Slide 1

Cosa si intende per ***Data Science***?

È un ambito di ricerca che cerca di utilizzare i dati per dare delle risposte a diversi tipi di problemi che ricadono sotto il nome di problemi di tipo scientifico, i quali riescono dare risposta a fenomeni di diversa natura.

Si pone attenzione allo svilluppo dei dati.

Nell’ambito della ***data science*** rientrano un insieme di discipline che cercano di estrarre, interpretare, visualizzare e capire le motivazioni per cui si verificano certi fenomeni.

Slide 2

Tra le tecniche più utilizzate ci sono qulle realtive al ***machine learning***, per poter classificare dati…

• Nell’ambito delle tecniche di *machine learning* ci sono ad esempio tecniche basate su *clustering*.

CLUSTER → individuazione di comunità di nodi o individui che hanno caratteristiche comuni, ad esempio in un social consiste a trovare persone che sono tra di loro amici o che condividono gli stessi interessi.

• ***Supervisioned learning*** → relazioni che intercorrono con altri individui.

• Si ha a che fare con dati di grandi dimensioni → **BIG DATA**

• Avendo a che fare con tantissimi dati ritorna molto utile la *Data Visualisation*, che assume rilevanza per riuscire a comprendere meglio i fenomeni.

Legati alle tecniche di *machine learning* che definiscono in modo preciso l’andamento di certi fenomeni, c’è la tecnica di rappresentazione visuale che rende meglio la comprensione del fenomeno.

• AI EXPLAINBILITY: un altro problema grosso legato alla possibilità di capire il motivo per cui un certo fenomeno si verifica, è legato al mondo del *machine learning*, per dare una “*explainability*” per cui un fenomeno si è verificato.

Slide 3

Come rendere percebile l’informazione che è contenuta all’interno di un grafo?

Con grafo si intende una struttura che ha dei vertici (rappresentati come cerchi/punti rappresentati su un piano) e gli archi che sono le rappresentazini di relazioni/legami presenti tra i nodi.

Si parla di grafi sparsi quando il numero di archi che collegano i nodi sono ridotti e si parla di grafi densi quando i coollegamenti tra i nodi sono tanti.

Poi c’è un particolare tipo di grafo che è l’albero, all’interno del quale, per definizione, non si possono presentare dei cicli.

La struttura matematica del grafo si presta molto bene a diverse tipi di rappresentazioni di fenomeni del mondo reale:

Immagine che contiene albero, esterni, posando

Descrizione generata automaticamenteOgni nodo è una persona e l’arco indica la relazione che c’è tra gli individui.

I legami possono essere diversi, magari un legame indica che un certo individuo è amico di un altro, un altro legame indica che la persona è sposata con l’altra…

Ci sono quindi diverse relazioni che possono collegare i vertici del grafo.

In questo caso si parla di grafi etichettati per cui si aggiungono delle etichette che forniscono sia informazioni sul tipo di legame che collega gli individui, sia sul tipo di individui che si possono avere.

Slide 4

Ai vari individui, ai vari vertici, ai vari legami si possono associare delle proprietà.

In questi casi si parla di grafi con le proprietà: ogni nodo può avere delle caratteristiche – non ci sono limitazioni sulle caratteristiche associabili ai nodi.

Solitamente si parla di associazioni chiave-valore, dove il valore può essere una stringa o un tipo categoriale, come ad esempio il tipo di lavoro (si tratta quindi di una scelta finita tra un insieme di dati).

I valori categoriali possono essere stringhe o numerici.

Immagine che contiene testo, nero

Descrizione generata automaticamenteQuesto tipolaogia di grafi è detta grafi con proprietà e servono tecniche di visualizzazione del grafo che permettano di mettere in luce sia la topologia del grafo, cioè com’è strutturato il grafo, sia quali sono gli attributi che sono stati associatti ai nodi e agli archi del grafo.

Slide 5

Perché è utile studiare i grafi con proprietà?

Oggigiorno sta diventando sempre più utilizzato nell’ambito dei social network, ad esempio molti sistemi come Twitter, Facebook..si basano sull’utilizzo di strutture come i grafi con proprietà per analizzare gli iscritti e le relazioni intercorrenti tra gli utenti.

Oltre a sapere le relazioni, per aziende come Facebook, è anche interessante analizzare cosa acquistano gli utenti per proporre determinate offerte a date categorie di utenti.

Slide 6

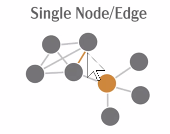
Si cerca di individuare delle visualizzazioni che rendono chiara la relazione che c’è tra i nodi.

Ci sono poi una serie di tecniche che considerano anche gli attributi che sono specificati.

Considerare anche gli attributi aggiunge un livello di difficoltà nel definire una visualizzazione perché questi attributi occupano (spazio) e anche perché si vogliono fare delle visualizzazioni che mettano bene in evidenza la presenza di quel dato attributo; proprio per interperatre meglio quel particolare dominio.

Slide 7

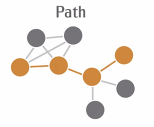
Quando si parla di tecniche di visualizzazione di un grafo ci sono diversi approcci alla visualizzazione che dipende dal tipo di task che si vuole fare sul grafo, come ad esempio identificare nodi o archi che hanno date caratteristiche.

Ad esempio:

voglio trovare un dato individuo che abita a Milano oppure le persone che sono tra loro sposate.

Si può essere ineteressati anche a operazioni di vicinato, ad esempio tutte le persone con cui si collabora:

si può avere una distanza 1 che indica quindi che i nodi direttamente connessi oppure più distanti.

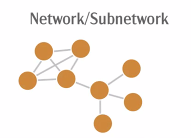


Si può poi essere inetressati al cammino che unisce due individui:

ad esempio c’è una distanza massimo 4 tra tutti gli individui del globo.

Dati due nodi si vuole sempre trovare il cammino più breve che li collega.

Si potrebbero voler trovare anche dei *cluster*, ad esempio trovare le persone che sono maggiormente amiche.

Con un grafo si possono anche rappresentare reti e sottoreti.

A seconda del tipo di task, del tipo di operazione, si definiscono delle operazioni differenti.

Slide 8

Le tecniche di visualizzazioni proposte dipendono da diversi fattori:

• la tipologia del grafo (se è un albero, un DAG [*Directional Acyclic Graph*] oppure è un grafo partizionato in livelli, oppure un grafo partizionati in k sottoparti…;

• la dimensione della rete → se voglio disegnare un grafo che ha un enorme quantità di nodi, la prima cosa che si vede è una nuvola di punti neri e non si capisce molto, non è chiaro.

• il tipo di operazione/task che si deve andare ad effettuare sul grafo.

Slide 9

• La prima tipologia di approcci è la *node link layouts*: questi hanno lo scopo di disegnare i vertici e gli archi del grafo. Quindi il principio di questa visualizzazione è quella di disegnare la base del grafo.

[ Poi c’è un altro tipo di approccio basato sulle tabelle ed un altro tipo ancora basato sulle forze. ]

L’idea è di migliorare la visualizzazione del grafo per evitare che ci siano nodi o archi che si sovrappongano e non rendono bene, non fanno appprezzare bene, la visualizzazione del grafo.

• Oltre queste tecniche “*driven*”, ci sono anche tecniche che considerano gli attributi e si disegna il grafo per mettere in evidenza gli attributi assunti dai nodi.

Si crea un partizionamento del piano in cui si possono avere delle categorie ed i nodi vengono messi in un apposita categoria. Gli archi vanno poi a collegarli.

*Attribute driven layouts*

Un’altra tipologia di visualizzazione è detta *Attribute-Driven Positioning*: si prende un grafo per cui si conoscono le coordinate geografiche e questo grafo può quindi essere visualizzato sul globo.

Quindi l’attributo dipende dalla posizione del nodo.

A differenza del *topology driven layout*, gli *attribute driven layout*/le visualizzazioni dipendono dal valore assunto dagli attributi.

Questa tipologia di grafo diventa poco scalabile nel momento in cui aumenta il numero di punti.

• Si ha anche un’altra tipologia di visualizzazione basata sulle matrici (*Tabular Layouts*) in cui si vanno a mettere i nodi sulle righe e sulle colonne.

• *Adjiaceny matrix*: l’incrocio tra righe e colonne vuol dire che esiste un legame tra quei due nodi.

Permette di visualizzare meglio le proprietà che ci sono sui vertici e sugli archi.

Questa visualizzazione non è ideale nel momento in cui si vogliono definire dei cammini, ma lo è quando si vogliono definire dei *cluster* di punti. → A partire dal nodo si può riuscire a capire la presenta di *cluster*.

• *Quits* ha l’idea di rappresentare le relazioni che esistono tra i nodi.

• *BioFabric* in cui si aggiungono dei collegamenti verticali nel momento in cui si ha lageme tra i due nodi:

Il limite è relativo alla dimensione del grafo, in quanto di questo bisogna visualizzzare tutto.

Ha però il vantaggio che utilizzando colori diversi si riesce a capire dove ci sono dei *cluster*.

• *Implicit Layouts*

Si possono applicare solo a strutture come alberi.

I figli sono gli elementi al primo livello.

Questo tipo di rappresentazione permette di mettere in luce la struttura dell’albero.

Sulla base di queste possibili visualizzazioni del grafo, è possibile anche combinare le varie tipologie.

Ad esempio si può mettere insieme la struttura topologica del grafo (*topology driven layout*) con gli attributi (*attribute driven layout*).

• *Integreted visualisation* → le proprietà dei nodi e degli archi dipendono da ciò che si vede nella tipologia.

• *Overloaded* → fa riferimento al fatto che se ad esempio si hanno dei *cluster* di dati, si può identificare meglio dove questi si trovano.

Le visualizzaizoni rimanenti:

*smal multiple* → si prende un frammento del grafo e lo si riproduce.

Sulle varie rappresentazioni che si definiscono, si possono voler definire delle operazioni, come unire i nodi e compattare le visualizzazioni; si può voler derivare nuovi attributi che si possono estrarre da alcune proprietà del grafo.

Il motivo per cui fare questo è che spesso è molto più semplice visualizzare i nodi che le proprietà.

Questa slide è un overview delle teniche di visualizzazione poposte.

Slide 10

È una rappresentazione di tipo *node link layout* in cui il grafo viene rappresentato sul piano.

Si immerge il grafo all’interno di un piano, per farlo ci sono diverse tecniche, come la tecnica delle forze.

Slide 11

Possono essere forze attrattive o repulsive.

I nodi sono collegati da “molle” che possono allargarsi o restringersi in modo da creare una situazione di equilibrio con le forze.

Slide 12

Forze attrattive: simili a quelle delle molle e servono per attrarre gli estremi di due nodi.

Forze di tipo repulsivo: come quelle delle particelle caricate elettricamente e si basano sulla legge di *Coulomb*.

Come si fa a definire la posizione nel piano di questi nodi, tenendo conto di queste forze?

L’ideale è stabilire uno stato di equilibrio in modo che c’è anche equilibrio non solo tra le particelle ma anche tra le forze presenti.

Si vuole riuscire, per ogni coppia di nodi tra loro vicini (esiste un arco che li collega), a creare una situazione di equilibrio in cui tutte le forze sono equilibrate.

Layout basati sulle forze devono essere spesso ricalcolati: sono tanto più precisi quanto più si itera il procedimento, in modo da arrivare al punto di equilibrio appunto.

Slide 13

Per riuscire a calcolare il punto di equilibrio c’è bisogno di algoritmi che hanno complessità **n** al cubo.

Si considerano i soli archi presenti tra i nodi.

A volte queste tecniche richiedono diverse iterazioni (come detto prima) per ottenere una rappresentazione ottimale e che non vada a ricadere nel ritrovamento dei minimi locali.

Slide 14

Come si rappresentano gli attributi?

Si utilizzano i colori per differenziare le proprietà dei nodi.

Intensità del colore per indicare la maggior importanza o meno delle proprietà.

Slide 15

Ci sono un enorme quantità di teniche per dire se la qualità della visualizzazione del grafo è buona.

Slide 16

Ci sono tante tenciche e librerie sviluppate per la visualizzazione dei grafi.

Slide 18

Con questa visualizzazione si vede abbastanza la presenza di gruppi/*cluster* che sono tra loro vicini e questo avviene sfruttando la presenza del colore che dà anche l’intensità del legame che lega i due nodi del grafo.

Slide 19

Lo spessore della linea indica l’intensità del legame tra i due individui.

Esprime la quantità di bevande che gli individui bevono nello stesso giorno.

Slide 20

La rappresentazione del grafo viene fatta tenendo conto del tipo di bevanda che ogni individuo preferisce.

Slide 21

Si tiene conto delle coordinate geografiche di ogni nodo.

Slide 22

Ci possono essere *cluster* a livello di nodi e a livello di archi; si possono definire delle visualizzazioni ad un livello di rsappresentazione più elevato e mettono in evidenza la presenza di relazioni tra i nodi.

Slide 30

L’algoritmo di Louvain è un algoritmo di clustering piuttosto veloce.