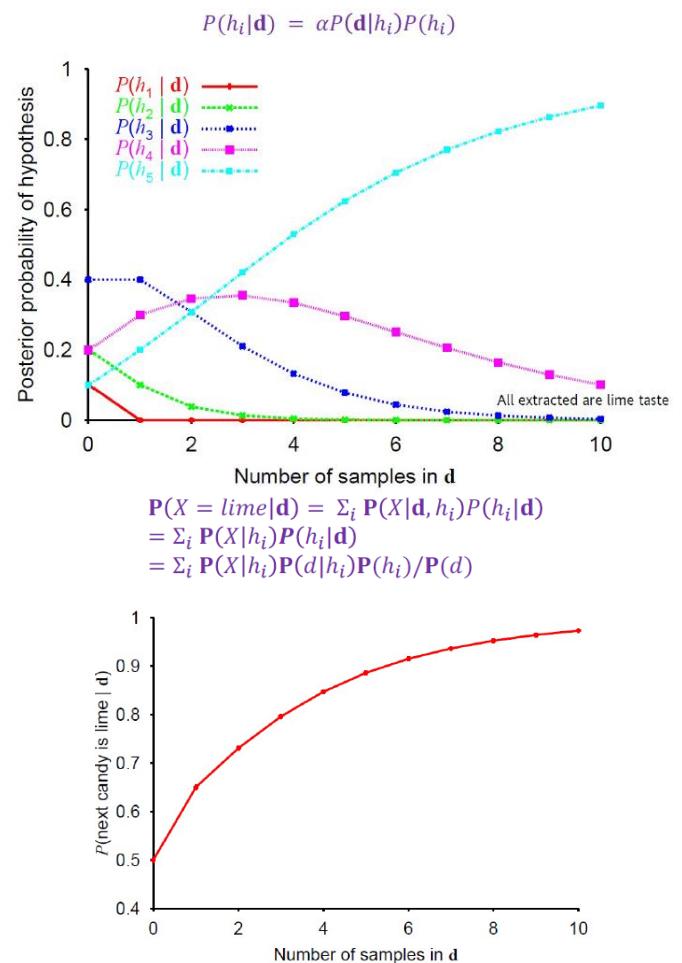
Dal grafico vediamo come cambiano le probabilità per le 5 ipotesi, considerando il vettore di 10 caramelle lime. Le probabilità cambiano all'aumentare delle caramelle estratte.

Se voglio conoscere la probabilità che la prossima caramella sia lime dato il vettore  $d$ , la curva di probabilità cresce all'aumentare dei campioni.

Le ipotesi vere alla fine hanno dominato la predizione bayesiana (caratteristica di questo tipo di apprendimento).

La predizione bayesiana è **ottimale**; tuttavia, il suo spazio delle ipotesi spesso è molto grande o addirittura infinito.

Il problema è dato dal numero di ipotesi, che possono essere molte e quindi abbiamo moltissimi valori, richiederanno molto tempo di computazione. Ciò che si fa è un compromesso, ovvero perdere precisione ma ottenere un calcolo più rapido, utilizzando l'approssimazione MAP.



## APPROXIMAZIONE MAP:

Invece di analizzare tutte le ipotesi qui andiamo a considerare solo l'unica ipotesi che massimizza il prodotto successivo.

Apprendimento **Maximum a posteriori (MAP)**:

$$h_{\text{map}} \text{ è } h_i \text{ che massimizza } P(h_i | d) \cong P(d | h_i)P(h_i)$$

Le predizioni con  $h_{\text{map}}$  sono approssimativamente bayesiane  $P(X | d) \cong P(X | h_{\text{map}})$ , trovare le ipotesi MAP è molto più semplice dell'apprendimento bayesiano.

Esempio:

Nell'esempio  $h_{\text{map}}=h_5$  dopo aver mangiato 3 caramelle a lime.

Quindi un agente MAP predirà che la quarta caramella sia lime con probabilità 1 (0,8 è invece la predizione bayesiana) all'aumentare dei dati si avvicina a quella bayesiana.

Entrambe le tecniche fanno uso della distribuzione a priori  $P(h_i)$  per ridurre la complessità mentre per le ipotesi deterministiche  $P(d | h_i)$  vale 1 se consistente, 0 altrimenti  $\rightarrow h_{\text{map}} = \text{l'ipotesi più semplice consistente con i dati}$ .

Apprendimento **Maximum a posteriori (MAP)**:

$$h_{\text{map}} \text{ è } h_i \text{ che massimizza } P(h_i | d) \cong P(d | h_i)P(h_i)$$

Equivale a minimizzare  $-\log_2 P(d | h_i) - \log_2 P(h_i)$  dove:

- $\log_2 P(h_i)$  equivale al numero di bit necessari a specificare l'ipotesi  $h_i$ ;
- $\log_2 P(d | h_i)$  numero di bit aggiuntivi richiesti per la specifica dei dati fissata l'ipotesi  $h_i$ .

L'apprendimento MAP sceglie  $h_i$  che più comprime i dati, detta anche **minimum description length (MDL)**; se consideriamo l'esempio di prima  $\log_2 P(h_5) = \log_2 1 = 0$  non serve alcun bit.

## APPROXIMAZIONE ML (MAXIMUM LIKELIHOOD):

Analizzando la formula del MAP, nei problemi reali molto spesso, tutte le ipotesi sono equiprobabili, quindi viene rimosso l'elemento  $P(h_i)$ . Questo è ragionevole quando non c'è motivo di preferire un'ipotesi rispetto ad un'altra.

Per dataset di grandi dimensioni, la distribuzione a priori  $P(h_i)$  diventa irrilevante. L'apprendimento con **massima verosimiglianza** (Maximum Likelihood) rappresenta una buona approssimazione dell'apprendimento bayesiano e di MAP:

- Si sceglie  $h_{ML}$  massimizzando  $P(d | h_i)$ , ottenendo in maniera semplice il migliore adattamento ai dati, **ipotesi di massima verosimiglianza**.

È identico al MAP per la distribuzione a priori, laddove però essa risulti **uniforme** (che è ragionevole se tutte le ipotesi sono della stessa complessità). ML rappresenta il metodo "standard" non bayesiano per l'apprendimento statistico.

## APPRENDIMENTO PARAMETRI ML NELLE RETI BAYESIANE:

Nelle reti bayesiane, abbiamo dei nodi con delle tabelle di probabilità condizionate dai valori dei genitori. Possiamo quindi effettuare un apprendimento dei valori di probabilità della rete, analizzando dei dati di training.

### Esempio:

Abbiamo un sacchetto da un nuovo produttore; frazione  $\theta$  di caramelle alla ciliegia? Qualsiasi  $\theta$  è possibile: continuum d'ipotesi  $h_\theta$  dove  $\theta$  è un **parametro** per questa famiglia di modelli semplici. È ragionevole adottare l'approccio ML (dato che i due gusti sono ugualmente probabili).

Supponiamo di scartare  $N$  caramelle,  $c$  alla ciliegia e  $\ell=N-c$  al lime.

Queste sono osservazioni indipendenti e identicamente distribuite, pertanto:

$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^N P(d_j|h_\theta) = \theta^c \cdot (1-\theta)^\ell$$

Massimizzandolo con riferimento a  $\theta$ , che risulta più facile per la verosimiglianza logaritmica:

$$\begin{aligned} L(\mathbf{d}|h_\theta) &= \log P(\mathbf{d}|h_\theta) = \sum_{j=1}^N \log P(d_j|h_\theta) = c \log \theta + \ell \log(1-\theta) \\ \frac{dL(\mathbf{d}|h_\theta)}{d\theta} &= \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell} = \frac{c}{N} \end{aligned}$$

Problema: alcuni eventi potrebbero avere valore 0 qualora non fossero stati osservati.

Per massimizzare, andiamo a fare la derivata rispetto a theta, eguagliando a 0, così da ottenere il valore di theta.

### PARAMETRI MULTIPLI:

L'incartamento rosso/verde dipende probabilisticamente dal sapore.

Probabilità (ad es.) di avere caramelle alla ciliegia nella carta verde:

$$\begin{aligned} P(F = \text{cherry}, W = \text{green} | h_{\theta, \theta_1, \theta_2}) &= P(F = \text{cherry} | h_{\theta, \theta_1, \theta_2}) P(W = \text{green} | F = \text{cherry}, h_{\theta, \theta_1, \theta_2}) \\ &= \theta \cdot (1 - \theta_1) \end{aligned}$$

$N$  caramelle,  $r_c$  caramelle alla ciliegia in carta rossa, ecc...

$$P(\mathbf{d}|h_{\theta, \theta_1, \theta_2}) = \prod_{j=1}^N P(d_j|h_{\theta, \theta_1, \theta_2})$$

$$P(\mathbf{d}|h_{\theta, \theta_1, \theta_2}) = \theta^c (1-\theta)^\ell \cdot \theta_1^{r_c} (1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell} (1-\theta_2)^{g_\ell}$$

$$\begin{aligned} L &= [c \log \theta + \ell \log(1-\theta)] \\ &\quad + [r_c \log \theta_1 + g_c \log(1-\theta_1)] \\ &\quad + [r_\ell \log \theta_2 + g_\ell \log(1-\theta_2)] \end{aligned}$$

Le derivate di  $L$  contengono solo i parametri rilevanti:

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \quad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c + g_c}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \quad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell + g_\ell}$$

Con dati completi, i parametri possono essere appresi separatamente.

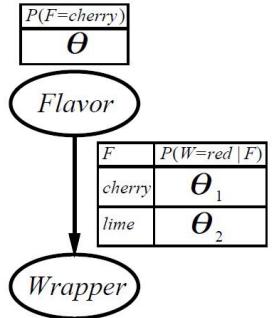
### MODELLI BAYESIANE NAIVE:

Rappresentano il modello di rete bayesiana comunemente usato nel machine learning, un solo nodo radice e tanti figli.

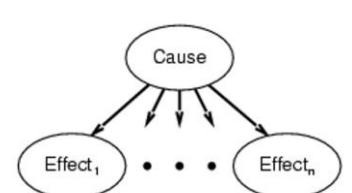
Viene fatta un'assunzione, dove gli effetti sono tutti indipendenti l'uno dall'altro (condizionalmente indipendenti).

La variabile  $C$  da predire è la radice, le variabili attributo  $x_i$  sono le foglie. Questo modello è ingenuo perché assume che gli attributi sono condizionalmente indipendenti.

Una predizione deterministica può essere ottenuta scegliendo le classi più probabili.

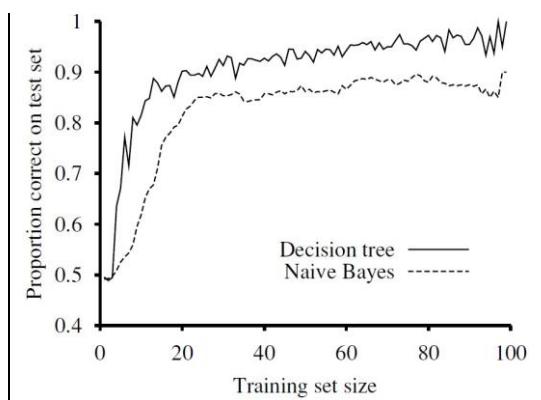


$$P(C|x_1, x_2, \dots, x_n) = \alpha P(C) \prod_i P(x_i|C)$$



Il loro comportamento è leggermente peggiore degli alberi di decisione.

Non ha nessuna difficoltà con i dati rumorosi, per  $n$  attributi booleani, ci sono  $2n+1$  parametri, non è richiesta nessuna ricerca per trovare  $h_{ML}$ .



### RIASSUNTO:

L'apprendimento bayesiano formula un apprendimento come una forma d'inferenza probabilistica, usando le osservazioni per aggiornare una distribuzione a priori attraverso le ipotesi.

L'apprendimento MAP seleziona una singola ipotesi più probabile, sfruttando i dati di training  $P(\mathbf{d}|h_i), P(h_i)$ .

Il metodo della massima verosimiglianza (ML) seleziona le ipotesi che massimizzano la verosimiglianza dei dati (uguale a MAP ma con una distribuzione a priori uniforme)  $P(\mathbf{d}|h_i)$ .

## 14. RETI NEURALI

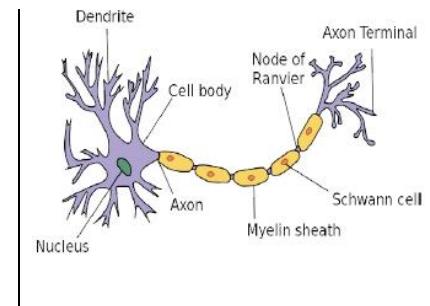
L'idea delle reti neurali è nata come modello che simulava il funzionamento dei neuroni nel cervello.

I circuiti collegati sono stati utilizzati per simulare il suo comportamento intelligente.

Il cervello è composto da neuroni:

- un corpo cellulare
- dendriti (ingressi)
- un assone (uscite)
- sinapsi (elaborano l'input e creano l'output), possono essere stimolate o inibite e possono cambiare nel tempo.

Quando la somma degli ingressi raggiunge una certa soglia, verrà inviato un impulso elettrico sull'assone.



### McCULLOCH-PITTS UNIT (RETI NEURALI):

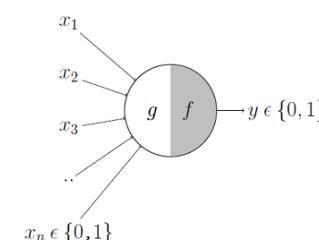
Nel 1943 McCulloch e Pitts proposero il primo modello di neurone ed era molto elementare.

Un'**unità McCulloch Pitts** prende come input valori booleani, li dà ad una funzione che calcola la somma su tutti gli input booleani e poi, il risultato della sommatoria, viene dato ad una funzione di attivazione  $g$  che ritorna 1 o 0 in base ad una soglia, quando il valore della sommatoria supera la soglia ritorna 1 altrimenti 0.

Una **rete neurale artificiale** è un **grafo diretto** di unità e collegamenti. Un link dall'unità  $i$  all'unità  $j$  propaga l'attivazione  $a_i$  dall'unità  $i$  all'unità  $j$ , ed ha un peso  $w_{i,j}$  ad essa associato.

$$g(x_1, x_2, x_3, \dots, x_n) = g(\mathbf{x}) = \sum_{i=1}^n x_i$$

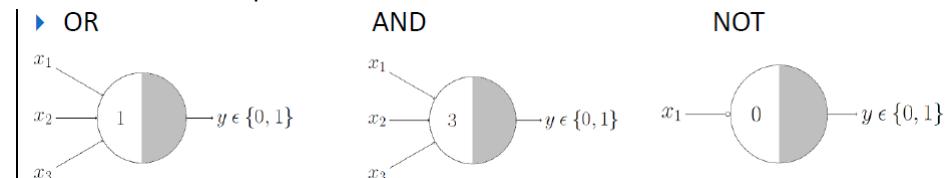
$$\begin{aligned} y = f(g(\mathbf{x})) &= 1 \quad \text{if} \quad g(\mathbf{x}) \geq \theta \\ &= 0 \quad \text{if} \quad g(\mathbf{x}) < \theta \end{aligned}$$



### IMPLEMENTAZIONE DI FUNZIONI LOGICHE COME UNITÀ:

Le unità di McCulloch Pitts sono una grossolana semplificazione eccessiva dei neuroni reali, ma il suo scopo è sviluppare la comprensione di ciò che possono fare le reti neurali di unità semplici.

Ogni **funzione booleana** può essere implementata come **reti McCulloch Pitts**:



### PERCETTRONE:

Il problema con i neuroni di McCulloch e Pitts sono:

- Modello troppo semplice, non si può modellare qualcosa di più complesso, siccome la rete non accetta dati continui (non booleani);
- Questo tipo di rete prevede dei pesi e delle soglie che sono definite assieme alla rete (non è possibile apprenderli e cambiarli).

L'evoluzione di questo modello è il **percettrone** che risolve diversi problemi precedenti:

- I valori sono continui, bipolare e multivalore;
- I valori dei pesi e delle soglie possono essere determinati analiticamente o mediante un algoritmo di apprendimento. La formula di aggiornamento dei pesi è la seguente:

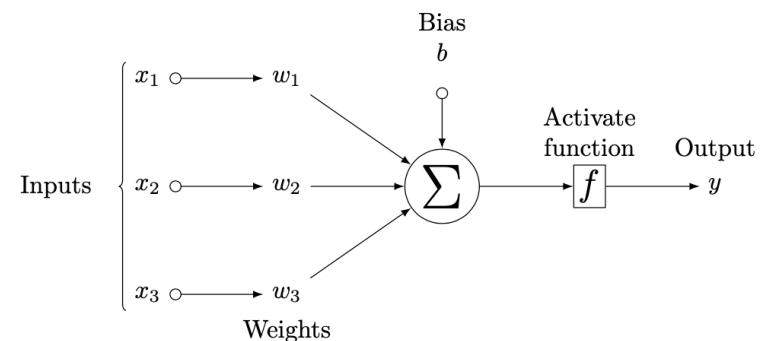
$$w_j^{(t)} \leftarrow w_j^{(t)} + \alpha x_j(t - y)$$

dove  $t$  l'output corretto e  $y$  la funzione di output della rete mentre  $\alpha$  è il fattore di apprendimento

Il Percettrone ha la seguente architettura:

L'input è  $x_1, \dots, x_n$  (non booleani) con i pesi associati  $w_1, \dots, w_n$ .

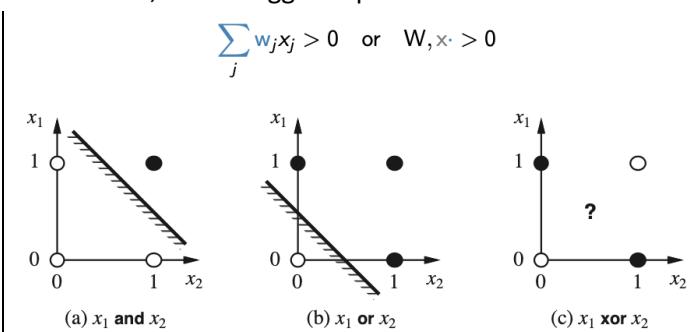
Una volta presi input e pesi, effettua una sommatoria pesata (moltiplicando ogni input per il peso e poi va a sommare), dopodiché il risultato viene dato ad una funzione di attivazione che produrrà l'output.



## ESPRESSIVITÀ DEI PERCETTRONI:

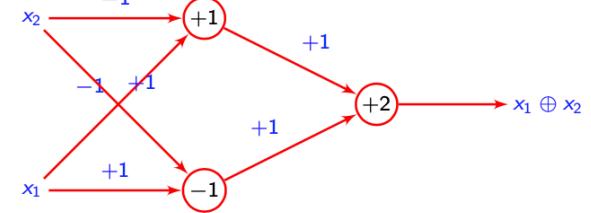
Uno dei problemi del **percettrone** è l'espressività, dove può rappresentare AND, OR, NOT, maggioranza, ecc. (separatori lineari), ma non lo XOR (e quindi nessun sommatore). Per risolvere tutto ciò, si sono aggiunti più strati.

Rappresenta un **separatore lineare** nello spazio di input:



Il seguente Perceptron multistrato può risolvere il problema dello XOR:

Lo strato intermedio è uno strato nascosto.



## RETI FEED-FORWARD (STRUTTURE DI RETI):

Quindi per affrontare il problema dell'espressività si vanno a creare le reti feed-forward, fondamentalmente si possono collegare tanti percetroni organizzandoli in layer. Si parte con dei layer di input che ricevono i dati  $x_n$  del problema e si aggiungono tanti strati dove all'interno ci sono sempre dei percetroni.

Si chiama **feed-forward** se la rete è aciclica, perché l'output di un percettrone viene dato in input al prossimo percettrone, in questo modo i dati non vengono riusati per altri calcoli, pertanto, si dice che sono reti "senza memoria".

Più formalmente, le reti feed forward sono organizzate in livelli (layers): una rete a  $n$ -livelli ha una partizione  $\{L_0, \dots, L_n\}$  dei nodi, in modo tale che gli archi colleghino solo i nodi al livello successivo.

$L_0$  è chiamato **livello di input** ed i suoi elementi **unità di input**, e  $L_n$  **livello di output** ed i suoi elementi **unità di output**.

Qualsiasi unità che non si trova nel livello di input o nel livello di output viene chiamata **hidden**.

## RETI RICORRENTI (STRUTTURE DI RETI):

Mentre, una rete neurale si chiama **ricorrente** se ha dei cicli, come se la rete abbia una memoria siccome alcuni percetroni possono usare i risultati di altri percetroni prendendo una decisione anche in base al risultato delle operazioni precedenti.

## PERCETTRONI SINGLE-LAYER:

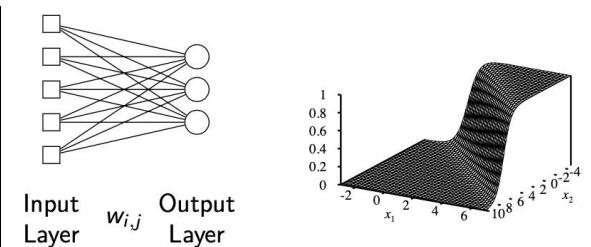
Il modello più semplice di una rete neurale è la **rete perceptron**, che è una rete feed forward con un solo layer che produce direttamente un valore di output. Una rete perceptron a strato singolo è chiamata semplicemente **perceptron**.

**NOTA:** il layer di output è un problema di classificazione non binaria multiclasse, siccome ogni nodo di output rappresenta una classe. Quindi un solo nodo è una classificazione binaria, con più nodi ci sono più classi.

Tutte le unità di ingresso sono collegate direttamente all'unità di uscita.

Le unità di output funzionano tutte separatamente, nessun peso condiviso, e sono trattate come la combinazione di  $n$  unità percetroni.

La **regolazione dei pesi** è la componente che fa cambiare il funzionamento della rete, esempio, sposta la posizione, l'orientamento e la pendenza della scogliera.



### Esempio reti neurali Feed-Forward:

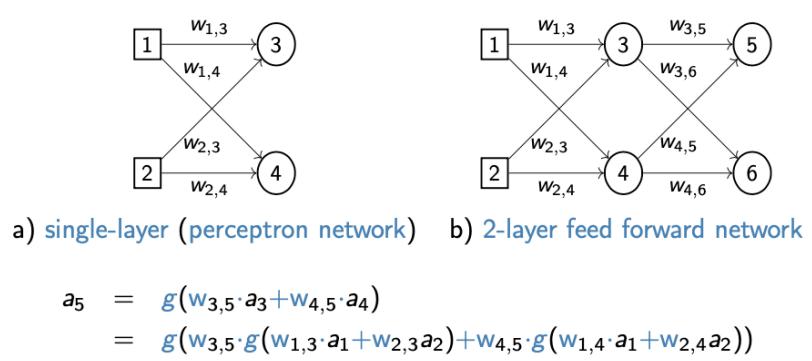
Feed forward network = una famiglia parametrizzata di funzioni non lineari:

Mostriamo due reti feed forward:

- Ad un livello;
- A due livelli.

Come mostrato, per determinare il valore del nodo 3 si calcola la funzione di attivazione con gli output dei nodi 1 e 2 con i rispettivi pesi  $w_{1,3}$  e  $w_{2,3}$ .

**NOTA:** la regolazione dei pesi cambia la funzione



## FUNZIONE DI ATTIVAZIONE:

Una **funzione di attivazione** viene aggiunta ad una rete neurale per aiutarla ad apprendere schemi complessi nei dati, essa converte l'output di un neurone in un'altra forma utilizzata come input per il neurone successivo.

Le funzioni di attivazione sono necessarie perché devono capire l'andamento dei dati e dovrebbero essere funzioni non lineari, perché con funzioni lineari tutta la rete neurale si semplifica in una funzione lineare e molti problemi non sono modellabili in questo modo, e devono mantenere l'output di un neurone a un certo intervallo secondo il nostro requisito.

**Caratteristiche desiderabili** di una funzione di attivazione sono le seguenti proprietà:

- **Gradiente non nullo:** Formiamo reti neurali utilizzando algoritmi basati su gradiente. Quindi, il gradiente deve essere non zero in tutti i punti di dominio. Se il gradiente tende a zero perderemo informazioni e l'apprendimento non funzionerà come deve (problema del *vanishing gradient*);
- **Centrato sullo zero:** L'uscita di una funzione di attivazione deve essere simmetrica a zero. Questo impedisce ai gradienti di spostarsi in una direzione particolare;
- **Costi computazionali:** Le funzioni di attivazione vengono applicate più volte. Dovrebbero essere computazionalmente poco costose per calcolarle insieme alla sua derivata;
- **Differenziabile:** Nell'apprendimento basato sul gradiente, dobbiamo calcolare il gradiente di attivazione funzioni. Pertanto, le funzioni di attivazione devono essere differenziabili.

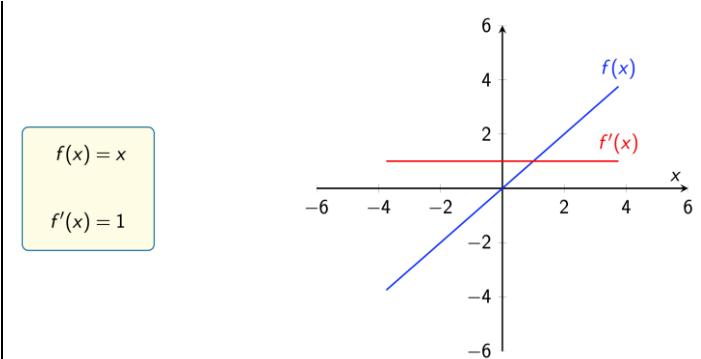
## FUNZIONE DI ATTIVAZIONE LINEARE:

Vantaggi:

- Il suo calcolo è semplice.
- La sua derivata è diversa da zero.

Svantaggi:

- L'output non è in un intervallo.
- Nessun risultato nella non linearità.



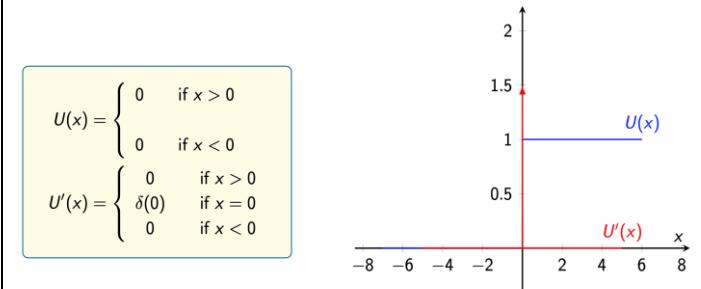
## FUNZIONE DI ATTIVAZIONE SOGLIA:

Vantaggi:

- Il suo calcolo è semplice.

Svantaggi:

- L'output non è in un intervallo.
- La sua derivata è zero tranne che all'origine.



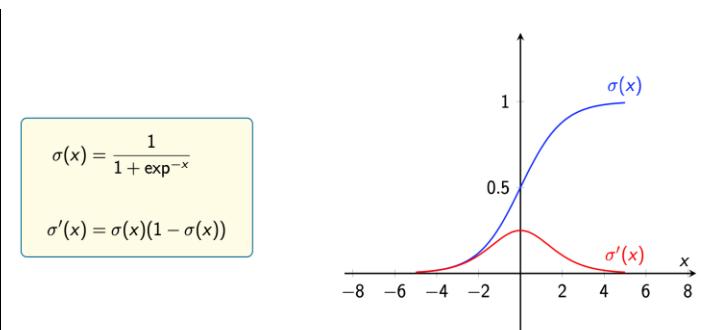
## FUNZIONE DI ATTIVAZIONE SIGMOIDE:

Vantaggi:

- L'output è nell'intervallo  $[0, 1]$ .
- È differenziabile.

Svantaggi:

- Satura e uccide i gradienti (lo porta a 0 siccome applica continuamente la derivata).
- Il suo output non è centrato sullo zero (ma su 0,5).



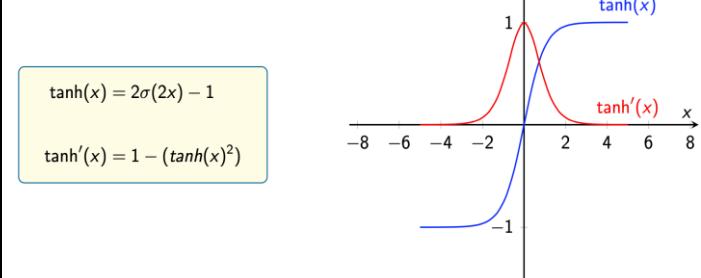
## FUNZIONE DI ATTIVAZIONE TANH:

Vantaggi:

- L'output è nell'intervallo  $[0, 1]$ .
- È differenziabile.
- Il suo output è centrato sullo zero.

Svantaggi:

- Satura e uccide i gradienti.



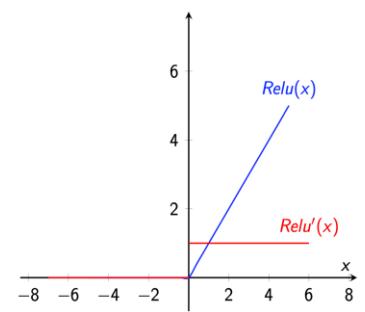
## FUNZIONE DI ATTIVAZIONE RELU:

Vantaggi:

- Il suo calcolo è semplice.
  - La sua derivata è diversa da zero quando  $x > 0$  (identità).
  - Non è computazionalmente costosa.
- Svantaggi:
- L'output non è in un intervallo.
  - Il gradiente svanisce quando  $x < 0$ .

$$\text{Relu}(x) = \max(0, x)$$

$$\text{Relu}'(x) = \begin{cases} 1 & \text{if } x > 0 \\ ? & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$



## FUNZIONE DI ATTIVAZIONE SOFTMAX:

Softmax è una forma più generalizzata del sigmoide, viene utilizzato nel livello di output delle reti neurali per problemi di classificazione multiclasse.

Esempio:

Sia  $x = [1.60, 0.55, 0.98]^\top$ .

Applicando softmax si ottiene  $a_i = [0.51, 0.18, 0.31]^\top$ .

$$a_i(x) = \frac{\exp^{z_i}}{\sum_k \exp^{z_k}}$$

## FUNZIONE DI LOSS:

L'obiettivo degli algoritmi di machine learning è costruire un modello (ipotesi) che possa essere utilizzato per stimare  $t$  in base a  $x$ . Sia il modello in forma di (che va a modellare il funzionamento del neurone):

$$h(x) = w_0 + w_1 x$$

L'obiettivo della creazione di un modello è scegliere parametri in modo che  $h(x)$  sia vicino a  $t$  per i dati di training  $x$ .

Abbiamo sempre bisogno di minimizzare la funzione di loss, ovvero la distanza tra la funzione reale e ciò che ha stimato sia il più piccolo valore possibile. Una funzione che è spesso usato è l'**errore quadratico medio**:

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (h(x_i) - t_i)^2$$

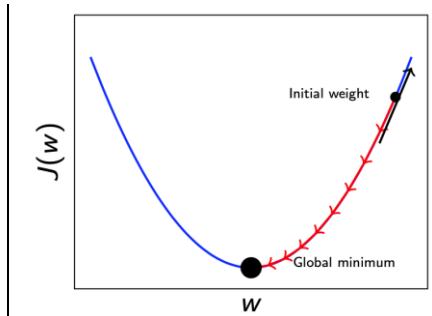
## DISCESA DEL GRADIENTE:

Per minimizzare la loss function, come nella regressione, determiniamo il gradiente e modifichiamo i valori dei pesi affinché la loss diminuisca; quindi, andiamo nella direzione opposta del gradiente.

La discesa del gradiente è di gran lunga la strategia di ottimizzazione più popolare, utilizzata al momento nell'apprendimento automatico e nel deep learning.

Il costo (errore) è una funzione dei pesi (parametri).

Vogliamo ridurre al minimo l'errore (punto *initial weight*), il gradiente restituisce la direzione per salire ma andiamo nella direzione opposta, scendendo passo per passo fino a che non raggiungiamo il minimo  $w$ .



Abbiamo la seguente ipotesi e dobbiamo adattare i dati di training:

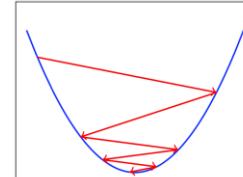
$$h(x) = w_0 + w_1 x$$

Usiamo una funzione di loss come **Errore quadratico medio**:

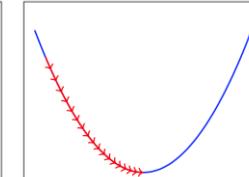
$$J(w) = \frac{1}{2m} \sum_{i=1}^m (h(x_i) - t_i)^2$$

Questa funzione di loss può essere ridotta al minimo utilizzando la discesa del gradiente con il **tasso di apprendimento**  $\alpha$ , con un valore di tasso alto passiamo da una curva all'altra mentre con un piccolo tasso si ha un lento aggiornamento dei pesi fino ad un minimo.

Big step size



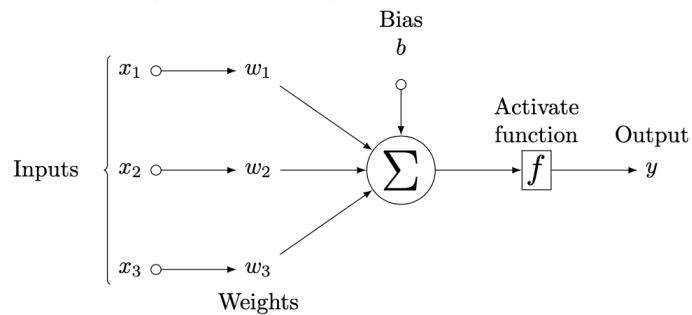
Small step size



$$\begin{aligned} w_0^{(t+1)} &= w_0^{(t)} - \alpha \frac{\partial J(w^{(t)})}{\partial w_0} \\ w_1^{(t+1)} &= w_1^{(t)} - \alpha \frac{\partial J(w^{(t)})}{\partial w_1}, \end{aligned}$$

## ALLENARE UN NEURONE CON ATTIVAZIONE SIGMOIDEA (REGRESSIONE):

Consideriamo il seguente singolo neurone:



Vogliamo addestrare questo neurone per ridurre al minimo la seguente funzione di costo:

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (h(x^i) - t^i)^2$$

Considerando la *funzione di attivazione sigmoidea*:

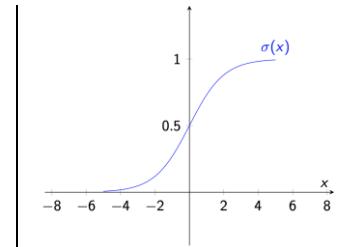
$$f(z) = \frac{1}{1+e^{-z}}$$

Vogliamo calcolare:

$$\frac{\partial J(w)}{\partial w_i}$$

Usando la **chain rule**, otteniamo ( $\alpha$  è il tasso di apprendimento):

$$\begin{aligned}\frac{\partial J(w)}{\partial w_j} &= \frac{\partial J(w)}{\partial f(z)} \times \frac{\partial f(z)}{\partial z} \times \frac{\partial z}{\partial w_j} \\ \frac{\partial J(w)}{\partial f(z^i)} &= \frac{1}{m} \sum_{i=1}^m (f(z^i) - t^i) \\ \frac{\partial f(z)}{\partial z} &= \frac{e^{-z}}{(1 + e^{-z})^2} = f(z)(1 - f(z)) \\ \frac{\partial z}{\partial w_j} &= x^j \\ w_j^{(t+1)} &= w_j^{(t)} - \alpha \frac{\partial J(w)}{\partial w_j}\end{aligned}$$



Vogliamo addestrare questo neurone per ridurre al minimo la seguente funzione di costo:

$$J(w) = \sum_{i=1}^m [-t^i \ln h(x^i) - (1 - t^i) \ln(1 - h(x^i))]$$

Calcolando i gradienti di  $J(w)$  rispetto a  $w$ , otteniamo:

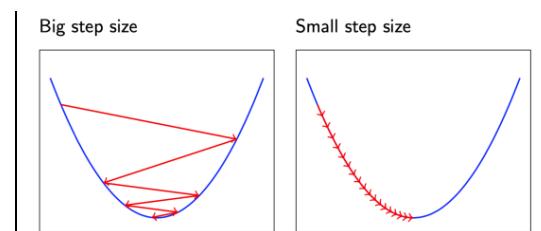
$$\nabla J(w) = \sum_{i=1}^m t^i x^i (h(x^i) - t^i)$$

L'aggiornamento del vettore di peso utilizzando la regola di discesa del gradiente risulterà:

$$w^{(t+1)} = w^{(t)} - \alpha \sum_{i=1}^m t^i x^i (h(x^i) - t^i)$$

### TUNING LEARNING RATE ( $\alpha$ ):

- Se  $\alpha$  è troppo alto, l'algoritmo diverge e il punto di minimo potrebbe saltarlo.
- Se  $\alpha$  è troppo basso, rallenta la convergenza dell'algoritmo e l'algoritmo richiede tanti passi per arrivare al minimo.



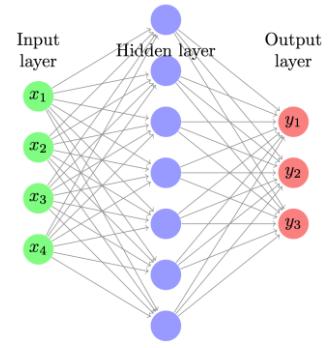
Una pratica comune consiste nel rendere  $\alpha_k$  una funzione decrescente del numero di iterazione  $k$ :

$$\alpha_k = \frac{c_1}{k + c_2} \quad \text{dove } c_1 \text{ e } c_2 \text{ sono due costanti.}$$

Le prime iterazioni causano grandi cambiamenti nella  $w$ , mentre le successive effettuano solo il fine-tuning. Più le iterazioni aumentano più il passo diminuisce siccome si suppone che ci si trovi vicino al minimo.

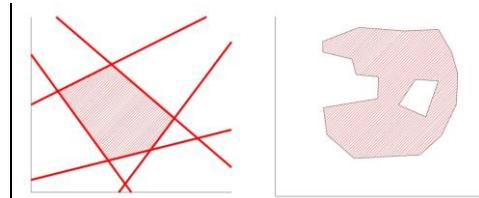
## DEEP FEED-FORWARD NETWORKS:

Soltanamente, le reti neurali sono formate da più layer, pertanto, i dati di input attraverseranno tutti questi perceptri.



## LA TOPOLOGIA OTTIMALE DELLE RETI:

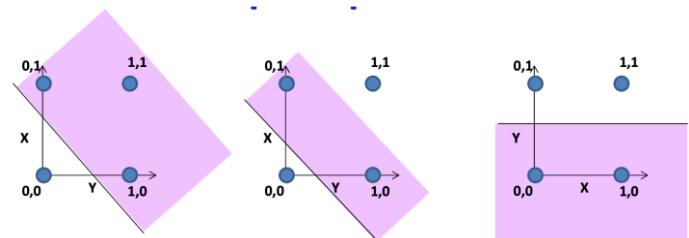
Il funzionamento delle reti neurali deep non è noto siccome ci sono tanti parametri che si aggiornano nella fase di training.  
Una rete di tale livello non fa altro che creare delle approssimazioni di funzioni.



## SUPERFICIE DECISIONALE DEL PERCEPTRON:

Ad esempio, definendo una rete a tre livelli, si riesce a definire un classificatore che assegna a tutti i punti dentro all'area rettangolare una determinata label e quelli al di fuori un'altra label.

Se bisogna definire una forma più complessa si necessita di più layer, andando a costruire una topologia di rete diversa.



## SCELTA DELLA TOPOLOGIA DI RETE:

Specificare la **topologia della rete** in base a:

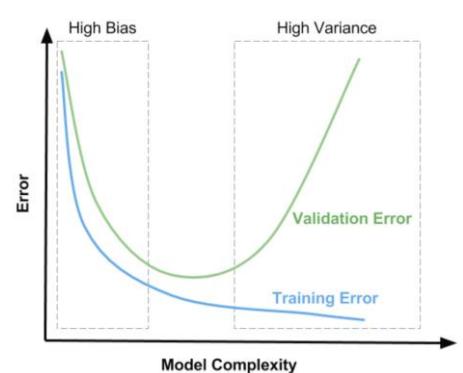
- #-strati
- #-nodi in ogni livello
- funzione di ogni nodo
- attivazione di ogni nodo

Ricordando che le reti neurali sono approssimatori universali e che bisogna specificare la funzione di loss, usando l'algoritmo discesa del gradiente per addestrare la rete.

Bisogna poi scegliere le funzioni di attivazione per i vari layer, ad esempio per l'Output layer (Linear, ReLU, Sigmoid, Softmax) e l'Hidden layers (Linear, ReLU, Sigmoid).

Un [approccio semplice](#) per la scelta della topologia di rete è per tentativi ed errori (**trial and error**). Suddividiamo i dati disponibili in tre parti: dati di training, dati di validation e dati di testing. Sceglio una topologia e addestriamo la rete utilizzando i dati di addestramento. Dopo l'addestramento, valutiamo la rete addestrata utilizzando i dati di convalida.

La parte più importante è la complessità del modello, siccome se il modello è troppo semplice va in underfitting (Bias) mentre troppo complesso va in overfitting (Varianza). Pertanto, bisogna trovare il giusto compromesso.



## ALGORITMO BACKPROPAGATION:

Per l'apprendimento su una rete con un solo layer si usa l'algoritmo di **backpropagation**, è sempre l'algoritmo basato sul gradiente solo che si contestualizza ad una rete con più layer.

Il training lo si può fare in diversi modi, esempio se abbiamo un training set di 1000 campioni possiamo usare il **batch size**, ovvero quanti dati vogliamo rendere disponibili per fare training ad ogni iterazione.

L'algoritmo prende un sottoinsieme di campioni, li da in input alla rete e calcola la loss, dopodiché a partire dalla loss va ad aggiornare il valore dei pesi e lo fa in maniera iterativa. Ripete finché la differenza dei pesi si azzera o perché il **numero di iterazioni** specificato viene raggiunto.

Le **epoch** sono quante volte viene considerato l'insieme di training, siccome è possibile usarlo più volte.

## Training a neural network

**Data:** A training set  $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$

**Result:** Weight matrices of the neural network

Initialize randomly weights in the network;

**while** not at trained **do**

Create a batch  $S_B \subseteq S$ ;

Let  $K \leftarrow |S_B|$ ;

**for**  $i \leftarrow 1$  **to**  $K$  **do**

    Give  $x_i$  to the network and  $\hat{y}_i$ ;

**end**

    Compute  $J(w)$ ;

    Compute  $\nabla_w J(w)$ ;

    Compute  $w^{t+1} \leftarrow w^t - \alpha \nabla_w J(w)$ ;

**end**

## BATCH SIZE:

La discesa del gradiente può essere utilizzata con batch di diverse dimensioni, esempio:

- $K = 1$  (Discesa stocastica del gradiente);
- $K \ll m$  (Discesa gradiente mini-batch);
- $K = m$  (Discesa gradiente batch).

### Esempio:

Sia la dimensione del training set  $m = 1000$ , nella discesa stocastica del gradiente, utilizziamo 1000 batch di dimensione 1.

Sia  $K = 50$ , in discesa gradiente mini-batch, utilizziamo 20 lotti di dimensione 50, quindi dopo 20 iterazioni si fa un'epoca.

Nella discesa del gradiente batch, utilizziamo un batch di dimensione 1000.

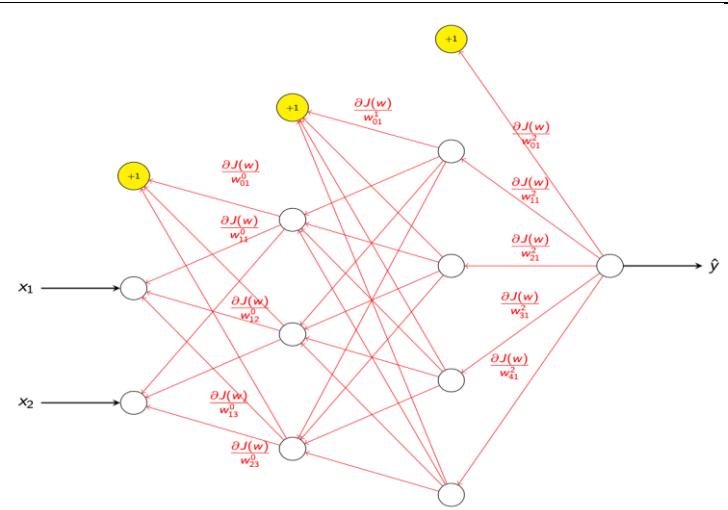
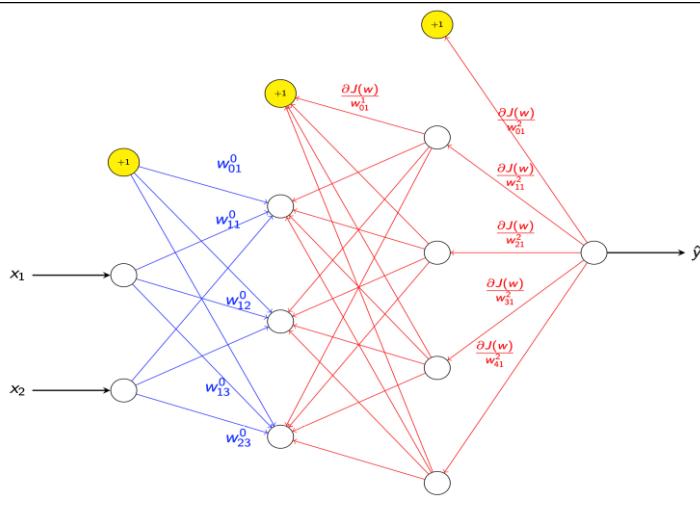
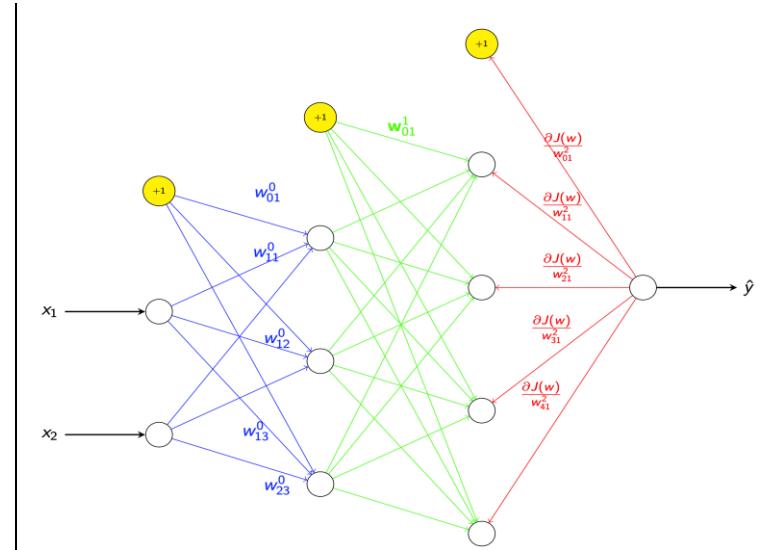
## BACKWARD PASS:

La parte importante dell'algoritmo è come aggiornare i pesi nella rete. I pesi in una rete multilayer si trovano a più livelli da 0 a  $n$ , ma la loss si trova all'ultimo layer. Quindi bisogna aggiornare i pesi facendo la derivata della loss rispetto ai pesi del layer in cui ci si trova, dopodiché si aggiornano i pesi, si passa al layer successivo e così via andando avanti fino all'origine.

Dopo aver calcolato la funzione di loss, dobbiamo calcolare  $\nabla_w J(w)$ .

Quindi per ogni peso  $w$ , utilizziamo la seguente regola per aggiornare quel peso.

$$w^{t+1} = w^t - \alpha \frac{\partial J(w)}{\partial w}$$

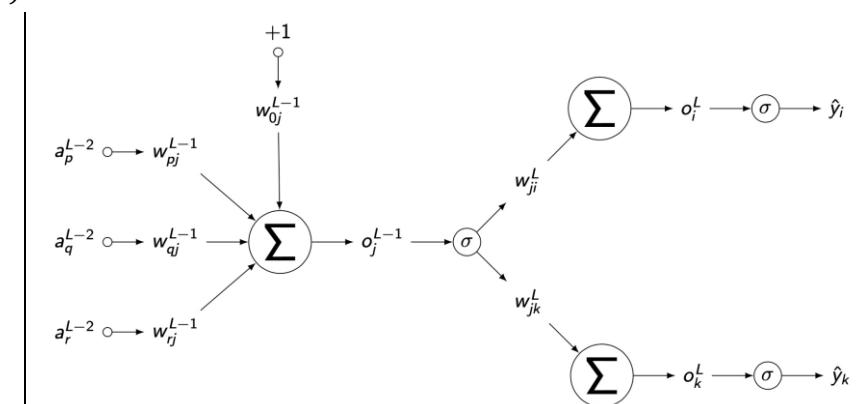


Usiamo la seguente notazione per il calcolo  $\nabla_w J(w)$ :

Supponiamo di avere una rete con  $L$  livelli, in cui l'ultimo livello ha nodi di output  $C$ .

Supponiamo che tutti i nodi utilizzano funzioni di attivazione del sigmoide.

Indichiamo l'output del  $j$ -esimo nodo nel layer  $L$ -esimo strato con  $a_j^L$ .



Dobbiamo calcolare  $\frac{\partial J(w)}{\partial w}$

$$\frac{\partial J(w)}{\partial w_{pj}^{L-1}} = \frac{\partial \sum_{s=1}^K \sum_{c=1}^C \ell(\hat{y}_{cs}, y_{cs})}{\partial w_{pj}^{L-1}} = \frac{\partial \sum_{c=1}^C \sum_{s=1}^K \ell(\hat{y}_{cs}, y_{cs})}{\partial w_{pj}^{L-1}} = \sum_{c=1}^C \frac{\partial \sum_{s=1}^K \ell(\hat{y}_{cs}, y_{cs})}{\partial w_{pj}^{L-1}}$$

Come calcolare  $\sum_{c=1}^C \frac{\partial \sum_{s=1}^K \ell(\hat{y}_{cs}, y_{cs})}{\partial w_{pj}^{L-1}}$ :

$$\sum_{c=1}^C \frac{\partial \sum_{s=1}^K \ell(\hat{y}_{cs}, y_{cs})}{\partial w_{pj}^{L-1}} = \sum_{c=1}^C \sum_{s=1}^K \frac{\partial \ell(\hat{y}_{cs}, y_{cs})}{\partial \hat{y}_{cs}} \frac{\partial \hat{y}_{cs}}{\partial o_c^L} \frac{\partial o_c^L}{\partial a_j^{L-1}} \frac{\partial a_j^{L-1}}{\partial o_j^{L-1}} \frac{\partial o_j^{L-1}}{\partial w_{pj}^{L-1}}$$

Abbiamo:

$$\sum_{c=1}^C \frac{\partial \sum_{s=1}^K \ell(\hat{y}_{cs}, y_{cs})}{\partial w_{pj}^{L-1}} = \sum_{c=1}^C \sum_{s=1}^K \frac{\partial \ell(\hat{y}_{cs}, y_{cs})}{\partial \hat{y}_{cs}} \frac{\partial \hat{y}_{cs}}{\partial o_c^L} \frac{\partial o_c^L}{\partial a_j^{L-1}} \frac{\partial a_j^{L-1}}{\partial o_j^{L-1}} \frac{\partial o_j^{L-1}}{\partial w_{pj}^{L-1}}$$

Otteniamo:

$$\begin{aligned} \frac{\partial \ell(\hat{y}_{cs}, y_{cs})}{\partial \hat{y}_{cs}} &=? \\ \frac{\partial \hat{y}_{cs}}{\partial o_c^L} &= \sigma(o_c^L) (1 - \sigma(o_c^L)) \\ \frac{\partial o_c^L}{\partial a_j^{L-1}} &= \frac{\partial \sum_l w_{lc}^L a_l^{L-1}}{\partial a_j^{L-1}} = w_{jk}^L \\ \frac{\partial a_j^{L-1}}{\partial o_j^{L-1}} &= \sigma(o_j^{L-1}) (1 - \sigma(o_j^{L-1})) \\ \frac{\partial o_j^{L-1}}{\partial w_{pj}^{L-1}} &= \frac{\partial \sum_l w_{lj}^{L-1} a_l^{L-2}}{\partial w_{pj}^{L-1}} = a_p^{L-2} \end{aligned}$$

## RIASSUNTO:

Multi-layer Perceptron (MLP) utilizza l'algoritmo di backpropagation dell'errore per propagare l'errore a tutti i livelli.

MLP utilizza la discesa del gradiente (stocastico/mini-batch) per aggiornare i pesi.

La funzione di loss contiene molti minimi locali e non vi è alcuna garanzia di convergenza.

Quanto bene apprende la MLP e come possiamo migliorarla?

Quanto bene si generalizzera MLP (dati di test esterni)?

## 15. DEEP LEARNING

I modelli di classificazione derivati per l'apprendimento supervisionato sono semplificazioni della realtà, siccome le semplificazioni si basano su certe ipotesi. Le ipotesi falliscono in alcuni situazioni, ad esempio, a causa dell'incapacità di stimare perfettamente i parametri del modello ML da limitati dati.

### TEOREMA "NO FREE LUNCH":

Qualsiasi problema che abbiamo, non esiste la possibilità di definire un modello che funziona per tutti i problemi. Vuol dire che non esiste una soluzione che funzionerà bene per tutti i tipi di problemi.

Formalmente afferma che:

*Nessun singolo classificatore funziona al meglio per tutti i possibili problemi.*

(Dal momento che dobbiamo fare ipotesi per generalizzare).

### DIFFERENZE TRA MACHINE LEARNING (ML) E DEEP LEARNING (DL):

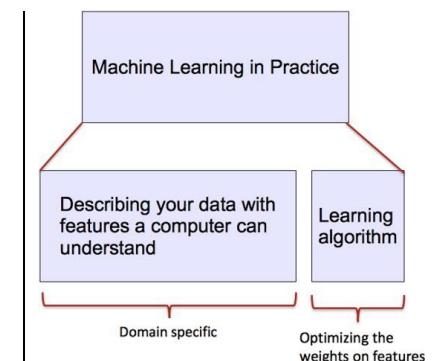
Una differenza importante riguarda le **features**.

I metodi di apprendimento automatico convenzionali si basano su **rappresentazioni di features progettate dall'uomo**. ML diventa solo l'ottimizzazione dei pesi per fare al meglio le predizioni.

Gli algoritmi visti fino ad ora si basano su caratteristiche che l'algoritmo utilizzerà per fare la classificazione. Chi va a progettare la rete neurale deve andare anche a definire quali sono le features più importanti per andare ad aiutare il processo di classificazione. Una volta definite le features, tutto il resto del processo è completamente automatico, il problema consiste poi ad andare ad ottimizzare i pesi in base a dati di esempio.

La parte che impatta sulla qualità del modello sono le features che dipendono dal tipo di problema che si va a risolvere.

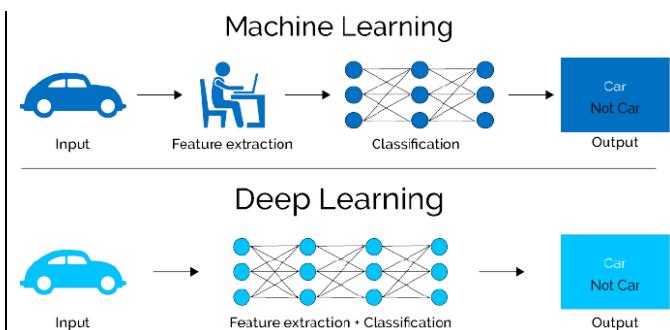
Nella pratica (come è definito dallo schema al lato), il task è dipendente dal dominio, il progettista deve avere conoscenza del dominio per capire quali sono le features utili per definire la funzione di approssimazione del modello, dopodiché si danno i dati all'algoritmo di apprendimento che viene eseguito per apprendere le caratteristiche del modello.



Il **deep learning (DL)** è un sottocampo di machine learning che utilizza più livelli per apprendere le rappresentazioni dei dati. DL è eccezionalmente efficace nell'apprendimento di pattern.

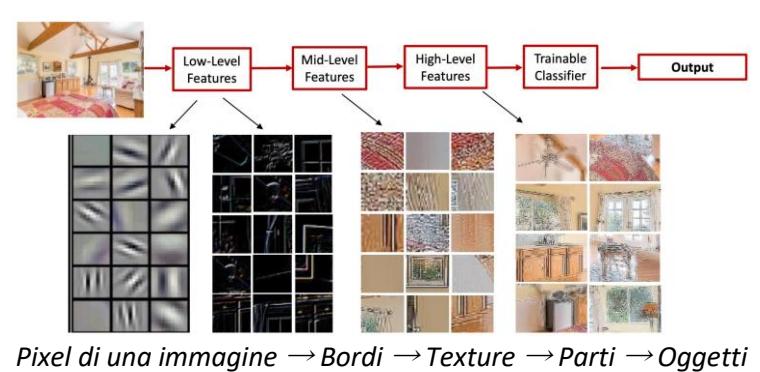
La caratteristica importante è che nel ML il programmatore deve fare l'estrazione delle features, mentre **nel DL la rete è in grado di estrarre in maniera automatica le features dai dati**, bisogna solo configurare la rete.

Le features vengono estratte dal modello grazie ai tanti strati, dove ognuno di esso si occupa di estrarre particolari caratteristiche dell'input.



### Esempio DL:

Il DL viene utilizzato molto nell'ambito della visione (delle immagini), perché è un dominio che presenta tantissime features. Anziché definirle a mano, meglio utilizzare una rete profonda che lo faccia in maniera automatica, riuscendosi ad adattare in base all'immagine di input. Si parte dall'immagine di input che può essere in formato RGB, si applicano dei filtri si ottengono diverse features dal più basso (bordi) al più alto livello (oggetti).



*Pixel di una immagine → Bordi → Texture → Parti → Oggetti*

Esiste però un vincolo per estrarre le feautes, ovvero c'è necessità di dare in input tantissimi dati. Nell'esempio sopra, per estrarre oggetti bisogna dare in input tantissime immagini contenente l'oggetto in questione. Un altro svantaggio sono i tempi richiesti per addestrare la rete, richiede un hardware molto potente.

DL fornisce un framework flessibile e di apprendimento per la rappresentazione di informazioni visive, testo, linguistiche, in più esso può apprendere in modo supervisionato e non.

## POTERE RAPPRESENTATIVO:

Le reti DL empiricamente funzionano bene ma non si sa il motivo del perché statisticamente si comporta meglio degli altri modelli, questo non è provabile del perché quell'algoritmo riconosce con una certa accuratezza, il funzionamento interno è sconosciuto all'utilizzatore, infatti, ciò che si fa è cambiare i parametri per vedere quale è la configurazione che si comporta meglio che ottiene le migliori prestazioni.

Infatti, andando ad analizzare il potere rappresentativo (potere espressivo) di una rete neurale semplice, le NN con almeno un livello nascosto sono **approssimatori universali**:

*Data una qualsiasi funzione continua  $h(x)$  e qualche  $\epsilon > 0$ , esiste una NN con uno strato nascosto (e con una ragionevole scelta di non linearità) descritto con la funzione  $f(x)$ , tale che  $\forall x, |h(x) - f(x)| < \epsilon$ . Cioè, le NN possono approssimare qualsiasi funzione continua complessa.*

Le NN utilizzano la mappatura non lineare degli input  $x$  agli output  $f(x)$  per calcolare confini decisionali complessi.

Praticamente si sta dicendo che con una rete ad un solo layer si riesce a classificare in maniera abbastanza precisa un qualsiasi problema di classificazione; pertanto, esiste  $h$  ma il problema è come si fa a trovarla, teoricamente esiste ma praticamente è difficile trovarla.

Il fatto che le NN profonde funzionino meglio è un'osservazione empirica. Matematicamente, le NN profonde hanno lo stesso potere rappresentativo di una NN ad un livello.

## INTRO ALLE RETI NEURALI:

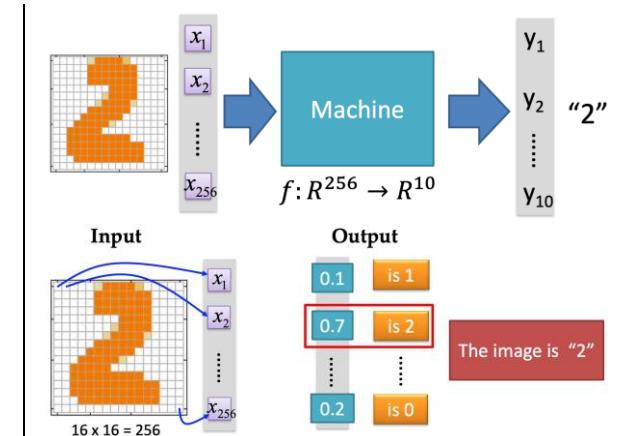
Uno dei casi più facili di classificazione è il riconoscimento di cifre da 0 a 9 da immagini di scrittura.

Supponiamo di avere il numero in una matrice 16x16 pixel, si può rappresentare l'immagine con un vettore di 256 valori, ogni  $x$  vale 0 se non è presente inchiostro 1 se è presente.

La rete neurale costruisce un modello che a partire da 256 valori booleani va a produrre un vettore di 10 valori (classificazione multiclasse).

In base al valore di probabilità si prende quello con probabilità maggiore, nell'esempio il numero è 2.

Questo mapping lo fa la funzione  $f$ , pertanto  $f$  è la nostra rete neurale.

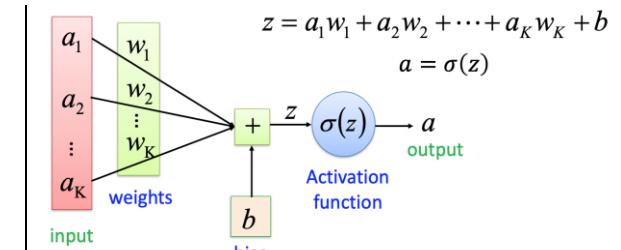


## ELEMENTI DI RETE NEURALE:

Le NN sono costituite da strati nascosti con neuroni (cioè unità di calcolo). Un singolo neurone mappa un insieme di input in un numero di output, o  $f: R^K \rightarrow R$ .

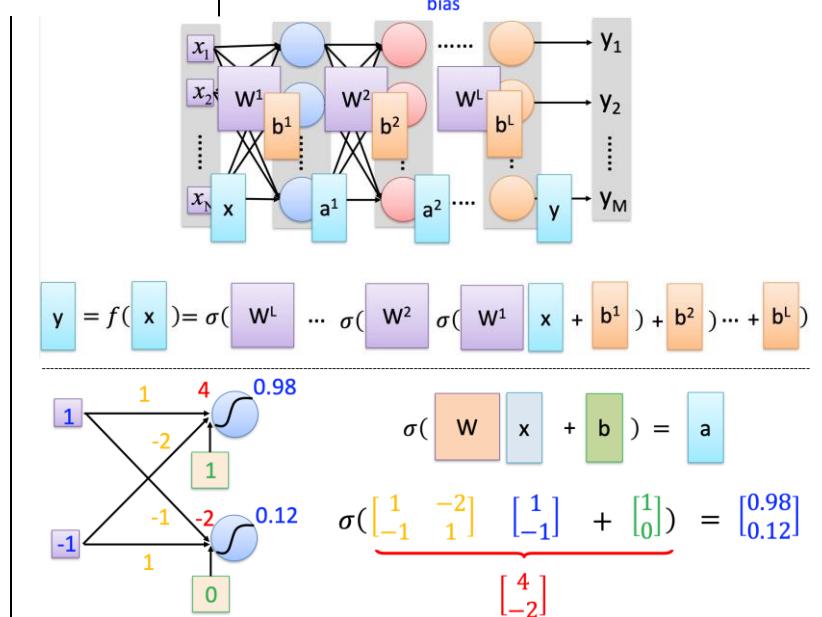
Ogni nodo di una rete semplice effettua i seguenti calcoli:

Fa una somma pesata dell'input con l'aggiunta di un valore (bias), il risultato viene passato ad una funzione di attivazione non lineare che va poi a produrre l'output.



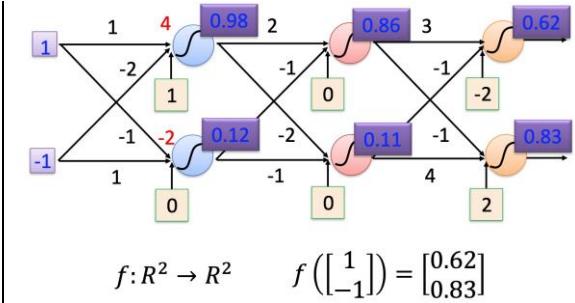
Nelle reti a più livelli si effettuano le stesse operazioni ma questa volta sulle matrici:

possiamo rappresentare il calcolo di  $h$ , come prodotto di una matrice per il vettore di input  $x$  più il vettore di bias.



## Esempio rete più livelli:

Abbiamo 6 neuroni (l'input non si conta), con una rete simile dobbiamo apprendere 26 parametri:



## SOFTMAX LAYER:

Nelle attività di **classificazione multiclasse**, il livello di output è in genere uno **strato softmax**, cioè impiega una **funzione di attivazione softmax**.

Avendo tanti nodi di output (classificazione multiclasse) bisogna decidere tra le possibili classi quella più probabile, ovvero quella che ha ottenuto il valore maggiore e questo viene fatto dalla funzione di attivazione softmax.

Questa funzione se abbiamo 3 opzioni (schema al lato) darà più probabilità al primo. Se invece venisse utilizzato uno strato con una funzione di attivazione sigmoidea come strato di output, le previsioni della NN potrebbero non essere facili da interpretare.

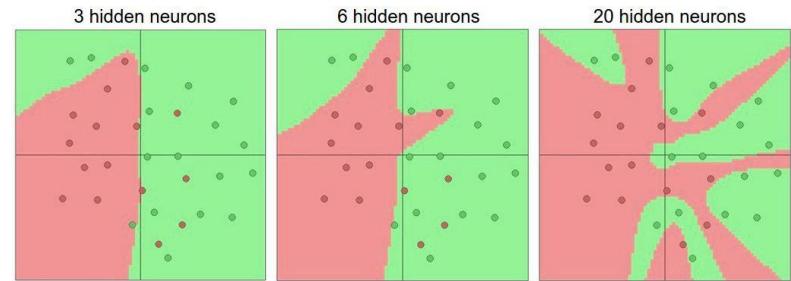
### A Layer with Sigmoid Activations

$$\begin{aligned} z_1 &\xrightarrow{\sigma} 0.95 & y_1 &= \sigma(z_1) \\ z_2 &\xrightarrow{\sigma} 0.73 & y_2 &= \sigma(z_2) \\ z_3 &\xrightarrow{\sigma} 0.05 & y_3 &= \sigma(z_3) \end{aligned}$$

## FUNZIONI DI ATTIVAZIONE:

Sono necessarie **attivazioni non lineari** per apprendere rappresentazioni di dati complessi (non lineari), altrimenti, NNs sarebbe solo una funzione lineare (come  $W_1 W_2 x = Wx$ ).

NN con un gran numero di strati (e neuroni) possono approssimare funzioni più complesse, più neuroni migliorano la rappresentazione (ma potrebbero andare in overfit).



## TRAINING NN:

Supponiamo una rete con 256 input, layer hidden e layer di output, si vogliono apprendere i parametri di questa rete che sono tutti i pesi lungo la rete incluso in bias.

Per fare questo bisogna effettuare una fase di training, ovvero dare in input al modello delle coppie che sono immagine del numero e la classe (label) per andare a determinare theta ( $\theta$ ) che sono tutti i parametri.

I **parametri** di rete  $\theta$  includono le matrici dei pesi e i vettori di bias per tutti i layer:

$$\theta = \{W^1, b^1, W^2, b^2, \dots, W^L, b^L\}$$

Spesso i parametri del modello  $\theta$  sono indicati come pesi.

Addestrare un modello per apprendere una serie di parametri  $\theta$  che sono ottimali (secondo un criterio) è una delle maggiori sfide nel ML.

Quando si trattano immagini si effettua una **fase di preelaborazione** dei dati per migliorare il training, in questo modo il modello dovrebbe apprendere meglio siccome le immagini sono tutte della stessa forma. Le preelaborazioni sono:

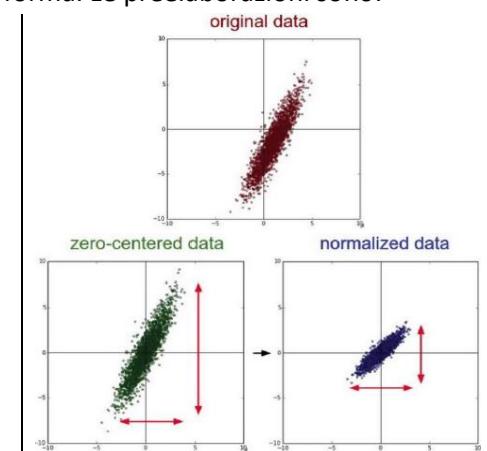
- **Sottrazione della media**, per ottenere dati incentrati sullo zero:

Sottrarre la media per ogni singola dimensione dei dati (feature)

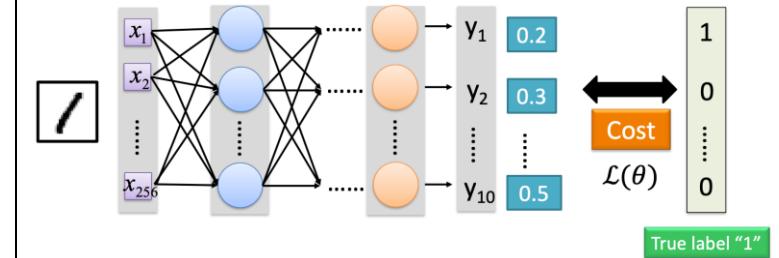
- **Normalizzazione**:

Dividi ogni feature per la sua deviazione standard, per ottenere una deviazione standard di 1 per ciascuna dimensione dei dati (feature).

Oppure, ridimensionare i dati all'interno dell'intervallo [0,1] o [-1, 1], ad esempio, le intensità dei pixel dell'immagine sono divise per 255 per ridimensionare in [0,1].



Quello che si vuole fare durante il training è che, ad esempio, se viene preso in input l'immagine contenente il numero 1 allora  $y_1$  deve avere il valore più alto. Così che la funzione softmax restituisce la giusta classe, questo lo si fa definendo la funzione di loss.



Una **funzione di loss** (funzione di costo/ obiettivo)  $\mathcal{L}(\theta)$  calcola la differenza (errore) tra la previsione del modello e l'etichetta vera. Per esempio, può essere errore quadratico medio, cross-entropy, eccetera.

Per un training set di  $N$  immagini, calcola la loss totale su tutte le immagini  $\mathcal{L}(\theta) = \sum_{n=1}^N \mathcal{L}_n(\theta)$

L'obiettivo è minimizzare la funzione di loss che può essere fatto con un algoritmo iterativo basato sulla discesa del gradiente e ciò che fa è andare a vedere, partendo dalla loss, come modificare i pesi affinché la loss stessa si riduca. Essendo che la loss dipende dai parametri (i pesi), siccome si è fatta prima classificazione e poi applicata la loss.

Nell'esempio di prima erano 26 parametri  $\theta$  e per tanto bisogna fare 26 calcoli di gradiente e vedere come aggiornare questi 26 pesi. Il gradiente fa vedere come la curva cresce e si va nella direzione opposta ovvero dove diminuisce la funzione.

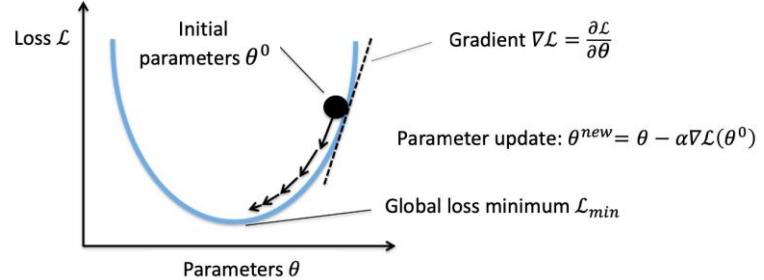
L'algoritmo è iterativo:

1. Inizialmente i pesi sono messi in maniera casuale;
2. Si va a calcolare il gradiente della loss rispetto ai parametri iniziali;
3. I pesi saranno aggiornati secondo la formula allo step 3 in figura;
4. Si ritorna allo step 2 ripetendo l'aggiornamento.

L'algoritmo si ferma quando si raggiunge un **minimo** nella speranza che sia globale e non solo locale.

- Steps in the **gradient descent algorithm**:

1. Randomly initialize the model parameters,  $\theta^0$
2. Compute the gradient of the loss function at the initial parameters  $\theta^0$ :  $\nabla \mathcal{L}(\theta^0)$
3. Update the parameters as:  $\theta^{new} = \theta^0 - \alpha \nabla \mathcal{L}(\theta^0)$ 
  - o Where  $\alpha$  is the learning rate
4. Go to step 2 and repeat (until a terminating criterion is reached)



## BACKPROPAGATION:

L'**algoritmo di backpropagation** permette di derivare la loss rispetto ai pesi. L'idea è di fare prima una procedura di forward propagation, ovvero dall'input va avanti fino ad arrivare all'output, dopodiché dopo ottenuto l'output si calcola la loss e, per ricavare la derivata della loss rispetto a tutti i pesi, si fa il backpropagation, cioè si fa un passaggio dalla fine verso l'inizio sfruttando la regola della chainrule per le derivate parziali che permettono di fare le derivate all'indietro.

Più formalmente:

- Per addestrare NN, la **forward propagation** (forward pass) si riferisce al passaggio degli input  $x$  attraverso i livelli nascosti per ottenere gli output del modello (previsioni)  $y$ , è calcolata la perdita  $\mathcal{L}(y, \hat{y})$ ;
- La **back propagation** attraversa la rete in ordine inverso, dagli output  $y$  verso gli input  $x$  per calcolare i gradienti della perdita  $\nabla \mathcal{L}(y, \hat{y})$ . La **chain rule** serve per calcolare le derivate parziali della funzione di perdita rispetto ai parametri  $\theta$  nei diversi strati della rete;
- Ogni aggiornamento dei parametri del modello  $\theta$  durante l'allenamento effettua un passaggio in avanti e uno indietro (ad es. per un batch di inputs);

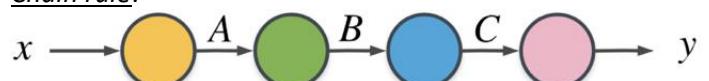
Il calcolo automatico dei gradienti (differenziazione automatica) è disponibile in tutte le librerie di deep learning. Semplifica notevolmente l'implementazione di algoritmi di deep learning, poiché evita di derivare le derivate parziali della funzione di perdita a mano.

Il **problema** è calcolare la derivata della loss che abbiamo alla fine della rete rispetto ai pesi presenti nella rete, cioè si vuole calcolare la derivata di  $y$  rispetto a  $x$ .

Questa derivata si può calcolare andando a trasformare la frazione (1.) come prodotto della derivata di  $y$  rispetto a  $C$ , per la derivata di  $C$  rispetto a  $B$ , per la derivata di  $B$  rispetto ad  $A$ , per la derivata di  $A$  rispetto a  $x$ .

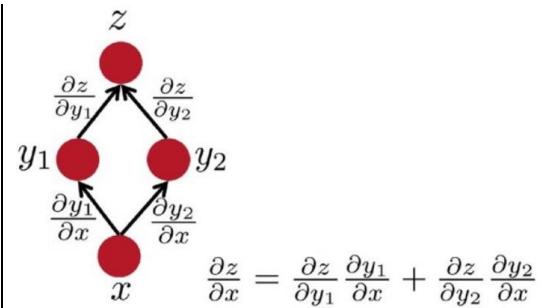
I valori  $x$  e  $y$  sono distanti tra loro e calcolare la derivata di  $y$  rispetto a  $x$  è impossibile, ma se lo si scomponete e si calcola in maniera diretta ogni frazione è possibile derivare la frazione iniziale.

### Chain rule:



$$(1.) \quad \frac{\partial y}{\partial x} = \frac{\partial y}{\partial C} \times \frac{\partial C}{\partial B} \times \frac{\partial B}{\partial A} \times \frac{\partial A}{\partial x}$$

Caso in cui abbiamo una diramazione nel grafo, pertanto la formula cambia in una somma di prodotti:



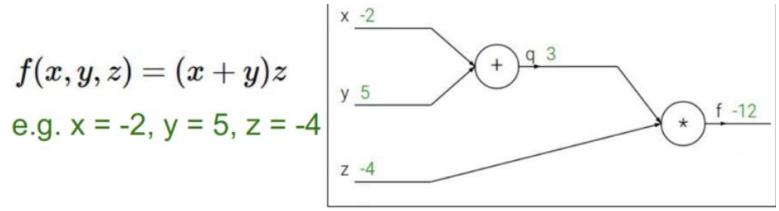
### Esempio:

Caso semplice con un grafo computazionale.

$x$  e  $y$  sono i due input che vengono prima sommati per produrre  $q=x+y$  e poi moltiplicato  $z$  per produrre  $f=q*z$ .

Si vuole capire come aggiornare i pesi per far cambiare i valori di  $f$ . Interessa fare la derivata di  $f$  rispetto a  $x$ , derivata di  $f$  rispetto a  $y$  e derivata di  $f$  rispetto a  $z$ .

Dalle formule nei riquadri, si può calcolare solo la derivata rispetto a  $x$  e a  $y$  (primo riquadro) e la derivata rispetto a  $q$  e a  $z$  (secondo riquadro), quindi abbiamo il modo di calare solo la derivata di  $f$  rispetto a  $z$  (terza frazione che si vuole). Le altre due frazioni per calcolarle bisogna ricavare il valore di  $f$  rispetto a  $q$  che sarebbe  $z$  (nell'esempio vale -4).



$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

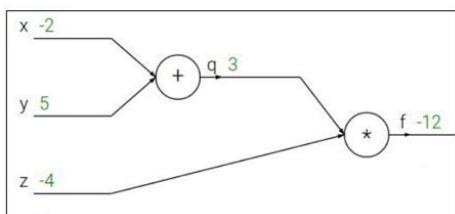
$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

$$\text{Want: } \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$$

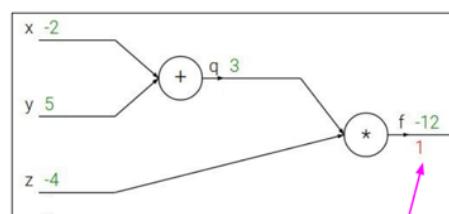
Applicando la chain rule è possibile ricavarsi tutto il necessario:

(il verde è il passo forward, mentre il rosso è la fase di backward)

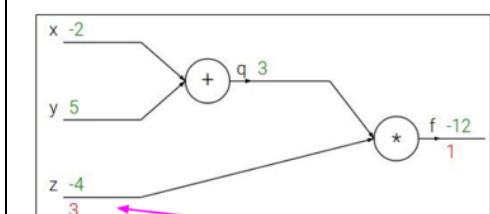
1.



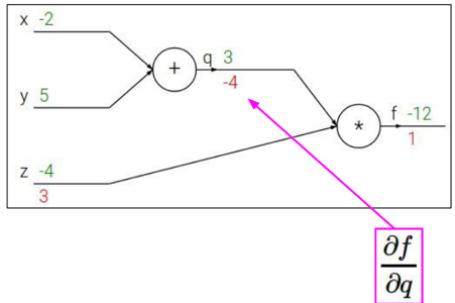
2.



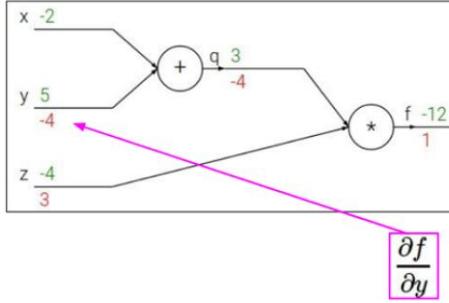
3.



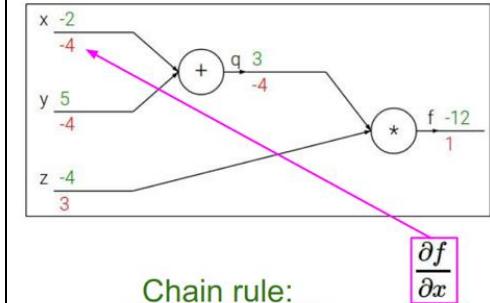
4.



5.



6.

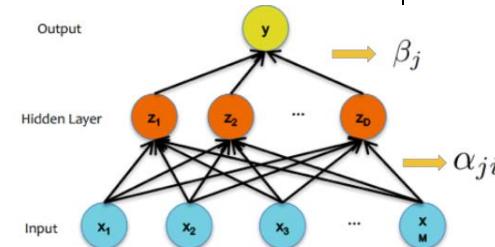


Riassumendo la rete fa tutto ciò rappresentato dai grafici a destra →

Hidden layer calcola la sommatoria, il risultato va dato alla funzione di attivazione sigmoide, viene calcolato l'output ed infine viene applicato su di esso la sigmoide.

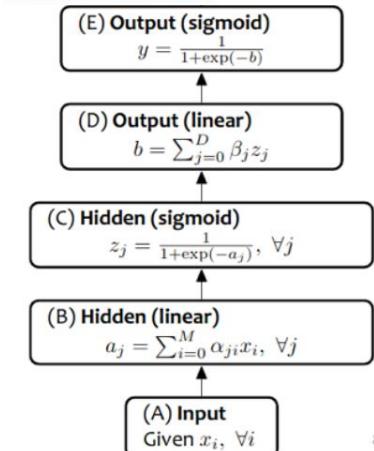
Abbiamo i pesi  $\alpha$  sui primi archi e i pesi  $\beta$  sugli altri archi.

La funzione di loss si suppone che sia  $J$  (presente a destra), bisogna calcolare la derivata della loss rispetto ai pesi  $\alpha$  e  $\beta$ , così da poterli aggiornare.



Assume loss is as follows:

$$J = y^* \log y + (1 - y^*) \log(1 - y)$$



Si fa il forward pass che sarebbe calcolare  $a_j, z_j, b, y$  per poi andare a calcolare la loss  $J$ .

Nel passo backward si parte dal calcolo della derivata della loss rispetto all'output, dopodiché è possibile calcolare la derivata della loss rispetto a  $b$ .

Ora è possibile calcolare la derivata della loss rispetto a  $\beta$  che è come il peso  $\beta$  deve essere aggiornato.

Poi si calcola la derivata della loss rispetto a  $z$ , e così via fino ad arrivare a calcolare le derivate della loss rispetto ai pesi che è il nostro obiettivo. Facendo poi il prodotto si calcola la derivata finale.

### Forward

$$J = y^* \log y + (1 - y^*) \log(1 - y)$$

$$y = \frac{1}{1 + \exp(-b)}$$

$$b = \sum_{j=0}^D \beta_j z_j$$

$$z_j = \frac{1}{1 + \exp(-a_j)}$$

$$a_j = \sum_{i=0}^M \alpha_{ji} x_i$$

### Backward

$$\frac{dJ}{dy} = \frac{y^*}{y} + \frac{(1 - y^*)}{y - 1}$$

$$\frac{dJ}{db} = \frac{dJ}{dy} \frac{dy}{db}, \frac{dy}{db} = \frac{\exp(-b)}{(\exp(-b) + 1)^2}$$

$$\frac{dJ}{d\beta_j} = \frac{dJ}{db} \frac{db}{d\beta_j}, \frac{db}{d\beta_j} = z_j$$

$$\frac{dJ}{dz_j} = \frac{dJ}{db} \frac{db}{dz_j}, \frac{db}{dz_j} = \beta_j$$

$$\frac{dJ}{da_j} = \frac{dJ}{dz_j} \frac{dz_j}{da_j}, \frac{dz_j}{da_j} = \frac{\exp(-a_j)}{(\exp(-a_j) + 1)^2}$$

$$\frac{dJ}{d\alpha_{ji}} = \frac{dJ}{da_j} \frac{da_j}{d\alpha_{ji}}, \frac{da_j}{d\alpha_{ji}} = x_i$$

$$\frac{dJ}{dx_i} = \frac{dJ}{da_j} \frac{da_j}{dx_i}, \frac{da_j}{dx_i} = \sum_{j=0}^D \alpha_{ji}$$

## DISCESA DEL GRADIENTE IN MINI BATCH:

Per quanto riguarda l'esecuzione pratica dell'algoritmo descritto, esistono vari modi. Una prima versione sarebbe andare a calcolare la perdita sull'insieme di dati di training (versione iniziale - vanilla), ma richiede di effettuare tutto il training di tutto il dataset per fare un aggiornamento dei pesi, esso è molto inefficiente.

La versione più utilizzata è la versione con **mini-batch**, dove si prende il dataset e si sceglie un sottoinsieme per sottoporlo al training, si calcola la loss e si aggiornano i pesi, si prende un altro sottoinsieme e si ripete il passaggio. Quando i sottoinsiemi sono finiti si finisce una epoca e si riparte con una epoca successiva.

Più formalmente:

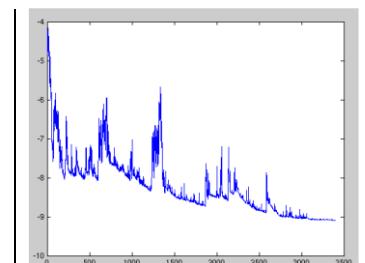
- Calcola la perdita  $\mathcal{L}(\theta)$  su un mini-batch di immagini, aggiorna i parametri  $\theta$  e ripeti finché non sono usate tutte le immagini;
- All'epoca successiva, mescola i dati di training e ripeti il processo precedente.

Il mini-batch porta ad un addestramento molto più veloce. Tipiche taglie di mini-batch: da 32 a 256 immagini. Funziona perché il gradiente di un mini-batch è una buona approssimazione del gradiente dell'intero training set.

## DISCESA DEL GRADIENTE STOCASTICO:

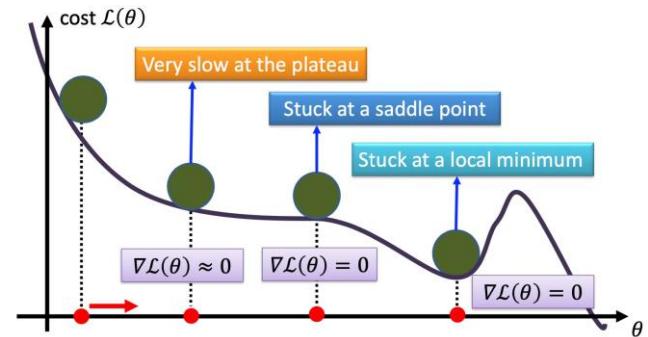
SGD utilizza mini-batch costituiti da un **unico esempio di input**. Ad esempio, una sola immagine di mini-batch. Pertanto, si fa il test, si fa il training di un solo campione, si calcola la loss e si aggiornano i pesi.

Sebbene questo metodo sia molto veloce, può causare fluttuazioni significative nella funzione. Pertanto, è meno comunemente usato, viene preferito GD minibatch.



## PROBLEMI CON LA DISCESA DEL GRADIENTE:

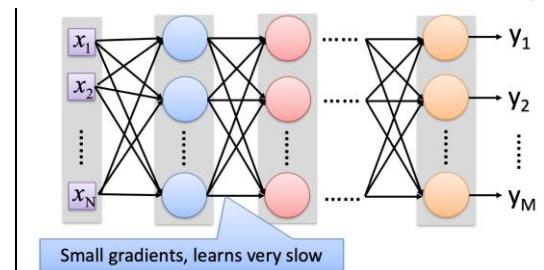
Oltre al problema dei minimi locali, l'algoritmo GD può essere molto lento ai **plateaus**, e può rimanere bloccato in **punti saddle**.



In alcuni casi, durante l'allenamento, i gradienti possono diventare o molto piccoli (**gradienti evanescenti**) o molto grandi (**esplosione del gradiente**).

Portano ad un aggiornamento molto piccolo (i pesi non vengono più aggiornati) o molto grande (variazione dei pesi eccessiva) dei parametri.

La soluzione è il cambio del tasso di apprendimento, attivazioni ReLU, regolarizzazione oppure le unità LSTM in RNN.



## REGOLARIZZAZIONE – WEIGHT DECAY:

Chiamata anche  **$L_2$  weight decay** perché fa la sommatoria dei quadrati dei pesi.

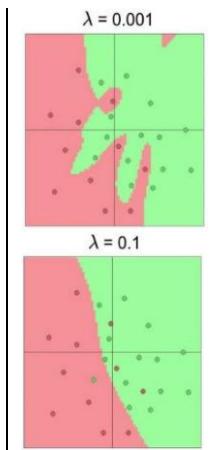
La tecnica del **decadimento del peso** viene utilizzata quando si verifica l'esplosione del gradiente, per **regolarizzare** il valore della loss. Alla loss function si aggiunge un termine di regolarizzazione che penalizza i pesi grandi:

$$\mathcal{L}_{reg}(\theta) = \mathcal{L}(\theta) + \lambda \sum_k \theta_k^2$$

Data loss      Regularization loss

Per ogni peso della rete, alla loss aggiungiamo il termine di regolarizzazione. Durante l'aggiornamento dei parametri in GD, ogni peso viene decaduto linearmente verso zero.

Il **coefficiente di decadimento del peso  $\lambda$**  determina quanto sia dominante la regolarizzazione durante il calcolo del gradiente.



Un'altra versione si chiama  **$L_1$  weight decay** perché fa la somma dei valori assoluti dei pesi:

$$\mathcal{L}_{reg}(\theta) = \mathcal{L}(\theta) + \lambda \sum_k |\theta_k|$$

Il decadimento del peso  **$L_1$**  è meno comune con NN. Spesso ha prestazioni peggiori di  **$L_2$** .

È anche possibile combinare  **$L_1$**  e  **$L_2$**  chiamata **elastic net regularization**:

$$\mathcal{L}_{reg}(\theta) = \mathcal{L}(\theta) + \lambda_1 \sum_k |\theta_k| + \lambda_2 \sum_k \theta_k^2$$

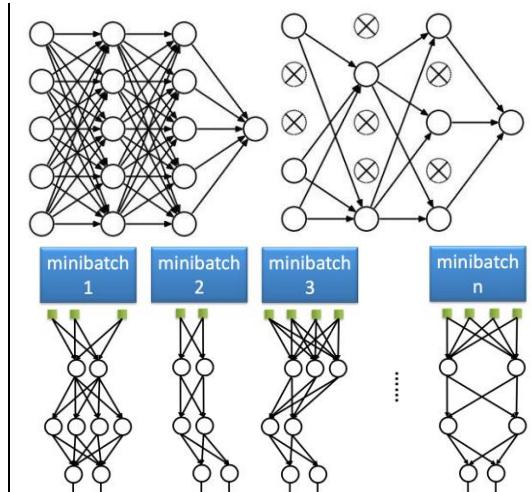
$$\mathcal{L}_{reg}(\theta) = \mathcal{L}(\theta) + \lambda_1 \sigma_k |\theta_k| + \lambda_2 \sigma_k \theta_k^2$$

## REGOLARIZZAZIONE – DROPOUT:

Questa tecnica rilascia casualmente le unità (insieme alle loro connessioni) durante il training.

Ciascuna unità viene spenta con un **tasso di dropout  $p$** , indipendente dalle altre unità.

È necessario scegliere l'iperparametro  $p$  (tuned), spesso, tra il 20% e il 50% delle unità sono dropped.



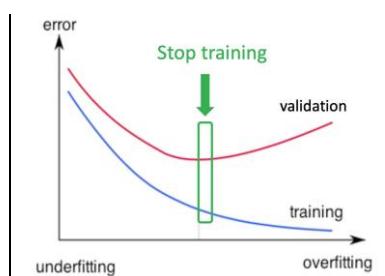
Il dropout può essere visto come una specie di **ensemble learning**, usando un mini-batch diverso per addestrare la rete. Ad esempio, la prima volta si fa training col minibatch 1 e così via... Quello che rimane sono i pesi, siccome i pesi appresi con un minibatch vengono aggiornati nel successivo. In questo modo riescono a catturare delle caratteristiche in maniera diversa in base a quanti nodi sono attivi.

## REGOLARIZZAZIONE – FERMATA ANTICIPATA:

Un'altra tecnica è l'**early-stopping**:

- Durante l'addestramento del modello, utilizzare un **set di validazione**.
- Stop quando l'accuracy (o la perdita) della convalida non è migliorata dopo  $n$  epoche, il parametro  $n$  viene chiamato **patience**.

Tutto ciò evita l'overfitting.



## NORMALIZZAZIONE BATCH:

I **livelli di normalizzazione batch** sono dei nodi che si vanno ad aggiungere alla rete (chiamati **layers BatchNorm**) per non far variare i dati rispetto a media e varianza, esso è un processo di normalizzazione.

Calcolano la media  $\mu$  e la varianza  $\sigma$  di un batch di dati di input e normalizzano i dati  $x$  su una media zero e una varianza unitaria →

$$\text{I.e., } \hat{x} = \frac{x - \mu}{\sigma}$$

Portano a convergenza più rapida del training, consentono un tasso di apprendimento più ampio.

## HYPER PARAMETRI TUNING:

Il training di NN (reti profonde) può comportare il settaggio di molti **iperparametri**, quelli più comuni sono:

- Il numero di layers ed il numero di neuroni per strato;
- Tasso iniziale del learning rate;
- Tasso di decadimento del learning rate (ad esempio costante);
- Genere di ottimizzatore.

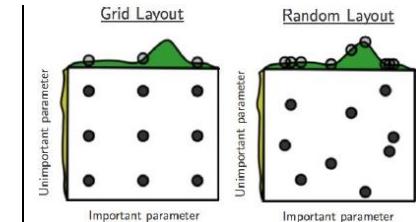
Altri iperparametri possono includere:

- Parametri di regolarizzazione (penalità (#, dropout rate);
- Batch size;
- Funzioni di attivazione;
- Funzione di loss.

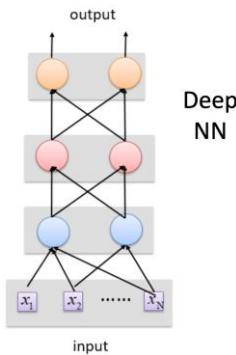
L'ottimizzazione degli iperparametri può richiedere molto tempo per NN grandi. Si parte dall'esperienza appresa della rete, poi bisogna fare degli aggiornamenti e ripetere l'esperimento per vedere come cambia, questo è costoso.

Per fare il Tuning dei iperparametri si possono anche eseguire delle tecniche che sono:

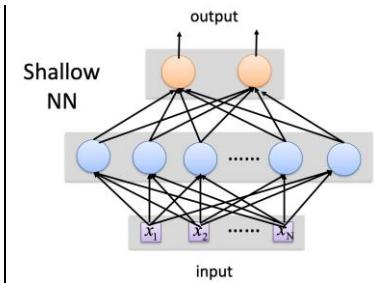
- **Grid search**: Controllare tutti i valori in un intervallo con un valore di step;
- **Random search**: Campiona casualmente i valori per il parametro. Spesso preferito alla ricerca nella griglia;
- **Ottimizzazione bayesiana degli iperparametri**: È un'area attiva di ricerca.



### DEEP VS SHALLOW NETWORKS:



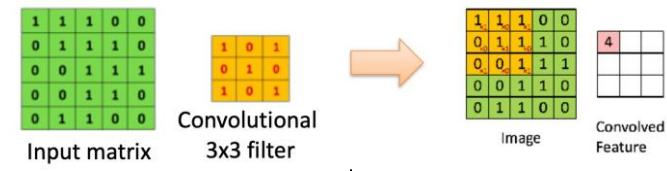
Le **reti più profonde** funzionano meglio di quelle poco profonde reti, ma solo fino a un certo limite, dopo un certo numero di strati, le prestazioni delle reti più profonde si stabilizzano.



### RETI NEURALI CONVOLUZIONALI (CNN):

Le **reti neurali convoluzionali** (CNN) sono state progettate principalmente per immagini.

La caratteristica principale delle CNN è che utilizzano, all'interno della rete (nei neuroni), un **operatore convoluzionale** per estrarre le feature.



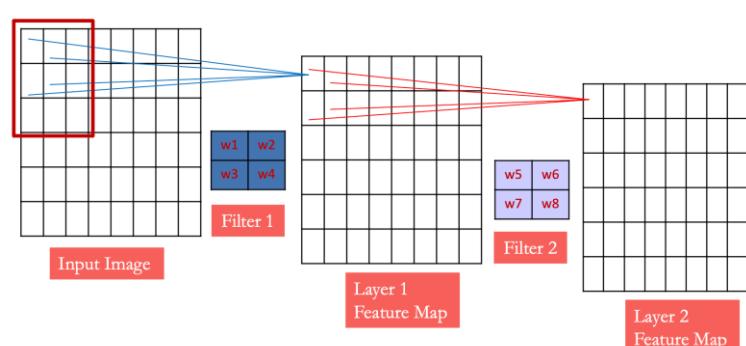
Un filtro convoluzionale scorre attraverso l'immagine. Quando i filtri convoluzionali vengono scansionati sull'immagine, essi catturano features utili, prima di basso livello come bordi, fino ad estrarre interi oggetti.

Una cosa importante è che le reti fully connected hanno uno svantaggio riguardante i pesi, essi erano matrici perché partivano da un nodo e andavano in tutti gli altri nodi. Nella realtà, questa dipendenza tra nodi non è sempre necessaria. Con le CNN si riesce ad avere condivisione di pesi tra i nodi e non tutti i nodi sono collegati al livello successivo. Ogni arco ha un peso e nel fully connected abbiamo una quantità di pesi da allenare molto elevata, mentre nelle CNN abbiamo un numero di parametri che si riduce perché gli archi non collegano tutti i nodi.

L'idea principale della rete CNN è l'applicazione del filtro convoluzionale che viene applicato ai dati di input che possono essere visti come delle matrici (rappresentano una immagine) e questo filtro viene fatto passare sopra alla matrice. Questo poi si ripete in base ad un parametro in base dei criteri.

Nelle CNN, le unità nascoste in un livello sono collegate solo a una piccola regione dello strato precedente (chiamato campo **ricettivo locale**).

La profondità di ciascuna **feature map** corrisponde al numero di filtri convoluzionali utilizzati in ciascuna strato.



Dopo aver applicato un operatore convoluzionale si può applicare un **pooling**, che serve per ridurre la quantità di dati, riducendo la dimensione della matrice iniziale:

- **Max pooling**: riporta l'output massimo all'interno di un rettangolo (vicinato);

- **Average pooling**: riporta l'output medio di un rettangolo vicino;

I livelli di pooling riducono la dimensione spaziale delle feature map. Ridurre il numero di parametri, prevenire overfitting.

1	3	5	3
4	2	3	1
3	1	1	3
0	1	0	4

Input Matrix

MaxPool with a  $2 \times 2$  filter with stride of 2

4	5
3	4

Output Matrix

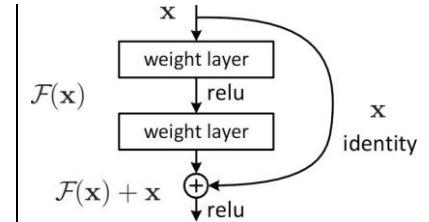
## CNN RESIDUALI (ResNet):

Una variante delle CNN sono le **residuali** che fanno avanzare delle informazioni da un layer ad un layer più avanti nella catena. È utile, siccome potrebbe capitare che da un layer al successivo ci sono delle grosse variazioni dovuto a delle perturbazioni.

Tra i vari processi di convoluzione e pooling potremmo perdere informazioni legate all'immagine originale, per prevenire questo portiamo avanti i dati che abbiamo in un certo layer.

### Esempio:

Immaginiamo di avere  $x$ , potrebbe capitare che i due strati cambiano il significato dei dati. Quindi, propaghiamo  $x$  in avanti, mitigando il problema del vanishing gradient durante il training.



## RETI NEURALI RICORRENTI (RNN):

Le **Recurrent NN** vengono utilizzate per modellare **dati sequenziali** e dati con input e output di lunghezza variabile, esempio video, testo, parlato, sequenze di DNA, dati di scheletri umani. Le RNN tengono in considerazione non solo  $x$  ma anche i dati precedenti, proprio per questo non vengono utilizzati per le immagini, siccome per esse si vuole sapere  $y$  solo dato  $x$ . La memoria degli input precedenti viene archiviata nello stato del modello ed influenzano le predizioni del modello.

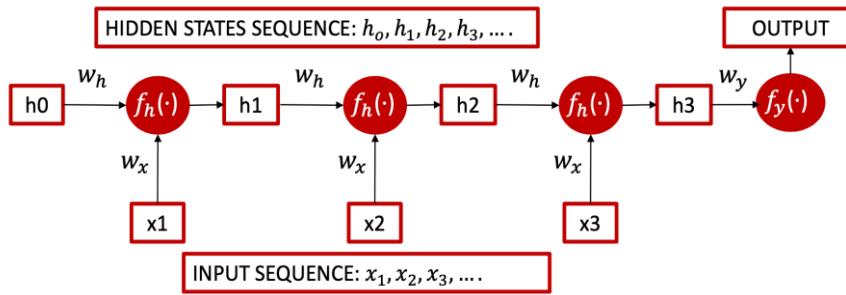
Le RNN sono più sensibili al **problema del gradiente evanescente** rispetto alle CNN, siccome vanno considerati i pesi per i dati precedenti.

Più formalmente, RNN usa lo stesso set di pesi  $w_h$  e  $w_x$  **in tutti i passi temporali**:

- Viene appresa una sequenza di **hidden state**  $\{h_1, h_2, \dots\}$  che rappresenta la memoria della rete;
- Lo stato hidden al passo  $t$ ,  $h(t)$ , viene calcolato in base al precedente hidden state  $h(t-1)$  e l'input al passo corrente  $x(t)$ , cioè:

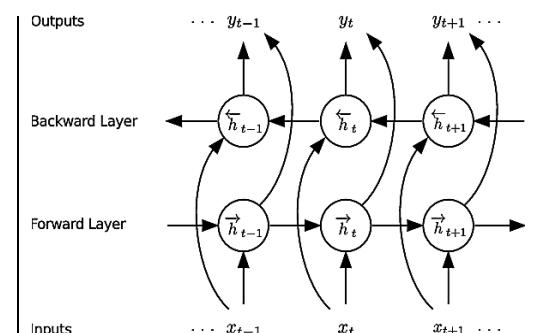
$$h(t) = f_h(w_h * h(t-1) + w_x * x(t))$$

- La funzione  $f_h(\cdot)$  è una funzione di attivazione non lineare, e.g., ReLU o tanh



## BIDIREZIONALE RNN:

Sono una evoluzione delle precedenti ed è una generalizzazione della precedente, cioè esse guardano non solo indietro ma anche in avanti. Quando analizzano un certo  $x_i$  tengono in considerazione sia il layer di forward che di backward.



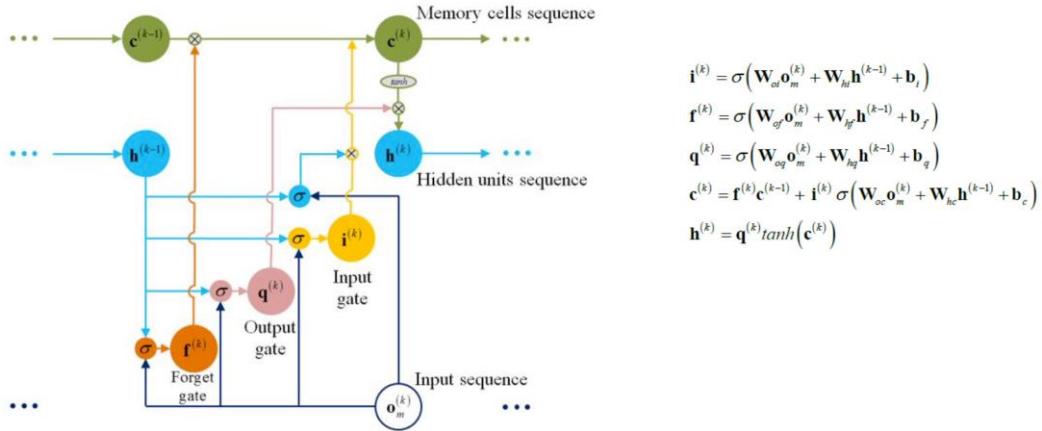
## LSTM RETI:

Le **Long Short-Term Memory** (LSTM) sono una variante di RNN, LSTM attenua il problema della fuga del gradiente/esplosione, la soluzione è una *Memory Cell*, aggiornata ad ogni passaggio della sequenza.

Tre porte controllano il flusso di informazioni da e verso la *Memory Cell*:

- *Input Gate*: protegge il passo corrente dagli input irrilevanti;
- *Output Gate*: impedisce al passo corrente di trasmettere informazioni irrilevanti ai passi successivi;
- *Forget Gate*: limita le informazioni passate da una cella alla prossima.

Il problema dei RNN è che si portano avanti delle informazioni ma alcune di esse vengono perse comunque nei vari layer successivi (perdita del contesto), invece con i gate possiamo decidere quanto far passare di contesto al layer successivo (maggior controllo sui dati tramite le porte).



$$\begin{aligned}\mathbf{i}^{(k)} &= \sigma(\mathbf{W}_{oi} \mathbf{o}_m^{(k)} + \mathbf{W}_{ih} \mathbf{h}^{(k-1)} + \mathbf{b}_i) \\ \mathbf{f}^{(k)} &= \sigma(\mathbf{W}_{of} \mathbf{o}_m^{(k)} + \mathbf{W}_{hf} \mathbf{h}^{(k-1)} + \mathbf{b}_f) \\ \mathbf{q}^{(k)} &= \sigma(\mathbf{W}_{oq} \mathbf{o}_m^{(k)} + \mathbf{W}_{hq} \mathbf{h}^{(k-1)} + \mathbf{b}_q) \\ \mathbf{c}^{(k)} &= \mathbf{f}^{(k)} \mathbf{c}^{(k-1)} + \mathbf{i}^{(k)} \sigma(\mathbf{W}_{oc} \mathbf{o}_m^{(k)} + \mathbf{W}_{hc} \mathbf{h}^{(k-1)} + \mathbf{b}_c) \\ \mathbf{h}^{(k)} &= \mathbf{q}^{(k)} \tanh(\mathbf{c}^{(k)})\end{aligned}$$