# Laboratorio di Apprendimento Automatico

Fabio Aiolli

Università di Padova

# Esempi di Applicazioni

- Web page Ranking
  - Quali documenti sono rilevanti per una determinata query? Quali sorgenti di informazione utilizzare per determinare la rilevanza di una pagina?
- Recommendation
  - Amazon, Netflix: come vendere più prodotti? Quali prodotti o altri libri/film suggerire agli utenti basandosi sulle loro scelte precedenti o su quelle di altri utenti "simili"? Grafi e facebook.
- Traduzione Automatica
  - Comprensione del testo usando un insieme di regole di un bravo linguista computazionale o usare esempi di traduzione (documenti ONU e/o UE)?

# 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, …) potrebbe essere utile predire l'informazione mancante basandosi sulle strategie che l'avversario ha usato nel passato (minacce, reazioni, ecc.).
- Bioinformatica
  - I macroarray 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, Detection of Failure, e molto altro ancora.

# Problemi di Apprendimento Automatico

- Classificazione Binaria
- Classificazione Multiclasse
- Ranking di istanze e di classi
- Clustering
- Regressione
- Novelty Detection
- Link Prediction

# Pipeline
# Apprendimento Supervisionato

- Analisi del problema
- Raccolta, analisi e preprocessing dei dati
- Studio delle correlazioni tra variabili
- Feature Selection/Weighting/Normalization
- Scelta del predittore e Model Selection
- Test

# Oggetti

- **Vettori**
  - p.e. Valori di pressione del sangue, battito cardiaco, altezza peso di una persona, utili ad una società assicurativa per determinare la sua speranza di vita
- **Stringhe**
  - p.e. Le parole di un documento testuale in ENR, o la struttura del DNA
- **Insiemi** e **Bag**
  - p.e. L'insieme dei termini in un documento, o consideriamo anche la loro frequenza
- **Array Multidimensionali**
  - p.e. Immagini e Video
- **Alberi** e **Grafi**
  - p.e. Struttura di un documento XML, o di una molecola in chimica
- …
- **Strutture composte**
  - p.e. una pagina web può contenere immagini, testo, video, tabelle, ecc.

# Natura dei Dati

- **Feature categoriche o simboliche**
  - Nominali [Nessun ordine]
    - p.e. per un auto: paese di origine, fabbrica, anno di uscita in commercio,  colore, tipo, ecc.
  - Ordinali [Non preservano distanze]
    - p.e. gradi militari dell'esercito: soldato, caporale, sergente, maresciallo, tenente, capitano)
- **Feature quantitative o numeriche**
  - Intervalli [Enumerabili]
    - p.e. livello di apprezzamento di un prodotto da 0 a 10
  - Ratio [Reali]
    - p.e. il peso di una persona

# Mapping Feature Categoriche

- Le feature categoriche si possono mappare in un vettore con tante componenti quanti sono i possibili valori della variabile
- Possibili valori della variabili:
  - Marca: Fiat [c1], Toyota [c2], Ford[c3]
  - Colore: Bianco [c4], Nero [c5], Rosso [c6]
  - Tipo: Economica [c7], Sportiva [c8]
- (Toyota, Rossa, Economica)->[0,1,0, 0,0,1, 1,0]

# Ripasso di Algebra Lineare

- Nozione di vettore, lunghezza (norma)
- Prodotto scalare e relazione con il l'angolo tra i vettori
- Distanze tra vettori
  - Nota: $||\mathbf{x} - \mathbf{y}||^2 = ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2\mathbf{x}^\top \mathbf{y}$
  - Se i vettori hanno stessa norma la distanza è equivalente a similarità indotta dal prodotto scalare
  - ovvero: $||\mathbf{x} - \mathbf{y}||^2 = \text{const} - 2\mathbf{x}^\top \mathbf{y}$
  - Altrimenti anche la lunghezza dei due vettori conta, non solo l'angolo!
- Similarità coseno e normalizzazione

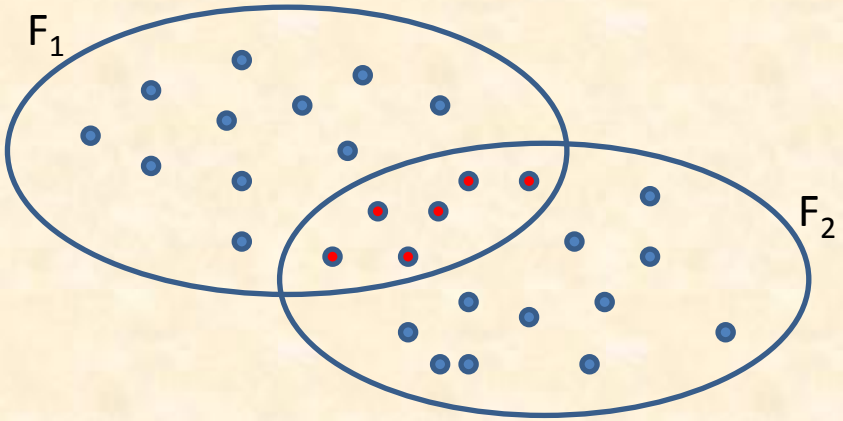
# Analisi delle correlazioni

- Per ogni feature (att:val)
  - Frequenza (IN), ovvero il numero di volte che compare un determinato valore per un determinato attributo
- Per ogni coppia (att1:val1) (att2:val2)
  - Frequenza congiunzione (AND), ovvero quante volte due coppie attributo-valore compaiono insieme negli esempi
  - Frequenza disgiunzione (OR), ovvero quante volte almeno una coppia attributo-valore compare negli esempi

# Misure di Correlazione

| | Caso Simbolico (insiemi) | Caso Generale |
|---|---|---|
| Prodotto Scalare | $\|a \cap b\|$ | $a^\top b$ |
| Dice | $\frac{2\|a \cap b\|}{\|a\|+\|b\|}$ | $\frac{2a^\top b}{\|\|a\|\|^2+\|\|b\|\|^2}$ |
| Jaccard | $\frac{\|a \cap b\|}{\|a \cup b\|}$ | $\frac{a^\top b}{\|\|a\|\|^2+\|\|b\|\|^2-a^\top b}$ |
| Overlap Coef. | $\frac{\|a \cap b\|}{min(\|a\|,\|b\|)}$ | $\frac{a^\top b}{min(\|\|a\|\|^2,\|\|b\|\|^2}$ |
| Coseno Asimmetrico | $\frac{\|a \cap b\|}{\|a\|^\alpha \|b\|^{(1-\alpha)}}$ | $\frac{a^\top b}{\|\|a\|\|^{2\alpha} \|\|b\|\|^{2(1-\alpha)}}$ |

# Correlazione tra features categoriche




$F_1$ = insieme di esempi che hanno una certa coppia (att1:val1)
$F_2$ = insieme di esempi che hanno una certa coppia (att2:val2)

$|F_1 \cap F_2|$ Intersezione

$\frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$ Jaccard

$\frac{2|F_1 \cap F_2|}{|F_1|+|F_2|}$ Dice

$\frac{|F_1 \cap F_2|}{min(|F_1|,|F_2|)}$ Overlap Coefficient

# Media e Deviazione Standard (Normalizzazione)

- E molto importante che gli esempi e le features siano ‘comparabili’ tra di loro

- Centramento (centering) degli esempi/features
- Normalizzazione STD

$$\hat{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad \sigma(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{N}}$$