Lezione 2

Corso	Riconoscimento e recupero dell'informazione per Bioinformatica
Data	@October 11, 2021 1:30 PM
Rifacimento	☑
Status	Completed
Tipo	Lezione

Rappresentazione

Cosa vuol dire fare rappresentazione = Voglio rappresentare gli oggetti in modo digitale, voglio una rappresentazione che il calcolatore possa capire.

Per fare rappresentazione devo prima fare campionamento.

Campionamento

Trovare una rappresentazione digitale dell'oggetto, raccogliere i dati.

Dato grezzo: il dato che esce direttamente dal sensore. Questo è già una rappresentazione, ma ovviamente migliorabile.

> *Esempio*: costruire un modello di determinati pesci.

> *Esempio*: supponendo di voler fare filogenesi, per rappresentazione intendo tirare fuori una sequenza (un file di testo) che mi indichino i componenti della sequenza. Da organismo → a sequenza: il campionamento consiste nel sequenziare.

> *Esempio*: classificazione del volto (3 diverse tipologie di sensori: fotocamer atradizionale, 3D e telecamera a infrarossi)

In entrambi i casi però l'obiettivo è: partire da un oggetto e arrivare ad una rappresentazione digitale.

Devo tener conto anche del fatto che l'informazione può essere o meno ***interessante*** (es. fotocamera a infrarosso per il riconoscimento dei volti).

Problemi da tenere in considerazione:

- **Frequenza di campionamento**

Il campionamento dipende strettamente dalla problematica (e tipicamente è deciso dall''esperto).

Il **problema vero e proprio**: possiamo avere rumore o info di scarto per le informazioni che ci arrivano dai dati grezzi.

Soluzione:
Elaborare i dati che vengono dal sensore. Estraggo delle feature e costruisco il pattern.

La **feature** è una misura che io faccio, dai dati che derivo. Il **pattern** è l'insieme di features.

Per risolvere un problema posso avere opzioni diverse, con pro e contro. Ogni volta dipende dalle cose di cui ho bisogno, capire qual è l'opzione corretta da applicare.

> Classificazioni più ricche possono essere più difficili da estrarre e da modellare.
>
> *esempio*: rappresentazione compatta per distinguere i pesci vs modellamento della forma dei pesci (non più uno spazio vettoriale)

Classificatore più semplice

Nearest Mean Classifier

Classifica in base alla media più vicina

La scelta della feature è chiaramente cruciale (deve essere rilevante, discriminante, misurabile)

Feature

È una misura che io effettuo e può essere di diversi tipi (discreta, continua, valori binari, valori nominali).

Costruzione del pattern

Mettere assieme le varie features.

Esempio: è diverso avere $[h, l]$ o $[l, h]$ o ancora $[h * l]$. **Il pattern è come metto assieme varie features.**

Tipi di Pattern

- **Vettori - dati vettoriali**

  Abbiamo un insieme prefissato di features, messe in ordine, fondamentale ma arbitrario.

  L'oggetto viene rappresentato come un punto in no spazio d-dimensionale, detto "spazio delle features", dove d è il numero delle features (con 2 lo posso vedere, con 3 lo posso immaginare,...)

  La scelta è cruciale e bisogna scegliere adeguatamente.

  Usando molte features, gli spazi diventano molto grandi e possiamo avere problemi: **Curse of dimensionality**

  Più sono le misure che faccio, migliore è la mia capacità di riconoscere l'oggetto: per il calcolatore è un po' diverso. Aggiungendo delle features il mio problema si incasina! ho problemi computazionali ma non solo:

  Ad *esempio*: usando 100 features, userei 100 oggetti per stimare 100 valori. Comincio ad avere dei problemi nella stima.

- **Sequenze**

  Si presentano in forma ordinata e sequenziale (uno dopo l'altro): l'ordine è fondamentale e non è arbitrario.

  Ad *esempio*: sequenze temporali, l'evoluzione e l'ordine sono dati dal tempo, oppure sequenze non temporali: nucleotidiche e amminoacidiche (ordine fondamentale ma non dato dal tempo)

- **I grafi**

  Rappresentano un insieme di nodi collegati da archi.

  Esempio: modellare le diverse parti di un corpo umano.

  Esempio: protein-protein interaction, oppure percorsi metabolici

- **Gli insiemi**

  Pattern estremamente non strutturato, una collezione non ordinata di dati a cardinalità variabile

Vettori vs altri pattern

Tipicamente pattern compressi sono più espressivi, ma molte tecniche si applicano meglio ai vettori

Preprocessing

- Concetto di scala
- Data standardization
- Data trasformation
  - riduzione della dimensionalità

Scala

Significatività relativa dei numeri.

Riconoscere la scala è fondamentale anche per capire la relazione tra due pattern.

Esempio: **distanza euclidea**

Soffre se le features sono a scala diversa: problema infatti è che spesso le variabili che descrivono un oggetto non sono nella sttessa scala

La distanza euclidea viene usata in tantissimi algoritmi. In che modo soffre?

Con 3 punti in uno spazio bidimensionale ad esempio:

Scala x : $[0-2]$

Scala y : $[0-20]$

A : $[1, 10]$

B : $[1, 12]$

C : $[2, 10]$

Calcolando la distanza euclidea:

$d(A, B) = 2$

$d(A, C) = 1$

A è più simile a C, ma sul grafico A è più simile a B quindi la risposta non è vera: A e C sono più lontani di A e B. Qual è il problema? **La differenza in una direzione sulle scale usate nel grafico sono troppo differenti.**

Come risolvere il problema della scala?

> **Data Standardization**
>
> Produce dati senza dimensionalità, facendo in modo che tutta la conoscenza su scala e locazione dei dati venga persa dopo la standardizzazione

Approcci di Data Standardization

Notazioni:

- **Dataset** X

$$X = \begin{bmatrix} x_{1,1} & x_{...} & x_{1,n} \\ x_{...,1} & x_{...} & x_{...} \\ x_{d,1} & x_{...} & x_{d,n} \end{bmatrix}$$

n punti in uno spazio d-dimensionale

$$A = \begin{bmatrix} 1 & 2 \end{bmatrix}$$
$$B = \begin{bmatrix} 4 & 5 \end{bmatrix}$$
$$C = \begin{bmatrix} 7 & 8 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \end{bmatrix}$$

- **Media** lungo la direzione j (j ccva da 1 a n)

$$\overline{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ji}$$

- **Range**: massima estensione delle feature

$$R_j = max_{x_{ji}} - min_{x_{ji}}$$

- **Deviazione standard**: dispersione rispetto alla media

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{ji} - \overline{x}_j)^2}$$