# Apprendimento Automatico

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

Sito web del corso
www.math.unipd.it/~aiolli/corsi/1617/aa/aa.html

# 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]

# Mapping Feature Continue

- In questo caso è decisamente più difficoltoso trovare un buon mapping
- Tipicamente, le feature vengono trasformate per ottenere valori ‘comparabili’ con le altre feature
  - Feature Centering
  - Feature Standardization

# Media e Deviazione Standard (Normalizzazione)

- E molto importante che gli esempi e le feature siano 'comparabili' tra di loro

- Centramento (centering) degli esempi/feature
  - $f(x) = x - \hat{x}$
- Normalizzazione STD
  - $f(x) = \frac{(x-\hat{x})}{\sigma(x)}$
- Oppure rescaling
  - $f(x) = \frac{(x-\hat{x}_{min})}{(\hat{x}_{max}-\hat{x}_{min})}$

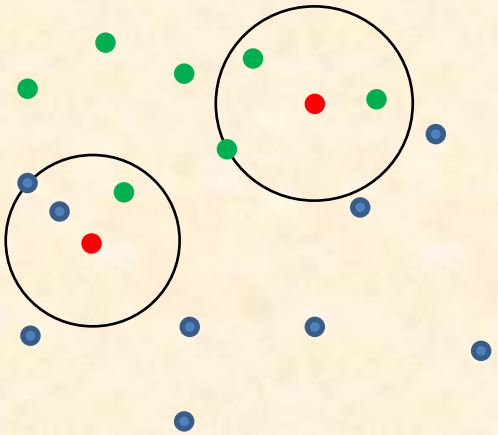
$$\hat{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad \sigma(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\hat{x})^2}{N}}$$

# Similarità e distanza

- 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 alla 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

# Algoritmo k-nn

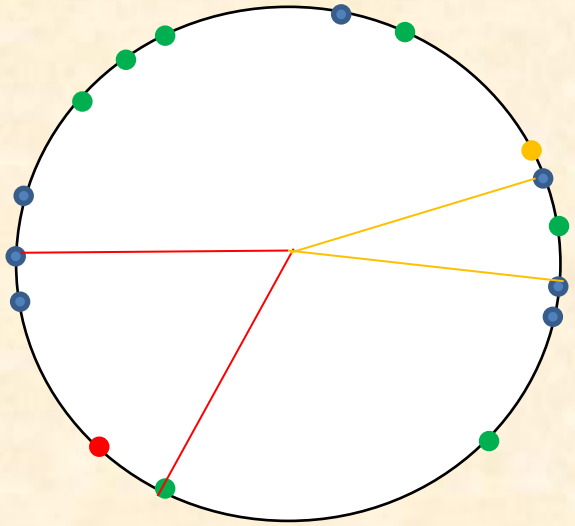
K-Nearest Neighbors è un algoritmo di classificazione in cui un esempio di test è classificato come la classe di maggioranza dei suoi k-vicini nel training set



Esempio 3k-NN in 2d usando la distanza Euclidea

# Algoritmo k-nn (normal)

K-Nearest Neighbors quando gli esempi stanno tutti in una palla di raggio unitario. La distanza diventa equivalente al prodotto scalare



Esempio 3k-NN in 2d con norma esempi unitaria

# Risorse

- In sklearn sono presenti molte funzionalità per il preprocessing
- http://scikit-learn.org/stable/modules/preprocessing.html

- Tra cui:
  - Standardization
  - Mean Removal
  - Variance Scaling
  - Scaling features to a range
  - Normalization
  - Binarization
  - Encoding categorical features
  - Imputation of missing values
  - …

# Scelta iper-parametri

- Come possiamo scegliere il numero di unità nascoste in una rete neurale?
- E il valore K in K-NN?
- O magari diverse misure di Guadagno per gli alberi di decisione? …

- La Model Selection è la fase di una pipeline di apprendimento supervisionato dove si vanno a individuare gli iper-parametri che stimiamo essere i migliori per il task