Laboratorio di Apprendimento Automatico

Fabio Aiolli
aiolli@math.unipd.it
Università di Padova

Informazioni

- Aula (16 ore) e Laboratorio (16 ore)
- Sempre il giovedi dalle 15:30 alle 17:00, cercheremo di alternare tra aula e laboratorio

- Contenuti:
  - Panoramica di applicazioni di ML
  - Formalizzazione di problemi ML
  - Casi studio e loro soluzioni

Esempi di Applicazioni

- Web page Ranking
  - Quali documenti hanno match con una determinata query? Quali sorgenti di informazione utilizzare per determinare la rilevanza di una pagina?
- Collaborative Filtering
  - 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"?
- Traduzione Automatica
  - Comprensione del testo usando un insieme di regole di un bravo linguista computazionale o usare esempi di traduzione quali documenti dell'ONU o della EU?

Esempi di Applicazioni

- Riconoscimento di Facce
  - Controllo degli accessi da registrazione video o fotografica. Quali sono le caratteristiche veramente rilevanti di una faccia?
- Named Entity Recognition (NER)
  - Il problema di identificare entità in una frase: posti, 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, …) predire l'informazione mancante basandosi sulle strategie che l'avversario ha usato nel passato.
- 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

Oggetti

- Vettori
  - p.e. Valori di pressione del sangue, battito cardiaco, altezza peso di una persona, 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 NER, o nella 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.

Rappresentazione 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 preserva 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 []
    - p.e. il peso di una persona

Ripasso di concetti base di Algebra Lineare

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

Esercizio

- Le tre opzioni seguenti ritornano lo stesso valore
  - Coseno di due vettori generici $\mathbf{x}_1$ and $\mathbf{x}_2$
  - Coseno di due vettori $v(\mathbf{x}_1)$ e $v(\mathbf{x}_2)$, dove $v(\mathbf{x})$ è un vettore di lunghezza normalizzata of $\mathbf{x}$.
  - Il prodotto scalare di vectori $v(\mathbf{x}_1)$ and $v(\mathbf{x}_2)$ normalizzati