



UNIVERSITÀ DEGLI STUDI DI TRENTO

Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in
Informatica

ELABORATO FINALE

TITOLO

Sottotitolo (alcune volte lungo - opzionale)

Supervisore
Prof. Alberto Montresor

Laureando
Giovanni De Toni

Anno accademico 2016/2017

Ringraziamenti

...thanks to...

Indice

Sommario	2
1 Metodi	4
1.1 Regressione Lineare	4
1.2 Modelli Lineari Generalizzati	5
1.3 Regolarizzazione	6
1.4 Implementazione	7
2 Analisi dei dati	9
2.1 Wikipedia	9
2.2 Influnet	10
2.3 Analisi	10
2.4 Data preprocessing	12
3 Risultati	13
3.1 Risultati del modello lineare	13
3.2 Risultati del modello di Poisson	14
3.3 Conclusioni	16
Bibliografia	16
A Grafici Modello Lineare	20
B Grafici Modello Poisson	21

Sommario

L'influenza è un'infezione respiratoria acuta causata principalmente da virus della famiglia *Orthomyxoviridae* che possono essere divisi in due ceppi, la varietà A e la varietà B. E' una patologia stagionale, che si presenta spesso durante i mesi invernali (nelle zone con clima temperato) ed è presente in tutto il mondo. Il vettore di trasmissione principale consiste nelle goccioline di muco e saliva, contenenti il virus, che vengono prodotte quando una persona infetta starnutisce o tossisce. Questo la rende una malattia che si diffonde facilmente e rapidamente, specialmente nel caso di zone molto affollate. I sintomi riscontrati spesso sono: febbre alta, tosse, emicrania, dolori articolari e malessere generale. Normalmente, l'infezione sparisce nel giro di una settimana, senza dover ricorrere a particolari cure mediche. In certe categorie a rischio però, se contratta, l'influenza può degenerare e portare anche alla morte [1].

Soltanto in Europa, il *Centro Europeo per il controllo delle malattie* (ECDC) indica come l'influenza stagionale causi da 4 ai 50 milioni di ammalati e circa 15000-70000 morti annuali a causa dell'infezione [2]. Globalmente, il numero di decessi a causa dell'influenza è di circa da 250000 a 500000 persone all'anno [1]. In Italia l'influenza colpisce mediamente ogni anno l'8% della popolazione [3]. Le categorie più colpite sono soprattutto le fasce di popolazione in età pediatrica (0-4 anni e 5-14 anni) con un incidenza cumulativa che decresce con l'aumentare dell'età. I casi severi e le complicanze sono più frequenti nei soggetti al di sopra dei 65 anni di età, oppure con condizioni di rischio ad esempio malattie cardiovascolari, respiratorie o immunitarie [3].

Essendo una patologia che può colpire la maggior parte della popolazione, essa è considerata un grave rischio per la sanità pubblica e per la collettività. Si stima che nei paesi industrializzati, epidemie influenzali possono generare alti livelli di assenteismo, sia lavorativo che scolastico, e una riduzione della produttività [1]. In Italia, uno studio ha stimato il costo totale delle epidemie influenzali nel periodo 1999-2008 che varierebbe da 15 a 20 miliardi di euro [4]. Si è anche stimato che la durata media di assenza dal posto di lavoro a causa dell'influenza è di circa 4,8 giorni. Inoltre, i costi diretti in media sono di circa 330 euro a persona (visite, diagnostica, farmaci) e possono salire a circa 3000-6000 euro in caso di ricovero ospedaliero. I costi sociali indiretti (inattività scolastica o lavorativa) ammontano invece a 1000 euro a persona [5].

Pur essendo una patologia facilmente curabile e prevenibile grazie all'uso degli appositi vaccini, l'influenza stagionale è un'infezione che non può essere sottovalutata. Infatti, epidemie e pandemie causate da questi virus possono essere molto gravi. Si pensi ad esempio alla cosiddetta *influenza spagnola*, una pandemia causata dal virus dell'influenza A sottotipo H1N1 che causò 500 milioni di casi globali e circa 50 milioni di morti totali tra il 1918 e il 1920 [6], addirittura più di quelli causati dalla peste nera del XIV secolo [7].

Attualmente, l'attività dei virus influenzali viene monitorata da alcuni centri facenti parte del *Global Influenza Surveillance and Response System* (GISRS), un network di sorveglianza globale sponsorizzato dall'WHO. Questi centri forniscono: informazioni sugli attuali ceppi circolanti, indicazioni per la produzione dei vaccini antiinfluenzali (su quali varietà focalizzarsi) ed esaminano e conservano campioni dei virus per scopi di ricerca [8].

Per quanto riguarda l'Italia, il Centro di Controllo delle Malattie (CCM) del Ministero della Salute sostiene che un componente fondamentale per il controllo dell'influenza (sia epidemica che pandemica) è la sorveglianza. Nel nostro paese esistono già programmi di monitoraggio dei livelli di ILI (Influenza-Like Illness), come Influnet. Influnet è il sistema nazionale di sorveglianza epidemiologica e virologica;

il suo compito è quello di stimare l'incidenza settimanale della sindrome influenzale (avvalendosi di dati raccolti da medici sentinella disseminati su tutto il territorio nazionale), in modo da rilevare la durata e l'intensità dell'epidemia [9]. Durante la stagione invernale vengono pubblicati anche dei bollettini settimanali, tramite il servizio FluNews, che illustrano l'evoluzione della situazione italiana [10]. Questi dati vengono anche condivisi sia con il WHO che con ECDC.

InfluNet fornisce informazioni molto importanti riguardo all'incidenza delle patologie influenzali sulla popolazione italiana, però i dati vengono spesso pubblicati con un certo ritardo (in media 2 settimane) rispetto all'arco di tempo che descrivono. Per attuare azioni efficaci e coordinare la distribuzione di materiale sanitario, produzione di vaccini etc. è necessario avere immediatamente a portata di mano dei dati sulla situazione.

Ci sono stati diversi sforzi per tentare di prevedere o stimare i livelli di incidenza di ILI all'interno della popolazione, sfruttando fonti di dati non convenzionali (cioè non direttamente le informazioni mediche) [11–15]. Normalmente, i dati che vengono sfruttati maggiormente sono quelli prodotti dai social media, ad esempio: messaggi di Twitter [15], page view di Wikipedia [11–13] e keywords di ricerca di Google [14]. Da questi studi emerge che attraverso l'utilizzo di tecniche di machine learning è possibile arrivare a delle stime dei livelli di ILI all'interno della popolazione, con settimane di anticipo rispetto ai metodi tradizionali.

L'obiettivo di questa tesi è cercare di replicare, per quanto possibile, alcune parti dei lavori precedentemente citati, per verificare se anche nella nostra penisola sia possibile effettuare, attraverso tecniche di machine learning, un'analisi attiva per la sorveglianza della diffusione di malattie influenzali. Lo studio che verrà utilizzato come base è la ricerca di David J. McIver e John S. Brownstein [11] in cui vengono delineate delle tecniche per l'utilizzo delle page view di Wikipedia per stimare il numero di malati settimanali negli Stati Uniti.

Il materiale di questo lavoro è suddiviso in capitoli ed ognuno di essi tratta una singola parte di tutto il processo svolto. Nel Capitolo 1 vengono descritti i metodi di machine learning che sono stati selezionati per procedere alla creazione del modello predittivo finale. Il Capitolo 2 fornisce una descrizione più dettagliata dei dati che sono stati utilizzati in questo progetto (page view di Wikipedia e bollettini di InfluNet). Inoltre, si definiscono anche i metodi usati per l'analisi degli stessi e alcune informazioni statistiche sulla composizione del dataset. Il Capitolo 3 e il Capitolo 4 presentano rispettivamente: i risultati dell'esperimento e le riflessioni finali su quello che gli esperimenti hanno evidenziato.

1 Metodi

I metodi che sono stati selezionati per creare i modelli utilizzati negli esperimenti di questo lavoro sono compresi nella categoria che in machine learning viene chiamata *apprendimento supervisionato*. Gli algoritmi appartenenti a questo gruppo cercano di produrre una funzione f_a che approssimi nel modo migliore un'altra funzione f_b ignota, sulla base di una serie di dati di esempio (o *osservazioni*) forniti. Attraverso l'utilizzo di questi dati di esempio, si suppone che l'algoritmo riesca ad accumulare abbastanza "esperienza" in modo da fornirci un'approssimazione della funzione f_b .

Dati n esempi di *training* (che corrispondono al nostro *dataset*), nella forma $\{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$ dove y_i è il valore associato all' i -esimo vettore di *feature* \mathbf{x}_i , l'algoritmo cerca di dedurre dalle osservazioni fornite una funzione $f : X \rightarrow Y$, dove X è lo spazio delle *feature* e Y è lo spazio del risultato (tale per cui $\forall y \in Y$). Questa funzione deve avere anche capacità di generalizzazione, cioè deve essere in grado di fare delle previsioni anche utilizzando come input delle osservazioni che non erano presenti nel dataset di training.

1.1 Regressione Lineare

Prevedere i livelli di ILI in Italia, durante la stagione influenzale, a partire dall'analisi delle voci di Wikipedia si configura come un problema che viene detto di *regressione*. Con questo termine si indica un'ampia classe di problemi il cui obiettivo è la modellazione di una relazione lineare tra una variabile dipendente y e una serie di variabili indipendenti x_1, x_2, \dots, x_n (chiamate anche *regressori*). Nel nostro caso, poichè dobbiamo predire una sola variabile dipendente y scalare si parla di *regressione lineare semplice*.

Più formalmente, dato un dataset $\{y_i, x_{i1}, x_{i2}, \dots, x_{in}\}_i^n$, la regressione lineare assume che la relazione che intercorre tra la variabile dipendente y_i e la variabili indipendenti $x_{i1}, x_{i2}, \dots, x_{in}$ sia lineare. La relazione viene modellata anche attraverso un variabile aleatoria ϵ per tenere conto di potenziale rumore presente nella dipendenza tra la variabile indipendente e i regressori. Il modello si presenta alla fine come:

$$y_i = \sum_{j=0}^k x_j \cdot \beta_j + \epsilon = \mathbf{x} \cdot \boldsymbol{\beta} + \epsilon \quad i = 0, 1, 2, \dots, n \quad (1.1)$$

Utilizzando il dataset di *training* dobbiamo "allenare" il modello in modo da stimare correttamente il valore del vettore dei pesi $\boldsymbol{\beta}$. La tecnica più semplice e anche più frequentemente utilizzata è il *metodo dei minimi quadrati* (*ordinary least square*) [16]. Esso consiste nell'assegnare ai vari β_i dei valori che minimizzino la somma degli scarti quadratici:

$$\min_{\boldsymbol{\beta}} \sum_{i=0}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 \quad (1.2)$$

Per ottenere ciò, è sufficiente porre tutte le derivate parziali rispetto ad β_i uguali a zero, in modo così da trovare il minimo della funzione. Risolvendo le equazioni è possibile arrivare ad ottenere una *closed-form solution* per $\boldsymbol{\beta}$ (cioè di una espressione matematica che può essere risolta in un numero finito di passi operazionali). Partendo dall'equazione (1.2), rappresentata nei passaggi successivi dopo

l'espansione del prodotto scalare $\mathbf{x}_i\boldsymbol{\beta}$, la dimostrazione è la seguente:

$$\sum_{i=0}^n \left(\sum_{j=0}^m (x_{ij}\beta_j) - y_i \right)^2 = 0 \quad (1.3)$$

$$\frac{\partial}{\partial \beta_k} \sum_{i=0}^n \left(\sum_{j=0}^m (x_{ij}\beta_j) - y_i \right)^2 = 0 \quad \forall k = 0, \dots, n \quad (1.4)$$

$$\sum_{i=0}^n \frac{\partial}{\partial \beta_k} \left(\sum_{j=0}^m (x_{ij}\beta_j) - y_i \right)^2 = 0 \quad (1.5)$$

$$\sum_{i=0}^n 2 \cdot \left(\sum_{j=0}^m (x_{ij}\beta_j) - y_i \right) \cdot x_{ik} = 0 \quad (1.6)$$

$$\sum_{i=0}^n \mathbf{x}_{ik} (\mathbf{x}\boldsymbol{\beta} - y_i) = 0 \quad (1.7)$$

In forma matriciale, la formula (1.7) può essere rappresentata come:

$$X^T \cdot (X\boldsymbol{\beta} - \mathbf{y}) = 0 \quad (1.8)$$

$$X^T X\boldsymbol{\beta} = X^T \mathbf{y} \quad (1.9)$$

$$\boldsymbol{\beta} = (X^T X)^{-1} X^T \mathbf{y} \quad (1.10)$$

Ovviamente esistono altri metodi per stimare il vettore dei pesi $\boldsymbol{\beta}$, che variano per complessità computazionale dei loro algoritmi, prerequisiti teorici e per la presenza o meno di una *closed-form solution*. Il modello di questo progetto utilizza come stimatore l'algoritmo chiamato *coordinate descent*, che permette di trovare minimo locale di una funzione in maniera iterativa.

1.2 Modelli Lineari Generalizzati

Un altro regressore utilizzato in alcuni lavori [11] per modellare la relazione tra le voci di Wikipedia e l'incidenza di ILI consiste in un *modello lineare generalizzato* [17] (o *generalized linear model*). Questi modelli assumono che la variabile dipendente y non segua una distribuzione normale, ma che possa essere distribuita come una qualsiasi variabile casuale della famiglia esponenziale (binomiale, poissoniana, gamma etc.). In questo lavoro di tesi verrà utilizzato un *modello lineare generalizzato di Poisson*.

Un modello lineare generalizzato ha bisogno di tre componenti per essere definito correttamente:

1. Una funzione di distribuzione f facente parte della famiglia esponenziale (in questo caso la funzione di densità di Poisson $p_\theta(x) = \frac{\theta^x}{x!} e^{-\theta}$);
2. Un predittore lineare $\eta = \mathbf{x}\boldsymbol{\beta}$, che tenga conto di tutte le variabili indipendenti x_i ;
3. Una funzione invertibile g detta *link function* (normalmente per Poisson viene utilizzato il logaritmo naturale). Questa funzione serve a trasformare il valore atteso della distribuzione $\mathbb{E}(y)$ nel predittore lineare η , cioè $g(\mathbb{E}(Y)) = \eta$.

Di seguito riporto la procedura per stimare correttamente il vettore dei pesi $\boldsymbol{\beta}$ del regressore lineare η . Dato un dataset con n vettori di feature x_i , un insieme di variabili dipendenti associate y_i , la probabilità di ottenere questo specifico dataset, dato un vettore di pesi $\boldsymbol{\beta}$ è:

$$p(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n; \boldsymbol{\beta}) = \prod_{i=1}^n \frac{e^{\boldsymbol{\beta}\mathbf{x}_i y_i}}{y_i!} e^{-\boldsymbol{\beta}\mathbf{x}_i} \quad (1.11)$$

Questo si può dedurre dalla funzione di densità della distribuzione di Poisson, con il parametro θ uguale a $e^{\beta x_i}$. Si può facilmente dimostrare che $\theta = e^{\beta x}$ utilizzando le definizioni precedenti dei modelli lineari generalizzati. Si può infatti notare che il valore atteso per una distribuzione di Poisson è proprio il parametro θ , quindi, per la definizione 3, $\mathbb{E}(y) = \theta = g^{-1}(\eta)$.

Per stimare correttamente il vettore dei pesi necessario viene utilizzata la tecnica della *maximum likelihood estimation*, che implica di identificare β in modo che la probabilità espressa da (1.11) sia massima. Prima di tutto, si riscrive (1.11) come una *funzione di verosimiglianza*, $\mathcal{L}(\theta|X, Y)$. Un funzione di verosimiglianza (o *likelihood function*) viene definita come:

$$\mathcal{L}(\theta|x) = p_\theta(x) = p_\theta(X = x) \quad (1.12)$$

Quindi, nel caso di una variabile aleatoria di Poisson, la funzione di verosimiglianza diventa:

$$\mathcal{L}(\beta\mathbf{x}|X, Y) = \prod_{i=1}^n e^{y_i\beta x_i} - e^{-\beta x_i} y_i! \quad (1.13)$$

Per semplificare i calcoli, spesso si utilizza una versione semplificata della funzione di verosimiglianza, la *log-likelihood function*, $l(\theta|X, Y)$:

$$l(\beta\mathbf{x}|X, Y) = \log \mathcal{L}(\beta\mathbf{x}|X, Y) \quad (1.14)$$

Eseguendo le dovute semplificazioni, la funzione di verosimiglianza da minimizzare (attraverso tecniche di *gradient descent*) diventa quindi:

$$\min_{\beta} - \sum_{i=1}^n (y_i\beta x_i - e^{\beta x_i}) \quad (1.15)$$

1.3 Regularizzazione

Come abbiamo precedentemente detto, utilizzando vari stimatori (*coordinate descent* e *maximum likelihood*) andiamo a stimare i valori dei pesi che andranno poi a formare il nostro regressore lineare. Nonostante tutto, questo metodi non sono privi di errori e possono incappare in problemi cosiddetti di *overfitting*, causati principalmente da un numero eccessivo di parametri rispetto al totale delle osservazioni. In questi casi, il modello potrebbe descrivere perfettamente i dati del dataset di training, ma non essere comunque abbastanza generale da prevedere ulteriori dati di test (si veda l'esempio mostrato in figura 1.1). All'interno delle funzioni da minimizzare precedentemente citate, in particolare (1.4) e (1.15), abbiamo inserito allora una penalità che ci permetterà di ottenere un modello meno sensibile all'*overfitting*.

Questa tecnica è detta *regularizzazione* (o *regularization*) ed implica l'aggiunta di un *termine di regularizzazione* $R(\beta)$ alla *loss function*. Nel caso dei regressori lineari il *termine di regularizzazione* imporrà una penalità sul vettore dei pesi che deve essere stimato.

$$\min_{\beta} \sum_{i=1}^n (\mathbf{x}_i\beta - y_i)^2 - \lambda R(\beta) \quad (1.16)$$

Il termine λ controlla l'importanza della regularizzazione.

La tecnica utilizzata in entrambi i modelli utilizzati in questo lavoro viene chiamata *LASSO*, acronimo per *Least Absolute Shrinkage and Selection Operator*. Essa serve a ridurre l'*overfitting* e ad applicare una selezione sulle varie feature disponibili. Infatti, un modello regularizzato con LASSO tenderà ad assegnare un peso positivo a poche feature, determinando così quali sono i regressori che contribuiscono in modo maggiore. Inoltre, la regularizzazione LASSO è indicati nei casi di multicollinearità, cioè quando due o più feature sono altamente correlate tra di loro (LASSO ne sceglie solo una ed evita di

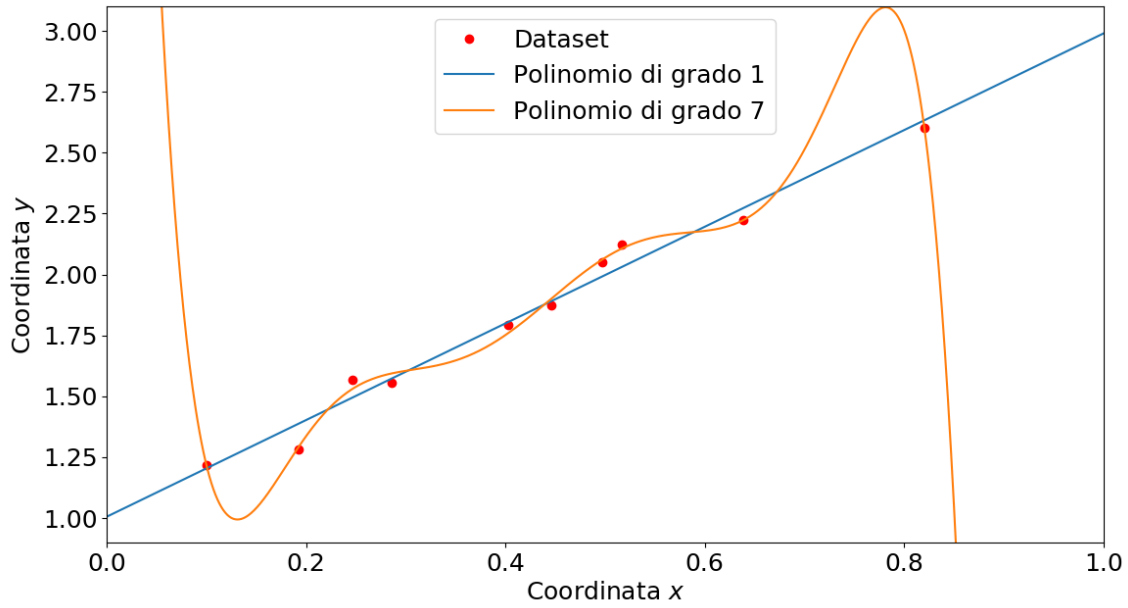


Figura 1.1: In questo esempio, stiamo cercando di trovare un polinomio che meglio approssima i punti del dataset. Il polinomio di grado 7, pur rappresentando esattamente i punti del dataset non ha capacità di generalizzazione, quindi nella pratica si tenderà ad utilizzare il polinomio di grado 1.

includere nel modello finale le altre, visto che non diminuirebbero l'entropia del modello). Nel caso di *LASSO*, il termine di regolarizzazione corrisponde alla norma 1, cioè $R(\beta) = \|\beta\|_1 = \sum_{i=0}^k \beta_i$. In questo caso, le due *loss function* (1.15) e (1.15) vengono ridefinite come:

$$\min_{\beta} \sum_{i=1}^n (x_i \beta - y_i)^2 + \lambda \|\beta\|_1 \quad (1.17)$$

$$\min_{\beta} - \sum_{i=1}^n (y_i \theta x_i - e^{\theta x_i}) + \lambda \|\beta\|_1 \quad (1.18)$$

Per stimare il parametro λ abbiamo utilizzato la tecnica di *cross-validation*. Essa consiste nel suddividere il dataset in k partizioni di uguale numerosità, $k - 1$ verranno utilizzate per il *training* mentre l'unica rimanente verrà utilizzata per il *testing*. La procedura viene poi ripetuta k volte in modo che ogni partizione abbia sia stata utilizzata come *testing set*. Attraverso l'utilizzo di questo metodo possiamo avere maggiori informazioni su come il modello finale si comporterà con dati nuovi (cioè quanto il modello è capace di generalizzare). Inoltre, questo metodo permette appunto di stimare certi parametri del modello, in modo che il loro valore massimizzi le performance dell'algoritmo (o minimizzino al meglio la *loss function*).

1.4 Implementazione

Per la realizzazione del codice, sono stati utilizzati due framework in Python che possiedono l'implementazione dei metodi sopracitati. La suite *scikit-learn* ha fornito il modello lineare semplice, regolarizzato con *LASSO* e validato attraverso *cross-validation*, chiamato *LassoCV*. Per il modello lineare generalizzato di Poisson è stato usato il framework *glmnet* che ha fornito il metodo *cvglmnet* (anch'esso con *LASSO+cross-validation*).

Entrambe le librerie utilizzano funzioni che sono leggermente diverse da quelle presentate precedentemente (in particolare (1.17) e (1.18)), che però sono essenzialmente equivalenti ai fini

dell'esperimento vero e proprio.

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^n \|\boldsymbol{\beta} x_i - y_i\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \quad (1.19)$$

$$\min_{\boldsymbol{\beta}} -\frac{1}{n} \sum_{i=1}^n (y_i \boldsymbol{\beta} x_i - e^{\boldsymbol{\beta} x_i}) + \lambda \|\boldsymbol{\beta}\|_1 \quad (1.20)$$

2 Analisi dei dati

2.1 Wikipedia

Wikipedia (<https://wikipedia.com>) è un'enciclopedia online, a contenuto libero e gratuito, lanciata da Jimmy Wales e Larry Sangers nel 2001 ed ora gestita dalla Wikimedia Foundation. Al suo interno sono presenti 45 milioni di voci in oltre 280 lingue raggruppate in diverse *categorie*. Wikipedia è anche il 5 sito più visitato al mondo. Soltanto la versione italiana di Wikipedia (<http://it.wikipedia.org>), che al momento (maggio 2017) conta circa 1400000 articoli [18], nel 2016 ha avuto una media giornaliera di 17 milioni di visite [19].

I dati che andremo ad utilizzare riguardano le *page view* di Wikipedia in italiano [20]. Con *page view* si intende il numero di visite che sono state effettuate su una determinata pagina di Wikipedia in un certo lasso di tempo. Essi sono dati grezzi: infatti indicano il numero di visite cumulative ricevute da una voce (quindi anche multiple visite dello stesso utente nella stessa ora) e non distinguono tra visite effettuate da esseri umani oppure effettuate da bot (spider, crawler, etc.). Queste informazioni sono comprensive soltanto delle visite effettuate tramite dispositivi desktop; le visualizzazioni effettuate da dispositivi mobile non sono presenti all'interno delle statistiche.

I file di log di Wikipedia che siamo andati ad analizzare per estrarre le informazioni sono così formati: ogni giorno vengono redatti una serie di file, con cadenza oraria, in cui vengono memorizzati per ogni voce di ogni progetto di Wikipedia le visite effettuate in quell'ora. Per esempio, il file *pagecounts-20071209-180000.gz* conterrà, per ogni voce di Wikipedia, le visite orarie del 09 Settembre 2007 effettuate dalle ore 18:00 fino alle 18:59. La struttura interna dei file è invece formata da quattro campi principali: una sigla identificativa del nome del progetto di Wikipedia, il nome della pagina, il numero di visite di quella pagina in quell'ora e la dimensione della pagina richiesta (in *byte*).

Per ottenere un elenco di tutte le voci è stato utilizzato il tool *PetScan* [21], una web-application che permette di trovare specifici articoli, immagini, categorie che soddisfano specifici requisiti. La tabella 2.1 indica le categorie selezionate per essere utilizzate come regressori del modello. Inoltre, per verificare poi la validità del modello finale, abbiamo anche estratto un altro dataset contenente tutte le voci di alcune categorie di Wikipedia prese in modo casuale.

Dataset principale	Dataset random
<i>Malattie infettive virali</i>	<i>Investigatori immaginari</i>
<i>Malattie infettive</i>	<i>Aziende di abbigliamento italiane</i>
<i>Epidemie</i>	<i>Software con licenza GNU LGPL</i>
<i>Virus</i>	
<i>Vaccini</i>	
<i>Segni clinici</i>	

Tabella 2.1: *Categorie di Wikipedia da cui sono state prese le voci per i vari dataset.*

Sono state scelte queste categorie poiché si cerca appunto una correlazione tra l'uso di Wikipedia e la percentuale di popolazione con patologie influenzali. Infatti, si assume che una persona malata tenda a cercare su internet, utilizzando motori di ricerca come Google, i propri sintomi (*Febbre, Influenza o Raffreddore*) e che nella maggior parte dei casi vada a controllare le voci di Wikipedia che trattano quei sintomi. Il dataset casuale è stato appunto definito per verificare che esista effettivamente questa

relazione e che non sia possibile costruire modelli equivalenti utilizzando voci scelte a caso. Infatti, il modello dovrebbe dare risultati coerenti soltanto utilizzando il dataset che comprende voci di categoria medica inerenti all'influenza.

Per raffinare l'elenco di voci, l'elenco è stato generato sottraendo tutti quegli articoli affini anche alla categoria *Patologie aviarie*, per arrivare ad un elenco contenente 468 feature [22]. Successivamente, sono state aggiunte le voci della *Pagina principale* (come parametro di controllo) e della voce di disambiguazione *Influenza*, per arrivare ad un totale finale di 470 feature. Per quanto riguarda il dataset casuale, possiamo contare un totale di 463 voci [23].

Poiché i dati di Influnet ci forniscono informazioni settimanali anche i dati di Wikipedia sono stati aggregati settimanalmente. Il dataset totale di Wikipedia copre un arco di tempo che va dal Dicembre 2007 al Maggio 2016, per un totale di circa 480 settimane.

2.2 Influnet

I dati riguardanti i livelli di ILI in Italia sono stati ottenuti attraverso l'analisi dei bollettini Influnet, che vengono pubblicati settimanalmente dal Ministero della Salute per tutta la durata del periodo influenzale. Questi bollettini forniscono dati e statistiche dettagliate sulla diffusione e sui casi di influenza. Essi presentano informazioni come l'incidenza nelle varie fasce d'età o il numero di assistiti totali sottoposti a controlli. Questi dati vengono raccolti da diversi medici sentinella disseminati in tutto il territorio nazionale.

Attualmente, i dati non sono disponibili in formati accessibili e analizzabili (come file *.csv*). Tutti i bollettini sono in formato PDF e tutte le informazioni contenute nelle varie tabelle sono state estratte mediante l'utilizzo del software *Tabula* (<http://tabula.technology/>). I file CSV ottenuti contengono i seguenti campi:

- **Settimana:** coppia anno-settimana che indica la fascia temporale a cui i dati si riferiscono;
- **Totale Medici:** totale dei medici sentinella che hanno partecipato al programma ed inviato i dati durante la settimana di riferimento;
- **Totale Casi:** totale dei casi di influenza rilevati;
- **Totale Assistiti:** totale degli assistiti coperti dai medici sentinella;
- **Incidenza Totale:** incidenza totale dell'influenza su 1000 assistiti;

I dati forniti da influnet spaziano dalla stagione influenzale 2003-2004 fino alla più recente 2016-2017. Per questo progetto abbiamo utilizzato le informazioni che comprendono stagione influenzale 2007-2008 fino alla 2015-2016, per un totale di 227 settimane. Per ogni stagione influenzale i dati coprono le settimane che vanno dalla numero 42 (metà Ottobre circa) fino alla numero 17 (metà Aprile circa dell'anno successivo). Inoltre, poiché la stagione influenzale 2008-2009 è stata più aggressiva del solito (a causa del virus H1N1), in quel caso abbiamo dati fino che spaziano dall'Ottobre 2008 fino all'Ottobre 2009.

2.3 Analisi

Presentiamo ora alcuni grafici che mostrano alcune informazioni riguardo al dataset di Wikipedia e al dataset di Influnet che andremo ad utilizzare allenare/testare il modello.

Abbiamo analizzato come cambia negli anni il volume totale delle visite effettuate sulle 470 voci selezionate. Inoltre, abbiamo osservato l'andamento delle varie stagioni influenzali. Come si può notare, negli anni il numero di visualizzazioni è calato drasticamente (si veda la figura 2.1). Questo potrebbe essere spiegato dal fatto che i dati non contano gli accessi effettuati tramite dispositivi mobile (tablet e smartphone) che oramai stanno diventando un mezzo privilegiato per accedere ad internet.

Riguardo ai dati ILI, si può notare come il picco influenzale si collochi in media alla quinta settimana della stagione con l'unica eccezione della stagione 2009-2010 che invece ha il suo picco nella quarantesima settimana (si veda la figura 2.2). L'incidenza delle patologie influenzali si attesta mediamente a 3.40 casi per 1000 assistiti, mentre il picco massimo è stato nel 2009-2010 con 12.92 casi e il picco minimo è stato nella stagione 2015-2016 con 6.14 casi per 1000 assistiti.

Per quanto riguarda statistiche generali sul dataset di Wikipedia: la voce con il maggior numero medio di visualizzazioni settimanali è quella del *Virus del papilloma umano* (circa 5560), mentre la voce *Stimmate (medicina)* è quella con il numero medio minore di visualizzazioni (si parla sempre del periodo Dicembre 2007 - Maggio 2016). Inoltre, il numero medio di visite settimanali fra tutte le voci è circa 215. Abbiamo anche analizzato il traffico di Wikipedia in italiano per comprendere da quale paese sono effettuate la maggior parte delle visite (per controllare se ci sono possibili problemi di "rumore" che possono incidere sulle prestazioni del modello). Ovviamente, il 90.8% del traffico viene effettuato dall'Italia, con punte trascurabili del 1.4% dalla Germania, 1.0% dagli Stati Uniti e di altri paesi minori [24].

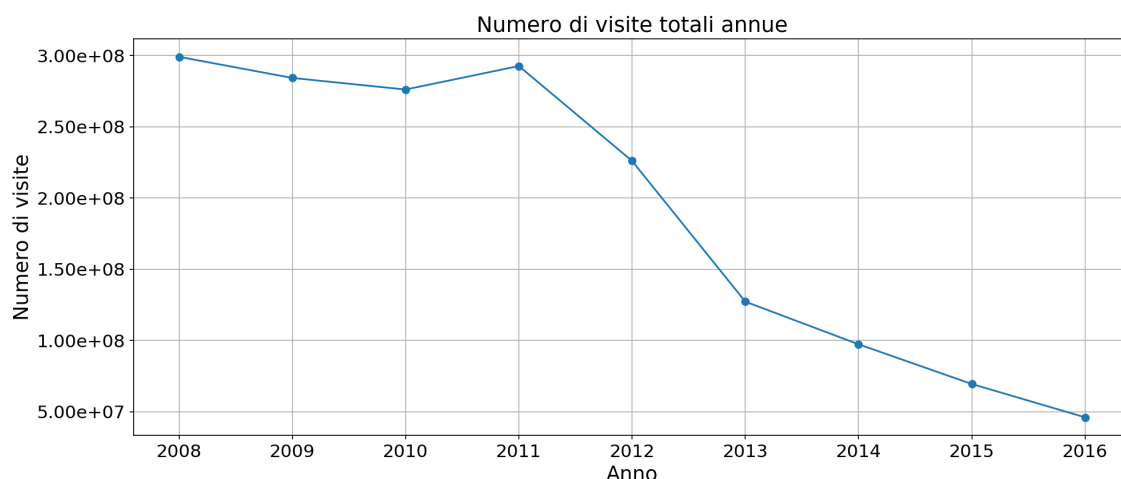


Figura 2.1: Andamento delle visite totali annue per le categorie di Wikipedia selezionate.

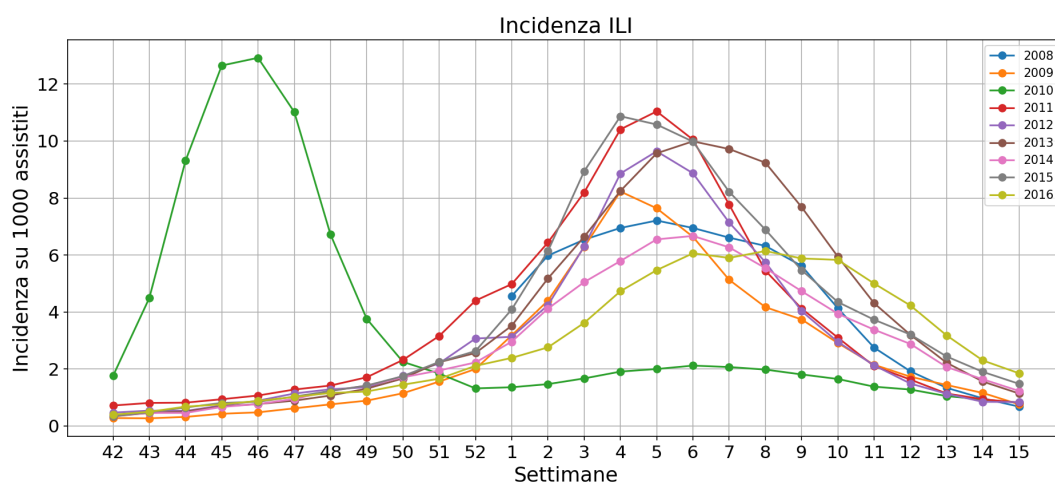


Figura 2.2: Questo grafico rappresenta i livelli di incidenza di patologie influenzali in Italia negli anni che vanno dal 2008 al 2016. L'incidenza viene mostrata solo per le settimane che sono state utilizzate per questo progetto.

2.4 Data preprocessing

Prima di poter utilizzare il dataset per allenare i vari modelli abbiamo dovuto effettuare delle attività di *preprocessing* per correggere alcune imperfezioni dei dati originali. Il primo problema che siamo andati ad affrontare riguarda la mancanza di dati affidabili riguardo il numero totale di visite di alcune voci di Wikipedia. Infatti, alcune delle voci selezionate sono state create successivamente al 2007 (ad esempio nel 2010) rendendo quindi impossibile ottenere dati corretti sul numero di visualizzazioni settimanali prima della loro data di creazione.

Per ovviare a questo problema, durante la creazione del dataset è stato effettuato un controllo sulla data di creazione di ogni voce (attraverso l'utilizzo di un'apposita API di Wikipedia) che permetteva di inserire dei valori identificativi *N/A* così da poter poi effettuare successivamente un'analisi più approfondita. Ad esempio, la pagina *Hepatitis_B_virus* è stata creata per la prima volta il 26 Aprile 2011, quindi, in tutti gli anni precedenti, la voce *Hepatitis_B_virus* possiederà dei valori *N/A*. Grazie a questi indicatori, si potrebbe andare ad agire su quelle voci, ad esempio stimando il numero di visite che quella pagina avrebbe ricevuto in quella specifica settimana (per semplicità, nel nostro dataset tutti i valori *N/A* sono stati sostituiti dal valore zero).

Un altro problema che siamo andati ad affrontare riguarda la diversa numerosità delle voci selezionate. Alcune voci infatti possiedono un volume di visite molto maggiore rispetto alle altre. Ad esempio, la pagina *Raffreddore* ha una media di 529 visite settimanali, contro le 2821 di una voce come *Febbre*. Nel caso di un modello lineare questo potrebbe causare un errato assegnamento dei valori dei pesi β poiché non si riesce a catturare l'importanza relativa delle varie feature.

Abbiamo proceduto quindi applicando una tecnica di *feature scaling*, che consiste nel *normalizzare* i dati del dataset entro uno specifico range, di solito $[0, 1]$ o $[-1, 1]$, così da eliminare le differenze di numerosità. Nel nostro caso abbiamo ridotto le feature nel range $[-1, 1]$ applicando questa formula per ogni feature:

$$\mathbf{x}' = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\max(\mathbf{x}) - \min(\mathbf{x})} \quad (2.1)$$

Dove \mathbf{x} è il vettore contenente tutti i valori di una specifica feature, $\bar{\mathbf{x}}$ è la media dei valori assunti dalla feature, $\max(\mathbf{x})$ indica il massimo valore assunto da \mathbf{x} e, rispettivamente, $\min(\mathbf{x})$ il minimo.

Alla fine, i dati per l'esperimento sono stati utilizzati nel seguente modo. Dato il dataset contenente le informazioni delle 9 stagioni influenzali in nostro possesso, esso è stato diviso in modo da ottenere un dataset di training e un dataset di test disgiunti tra loro. Il dataset di test conterrà i valori di una sola stagione influenzale, mentre il dataset che verrà usato per allenare il modello conterrà le informazioni delle rimanenti. La procedura è stata eseguita per ogni stagione influenzale, così da simulare una valutazione per *cross-validation* e così da ottenere una stima su come il modello si comporterà una volta allenato con tutto il dataset.

3 Risultati

3.1 Risultati del modello lineare

Per valutare le performance del modello lineare, oltre ai sistemi classici come l'*errore quadratico medio*, abbiamo cercato di misurare con che precisione il modello tende a predire il picco influenzale, cioè la settimana con la massima incidenza della patologia nella popolazione italiana. Come si può notare dalla tabella 3.1, nella metà delle stagioni influenzali il modello lineare predice il picco entro una settimana indicata dai dati Influnet; inoltre, nelle stagioni 2008-2009, 2011-2012 e 2014-2015 il picco è previsto correttamente. Negli altri quattro casi il modello lineare tende ad anticipare la settimana del picco (stagioni 2007-2008, 2010-2011 e 2013-2014).

Il modello lineare possiede una buona capacità predittiva anche nel caso della stagione influenzale 2009-2010, che fu particolarmente grave a causa della diffusione di un'epidemia causata dalla varietà H1N1 del virus influenzale. Il picco in questo caso è spostato verso le ultime settimane del 2009 (dalla numero 42 alla 52) e non corrisponde al classico andamento della patologia (si nota molto bene dalla figura 2.2). Inoltre, quell'anno c'era stata una grande copertura mediatica riguardo all'epidemia, cosa che avrebbe potuto in qualche modo aggiungere rumore ai dati del dataset. Nonostante ciò, il modello riesce a prevedere il picco della stagione correttamente (con uno scarto di una settimana).

Abbiamo anche analizzato quali feature siano più importanti per la previsione dei valori di incidenza. Ci siamo soffermati su quelle voci di Wikipedia che sono presenti più volte in tutti i modelli e quali di esse avessero il più grande peso associato. Attraverso l'utilizzo di LASSO abbiamo potuto ridurre considerevolmente il numero di feature che effettivamente giocano un ruolo nella previsione dell'incidenza, infatti, da 470 feature i modelli ne andavano ad utilizzare in media 111 (con utilizzare si intende che veniva assegnato loro un peso $\beta_i \neq 0$).

La tabella 3.2 mostra le prime 5 feature con peso medio maggiore e in quanti modelli esse siano presenti. Come si può notare, sono voci di Wikipedia legate ai sintomi influenzali (la voce con più alto peso medio riguarda la patologia *Febbre*). Poiché quasi nel 100% dei modelli abbiamo le stesse feature, e poiché esse danno il contributo più grande nel determinare il livello di incidenza influenzale, possiamo supporre che anche l'andamento delle pageview di queste voci rispecchi l'andamento dei dati Influnet e che quindi pageview ed Influnet siano in qualche modo correlati.

Questa relazione è facilmente visibile ad esempio nella stagione influenzale 2012-2013, come ci mostra la figura 3.1. Se osserviamo l'andamento delle pageview di alcune delle voci della tabella 3.2 rispetto all'incidenza ILI, si può notare come il loro valore inizi ad aumentare esattamente nello stesso periodo in cui anche l'incidenza della patologia aumenta. Questa relazione si può osservare anche nelle stagioni influenzali 2009-2010 e 2011-2012.

Per verificare che questa relazione tra le pageview e l'incidenza ILI in Italia si presenti solo con pagine legate all'ambito medico, in particolare con pagine legate alla patologia influenzale, abbiamo provato ad utilizzare il dataset formato da voci estratte casualmente da Wikipedia per allenare il modello lineare, con le stesse modalità con cui abbiamo usato quello reale. Come ci aspettavamo, utilizzando le pagine casuali, il modello non riesce ad effettuare previsioni coerenti con Influnet. In certe stagioni influenzali, il modello non riesce neppure a convergere ad una soluzione e molto spesso alla quasi totalità delle 470 feature non viene assegnato nessun peso β_i diverso da zero. Ciò significa che probabilmente non esiste una relazione lineare tra le pageview del dataset casuale con l'incidenza Influnet, a differenza di ciò che accade con il dataset formato da pagine di categoria medica.

Osservando i grafici successivi (figura A), si osserva come il modello lineare sia più efficace nelle stagioni influenzali che vanno dalla 2008-2009 fino alla 2012-2014, mentre si comporta relativamente peggio nelle stagioni 2007-2008, 2014-2015 e 2015-2016. Le basse prestazioni negli ultimi possono essere spiegati nel seguente modo:

- Nel caso della stagione 2007-2008, alcune delle voci utilizzate dal modello non erano ancora presenti come pagine di Wikipedia, quindi il valore delle pageview settimanali per la stagione 2007-2008 per queste feature non è disponibile. Infatti, su 470 feature, 184 sono "nulle", mentre delle 51 feature utilizzate dal modello 18 di queste hanno valori sempre uguali a zero nel dataset che copre l'arco di tempo 2007-2008. Probabilmente, a causa di ciò il modello ha stimato valori di incidenza ILI più bassi del normale, poichè non aveva sufficienti informazioni per dare una previsione corretta;
- Nel caso delle stagioni 2014-2015 e 2015-2016, la causa può essere imputata principalmente alla mancanza di una variazione significativa dei valori delle pageview per le voci selezionate dal modello. In sostanza, a differenza del 2012-2013, in cui l'andamento delle voci seguiva quello dell'incidenza ILI, nel 2014-2015 e 2015-2016 ciò non succede. Questo potrebbe essere causato dalla mancanza nel dataset di informazioni riguardanti gli accessi tramite dispositivi mobile che, se aggiunte, potrebbero migliorare le prestazioni. Infatti, da maggio 2015 il numero delle visite alle pagine di Wikipedia effettuate da dispositivi mobile ha superato il numero delle visite effettuate da desktop [25].

3.2 Risultati del modello di Poisson

Nonostante nel paper di riferimento [11] l'utilizzo di un modello generalizzato si sia rivelato efficiente e adatto a stimare i livelli di ILI, per l'Italia il modello di Poisson non ha mostrato significativi miglioramenti rispetto al modello lineare, anzi in certi casi ha mostrato un significativo calo di prestazioni. Come mostra la tabella 3.1, pur riuscendo a stimare il picco influenzale (con uno scarto di una settimana) in cinque stagioni su nove, il modello tende in alcuni casi a sovrastimare in maniera abbondante l'incidenza del picco.

Riguardo invece alle feature utilizzate, osservando la tabella 3.2, si nota come il modello di Poisson tenda a dare maggior peso a voci di Wikipedia che poco hanno a che fare con i sintomi della patologia influenzale. Inoltre, i pesi assegnati hanno un'entità molto minore rispetto a quelli del modello lineare. Inoltre l'andamento delle feature rispetto al valore dell'incidenza ILI non ha una corrispondenza così netta come nel caso del modello lineare (figura 3.2).

Il modello di Poisson è anche poco resistente agli outliers presenti nel dataset di Wikipedia. Infatti, come si nota dai grafici in figura B, nella quarantacinquesima settimana del 2010, nella sesta settimana del 2011 e nella terza e nona settimana del 2014, il modello stima dei valori di incidenza ILI molto alti rispetto al modello lineare. Questi sono per lo più causati da voci di Wikipedia che in quelle specifiche settimane hanno ricevuto un totale di visite che si discosta in maniera evidente dalla loro media settimanale (probabilmente a causa di alcuni bot).

Anche per il modello di Poisson abbiamo provato ad utilizzare il dataset composto da voci causali e i risultati sono stati gli stessi di quelli ottenuti con il modello lineare. Il modello generalizzato non riesce a prevedere con sufficiente accuratezza né i picchi influenzali e né il normale andamento dell'incidenza ILI se vengono utilizzate pagine casuali.

Stagione Influenzale	Picco Influnet (Settimana)	Picco M (Settimana)	Picco Influnet	Picco M	MSE
Modello Lineare					
2007-2008	5	1	7.21	1.64	11.00
2008-2009	4	4	8.23	6.64	4.43
2009-2010	46	45	12.92	10.89	5.02
2010-2011	5	4	11.04	14.80	7.35
2011-2012	5	5	9.64	7.99	1.29
2012-2013	6	5	9.99	11.95	2.67
2013-2014	6	3	6.67	13.80	7.38
2014-2015	4	4	10.87	5.11	7.18
2015-2016	8	5	6.14	5.59	2.55
Modello di Poisson					
2007-2008	5	1	7.21	1.01	13.02
2008-2009	4	4	8.23	5.78	3.31
2009-2010	46	45	12.92	1090.86	44764.56
2010-2011	5	6	11.04	59.16	147.97
2011-2012	5	5	9.64	8.57	1.27
2012-2013	6	8	9.99	8.45	6.25
2013-2014	6	3	6.67	16.48	6.15
2014-2015	4	5	10.87	4.95	9.73
2015-2016	8	7	6.14	5.23	2.64

Tabella 3.1: Tabella indicante l'MSE e varie informazioni sui picchi influenzali nelle varie stagioni. Con la sigla IN si intende il valore indicato dai dati Influnet, mentre con M si indica il valore predetto dai modelli.

Feature	Peso Medio	Modelli in cui è presente
Modello Lineare		
Febbre	13.11	9/9
Influenza	4.71	9/9
Bronchite	4.64	9/9
Influenzavirus_A_sottotipo_H1N1	3.98	8/9
Polipnea	3.01	9/9
Modello di Poisson		
Pagina principale	0.52	9/9
Febbre ricorrente	0.07	9/9
Vaccino per la febbre tifoide	0.06	2/9
Stimate_(medicina)	0.05	8/9
Virus dell'encefalite di Murray Valley	0.05	2/9

Tabella 3.2: Tabella indicante le feature utilizzate dai modelli ordinate per il loro peso medio.

3.3 Conclusioni

La nostra ipotesi iniziale può considerarsi parzialmente confermata da questi esperimenti. I risultati espressi precedentemente mostrano come l'analisi e la stima dell'incidenza influenzale in Italia sia possibile. Wikipedia è infatti una fonte di dati ed informazioni affidabile che ci permette di stimare con un certo grado di accuratezza la percentuale di popolazione che verrà colpita dalla patologia. Questi dati sono inoltre liberamente e facilmente accessibili ed in un formato facile da analizzare. Tutte queste caratteristiche permettono di poter produrre delle stime di incidenza ILI in Italia in brevissimo tempo, senza incorrere in quelle due-tre settimane di ritardo che i metodi tradizionali di InFluNet comportano. Questo approccio potrebbe essere utilizzato anche per prevedere l'incidenza di altre malattie, sarebbe sufficiente analizzare il traffico su altre voci di Wikipedia, inerenti al morbo di interesse.

Il modello lineare ha evidenziato come esista una relazione tra le pageview delle voci di Wikipedia (che descrivono sintomi e caratteristiche della patologia influenzale) e i dati di incidenza della malattia forniti da InFluNet. Tutto questo pur analizzando un sistema (quello italiano) che presenta sostanziali differenze rispetto al sistema delineato dallo studio di riferimento per questo progetto. Infatti, Wikipedia nella versione inglese possiede un numero maggiore di voci legate all'ambito *Influenza* e un volume molto maggiore di traffico sulle varie pagine (che riduce l'eventuale rumore causato visite "non umane").

Ovviamente, ci sono diverse migliorie che si potrebbero attuare per rendere più efficace il nostro approccio, che però lasceremo per sviluppi futuri. Come prima cosa sarà necessario incorporare nel dataset di Wikipedia i dati riguardanti gli accessi effettuati da dispositivi mobile. Come già affermato nei capitoli precedenti, nei dati utilizzati per questo progetto essi non sono stati considerati. Fortunatamente, negli ultimi due anni Wikipedia si è attivata per fornire dei dataset più rifiniti di quelli che attualmente possiede per l'intervallo 2007-2016. Infatti, essa ha iniziato a offrire dataset sempre più precisi, che comprendono: le visite effettuate da mobile (smartphone, tablets etc.), le visite standard effettuate da desktop e le visite effettuate dall'applicazione mobile di Wikipedia. Inoltre, su questi nuovi dataset viene effettuata anche un'attività di filtraggio per eliminare tutto quel traffico non umano che è presente sulle voci di Wikipedia (spider, crawlers etc.). L'utilizzo di un dataset più raffinato e meno "rumoroso" migliorerebbe sicuramente le prestazioni e ridurrebbe le imprecisione dei modelli.

Sarebbe anche interessante combinare i dati di Wikipedia con un'altra fonte di informazioni per migliorare le capacità predittive dei nostri modelli. Ad esempio, i dati delle prescrizioni mediche potrebbero essere di grande aiuto per stimare i livelli di incidenza ILI in Italia. Intuitivamente, ad un aumento del consumo di certi farmaci corrisponde anche un aumento della popolazione che assume quei farmaci per far fronte a determinati sintomi, ad esempio quelli influenzali. Esistono già servizi che permettono di ottenere queste informazioni, ad esempio *OpenPrescribing* ([https:// openprescribing.net/](https://openprescribing.net/)) è permette di ottenere informazioni sulle prescrizioni mediche effettuate in Inghilterra. Alcune nazioni invece mettono già a disposizione dati simili; ad esempio, la Danimarca fornisce dati sui rimborsi delle prescrizioni mediche ed essi sarebbe una fonte di informazioni ottima per monitorare l'evolversi delle varie patologie nel paese [26].

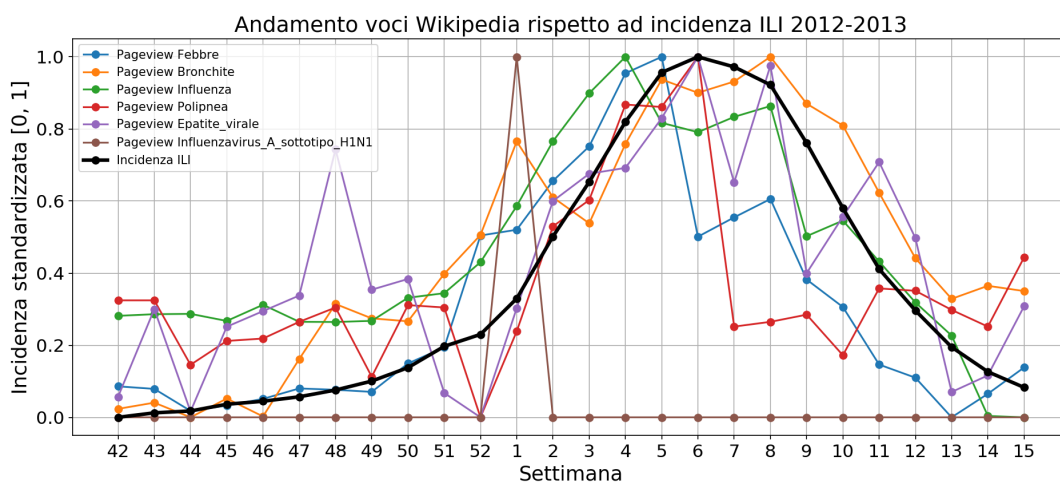


Figura 3.1: *Variazione del valore di alcune feature rispetto all'incidenza ILI (i valori per essere comparati sono stati normalizzati nell'intervallo $[0,1]$).*

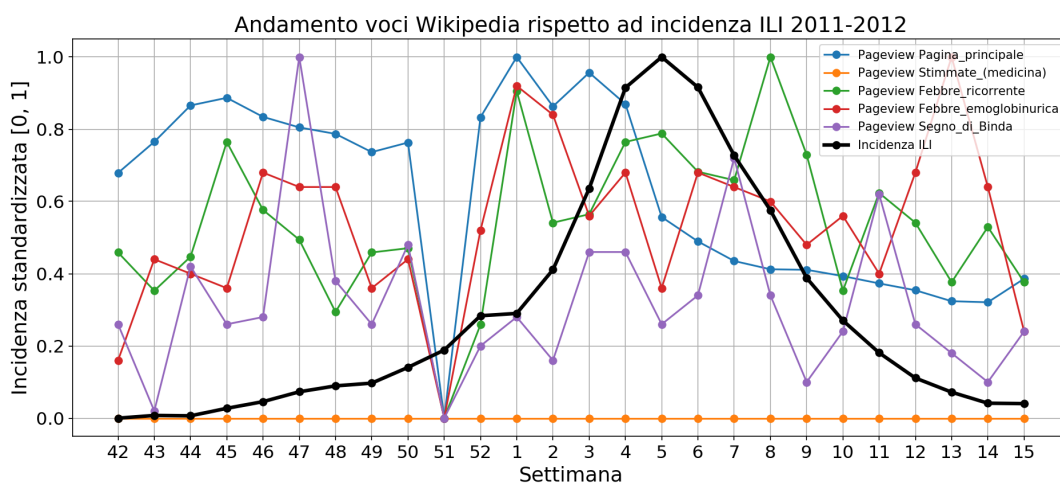


Figura 3.2: *Variazione del valore delle prime 5 feature rispetto all'incidenza ILI (i valori per essere comparati sono stati normalizzati nell'intervallo $[0,1]$).*

Bibliografia

- [1] Who - influenza (seasonal). <http://www.who.int/mediacentre/factsheets/fs211/en/>, November 2016. ultimo accesso 22/08/2017.
- [2] Ecdc - factsheet about seasonal influenza. <https://ecdc.europa.eu/en/seasonal-influenza/facts/factsheet>, 2017. ultimo accesso 26/08/2017.
- [3] Epicentro - aspetti epidemiologici. <http://www.epicentro.iss.it/problemi/influenza/epidItalia.asp>. ultimo accesso 22/08/2017.
- [4] PL Lai et al. Burden of the 1999-2008 seasonal influenza epidemics in italy: Comparison with the h1n1v (a/california/07/09) pandemic. *Hum Vaccin 7 Suppl.*, pages 217–225, 2011.
- [5] Aurelio Sessa, C Lucioni, Gaetano D’Ambrosio, and Germano Bettoncelli. Economic evaluation of clinical influenza in italy. 7:14–20, 01 2005.
- [6] Jeffery K Taubenberger and David M Morens. 1918 influenza: the mother of all pandemics. *Rev Biomed*, 17:69–79, 2006.
- [7] C.W. Potter. A history of influenza. *Journal of Applied Microbiology*, 91(4):572–579, 2001.
- [8] Who - influenza, surveillance and monitoring. http://www.who.int/influenza/surveillance_monitoring/en, 2016. ultimo accesso 25/08/2017.
- [9] Influnet. <http://www.iss.it/ifu>, 2013. ultimo accesso 25/08/2017.
- [10] Flunews - rapporto settimanale. <http://www.epicentro.iss.it/problemi/influenza/FluNews.asp>, 2017. ultimo accesso 25/08/2017.
- [11] David J. McIver and John S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLOS Computational Biology*, 10(4):1–8, 04 2014.
- [12] Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. Forecasting the 2013–2014 influenza season using wikipedia. *PLOS Computational Biology*, 11(5):1–29, 05 2015.
- [13] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. Global disease monitoring and forecasting with wikipedia. *PLOS Computational Biology*, 10(11):1–16, 11 2014.
- [14] Google flu trends. <https://www.google.org/flutrends/about/>, 2014. ultimo accesso 26/08/2017.
- [15] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLOS ONE*, 6(5):1–10, 05 2011.
- [16] Clara Dismuke and Richard Lindrooth. Ordinary least squares. *Methods and Designs for Outcomes Research*, 93:93–104, 2006.
- [17] J. A. Nelder and R. J. Baker. *Generalized Linear Models*. John Wiley & Sons, Inc., 2004.

- [18] Wikipedia statistics - charts italian. <https://stats.wikimedia.org/EN/ChartsWikipediaIT.htm>, 2017. ultimo accesso 26/08/2017.
- [19] Analisi visualizzazioni totali it.wikipedia.org. <https://tools.wmflabs.org/siteviews/?platform=all-access&source=pageviews&agent=user&range=last-year&sites=it.wikipedia.org>, 2017. ultimo accesso 26/08/2017.
- [20] Page view statistics for wikimedia projects. <https://dumps.wikimedia.org/other/pagecounts-raw/>, 2017. ultimo accesso 26/08/2017.
- [21] Petscan. <https://petscan.wmflabs.org/>, 2017. ultimo accesso 01/09/2017.
- [22] Petscan categorie dataset. <https://petscan.wmflabs.org/?psid=983513>, 2017. ultimo accesso 01/09/2017.
- [23] Petscan categorie random. <https://petscan.wmflabs.org/?psid=1190210>, 2017. ultimo accesso 01/09/2017.
- [24] Page views per wikipedia language. <https://stats.wikimedia.org/wikimedia/squids/SquidReport-PageViewsPerLanguageBreakdown.htm>, 2017. ultimo accesso 06/09/2017.
- [25] Page views mobile and desktop. <https://stats.wikimedia.org/EN/SummaryIT.htm>, 2017. ultimo accesso 06/09/2017.
- [26] Sigrun Alba et al. Johannesdottir. Existing data sources for clinical epidemiology: The danish national database of reimbursed prescriptions. *Clinical Epidemiology* 4, page 303–313, 2012.

Allegato A Grafici Modello Lineare

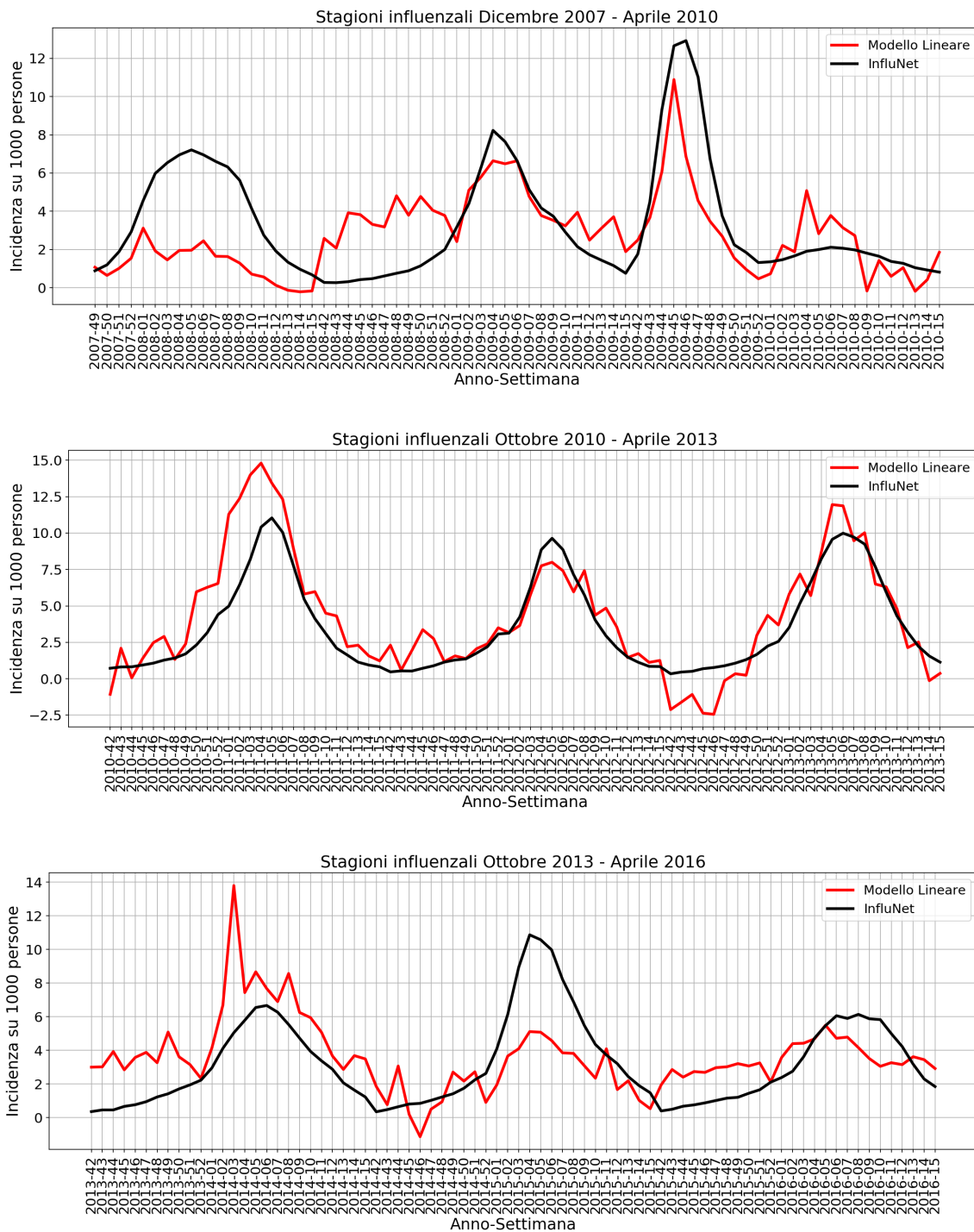


Figura A.1: Confronto tra le curve di incidenza ILI e le curve generate dal modello lineare.

Allegato B Grafici Modello Poisson

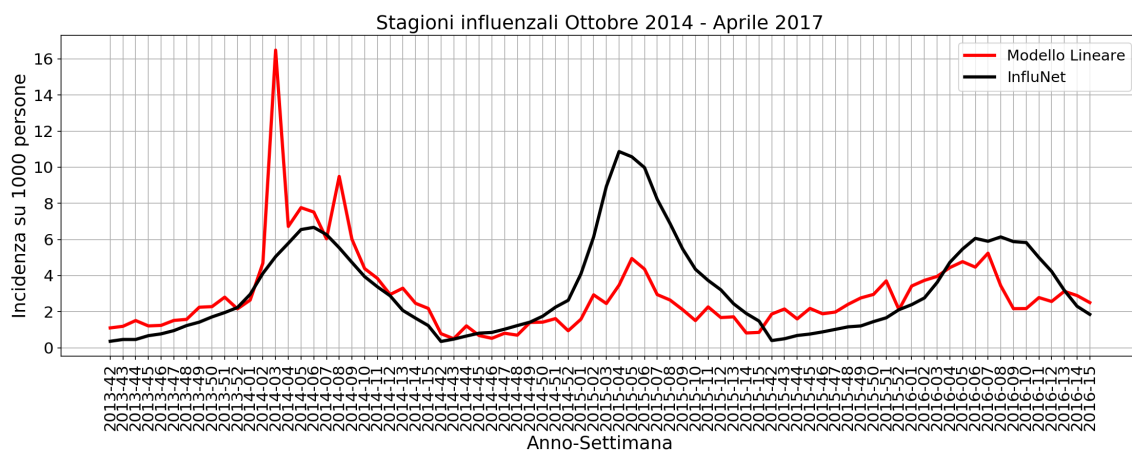
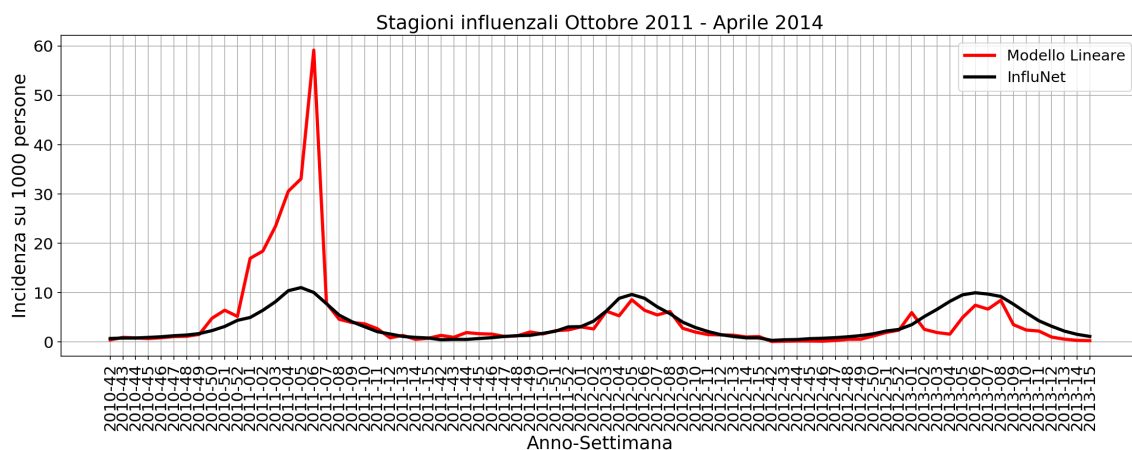
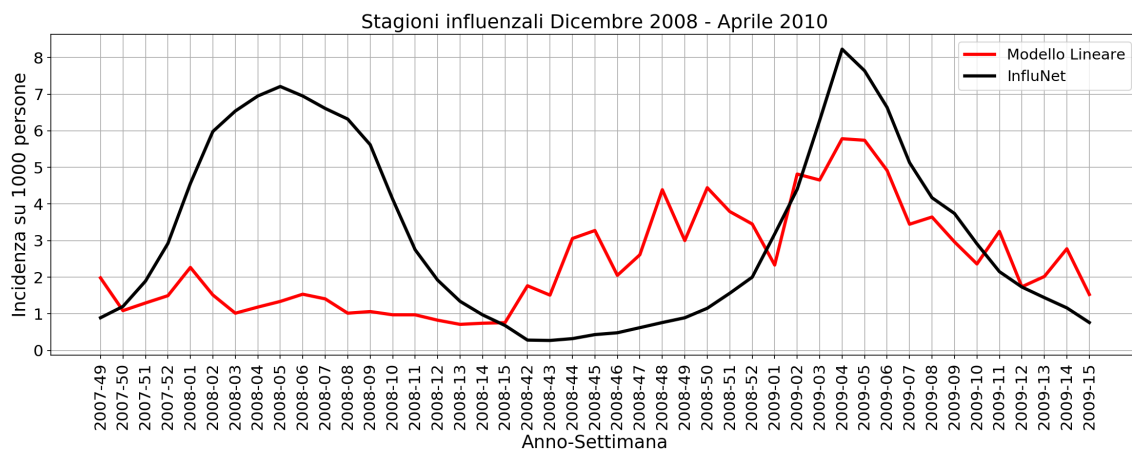


Figura B.1: Confronto tra le curve di incidenza ILI e le curve generate dal modello di Poisson.