Apprendimento Automatico (Intro)

Fabio Aiolli
www.math.unipd.it/~aiolli

Sito web del corso
www.math.unipd.it/~aiolli/corsi/1718/aa/aa.html

Orario

- ✓ 40 ore di lezione in aula (5cfu)
- ✓ 8 ore di laboratorio (1cfu)

- ✓ AULA: Lunedì dalle 12:30 alle 14:30 e Mercoledì dalle 14:30 alle 16:30, Aula 1BC50 (Torre Archimede)
- ✓ LABORATORI: Da definire (verso la fine del corso)

Testi e Esame

- T. Mitchell, "Machine Learning", McGraw Hill, 1998 (disponibile in biblioteca)
- E. Alpaydin, "Introduction to Machine Learning", Cambridge University Press (liberamente scaricabile)
- Slides e altro materiale presentato nel corso

- Esame 1/4 progetto (attività) e 3/4 scritto (+ orale opzionale)

Attività del corso…

Non necessariamente un reale progetto!

Consisterà nell'applicazione di tecniche di machine learning su uno o più dataset presi da kaggle https://www.kaggle.com o raccolti su un repository nostro. La valutazione sarà incrementale (a cadenza mensile).

E' premiata ogni attività collaborativa (tramite gruppo google)

- descrizione di nuovi task o datasets
- relazione su attività svolte e risultati ottenuti
- revisione di attività svolte da colleghi

P.e. Movie recommendation

https://www.kaggle.com/c/movie

Per info iscriversi a.s.a.p. al gruppo google ML18PD
https://groups.google.com/forum/#!forum/ml18pd

In alternativa è sempre possibile fare un normale progetto dopo la fine del corso.

Da dove cominciare..
per non rimanere indietro

Algebra Lineare (vettori e matrici):

- Youtube ('linear algebra for machine learning')

Probabilità:

- Appendice A (Alpaydin)

Python:

- http://www.python.it/doc/
- Aiolli. *Appunti di programmazione (scientifica) in Python*, Esculapio 2013
- Guardare `numpy`, `scipy`, `matplotlib`, `scikit-learn`: tutti disponibili nel pacchetto python(x,y)
- Avremo un laboratorio dedicato nelle prossime settimane

Domande chiave..
(a cui tenteremo di dare risposte)

Quando e perché è utile un approccio basato su Machine Learning?

Come si può apprendere?

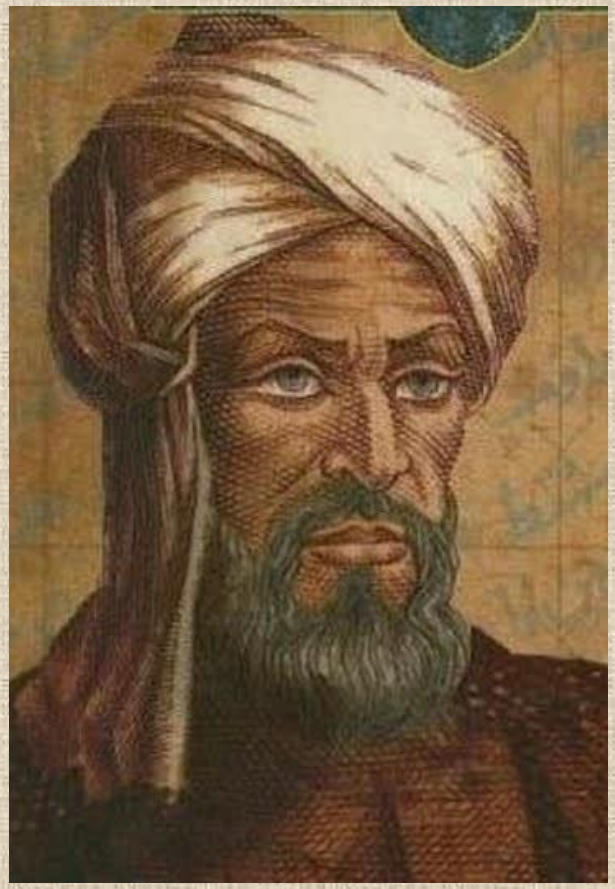
Ma la più importante di tutte: Possiamo veramente apprendere?

Algoritmo

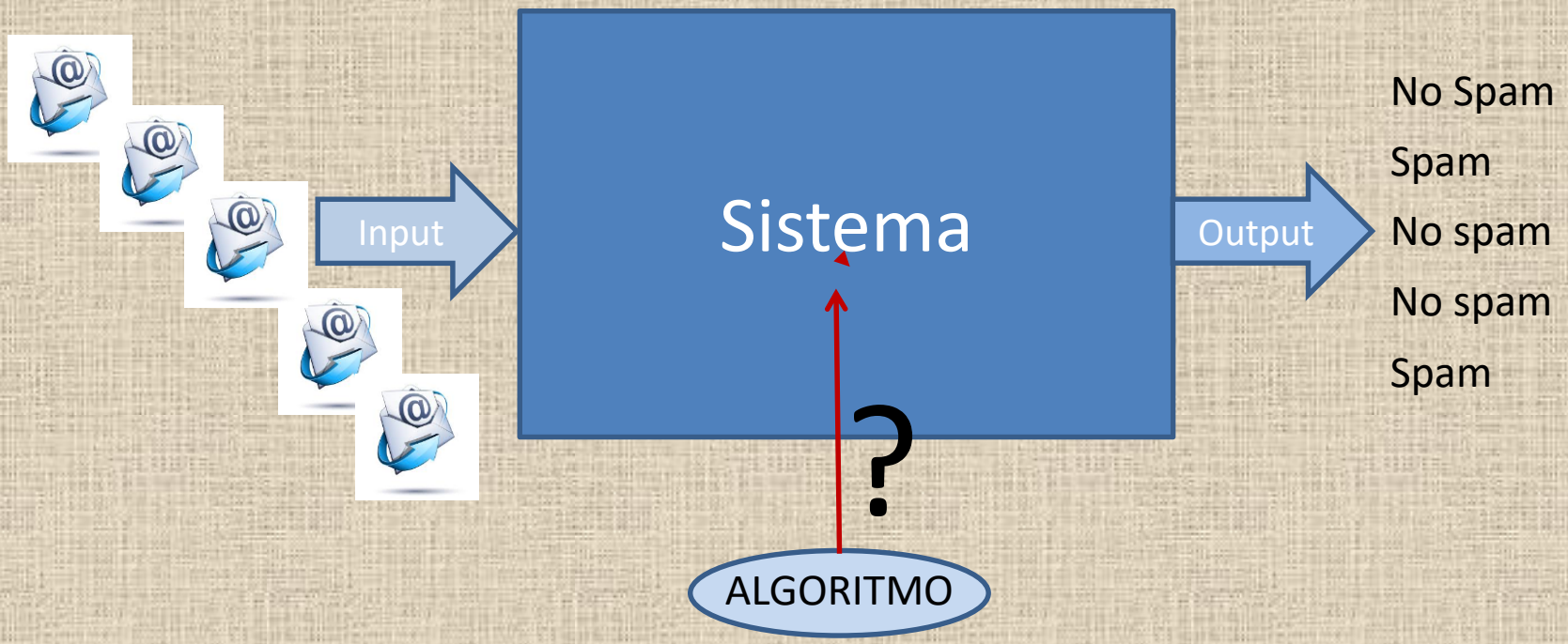
WIKIPEDIA: Un **algoritmo** è un procedimento che risolve un determinato problema attraverso un numero finito di passi elementari. Il termine deriva dalla trascrizione latina del nome del matematico persiano al-Khwarizmi, che è considerato uno dei primi autori ad aver fatto riferimento a questo concetto. L'algoritmo è un concetto fondamentale dell'informatica, anzitutto perché è alla base della nozione teorica di calcolabilità: un problema è calcolabile quando è risolvibile mediante un algoritmo. Inoltre, l'algoritmo è un concetto cardine anche della fase di programmazione dello sviluppo di un software: preso un problema da automatizzare, la programmazione costituisce essenzialmente la traduzione o codifica di un algoritmo per tale problema in programma, scritto in un certo linguaggio, che può essere quindi effettivamente eseguito da un calcolatore.

Più formalmente…

«una sequenza ordinata e finita di passi (operazioni o istruzioni) elementari che conduce a un ben determinato risultato in un tempo finito»



Quale algoritmo ?



Perché no un approccio algoritmico?

Per uno o più dei seguenti motivi:

- ✓ Impossibilità di formalizzare esattamente il problema
- ✓ Rumore e/o incertezza
- ✓ Alta complessità della soluzione
- ✓ Inefficienza della soluzione
- ✓ Mancanza di conoscenza ‘compilata’ sul problema da risolvere

Quando è importante l'apprendimento?

Quando il sistema deve:

- ✓ Adattarsi all'ambiente in cui opera (anche personalizzazione automatica)
- ✓ Migliorare le sue prestazioni rispetto ad un particolare compito
- ✓ Scoprire regolarità e nuova informazione (conoscenza) a partire da dati empirici

Dati Vs. Conoscenza

In ML, si studiano metodi per trasformare l'informazione empirica (presente nei dati) in nuova conoscenza

Grazie all'evoluzione dei computer e le reti, i dati oramai sono presenti ogni dove, e abbondanti!

- ✓ Scontrini di una catena di supermercati,
- ✓ Contenuto di pagine web,
- ✓ E-commerce
- ✓ Transazioni bancarie, ecc.

Assunzione fondamentale

Esiste un processo (stocastico) che spiega *i dati che osserviamo. Magari non ne conosciamo i dettagli, ma esiste!*

P.e. comportamento sociale non è puramente casuale

Lo scopo dell'apprendimento è costruire buone (o meglio, utili) approssimazioni di questo processo.

Il main goal del ML

L'obiettivo finale del ML è definire dei criteri di performance e ottimizzarli usando i dati o esperienza pregressa

I modelli saranno definiti su parametri che vogliamo apprendere ottimizzando un dato criterio. Usando dati e, eventualmente, conoscenza a priori sul dominio

- ✓ Modelli Predittivi (predizioni sul futuro)
- ✓ Modelli Descrittivi (ottenere nuova conoscenza)

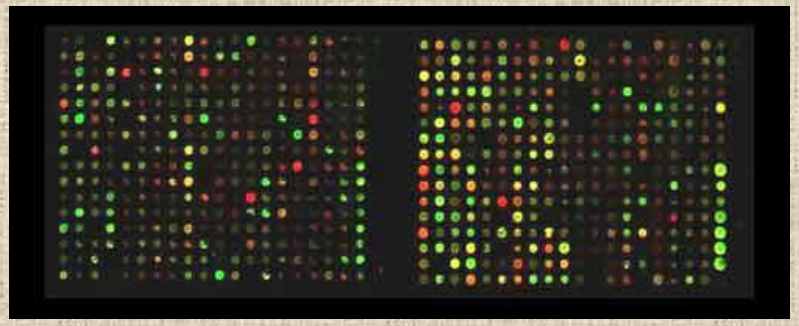
Problemi (Task) tipici di Apprendimento Automatico

- ✓ Supervised Learning
  - ✓ Classificazione Binaria
  - ✓ Classificazione Multiclasse
  - ✓ Regressione
  - ✓ Ranking di istanze e di classi
- ✓ Unsupervised Clustering
  - ✓ Novelty Detection
  - ✓ Clustering
  - ✓ Associazioni (Basket Analysis)
- ✓ Reinforcement Learning

Esempi di Applicazioni (1)

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE



Esempi di Applicazioni (2)

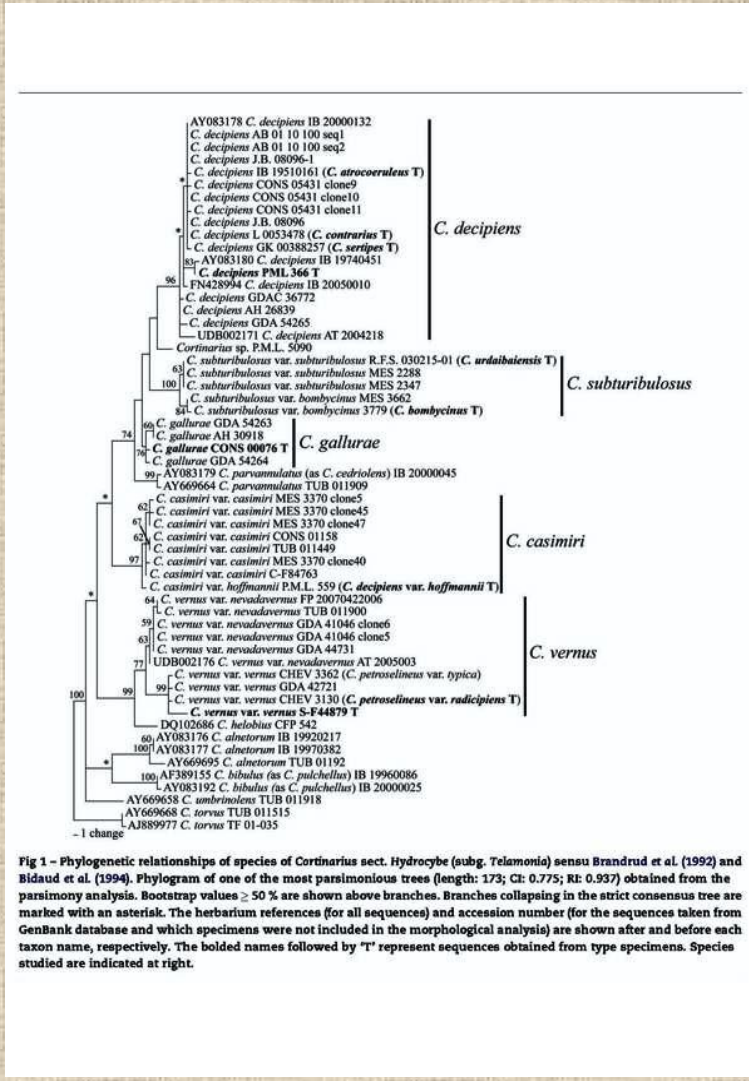
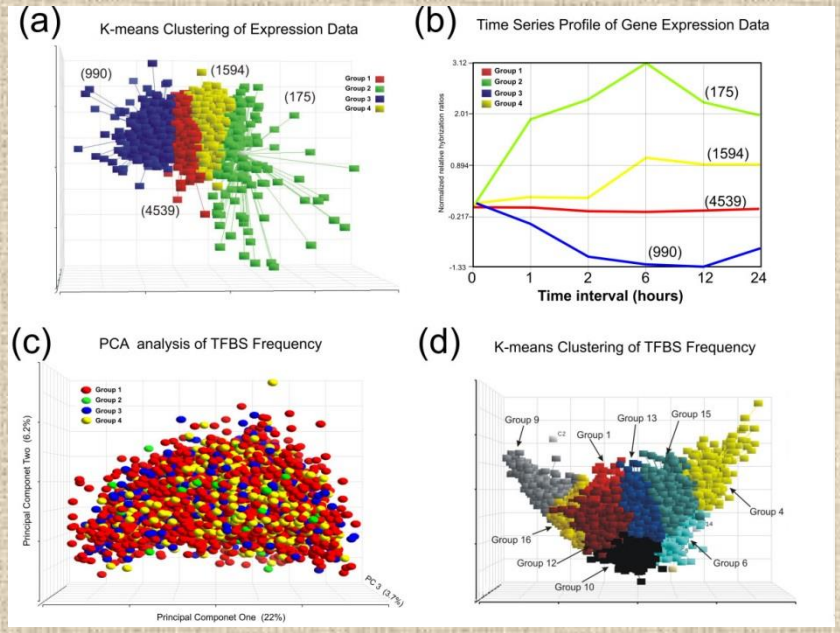
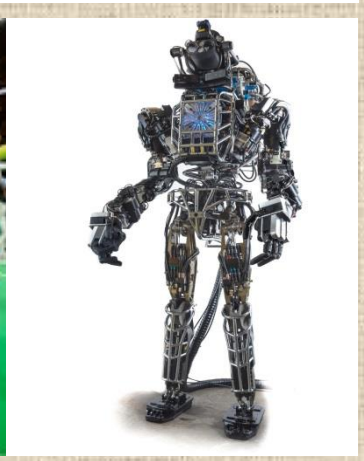
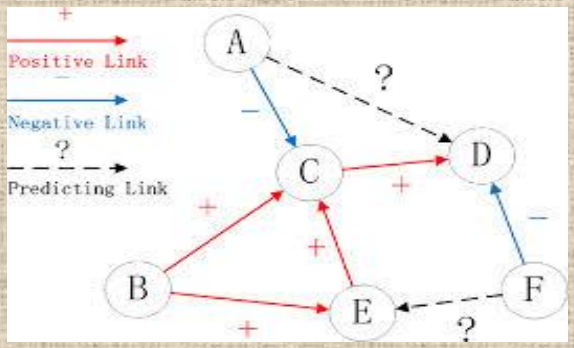


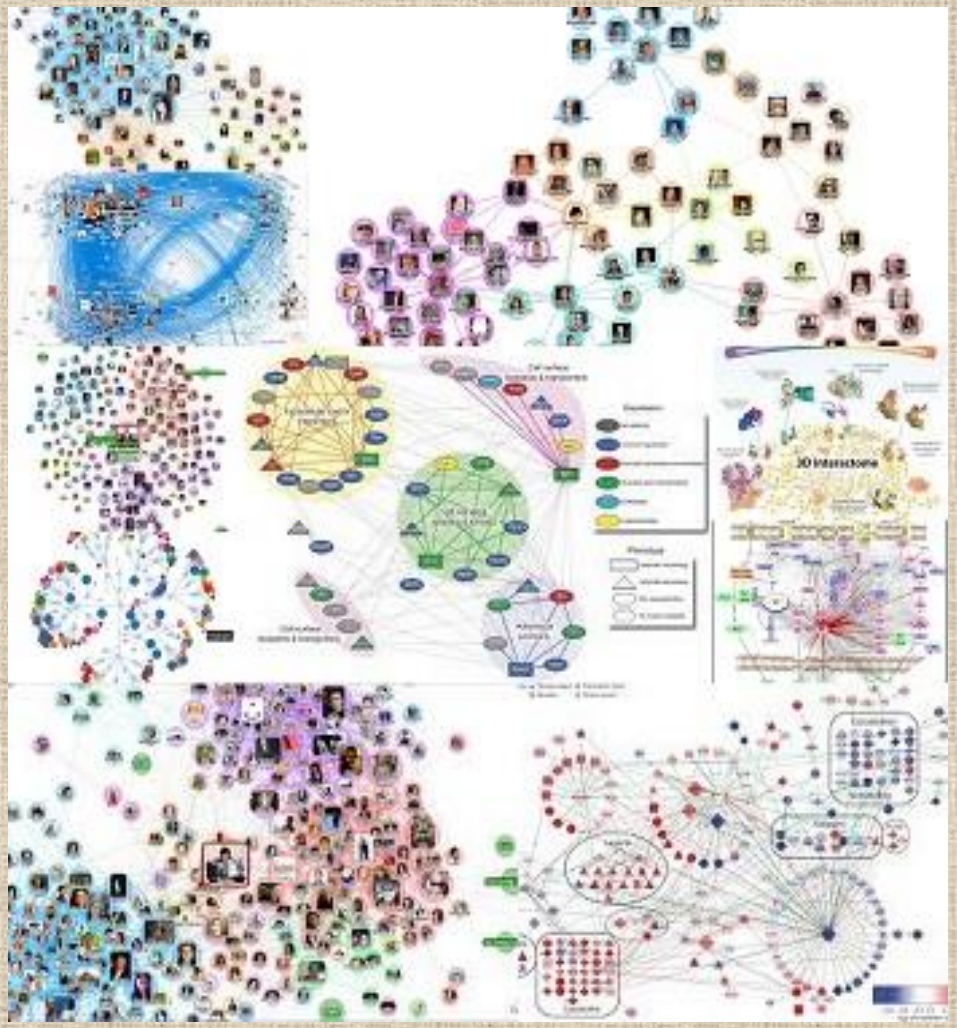
Fig 1 – Phylogenetic relationships of species of Cortinarius sect. *Hydrocybe* (subg. *Telamonia*) sensu Brandrud et al. (1992) and Bidaud et al. (1994). Phylogram of one of the most parsimonious trees (length: 173; CI: 0.775; RI: 0.937) obtained from the parsimony analysis. Bootstrap values ≥ 50 % are shown above branches. Branches collapsing in the strict consensus tree are marked with an asterisk. The herbarium references (for all sequences) and accession number (for the sequences taken from GenBank database and which specimens were not included in the morphological analysis) are shown after and before each taxon name, respectively. The bolded names followed by 'T' represent sequences obtained from type specimens. Species studied are indicated at right.





Esempi di Applicazioni (3)





Esempi di Applicazioni

- Riconoscimento di Facce
  - Controllo degli accessi da registrazioni video o fotografiche. Quali sono le caratteristiche veramente rilevanti di una faccia?
- Named Entity Recognition
  - Il problema di identificare entità in una frase: luoghi, titoli, nomi, azioni, ecc. Partendo da un insieme di documenti già marcati/taggati
- Classificazione di documenti
  - Decidere se una email è spam o meno, dare una classificazione ad un documento tra un insieme di topic (sport, politica,hobby, arti, ecc.) magari gerarchicamente organizzati

Esempi di Applicazioni

- Giochi e Profilazione Avversario
  - Per alcuni giochi ad informazione incompleta (giochi di carte, geister, risiko, …) vogliamo predire l'informazione mancante basandosi sulle strategie che l'avversario ha usato nel passato (minacce, reazioni, ecc.).
- Bioinformatica
  - I microarray sono dispositivi che rilevano l'espressione genica da un tessuto biologico. E' possibile a partire da questi determinare la probabilità che un paziente reagisca in modo positivo ad una certa terapia? …
- Speech Recognition, Handwritten Recognition, Social Network Analysis, e molto altro ancora.