Costruzione di un metodo per la stima delle interazioni in un data set

Pastori Simone

Abstract

In questo report propongo un metodo per l'analisi di una serie di eventi volto a stimare l'interdipendenza o meno di tali dati, e, in caso negativo, a valutare se la dipendenza segua un modello di tipo attrattivo o repulsivo. Nel fare ció introduco metodi per la creazione di dati che seguano tali modelli. Infine applico a titolo d'esempio questa metodologia a diversi data set.

Introduzione: modellazione del comportamento degli eventi

Considerati N eventi $\{x_i\}$ si vuole verificare la tipologia dell'interazione tra di essi. Nel caso di indipendenza reciproca, gli eventi avranno un'uguale probabilitá di verificarsi, quindi i punti si disporranno randomicamente. Nel caso di dipendenza, invece, la posizione dei dati sará in qualche modo determinata dalla posizione dei precedenti.

Voglio quindi creare una procedura poter determinare se l'influenza dei punti precedenti porti a un'attrazione degli eventi successivi, o a una repulsione. Per meglio visualizzare questa differenza, ho implementato due modelli che simulino data set conformi a tali comportamenti: (Il risultato é presentato in Figura 1)

- In un'interazione repulsiva perfetta, la distanza tra un evento e il successivo é costante. Piú spesso, l'interazione tra gli eventi puó presentare dipendenze da altri fattori rispetto alla posizione dei dati come incertezze di misura o rumore composto di dati indipendenti. Ho pertanto supposto che, in un set di eventi sottostante queste ipotesi, le distanze reciproche seguissero un andamento conforme a una gaussiana centrata nel "valore vero" che risulterebbe dalla loro misura nel caso in cui la dipendenza fosse perfetta. Ho, quindi, estratto le distanze tra i dati casualmente secondo questa funzione, sommandole reciprocamente per ottenere il set di dati. Aumentando la dispersione della distribuzione gaussiana, posso ritrovare il caso di indipendenza. Posso inoltre simulare il rumore nei dati generandone una parte in modo random uniforme.

- In un' interazione attrattiva, la probabilitá di un evento di verificarsi é maggiore nell'intorno degli eventi precedenti. Mi é sembrato pertanto naturale generare questi dati estraendoli casualmente con una funzione di probabilitá avente un massimo nell'intervallo in cui si verifica un cluster. Ho quindi simulato un set di dati generati randomicamente con densitá di probabilitá gaussiana per ogni cluster, che ho poi unito per creare l'intervallo di eventi. Come prima, aumentando la dispersione della probabilitá si ritrova il caso di eventi indipendenti. Il rumore puó essere, ancora una volta, simulato aggiungendo agli eventi un data set minore di eventi generati in modo random uniforme.

Ovviamente, il modello nullo, cioé il modello di non interazione, é semplicemente rappresentabile con un insieme di eventi estratti casualmente con probabilitá uniforme

Definizione della procedura

Posso ora definire la procedura per discriminare il comportamento di un data set di eventi. Il metodo che ho costruito si suddivide in due parti:

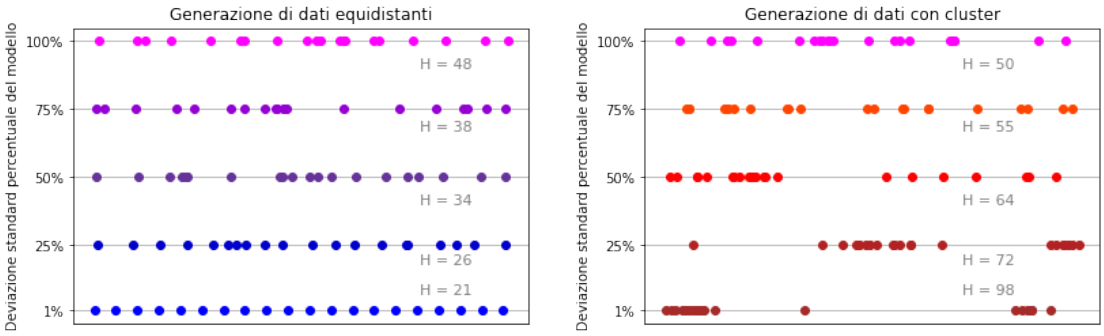

Figura 1 Dati generati con i due modelli di interazione. A destra il caso di clustering e a sinistra il caso di dati equidistanti. All'aumentare della deviazione corrisponde anche un aumento del rumore. Per valori massimi di deviazione i modelli vengono ricondotti al modello nullo (indicato in magenta). H é il valore della Hopkins Statistic effettuata sui set dai quali questi dati sono stati estratti

A. Determinazione dell'indipendenza tra i dati

Nella prima parte del test si verifica se il set di eventi sia o meno compatibile con il modello nullo. Per costruirlo, mi sono basato sull'ipotesi che i dati indipendenti seguano un processo di Poisson. Pertanto ho supposto, avallato dall'osservazione dei data set da me generati, che, campionando le distanze reciproche δx tra gli eventi, le loro frequenze seguissero un'andamento di tipo esponenziale negativo.

Eseguendo quindi un fit tra le frequenze campionate e una funzione esponenziale definita come

$$a^{-bx} \times c + d$$

(dove a, b, c, e d sono parametri ottenibili tramite il fit) posso verificare, tramite test di χ^2, quanto i dati possano essere interpolati da tale funzione. Se il p value cosí ottenuto é superiore a 0.05, l'interpolazione é ragionevolmente confermata, e quindi anche l'ipotesi di indipendenza.

Per eseguire il test é necessario scegliere il numero di campionamenti da svolgere sulle δx. Dato che all'aumentare della quantitá N di dati da analizzare aumenta il rischio di sottocampionamento, ho scelto un numero di bin dipendente da N. In particolare, (facendo riferimento a [1] in appendice, e a seguito di numerose prove per massimizzare la probabilitá sul modello nullo) , ho scelto una dipendenza lineare da $N^{\frac{2}{5}}$.

Punti deboli di questo tipo di test sono la tendenza a essere molto selettivo, e il peggioramento della precisione al variare della quantitá di dati analizzati, come mostrato in seguito. É inlotre basato sull'ipotesi che l'indipendenza sia condizione necessaria e sufficiente per ottenere un procedimento di Poisson, che, seppur applicabile in molti casi, non rappresenta la realtá

B. Determinazione del tipo di interazione tra i dati

Una volta ottenuto un risultato negativo (di non indipendenza) nella prima fase, occorre determinare la tipologia di interazione. Per questa seconda parte ho deciso di adottare un calcolo della Hopkins Statistic. Il valore H cosí computato dovrebbe segnalare una tendenza di clustering nei dati all'avvicinarsi a 1, e una tendenza all'equidistanza per risultati vicini a 0.

Questa statistica puó inoltre fornire una conferma del risultato positivo nella prima parte della procedura, poiché un valore di H vicino a 0.5 implica indipendenza nei dati considerati.

Ció nonostante, questa puó essere solo una conferma di quanto ottenuto con il test di χ^2, in quanto sono entrambi basati sullo stesso modello nullo.

Una limitazione é data dal fatto che questo test considera un campione di un decimo dei dati forniti, rendendo di conseguenza inconclusivi i test su ordine 1 dati.

Applicazione della procedura sui modelli, visualizzazione della dipendenza del risultato dalla quantitá di eventi *N*

Avendo definito la procedura, posso osservare il risultato sui modelli di indipendenza, interazione attrattiva, e interzione repulsiva precedentemente implementati. Posso inoltre visualizzare l'accuratezza della procedura al variare della dimensione dei dati forniti.

Dato che i test sono effettuati su dati pseudorandom, ho eseguito 100 misure di χ^2 e H, valutandone la media e la deviazione standard. I risultati di questa analisi sono visualizzati in Figura 2.

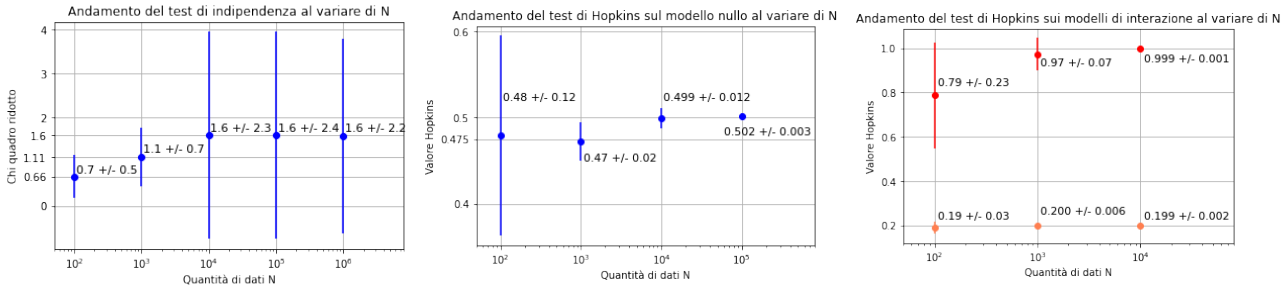

Figura 2 Test della procedura con numero di dati variabile. A sinistra il test di chi quadro per l'indipendenza dei dati, al centro il test di Hopkins su dati generati randomicamente, e a destra il test di Hopkins sui modelli di interazione (in rosso attrattiva, in arancione repulsiva).

La precisione del test di χ^2 sembra diminuire con l'aumentare dei dati, stabilizzandosi per numeri maggiori di 10^4.

Bisogna notare che anche il p value decresce con l'aumentare di N (dato che i gradi di libertá crescono con N). Inoltre il suo valore per il test applicato al modello nullo raramente supera il 40%.

Ho deciso pertanto, data anche la variazione di $\tilde{\chi}^2$, che fosse piú opportuno giudicare il p value dei futuri test su modelli non nulli come la compatibilitá del loro $\tilde{\chi}^2$ con il valore medio ottenuto dal test su N dati random (tramite t di student, se i valori di $\tilde{\chi}^2$ sono raffrontabili). Infine, questo test rigetta completamente i modelli di clustering e attrazione, con un p value prossimo a 0

La precisione della Hopkins statistic aumenta con N, ottenendo un'incertezza di grandezza 0.1% per $N = 5$ su dati indipendenti e per $N = 4$ per dati dipendenti.

Contrariamente a quanto aspettato, il valore di H per dati completamente equidistanti si assesta intorno a 0.20. Ció é giustificabile a causa della natura randomica del metodo stesso.

Come osservabile in Figura 1, valori di H che si discostino dagli estremi 0.2 e 0.99 sono comunque indicativi di un comportamento di clustering o di repulsione, in caso di maggior rumore o deviazione dal modello di dipendenza "pefetta".

Applicazione a data set reali

Applico ora la procedura costruita a quattro data set per verificarne le interazioni.

- **Testo: Frankenstein, di Mary Shelley**. Gli eventi sono qui rappresentati dalle ricorrenze di due parole: l'articolo "The" e la congiunzione "and". Queste parole ricorrono circa 3000 volte nel testo, quindi i risultati vanno paragonati con il valore di $\tilde{\chi}^2$ ottenuto sul modello nullo per N di ordine $3 : 1.1 \pm 0.7$

  Per la prima serie di eventi ottengo $\tilde{\chi}^2 = 1.3$. Calcolando $t = 0.29$ ottengo un p value di 77.18%. Posso quindi affermare che gli eventi siano indipendenti. Ció é confermato calcolando H, che viene pari a 0.48 ± 0.01.

  Per la seconda serie di eventi ottengo $\tilde{\chi}^2 = 4.4$. Calcolando $t = 4.7$ ottengo un p value inferiore a 0.01. Considero quindi gli eventi dipendenti. Ottengo un valore $H = 0.40 \pm 0.02$, che suggerisce una tendenza dei dati ad essere equidistanti, seppur non pronunciata.

Una possibile spiegazione di ció puó essere dovuta al fatto che, mentre la posizione dell'articolo é dettata dalla grammatica in modo indipendente, il posizionamento di congiunzioni ripetute e ravvicinate appesantisce il fluire del discorso, e viene pertanto evitato

- **Terremoti avvenuti dall'anno zero**. Gli eventi sono i giorni in cui si sono verificati. Ho pertanto trasformato il dataset, riportante anno, mese e giorno, nei soli giorni.

  Per l'analisi di questi dati é piú intuitivo partire dal calcolo di H. Si ottiene infatti:

  $H = 0.93 \pm 0.01$ per tutto il data set

  $H = 0.83 \pm 0.07$ per i terremoti dal 1800 ad oggi

  $H = 0.57 \pm 0.03$ per i terremoti dal 1900 ad oggi

  $H = 0.53 \pm 0.03$ per i terremoti dal 2000 ad oggi

  Ció si puó spiegare realizzando che la densitá dei terremoti nel data set cresce nettamente negli ultimi anni, a causa probabilmente dei migliori metodi di rilevazione e archiviazione di tali dati. L'algoritmo interpreta questa discrepanza, valutabile principalmente dall'anno zero ad oggi, come la presenza di un cluster verso gli anni 2000. Ció nonostante, sull'ultimo intervallo, ottengo $\tilde{\chi}^2 = 3.7$, che é incompatibile con il valore del modello nullo (per $N = 3$). Ció puó essere dovuto alla selettivitá del metodo, oppure al fatto che tali dati, pur essendo indipendenti, non abbiano distibuzione di δx esponenziale.

- **Genoma dell'Escherichia Coli**. Gli eventi sono rappresentati dalla posizione di inizio dei geni. Ottengo $\tilde{\chi}^2 = 25.1$, con un p value tramite t inferiore a 0.01. Considero quindi gli eventi dipendenti.

  Ottengo quindi $H = 0.38$, che indicherebbe un' interazione repulsiva. Questo puó essere spiegato ipotizzando che i geni abbiano lunghezze confrontabili. La non "perfetta" repulsivitá puó essere dovuta alla varianza delle lunghezze e alla presenza di geni sovrapposti sul genoma.

- **Orario dei post su Reddit**. Sono tre data set separati: gli eventi sono gli orari ai quali tre utenti hanno scritto dei post su Reddit. Ottengo dal test χ^2 un p value prossimo a 0. Gli eventi sono quindi dipendenti.

  Calcolo, per ogni data set, $H = 0.99$, $H = 0.97$ e $H = 0.97$. Sembra quindi che i dati abbiano una forte tendenza a clusterizzare. Posso spiegare questo risultato supponendo che gli utenti, quando in condizione di navigare sul network e scrivere un post, non si limitino a scriverne solo uno. Questo andamento puó anche essere imputato al ritmo della giornata, che regola il tempo libero utile alla navigazione sul social disponibile agli utenti.

APPENDICE

1. Campionamenti per test χ^2: https://kb.palisade.com/index.php?pg=kb.pageid=57