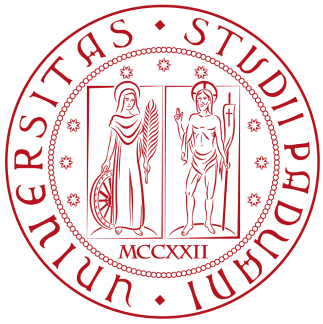
Università degli Studi di Padova

Corso di Laurea in Scienze Statistiche e Tecnologie Informatiche



Modelli Bayesiani Non Parametrici basati sul Processo di Dirichlet

Relatore

Livio Finos

Dipartimento di Scienze Statistiche

Co-relatore

Dario Solari

Candidata

Sally Paganin

Anno Accademico 2011/2012

Indice

Introduzione

Negli ultimi anni i modelli bayesiani di tipo non parametrico hanno ottenuto sempre più attenzione nelle comunità statistiche e in ambiti applicativi relativamente a problemi di classificazione, quali l'apprendimento automatico, l'analisi testuale, la genetica e altro ancora. Uno dei modelli più "popolari" si basa su un particolare processo stocastico, denominato *processo di Dirichlet*.

Il presente lavoro vuole presentarsi come un approfondimento relativo a questo tipo di modelli, esaminando in particolare modo il modello bayesiano non parametrico basato sul processo di Dirichlet.

Nel primo capitolo introdurremo i modelli probabilistici bayesiani in generale, come strumento alternativo nell'analisi dei gruppi. Faremo anche un breve cenno al concetto di osservazioni *interscambiabili* in quanto costituiscono una delle ipotesi alla base delle applicazioni pratiche.

Proseguiremo poi illustrando il processo di Dirichlet, un processo stocastico utilizzato spesso come distribuzione di base di questi modelli. Nell'illustrare le proprietà del processo seguiremo la struttura dell'articolo proposto da (2010).

Nel terzo capitolo descriveremo la struttura generale dei modelli bayesiani non parametrici basati sul processo di Dirichlet. Inoltre proporremo un modello specifico utilizzato nell'ambito dell'analisi testuale.

Il quarto capitolo sarà dedicato ad un applicazione pratica del modello proposto in precedenza, utilizzando il codice MATLAB[1] sviluppato da Yee Whye Teh, su una raccolta di dati estratti da Twitter.

[1]Il codice è scaricabile liberamente da http://www.gatsby.ucl.ac.uk/ ywteh/research/software.html

Capitolo 1

Approccio Bayesiano al clustering

Un problema ricorrente nell'ambito della statistica applicativa, è quello del cosiddetto *clustering* (o analisi dei gruppi): dato un insieme di dati, si vuole far emergere da essi gruppi di unità statistiche simili tra loro e dissimili da quelle degli altri gruppi (si cerca l'*internal cohesion and external isolation*, (1971)). Sulla base delle variabili osservate, ci si pone quindi l'obbiettivo di individuare i diversi gruppi (se esistono) e la loro numerosità. A tale scopo, sono state sviluppate diverse metodologie, come algoritmi di partizione (K-medie) o metodi gerarchici (legame singolo, legame completo, criterio di Ward), ma anche approcci di tipo probabilistico come i modelli di mistura.

1.1 Modelli di mistura

Dato un insieme di dati $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, questo tipo di modelli intende fornire una descrizione di come l'insieme di dati possa essere stato generato: si suppone che le istanze del dataset siano state prodotte a partire da una mistura di distribuzioni di probabilità. I gruppi sono quindi pensati come oggetti appartenenti ad una stessa distribuzione, che si differenziano tra loro in base ai diversi valori assunti dai parametri che caratterizzano tale distribuzione (e.g. Gaussiana con parametri la media e la varianza) e aventi una loro probabilità di essere rappresentati. Definendo le variabili, ad ogni osservazione x_i associamo una variabile casuale z_i indicante il gruppo di appartenenza, ovvero una componente di mistura di parametri θ_{z_i}. Dato

il vettore delle proporzioni di mistura $\pi = \{\pi_{z_i}\}_{z_i=1}^{K}$, il processo generativo descritto dal modello è quindi il seguente: selezionato uno dei K gruppi con probabilità data da π_{z_i}, generiamo x_i dalla componente di mistura corrispondente, che sarà parametrizzata da θ_{z_i}. Possiamo quindi definire la probabilità del nostro dataset, dati i parametri della distribuzione, come

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{n} \sum_{z_i=1}^{K} \pi_{z_i} p(x_i|z_i, \theta_{z_i}) \tag{1.1}$$

dove $p(x_i|z_i, \theta_{z_i})$ indica la densità di probabilità della componente associata al gruppo indicato da z_i . In quanto modello di tipo parametrico, possiamo procedere con l'usuale approccio di stima dei parametri, basato sulla verosimiglianza. Si tratta cioè di trovare una stima di θ tale da massimizzare la probabilità dei dati.

$$\theta^{SMV} = \arg\max_{\theta} p(\mathbf{x}|\theta) \tag{1.2}$$

Per il calcolo della stima di massima verosimiglianza in un modello di mistura, si sfrutta l'algoritmo di *expectation-maximization* (EM). Esso prevede fondamentalmente due passi:

1. Utilizzando delle stime iniziali dei parametri, si valuta, per ogni punto, la probabilità che esso appartenga ad un certo gruppo, condizionata rispetto ai valori correnti dei parametri:

$$P(z_i = k|x_i) = \frac{\pi_k P(x_i|\theta_k)}{\sum_l \pi_l P(x_i|\theta_l)} \tag{1.3}$$

   in cui le z_i sono variabili indicatrici latenti tali che $z_i = k$ esprime l'appartenenza dell'osservazione i al gruppo k.

2. Con le probabilità calcolate al passo precedente si aggiornano le stime.

Questi due passi vengono iterati fino alla convergenza e **?** dimostrarono che la verosimiglianza aumenta ad ogni iterazione. Nel contesto del clustering, la stima ottenuta dei parametri delle distribuzioni definisce i gruppi. Immaginiamo di avere un modello di mistura di distribuzioni normali: ogni cluster è definito da una distribuzione di probabilità $P(x_i|\theta_{z_i}) \sim N(\mu_{z_i}, \Sigma_{z_i})$, e sarà caratterizzato da una certa media e una certa varianza. Stimando i parametri, stimiamo la *forma* dei cluster: la media rappresenta il punto attorno al quale i dati si distribuiscono, mentre la varianza rappresenta il modo in cui

si distribuiscono. Un modello del genere risulta molto più flessibile nell'individuazione dei gruppi, rispetto a soluzioni basate sulla distanza (algoritmo delle K-medie e simili). Nonostante le fondamenta teoriche di questo tipo di modelli siano eccellenti, nel determinare il numero dei gruppi si incorre facilmente nel problema dell'eccessivo adattamento (overfitting), in quanto all'aumentare del numero K di gruppi, aumentano i parametri da stimare. Più un modello è complesso, meglio si adatta ai dati, ma risulta poi non generalizzabile: è necessario quindi trovare un equilibrio tra complessità del modello e la sua generalità. Nella scelta di un modello si ricorre generalmente a metodi che cercano appunto di misurarne la bontà, in termini di complessità e generalità, come ad esempio la validazione incrociata.

Una soluzione elegante al problema della sovrastima, prende in considerazione un approccio di tipo Bayesiano, in cui i parametri e le proporzioni di mistura sono essi stessi delle variabili casuali, definite da una distribuzione a priori. Quindi, dato un insieme di osservazioni $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, si suppone che esse provengano da una distribuzione $p(\mathbf{x}|\theta)$ in cui il parametro θ è sconosciuto, e ha distribuzione $p(\theta)$. Usando la regola di Bayes, si calcola la distribuzione a posteriori, cioè la distribuzione di θ condizionata rispetto i dati:

$$p(\theta|\mathbf{x}) = \frac{p(\theta)p(\mathbf{x}|\theta)}{\int p(\theta)p(\mathbf{x}|\theta)\,d\theta} = \frac{p(\theta)p(\mathbf{x}|\theta)}{p(\mathbf{x})} \tag{1.4}$$

ottenendo così una regola di aggiornamento per i parametri. Solitamente, viene specificata una funzione di verosimiglianza per i dati, $L(\mathbf{x}|\theta)$. La (1.4) può essere riscritta come:

$$p(\theta|\mathbf{x}) = \frac{p(\theta)L(\mathbf{x}|\theta)}{L(\mathbf{x})} \tag{1.5}$$

in cui $L(\mathbf{x})$ rappresenta la funzione di verosimiglianza marginale. Solitamente questa quantità non si riesce ad ottenere in quanto spesso comporta difficoltà computazionali non trascurabili. Per ovviare questa difficoltà, ci si limita a considerare solamente il numeratore, in modo che la distribuzione a posteriori dei parametri risulti proporzionale al prodotto tra la distribuzione a priori e la funzione di verosimiglianza:

$$p(\theta|\mathbf{x}) \propto p(\theta)L(\mathbf{x}|\theta) \tag{1.6}$$

I parametri vengono poi stimati sfruttando gli usuali metodo dell'inferenza bayesiana, come ad esempio i metodi Monte Carlo basati sulle catene di Markov (MCMC) (si veda (2003)). Un modello di questo tipo risulta alquanto flessibile dato che i parametri associati ai dati sono delle variabili casuali. In questo modo non si incorre nel problema dell'overfitting; non c'è motivo perciò, di doversi preoccupare della complessità del modello in termini di numero di parametri da stimare. Per questo motivo, il modello di tipo gerarchico è uno dei modelli fondamentali della statistica bayesiana. L'idea alla base di un modello gerarchico, è definire delle distribuzioni a priori per i parametri, le quali a loro volta possono introdurre nuovi parametri, comunemente indicati come "iperparametri"; anche gli iperparametri possono essere definiti da una distribuzione a priori parametrica, e tale costruzione può ripetersi ricorsivamente, definendo una gerarchia. L'impostazione gerarchica permette di gestire facilmente problemi di classificazione in cui sono coinvolti più insiemi di dati, e vogliamo che i cluster nei diversi insiemi siano condivisi.

Nei modelli bayesiani citati, rimane comunque un elemento che necessita di essere fissato: il numero K dei cluster. Solitamente si usa stimare più modelli per differenti valori di K e sfruttare poi tecniche di selezione o adattamento, ma si tratta di approcci che comportano difficoltà computazionali non trascurabili. Nella prossima sezione, verrà illustrata una tipologia di modelli che ci permettono di ovviare il problema della determinazione dei cluster: i modelli bayesiani non parametrici.

1.2 I modelli bayesiani non parametrici

In generale, un *modello bayesiano non parametrico* è un modello bayesiano dotato di spazio parametrico di dimensione infinita. Tipicamente, dato un problema di classificazione, lo spazio parametrico viene scelto come un'insieme di tutte le possibili soluzioni. Se consideriamo, ad esempio, un problema di regressione, tale spazio può essere rappresentato dall'insieme delle funzioni continue, mentre in un problema di stima della densità considereremmo l'insieme di tutte le densità. Un modello parametrico considera solo un insieme finito di tutte le possibili soluzioni, ed è quindi caratterizzato da un numero fissato di parametri; in particolare il modello bayesiano definisce una distribuzione a priori e una distribuzione a posteriori su un singolo e fissato

spazio parametrico. Affidandoci ad un approccio di tipo non parametrico, viene definito uno spazio parametrico non fissato: si considera sempre un sottoinsieme delle possibili soluzioni, ma tale sottoinsieme può variare secondo la grandezza del campione di dati, appunto perché lo spazio parametrico ha dimensione infinita. La maggior parte dei modelli non parametrici può essere derivata a partire da un modello parametrico standard, portando ad infinito il numero di parametri (i.e mistura infinita di Gaussiane). Generalmente, quando parliamo di spazio parametrico "infinito" in un contesto non parametrico, intendiamo definire uno spazio "finito ma illimitato".

In un problema di clustering, in cui ad ogni cluster è associata una distribuzione di probabilità parametrica, la dimensione dello spazio parametrico è direttamente collegata al numero di cluster; considerare uno spazio parametrico di dimensione infinita significa poter assumere un numero potenzialmente illimitato di gruppi. Tale impostazione ci permette di stimare sia il numero delle componenti in modello di mistura, sia i parametri delle componenti individuali, senza dover confrontare i modelli in maniera esplicita. Esemplifichiamo il concetto, partendo da un generale modello di mistura finito: esso definisce una funzione di densità per un insieme di dati x della forma $p(x) = \sum_{k=1}^{K} \pi_k p(\theta_k|x)$, dove π_k è la proporzione di mistura e θ_k i parametri associati alla componente k. La densità può essere riscritta anche in forma non standard, come un integrale: $p(x) = \int p(x|\theta)G(\theta)\, d(\theta)$, dove $G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$ è una distribuzione discreta di mistura che incapsula tutti i parametri del modello di mistura, mentre δ_{θ_k} rappresenta una distribuzione di Dirac centrata su θ (un punto). Le misture nei modelli bayesiani non parametrici, sono invece distribuzioni costituite da una quantità di punti illimitata ma numerabile:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \tag{1.7}$$

Sulla base di tali distribuzioni, nascono i modelli di mistura con numero infinito di componenti. Nell'applicazione pratica però, abbiamo un dataset finito e di conseguenza utilizzeremo nella stima del modello, un numero finito (ma variabile) di componenti di mistura per rappresentare i dati, in modo che ogni osservazione sia associata ad esattamente una componente, ma ad ogni componente possano corrispondere osservazioni multiple. Volendo poi definire il modello con l'aggettivo "bayesiano", dobbiamo definire per G una distribuzione a priori. Le funzioni e le misure casuali, e più in generale

le distribuzioni di probabilità su un oggetto casuale di dimensioni infinite, sono detti *processi stocastici*. Il particolare processo stocastico che vogliamo prendere in considerazione nella definizione e nell'utilizzo di un modello non parametrico, è il *processo di Dirichlet*, del quale vedremo le caratteristiche nel prossimo capitolo. Prima però vogliamo soffermarci brevemente su uno dei concetti alla base non solo dei modelli parametrici, ma dei bayesiani in generale.

1.3 Cenni sulle basi teoriche

L'assunzione di base di tutti i metodi bayesiani riguarda il parametro associato alle osservazioni: si assume che esso sia una variabile casuale. Tale ipotesi è risultata l'oggetto di diverse critiche. Possiamo addirittura affermare che essa rappresenti il "cuore" del dibattito tra bayesiani e non-bayesiani, che da tempo ha diviso la comunità statistica. Nonostante i dibattiti, esiste un tipo talmente generale di osservazioni per le quali l'esistenza di tale variabile casuale può essere derivata matematicamente: le osservazioni *interscambiabili*. Per questo tipo di osservazioni, l'assunzione bayesiana riguardante l'esistenza di un parametro con distribuzione casuale non è una prerogativa del modello, ma una conseguenza matematica delle proprietà dei dati. Ci limitiamo quindi a definire in questa sezione, il concetto di *interscambiabilità*, in quanto risulterà una delle assunzione di base nei modelli che esporremo.

Formalmente, una sequenza di variabili $X_1, X_2, \ldots, X_n$ definite sullo stesso spazio di probabilità $(\mathcal{X}, \Omega)$ è *interscambiabile* se la loro distribuzione congiunta è invariante alla permutazione delle variabili stesse. Cioè, se P è la distribuzione congiunta e σ rappresenta qualsiasi permutazione di $\{1, \ldots, n\}$, allora

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_{\sigma(1)}, X_2 = x_{\sigma(2)}, \ldots, X_n = x_{\sigma(n)}) \tag{1.8}$$

Una sequenza infinita di variabili $X_1, X_2, \ldots$ è *infinitamente interscambiabile* se $X_1, \ldots, X_n$ se è interscambiabile per *ogni* $n \geq 1$. Per comodità, d'ora in avanti ci limiteremo a scrivere "interscambiabile" intendendo "infinitamente interscambiabile". Tipicamente nelle applicazioni statistiche, si assume che osservazioni godano della proprietà di interscambiabilità: significa che le

corrispondenti variabili non dipendo dal proprio indice, nonostante possano avere una qualche forma di dipendenza tra loro. Considerare le variabili interscambiabili è differente dall'assumerle indipendenti e identicamente distribuite (iid). L'interscambiabilità è un'ipot molte più debole: le variabili iid sono automaticamente interscambiabili. Piuttosto possiamo interpretare l'interscambiabilità per le variabili, con il significato di"*condizionatamente* indipendenti e identicamente distribuite", in cui il "condizionatamente" si intende rispetto al parametro latente della distribuzione di probabilità sottostante. Se infatti, nel contesto descritto, θ parametrizza la distribuzione sottostante, e viene assunta una distribuzione a priori su tale parametro, allora la distribuzione marginale risultante di $X_1, X_2, \ldots$ rispetto a θ risulterà ancora interscambiabile. Un risultato fondamentale accreditato a (1931), dimostra che risulta vero anche il contrario. Cioè, se $X_1, X_2, \ldots$ è una sequenza (infinitamente) interscambiabile, allora esiste un qualche modello parametrico $P(X_i|\theta)$ con un qualche parametro θ tale che:

$$P(X_1, \ldots, X_n) = \int_\Theta P(\theta) \prod_{i=1}^{n} P(X_i|\theta)\, d\theta \tag{1.9}$$

In parole povere, il teorema di De Finetti afferma che se una sequenza è interscambiabile allora ogni suo sottoinsieme finito può essere considerato un campione casuale del modello $P(X_i|\theta)$ e quindi esiste una distribuzione a priori per θ che giustifica l'approccio bayesiano. L'assunzione di interscambiabilità sembra implicare automaticamente l'esistenza di un modello bayesiano con θ parametro latente casuale. In questo senso, l'interscambiabilità costituisce parte delle fondamenta della statistica bayesiana.

Capitolo 2

Processo di Dirichlet

Il *processo di Dirichlet* (Dirichlet Process, DP) è un processo stocastico dal quale deriva una classe di distribuzioni di probabilità largamente usata come priori nei modelli statistici bayesiani di tipo non parametrico, in particolare nei modelli di mistura basati sul processo di Dirichlet (conosciuti anche come modelli di mistura infiniti). Generalmente possiamo considerare i processi stocastici come distribuzioni su spazi di funzioni: nel caso del DP, si parla di distribuzione su misure di probabilità, le quali sono funzioni con caratteristiche tali da permetterci di interpretarle come distribuzioni su uno qualche spazio probabilistico. In questo senso il DP è una "distribuzione di distribuzioni". Vedremo più avanti che le distribuzioni estratte sono di tipo di discreto, ma non rappresentabili attraverso un numero finito di parametri, perciò il modello che ne deriva è di tipo non parametrico. Il processo di Dirichlet venne formalizzato da (1973), mentre il nome deriva dal fatto che le sue distribuzioni marginali finite sono distribuite come una variabile casuale di Dirichlet.

2.1 Definizione

Sia $(\Theta, \mathcal{B}, G)$ uno spazio di probabilità, in cui Θ è lo spazio campionario, $\mathcal{B}$ la σ-algebra di Borel dei sottoinsiemi di Ω e G una misura di probabilità. Sia H una distribuzione di probabilità su tale spazio, e α un numero reale positivo. Allora G si distribuisce secondo un processo di Dirichlet se, per ogni partizione finita e misurabile $(A_1, A_2, \ldots, A_r)$ di Θ, il vettore casuale $(G(A_1), G(A_2), \ldots, G(A_r))$ è distribuito come una variabile casuale

di Dirichlet con parametri di distribuzione $(\alpha H(A_1), \alpha H(A_2), \ldots, \alpha H(A_r))$, ovvero

$$(G(A_1), \ldots, G(A_r)) \sim Dir((\alpha H(A_1), \ldots, \alpha H(A_r))) \tag{2.1}$$

in cui $Dir()$ indica la distribuzione di Dirchlet (vedi appendice A), e scriviamo $G \sim DP(\alpha, H)$.

I parametri H e α hanno un ruolo abbastanza intuitivo nella definizione del DP; H é detta distribuzione di base ed è la media del DP, mentre α è il parametro di concentrazione e può essere pensato come una varianza inversa. Dalla definizione del processo e dalle proprietà della distribuzione di Dirichlet, abbiamo infatti:

$$G(A) \sim Beta(\alpha G(A), \alpha(1 - G(A)), \quad \forall A \in \mathcal{B} \tag{2.2}$$

Ricordando i momenti della distribuzione beta, si può facilmente verificare che:

$$E[G(A)] = H(A) \qquad V[G(A)] = \frac{H(A)(1 - H(A))}{(\alpha + 1)} \tag{2.3}$$

Notiamo perciò che, maggiore è il parametro α, minore è la varianza, e quindi il processo concentrerà la maggior parte della propria massa attorno alla sua stessa media. Il parametro di concentrazione è definito anche *parametro di forza*, in riferimento alla "forza" della priori quando il DP è utilizzato come distribuzione a priori in un modello Bayesiano non parametrico. Dal momento che α descrive la concentrazione di massa intorno alla media del processo, per $\alpha \to \infty$ avremo $G(A) \to H(A)$ per ogni misurabile A, cioè $G \to H$ puntualmente. Mettiamo in luce che non risulta affatto equivalente scrivere $G \to H$. Come anticipato, le estrazioni da un DP risultano essere distribuzioni di tipo discreto con probabilità uno, anche se H è continua. Perciò G e H non hanno bisogno di essere assolutamente continue una rispetto all'altra.

2.2 Distribuzione a posteriori

Considerato l'utilizzo del processo di Dirichlet in una cornice bayesiana, risulta utile analizzarne la distribuzione a posteriori. Dato che G è una distribuzione casuale, possiamo estrarre dei campioni da G stessa. Sia quindi $\theta_1, \ldots, \theta_n$ una sequenza di campioni indipendenti da G. Ricordiamo che

essendo G una distribuzione su uno spazio Θ, i θ_i assumeranno valori in tale spazio. Siamo quindi interessati alla distribuzione a posteriori di G dati i valori osservati di $\theta_1, \ldots, \theta_n$. Sia $A_1, \ldots, A_r$ una partizione misurabile di Θ, e sia $n_k = \#\{i : \theta_i \in A_k\}$ il numero di valori osservati in A_k. Dalla (2.1) e dalla connessione tra la distribuzione di Dirichlet e quella multinomiale abbiamo:

$$(G(A_1), \ldots, G(A_r))|\theta_1, \ldots, \theta_n \sim Dir(\alpha H(A_1)+n_1, \ldots, \alpha H(A_r)+n_r) \quad (2.4)$$

per tutte le partizioni finite e misurabili. Significa quindi che la distribuzione a posteriori di G è anch'essa un processo di Dirichlet, con parametro di concentrazione $\alpha+n$. Per calcolare la distribuzione di base ci basta calcolare il valore atteso del processo, come fatto in precedenza. Abbiamo che

$$P(G(A)|\theta_1, \ldots, \theta_n) \sim Beta(\alpha H(A) + n_k, (\alpha + n) - (\alpha H(A) + n_k)) \quad (2.5)$$

e perciò

$$E(G(A)|\theta_1, \ldots, \theta_n) = \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n} \quad (2.6)$$

dove δ_i rappresenta un punto di massa localizzato in θ_i e $n_k = \sum_{i=1}^{n} \delta_i(A_k)$. In altre parole il DP fornisce una famiglia di distribuzioni a priori (sulle distribuzioni) coniugate che è *chiusa* rispetto agli aggiornamenti a posteriori dei parametri, date le osservazioni. Riscrivendo la distribuzione a posteriori, abbiamo

$$G|\theta_1, \ldots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right) \quad (2.7)$$

Possiamo notare che la distribuzione di base nella a posteriori è una media pesata tra la distribuzioni di base a priori H e la distribuzione empirica $\frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}$. Il peso associato alla distribuzione di base a priori è proporzionale al parametro di concentrazione α, mentre la distribuzione empirica è pesata proporzionalmente al numero di osservazioni n. Vedremo tra poco che la distribuzione di base a posteriori è anche la distribuzione predittiva di θ_{n+1} dati $\theta_1, \ldots, \theta_n$. Infatti per $\alpha \to 0$, la priori diventa non informativa, nel senso che che la distribuzione predittiva è data soltanto dalla distribuzione empirica; d'altra parte, se il numero delle osservazioni diventa molto grande, $n \gg \alpha$, la distribuzione a posteriori è dominata da quella empirica, la

quale è a sua volta una stretta approssimazione della vera distribuzione di fondo. In ciò consiste la proprietà di consistenza del processo di Dirichlet: la distribuzione a posteriori del processo si avvicina alla vera distribuzione sottostante. Vediamo ora come ricavare la distribuzione predittiva.

2.3 Distribuzione predittiva

Consideriamo ancora di estrarre da $G \sim DP(\alpha, H)$ una sequenza i.i.d. $\theta_1, \theta_2, \cdots \sim G$. Vogliamo trovare la distribuzione predittiva per θ_{n+1} condizionata da $\theta_1, \ldots, \theta_n$ e marginalizzata (integrata) rispetto a G. Dato $\theta_{n+1}|G, \theta_1, \ldots, \theta_n \sim G$, per ogni misurabile $\mathcal{A} \subset \Theta$, abbiamo

$$\begin{aligned} P(\theta_{n+1} \in \mathcal{A}|\theta_1, \ldots, \theta_n) &= E[G(\mathcal{A})|\theta_1, \ldots, \theta_n] \\ &= \frac{\alpha}{\alpha + n} H(\mathcal{A}) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}(\mathcal{A}) \end{aligned} \tag{2.8}$$

dove l'ultimo passaggio segue dalla distribuzione di base a posteriori date le prime n osservazioni. Integrando poi rispetto a G abbiamo:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} H + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i} \tag{2.9}$$

Perciò la distribuzione di base a posteriori dati $\theta_1, \ldots, \theta_n$ è anche la distribuzione di previsione di θ_{n+1}. La sequenza di distribuzioni predittive (2.9) per $\theta_1, \theta_2, \ldots$ può essere interpretata in termini di un semplice modello d'urna, conosciuto come *Pòlya urn model*. Consideriamo ogni valore di Θ come un diverso colore, e le estrazioni $\theta \sim G$ come palline il cui valore di estrazione è il colore. Inoltre abbiamo un'urna che contiene le palline già estratte precedentemente. All'inizio, ovviamente, l'urna è vuota e quindi estraiamo un colore dalla distribuzione di base H, i.e. $\theta_1 \sim H$, dipingiamo una pallina di quel colore, e la mettiamo nell'urna. Nei passi successivi, tipo all' $n+1$-mo, avremo due possibilità: possiamo, con probabilità $\frac{\alpha}{\alpha+n}$ prendere un nuovo colore (cioè estrarre $\theta_{n+1} \sim H$) dipingere una pallina con quel colore e metterla nell'urna, oppure, con probabilità $\frac{n}{\alpha+n}$, estrarre casualmente dall'urna una pallina (estraggo θ_{n+1} dalla distribuzione empirica), dipingere una nuova pallina dello stesso colore e rimettere entrambe le palline nell'urna.

2.4 Clustering e processo del ristorante cinese

La formula (2.9) mostra non solo che un'estrazione da G ha una probabilità positiva di assumere un valore uguale ad una delle precedenti estrazioni, ma che vi è un effetto positivo di rafforzamento: quanto più spesso un valore è estratto, tanto è più probabile che venga estratto in futuro. Si tratta di un fenomeno per cui "i ricchi si arricchiscono" derivanti dalla proprietà di clustering (raggruppamento) del processo di Dirichlet. Per rendere questa proprietà esplicita, ci risulta utile introdurre un insieme di nuove variabili che rappresentano i valori distinti dei punti. Siano quindi $\theta_1^*, \ldots, \theta_m^*$ i valori unici tra $\theta_1, \ldots, \theta_n$, e n_k il numero di ripetizioni di θ_k^*. La distribuzione predittiva può essere scritta in maniera equivalente come:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} H + \frac{1}{\alpha + n} \sum_{k=1}^{m} n_k \delta_{\theta_k^*} \tag{2.10}$$

Possiamo notare che i valori θ_k^* saranno ripetuti da θ_{n+1} con una probabilità proporzionale a n_k, cioè il numero di volte in cui sono già stati osservati. Maggiore è n_k, maggiore è la probabilità che esso cresca. Si tratta di un fenomeno in cui i "i ricchi si arricchiscono", ovvero in cui i gruppi molto grandi (consideriamo come gruppo un insieme di θ_i con valori identici θ_k^*) crescono con velocità maggiore. Possiamo procurarci una visione più profonda della proprietà di raggruppamento del DP, dando uno sguardo alla partizioni indotte dal clustering. I valori unici di $\theta_1, \ldots, \theta_n$ inducono alla partizione dell'insieme $[n] = \{1, \ldots, m\}$ in gruppi tali che, all'interno di ogni gruppo, ad esempio il gruppo k, i θ_i assumono lo stesso valore θ_k^*. Dato che la sequenza di $\theta_1, \ldots, \theta_n$ è casuale, anche la partizione di $[n]$ è casuale. Tale partizionamento casuale incapsula di fatto tutte le proprietà del DP. La distribuzione delle partizioni viene comunemente definita come *Chinese restaurant process* (CRP), tradotto "processo del ristorante cinese". Si tratta di una metafora in cui supponiamo di avere un ristorante cinese con un numero illimitato di tavoli, ad ognuno dei quali può sedersi un numero infinito di clienti. Il primo cliente entra nel ristorante e si siede al primo tavolo. Il secondo cliente entra e può decidere se sedersi con il primo cliente, oppure da solo ad un nuovo tavolo. In generale, l'$n + 1$-mo cliente può scegliere se unirsi ad un tavolo k già occupato, con una probabilità proporzionale al numero n_k di clienti già seduti a tale tavolo, oppure sedersi ad un nuovo tavolo

con una probabilità proporzionale ad α. Identificando i clienti con numeri interi $1, 2, \ldots$ e i tavoli come gruppi, dopo che n clienti hanno preso posto ad un tavolo, essi definiscono un partizionamento di $[n]$ con una distribuzione che è la stessa di quella esposta sopra. Un altro aspetto da considerare, è la distribuzione del numero dei gruppi tra le n osservazioni. Notiamo che per $i \geq 1$, l'osservazione θ_i assume un nuovo valore (creando quindi un nuovo gruppo) con probabilità $\frac{\alpha}{\alpha+i-1}$ indipendentemente dal numero di gruppi tra i θ precedenti. Calcolando la media e la varianza della distribuzione si può mostrare che, per $n > \alpha \gg 0$, tali quantità risultano approssimabili con $\alpha \log(1 + \frac{n}{\alpha})$: il numero di gruppi cresce in scala logaritmica rispetto al numero di osservazioni. Questa lenta crescita è in accordo con il fenomeno dei "ricchi si arricchiscono": ci aspettiamo infatti di avere dei gruppi molto grandi, perciò il numero di gruppi m deve essere minore del gruppo di osservazioni n. Inoltre il parametro α controlla i gruppi in maniera diretta: un α molto grande indica un gran numero di gruppi a priori. Tale osservazione risulterà utile nell'applicazione del DP nei modelli di mistura. Per comprendere questo tipo di modelli, dobbiamo analizzare un ultima proprietà del processo di Dirichlet: la proprietà di discretezza.

2.5 Costruzione stick-breaking

Abbiamo già notato che le estrazioni da un processo di Dirichlet, $G \sim Dir(\alpha, H)$, sono discrete con probabilità uno. Abbiamo anche visto che tali estrazioni hanno una probabilità positiva di assumere lo stesso valore. Significa quindi che possiamo considerare la distribuzione di G come una somma pesata di punti di massa, i.e. una distribuzione discreta. Possiamo averne una rappresentazione più esplicita, fornendo una definizione costruttiva del DP, chiamata *stick-breaking construction* introdotta da (1994).

Date delle sequenze indipendenti di variabili casuali i.i.d. $(\beta_k)_{k=1}^{\infty}$ e $(\theta_k^*)_{k=1}^{\infty}$

$$\beta_k \sim Beta(1, \alpha) \qquad \theta_k^* \sim H$$

si definisce G come segue:

$$\pi_k \sim \beta_k \prod_{l=1}^{k-1}(1 - \beta_k) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \tag{2.11}$$

(1994) dimostrò che G, così definita, si comporta come una misura di probabilità distribuita secondo un processo di Dirchlet, $DP(\alpha, H)$. La costruzione di π può essere compresa metaforicamente come segue. Iniziamo con un bastoncino di lunghezza unitaria, lo spezziamo ad altezza β_1 e definiamo come π_1 la lunghezza del bastoncino che abbiamo appena rotto. Ora spezziamo ricorsivamente l'altra porzione ottenendo π_2, π_3 e così via. Notiamo che la sequenza $\pi = \{\pi_1, \pi_2, \dots\}$ soddisfa $\sum_{k=1}^{\infty} \pi_k = 1$ con probabilità uno, e possiamo quindi considerare π come una misura di probabilità su interi positivi. Per convenienza, si usa indicare definire tale distribuzione come $\pi \sim GEM(\alpha)$ (GEM sta per Griffiths, Engen, MacCloskey). La semplicità di questa rappresentazione ha portato ad una varietà di estensioni, insieme a nuove tecniche di inferenza, del processo di Dirichlet.

Capitolo 3

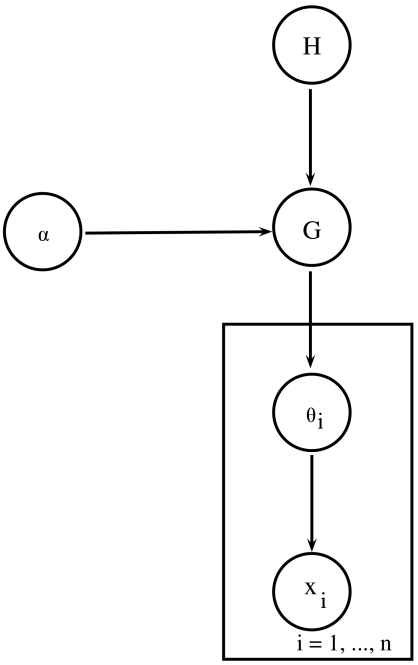
Applicazioni del DP nei modelli probabilistici

Abbiamo illustrato il processo di Dirichlet e le sue proprietà, in quanto esso rappresenta la distribuzione a priori maggiormente usata, nel contesto dei modelli bayesiani non parametrici. Nonostante il fatto che le estrazioni da un DP siano distribuzione di tipo discreto possa da una parte sembrare un limite, dall'altra parte tale proprietà di discretezza risulta fondamentale in alcuni campi applicativi. Vedremo infatti in questo capitolo, dopo una descrizione generale dei modelli derivanti, un esempio specifico collocato nell'ambito dell'analisi testuale.

3.1 Modello di mistura basato sul DP

Dato un insieme di osservazioni $\{x_1, \ldots, x_n\}$, associamo ognuna di essere a dei parametri latenti, $\{\theta_1, \ldots, \theta_n\}$, i quali sono dotati di una distribuzione a priori G distribuita secondo un processo di Dirichlet $DP(\alpha, H)$. Ogni θ_i quindi, rappresenta un estrazione indipendente e di forma identica da tale processo, mentre ogni x_i ha distribuzione $F(\theta_i)$ parametrizzata appunto da θ_i. Riassumendo, in un modello basato sul DP vengono definite le seguenti distribuzioni condizionate:

$$\begin{aligned} x_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G|\alpha, H &\sim DP(\alpha, H) \end{aligned} \tag{3.1}$$



Possiamo vedere nella figura appena sopra il corrispondente modello grafico rappresentate le dipendenze tra le variabili. Come abbiamo visto nel capitolo precedente, essendo G una distribuzione discreta, i diversi θ_i possono assumere gli stessi valori, e il modello appena esposto può essere visto come un modello di mistura, in cui le osservazioni x_i con gli stessi valori delle variabili θ_i provengono dalla stessa componente di mistura, e appartengono quindi allo stesso cluster. Possiamo adeguare il modello all'usuale rappresentazione per i modelli di mistura, usando la rappresentazione del processo di Dirichlet basata sulla costruzione "stick-breaking". Detoniamo quindi con z_i la variabile di assegnamento al gruppo, la quale assume valore k con probabilità π_k. Allora la 3.1 può essere espressa equivalentemente come

$$\begin{aligned} \boldsymbol{\pi}|\alpha &\sim GEM(\alpha) & \theta_k^*|H &\sim H \\ z_i|\boldsymbol{\pi} &\sim Mult(\pi) & x_i|z_i, \{\theta_k^*\} &\sim F(\theta_{z_i}^*) \end{aligned} \tag{3.2}$$

con $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$ e $\theta_i = \theta_{z_i}^*$. Nella terminologia dei modelli di mistura, $\boldsymbol{\pi}$ rappresenta il vettore delle proporzioni di mistura, i θ_k^* sono i parametri dei gruppi, $F(\theta_k^*)$ è la distribuzione dei dati del gruppo k, mentre H rappresenta la distribuzione a priori sui parametri dei gruppi.

Il modello basato sul processo di Dirichlet, rientra nella classe dei modelli bayesiani non parametrici: possiamo pensarlo come un modello di mistura con un numero di gruppi numerabile e illimitato. Tuttavia, data la decrescita esponenzialmente rapida dei π_k, solo un piccolo numero di gruppi verrà usato per modellare i dati a priori (abbiamo infatti visto prima che il numero atteso di componenti usate a priori, è proporzionale logaritmicamente al numero di osservazioni). A differenza dei modelli di mistura finiti, il numero di gruppi non è fissato, e può essere ricavato dai dati tramite gli usuali metodi di inferenza bayesiana (MCMC). L'equivalente operazione nei modelli di mistura finiti comporta l'utilizzo di metodi che risultano ricchi di difficoltà. Per questo motivo i modelli infiniti provvedono ad una convincente alternativa all'approccio tradizionale, in quanto rimangono comunque collegati ai modelli finiti. Partendo da una sequenza di modelli di mistura finiti, possiamo derivare il modello basato sul processo di Dirichlet come *limite* di questa sequenza, cioè portando ad infinito il numero di componenti di mistura.

Supponiamo di avere K componenti di mistura (equivalentemente, K gruppi). Sia poi $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ il vettore delle proporzioni di mistura. Definiamo per $\boldsymbol{\pi}$ una distribuzione a priori di Dirichlet con parametri simmetrici $(\alpha/K, \ldots, \alpha/K)$. Sia θ_k^* il vettore di parametri associati alla componente di mistura k, con distribuzione a priori H. Estrarre un osservazione x_i dal modello di mistura appena descritto, significa scegliere una specifica componente di mistura k con probabilità dalle proporzioni in π_k; denotiamo quindi con z_i tale componente. Abbiamo perciò il seguente modello:

$$\begin{aligned} \boldsymbol{\pi}|\alpha &\sim Dir(\alpha/K, \ldots, \alpha/K) & z_i|\pi &\sim \boldsymbol{\pi} \\ \theta_k^*|H &\sim H & x_i|z_i, \{\theta_k^*\} &\sim F(\theta_k^*) \end{aligned} \tag{3.3}$$

Sia $G^K = \sum_{k=1}^K \pi_k \delta_{\theta_k^*}$. (2002) hanno dimostrato che, per ogni misurabile funzione f integrabile rispetto a G abbiamo, per $L \to \infty$

$$\int f(\theta)\, dG^K(\theta) \xrightarrow{\mathcal{D}} \int f(\theta)\, dG(\theta) \tag{3.4}$$

Di conseguenza la distribuzione marginale indotta sulle osservazioni $x_1, \ldots, x_n$ si avvicina a quella di un modello di mistura basato sul processo di Dirichlet.

3.2 Estensione gerarchica

L'impostazione bayesiana permette l'estensione dei modelli di mistura, in modelli gerarchici: lo stesso vale per i modelli di mistura basati sul processo di Dirichlet. L'estensione del DP in "livelli di gerarchie", ci permette di affrontare anche problemi di classificazione in cui abbiamo a disposizione più insiemi di dati e vogliamo che i cluster nei diversi insiemi siano condivisi. L'idea base di tale estensione, consiste nel considerare come distribuzione di base di un processo di Dirichlet, $G \sim DP(\alpha, G_0)$ un altro processo di Dirichlet, $G_0 \sim DP(\gamma, H)$. Questa costruzione ricorsiva costringe G ad avere un supporto discreto determinato da G_0, e viene comunemente definita come *processo di Dirichlet gerarchico*.

Consideriamo J insiemi di osservazioni. Denotiamo con x_{ij} l'osservazione i-ma del j-mo gruppo e con θ_{ij} il parametro latente associato. Per ogni insieme j definiamo un processo di Dirichlet G_j. Abbiamo quindi una collezione indicizzata di processi $\{G_j\}$ definiti su di un comune spazio di probabilità. Il processo gerarchico lega, da un punto di vista probabilistico, queste misure casuali, permettendo loro di condividere la distribuzione di base e assumendo casuale tale distribuzione abbiamo:

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H) \\ G_j|\alpha, G_0 &\sim DP(\alpha, G_0) \qquad j = 1, \ldots, J \end{aligned} \tag{3.5}$$

in cui G_0 è il processo Dirichlet "padre" dei diversi G_j, e la sua distribuzione di base H è la distribuzione a priori dei parametri θ_{ij}. In questo modo è possibile la condivisione dei punti massa tra le misure casuali G_j in quanto ognuna di esse eredita l'insieme di atomi dallo stesso processo. Possiamo estendere il modello nella 3.3 nel seguente modo:

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H) \\ G_j|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ x_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned} \tag{3.6}$$

Per completezza, si riporta la rappresentazione del modello basata sulla costruzione "stick-breaking" del processo di Dirichlet, senza dimostrare i

diversi passaggi, per i quali si rimanda a (2010). Dati

$$\theta_k^*|H \sim H \qquad\qquad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^*} \tag{3.7}$$

con $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \sim GEM(\gamma)$. Essendo $G_j \sim DP(\alpha, G_0)$ avrà lo stesso supporto di G_0 sui punti $\boldsymbol{\theta} = (\theta)_{k=1}^{\infty}$ e può essere rappresentato come

$$G = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^*} \tag{3.8}$$

Sia $\boldsymbol{\pi_j} = (\pi_j k)_{k=1}^{\infty}$; (2010) mostrano che $\boldsymbol{\pi_j}$ si distribuisce come un processo di Dirichlet, $DP(\alpha, \boldsymbol{\beta})$ e danno la seguente rappresentazione del modello in 3.6:

$$\begin{aligned} \boldsymbol{\beta}|\gamma &\sim GEM(\gamma) & \theta_k^*|H &\sim H \\ \boldsymbol{\pi_j}|\alpha, \beta &\sim DP(\alpha, \beta) & & \\ z_{ji}|\boldsymbol{\pi_j} &\sim \boldsymbol{\pi_j} & x_{ji}|z_{ji}, \{\theta_k^*\} &\sim F(\theta_{z_{ji}}) \end{aligned} \tag{3.9}$$

dove z_{ji} è la variabile di assegnamento al gruppo associata all'osservazione x_{ji}, mentre θ_k^* è il parametro che caratterizza la componente di mistura k, e di conseguenza anche il cluster. Analogamente al modello in 3.1, le osservazioni x_{ji} associate a valori uguali di θ_{ji} saranno generate dalla stessa distribuzione, anche se tali osservazioni appartengono a gruppi diversi.

Per riuscire a comprendere la natura precisa della condivisione, risulta d'aiuto considerare una rappresentazione analoga al CRP, ovvero il "Chinese Reastaurant Franchise" (CRF). Nel CRF la metafora generativa del ristorante cinese, viene estesa ad un franchise di ristoranti cinesi, uno per ogni indice j. I clienti del j-mo ristorante si siedono ai tavoli nella stessa maniera descritta nel CRP, e ciò accade in maniera indipendente in ogni ristorante. L'associazione tra ristoranti avviene per mezzo di un vasto menù condiviso dai diversi ristoranti. Il primo cliente a sedersi in un tavolo di un ristorante sceglierà un piatto dal menù, e tutti i seguenti clienti che siederanno a quel tavolo condivideranno tale piatto. I diversi tavoli nei vari ristoranti potranno servire lo stesso piatto. In questa impostazione i ristoranti corrispondono ai diversi gruppi, e i clienti corrispondono alle osservazioni x_{ji}. L'insieme dei clienti condivide un menù in cui $\theta_1^*, \ldots, \theta_k^*$ sono i piatti. Introduciamo una

variabile indicatrice per i piatti, ψ_{jt}, la quale rappresenta il piatto servito al tavolo t nel ristorante j. Un cliente x_{ji} entra nel ristorante j e sceglie di sedersi al tavolo t_{ji} con gli n_jt cliente con una probabilità uguale a $\frac{n_{jt}}{\alpha+i-1}$ e condividere il piatto ψ_{jt}; oppure sceglie di sedersi ad un nuovo tavolo, t^{new} con probabilità $\frac{\alpha}{\alpha+i-1}$ è ordinare un nuovo piatto (cluster) ψ_{jt}^{new} dal menù generale. All'interno di ogni ristorante i clienti si siedono ai tavoli nella stessa maniera descritta nel CRP, definendo un partizionamento descritto dalla distribuzione condizionata dei θ_{ji}:

$$\theta_{ji}|\theta_{j1},\ldots,\theta_{j,i-1},\alpha,G_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{\alpha+i-1}\delta_{\psi_{jt}} + \frac{\alpha}{\alpha+i-1}G_0 \tag{3.10}$$

Il secondo processo di Dirichelt definisce come viene scelto un nuovo piatto una volta che il cliente nel ristorante j si siede al nuovo tavolo t^{new}: può scegliere un piatto θ_k^* già ordinato da m_k tavoli degli altri ristoranti con probabilità $\frac{m_k}{\sum_{l=1}^k m_l+\gamma}$, oppure può ordinare un nuovo piatto $\theta_k^{new} \sim H$ con probabilità $\frac{\gamma}{\sum_{l=1}^k m_l+\gamma}$. Il secondo processo completa la descrizione del partizionamento indotto dal CRP in ogni insieme di dati, permettendo ai cluster di essere condivisi definendo la distribuzione di condizionata della variabile indicatrice ψ_{ji} associata a θ_{ji} come:

$$\psi^{new}|\boldsymbol{\psi},\gamma,H \sim \sum_{k=1}^{K} \frac{m_k}{\sum_{l=1}^k m_l+\gamma}\delta_{\theta_k^*} + \frac{\gamma}{\sum_{l=1}^k m_l+\gamma}H \tag{3.11}$$

3.3 Esempio nell'analisi testuale

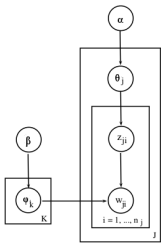
Finora abbiamo parlato dei problemi di clustering in maniera teorica Abbiamo anche visto come, in generale, un modello basato sul processo di Dirichlet ci permette di separare in cluster più insiemi di osservazioni mantenendo, allo stesso tempo, i cluster tra loro collegati. Vediamo ora come questa proprietà possa essere sfruttata in un ambito applicativo, come ad esempio l'analisi testuale.

I modelli bayesiani sono uno strumento largamente diffuso nella classificazione di testi e documenti, e sono comunemente definiti in letteratura come *topic models*. Tali modelli nascono con lo scopo di individuare dei cluster astratti all'interno di un documento, cioè, in parole povere, individuare gli argomenti principali con i quali poterlo descrivere. Tali argomenti

vengono definiti *topic*. Secondo la letteratura al riguardo (vedi (1986)), i documenti sono pensati come una "bag of words": significa che l'ordine delle parole non è caratterizzante. Si può quindi assumere per le parole l'ipotesi di interscambiabilità, che giustifica l'utilizzo di modelli bayesiani. L'utilizzo di tali modelli non si riduce soltanto all'individuazione dei topic, ma molto spesso, data una raccolta di documenti (definita come *corpus*), ci si pone l'obiettivo di modellare tale raccolta in modo da permettere che i topic siano condivisi tra i documenti del corpus. Secondo quanto esposto precedentemente, possiamo subito immaginare di poterci servire in qualche modo di un modello basato sul processo di Dirichlet, in particolar modo di una gerarchia di processi. Abbiamo visto anche che tali modelli non parametrici possono essere concepiti come estensione (o evoluzione) di un modello parametrico spingendo all'infinito il numero di cluster. Lo stesso vale nel topic modeling: partendo da un usuale modello parametrico possiamo estenderlo alla sua versione non parametrica applicando il processo di Dirichlet.

Il modello parametrico tipico dell'analisi testuale è il *Latent Dirichlet allocation* (LDA) introdotto da (2003). Esso si propone come modello probabilistico generativo di un insieme di documenti (corpus). L'idea base consiste nel rappresentare ogni documento come un modello di mistura finito, in cui le diverse componenti di mistura sono rappresentative di uno specifico topic, e le proporzioni di mistura sono estrazioni da una distribuzione a priori dei topic nel documento. Inoltre, date tali proporzioni, ogni parola nel documento è una estrazione indipendente dal modello di mistura. Supponiamo di avere un corpus costituito da J documenti e supponiamo un numero di topic K per il corpus.

Ogni documento è costituito quindi da n_j parole e ogni parola è indicata con w_{ji}. Ad ogni parola è associata una variabile indicatrice z_{ji}, tale che $z_{ji} = k$ indica che la parola w_{ji} appartiene al topic k. I topic hanno una distribuzione a priori multinomiale di parametro θ_j, mentre ogni parola w_{ji} ha una distribuzione a priori $F(\varphi_{z_{ji}})$. Solitamente $F(\varphi_{z_{ji}}) \sim Mult(\varphi_k)$, dato che definisce la distribuzione delle parole nel topic indicato da $z_{ji} = k$. A loro volta, i parametri θ_j e φ_k hanno una distribuzione a priori di Dirichlet con rispettivi iperparametri $\alpha = \alpha_1, \ldots, \alpha_K$ e $\beta = \beta_1, \ldots, \beta_V$, dove V è il numero di parole del vocabolario. Riportiamo sotto il modello grafico relativo, e un riassunto delle variabili e dei parametri.



$$
\begin{aligned}
\theta_j|\alpha &\sim Dir(\alpha) \quad \alpha = \alpha_1, \ldots, \alpha_K \\
\varphi_k|\beta &\sim Dir(\beta) \quad \beta = \beta_1, \ldots, \beta_V \\
z_{ji}|\theta_j &\sim Mult(\theta_j) \\
w_{ji}|z_{ji}, \varphi_k &\sim Mult(\varphi_k)
\end{aligned} \tag{3.12}
$$

Il modello LDA assume il seguente processo generativo per ogni documento j in un corpus:

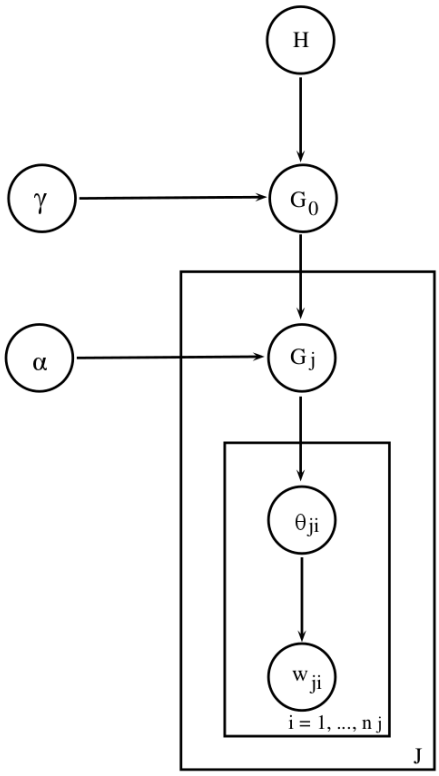
1. estraggo $\theta_j \sim Dir(\alpha)$, per $j = 1, \ldots, J$;
2. estraggo $\varphi_k \sim Dir(\beta)$, per $k = 1, \ldots, K$;
3. per ogni parola w_{ji}, per $i = 1, \ldots, n_j$:
   - estraggo un topic $z_{ji} \sim Mult(\theta_j)$;
   - estraggo una parola $w_{ji} \sim Mult(\varphi_{z_{ji}})$.

Possiamo notare come, attraverso tale processo, le parole possano provenire da differenti topic e come si riesca, già a livello parametrico, a permettere la condivisione dei topic tra i diversi documenti. Tuttavia è necessario fissare il

numero di topic a priori, ed è questo limite che vogliamo superare estendendo il modello tramite il processo di Dirichlet. Dato che ogni documento ha delle specifiche componenti di mistura, abbiamo bisogno di processi Dirichlet multipli, uno per ogni documento. Rimane da affrontare il problema della condivisione delle componenti di mistura, precisamente il problema che una gerarchia di processi ci permette di risolvere.

L'estensione del modello LDA basata sul processo gerarchico di Dirichlet, assume la seguente forma. Ricordando che i topic sono definiti come distribuzioni di probabilità parametriche su un dato vocabolario, supporre un numero di topic infinito significa supporre un numero illimitato di possibili vettori di parametri. Volendo estendere il modello LDA, manteniamo una distribuzione multinomiale per le parole nei topic: questi ultimi sono quindi vettori di probabilità multinomiali. Modellando ogni documento j con un processo di Dirichlet, ogni parola w_{ji} sarà perciò associata ad un topic estratto da una misura casuale G_j, con $G_j \sim Dir(\alpha, G_0)$. Possiamo interpretare un'estrazione da G_j come uno specifico vettore di probabilità multinomiale proveniente da un insieme potenzialmente infinito oppure, equivalentemente, uno specifico topic estratto da un insieme illimitato di topic. Per far sì che i documenti possano condividere lo stesso insieme di topic, consideriamo la misura di base stessa, G_0, come una misura casuale estratta da un processo di Dirichlet, $DP(\gamma, H)$. Data una qualsiasi distribuzione a priori H per l'insieme dei topic sul vocabolario, G_0 descrive una raccolta illimitata di questi vettori, definisce cioè l'insieme di tutti i topic che saranno presenti nel corpus. Riassumendo, con l'aiuto di un modello grafico, abbiamo:

$$\begin{aligned} G_0|\gamma, H &\sim DP(\alpha, H) \\ G_j|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ w_{ji}|\theta_{ji} &\sim Mult(\theta_{ji}) \end{aligned} \tag{3.13}$$



L'estensione gerarchica del modello LDA, assume quindi il seguente processo generativo:

1. per ogni documento j estraggo G_j da un processo di Dirichlet usando G_0 come misura di base, selezionando così il sottoinsieme di topic che sarà presente nel documento;

2. per ogni parola w_{ji}

   - estraggo da G_j uno specifico vettore di probabilità θ_{ji} (topic);
   - genero la parola $w_{ji} \sim Mult(\theta_{ji})$.

Possiamo sfruttare la metafora del CRF citata in precedenza per capire meglio la dinamica esposta. Nella nuova impostazione, ogni documento corrisponde ad un ristorante in cui i clienti corrispondono alle parole. L'insieme dei documenti condivide un menù di topic. Le parole in ogni documento sono divise in gruppi, ognuno dei quali condivide un tavolo, e ogni tavolo

è associato ad un topic (i piatti nella metafora); di conseguenza le parole sedute ad ogni tavolo sono associate al topic del tavolo. L'associazione di una parola ad un topic avviene quindi nella stessa maniera descritta in precedenza.

Capitolo 4

Applicazione e conclusioni

Esponiamo ora un'applicazione pratica del modello esposto nel capitolo precedente, utilizzando come dataset una raccolta di dati ottenuta da Twitter. Per eseguire tale applicazione è stato utilizzato il codice MATLAB[1] sviluppato da Yee Whye Teh che implementa i modelli gerarchici basati sul processo di Dirichlet.

4.1 Twitter

Twitter è un servizio gratuito di social network e microblogging, che permette ai propri utenti di comunicare tramite messaggi di testo con una lunghezza massima di 140 caratteri chiamati "tweets", letteralmente "cinguettii". Il servizio è costruito totalmente su architettura Open Source e fu ideato nel marzo del 2006 da Jack Dorsey. La meccanica di Twitter viene descritta in maniera semplice da (2009) in una articolo del Time:

> Come un social network, Twitter ruota intorno al principio dei seguaci (followers). Quando si sceglie di seguire un altro utente di Twitter, i tweet di tale utente vengono visualizzati in ordine cronologico inverso, sulla home page di Twitter. Se seguite 20 persone, si vedrà una miscela di tweet scorrere la pagina: aggiornamenti sui cereali per la colazione, nuovi link, consigli musicali, ma anche riflessioni sul futuro dell'istruzione.

[1] Il codice è scaricabile liberamente da http://www.gatsby.ucl.ac.uk/ ywteh/research/software.html

Twitter non è un vero e proprio social network, piuttosto un *information network*. Citando uno degli sviluppatori, "Twitter is for news. Twitter is for content. Twitter is information". Le informazioni vengono infatti condivise e aggregate tematicamente per mezzo dei *tag* (@) e degli *hastag* (#). I primi servono a coinvolgere altri utenti nella conversazione, mentre gli hashtag sono delle "etichette" che gli utenti possono inserire nei propri tweet per far capire che stanno parlando di un determinato argomento. Quando un hashtag è usato contemporaneamente da molte persone, esso diventa un vero e proprio tema di tendenza che può essere seguito e scoperto da chiunque, tanto che nell'ultimo anno è stato aggiunto nella home page di Twitter un post riportante i topic più popolari.

4.2 I dati

Il dataset raccolto è costituito da 218737 tweets il cui argomento comune è la politica, partiti e personaggi politici in particolare. Le chiavi di ricerca usate sono riportate nella tabella B.1, con accanto il numero di tweets corrispondenti recuperati. Possiamo vedere che per il recupero dei tweet, sono stati usati principalmente i nomi dei partiti politici, i nomi dei rispettivi rappresentanti e anche i tag del profilo Twitter ufficiale dei rappresentanti. I tweet raccolti fanno riferimento al periodo compreso tra l'01/03/2012 e il 05/04/2012: in tabella B.2 è riportato il numero di tweets per ogni giorno. I diversi tweet sono poi stati accorpati secondo query e data, ottendo un numero di 1338 "macro-tweet".

Nell'analisi testuale risulta fondamentale la scelta del numero di parole uniche del vocabolario da mantenere nei documenti: articoli, pronomi, preposizioni e simili risultano inutili al fine della classificazione. Sono tutti termini che appaiono frequentemente e la loro classificazione risulterebbe poi non informativa; lo stesso vale per i termini poco frequenti. Nella prima fase di scrematura sono state eliminate le parole che apparivano meno di 10 volte, ottenendo un vocabolario di 15317 parole; si è poi ricorsi al calcolo di una funzione comunemente usata nel campo delle applicazioni testuali per misurare l'importanza di un termine rispetto ad un documento o ad una collezione di documenti, la funzione peso tf-idf (term frequency - inverse document frequency). L'idea alla base di questa funzione è di dare più importanza ai termini che compaiono nel documento, ma che in generale sono

poco frequenti: essa aumenta proporzionalmente al numero di volte che il termine è contenuto nel documento, ma cresce in maniera inversamente proporzionale con la frequenza del termine nella collezione. La funzione può essere scomposta in due fattori: il primo è una matrice di frequenze di ogni termine in un particolare documento

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|} \tag{4.1}$$

dove n_{ij} è il numero di occorrenze del termine t_i nel documento d_j. Il denominatore invece rappresenta il numero di termini presenti nel documento d_j. Il secondo fattore misura l'importanza generale del termine nella collezione:

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{4.2}$$

dove $|D|$ è il numero di documenti nella collezione, mentre il denominatore è il numero di documenti che contengono il termine t_i. Abbiamo quindi che

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \tag{4.3}$$

In base ai pesi calcolati abbiamo ridotto ulteriormente il numero di parole. Solitamente nell'analisi testuale sono diverse le necessità e le motivazioni che conducono alla scelta del *quanto* ridurre il vocabolario: nel nostro caso si è scelto di mantenere 606 vocaboli, a causa dei tempi ristretti per la successiva elaborazione. Sono stati poi eliminati i tweet più piccoli che, a causa della riduzione del vocabolario, risultavano nulli; il conteggio finale è stato di 1217 tweet. I termini sono rappresentati in figura B.1 per mezzo di un grafico di tipo *wordcloud*. Le parole sono rappresentate secondo la loro frequenza: le parole più grandi, sono quelle che compaiono con una frequenza maggiore. Possiamo subito notare che i termini usati come chiavi di ricerca, ovviamente, hanno delle frequenze alte; tale aspetto sarà da tenere in considerazione nell'analisi dei cluster.

4.3 Applicazione

La stima del modello per i dati, è stata possibile grazie all'utilizzo del codice MATLAB indicato, seguendo gli esperimenti presentati in (2004). I risultati sono poi stati importati e analizzati tramite il software statistico

R. Per la stima del modello abbiamo assunto come distribuzione a priori H delle parole nel vocabolario, una distribuzione di Dirichlet simmetrica di parametro $1/606 = 0.001650165$. Abbiamo supposto anche una distribuzione a priori anche per i parametri di concentrazione, $\gamma \sim Gamma(1,1)$ e $\alpha \sim Gamma(1,1)$.

Sono stati stimati 495 cluster, ovvero 495 componenti di mistura con le relative proporzioni di mistura. Di seguito riportiamo i cluster con una proporzione di mistura maggiore dello 1%. Per ogni parola all'interno del cluster, riportiamo anche la rispettiva probabilità di estrazione, ricordando che il modello prevede una distribuzione per le parole nei cluster di tipo multinomiale.

Gruppo 1, proporzione di mistura = 0.07289192

beppegrillo	movimentocinquestelle	omezia
0.0826580227	0.8910048622	0.0016207455
epurazioni	101	informativo
0.0006077796	0.0010129660	0.0022285251
nuti	# amianto	gioè
0.0002025932	0.0079011345	0.0016207455
cifarelli	barano	putti
0.0018233387	0.0018233387	0.0022285251
# consigliomilano	tavolazzi	ciclabili
0.0008103728	0.0034440843	0.0010129660

Gruppo 2, proporzione di mistura = 0.05562391

# opencamera	futuroelibertàperlitalia	# opensenato
0.9856590030	0.0045526975	0.0081948555
# cooperazione	autocertificazione	castagnetti
0.0006829046	0.0006829046	0.0002276349

Gruppo 3, proporzione di mistura = 0.08256857

comportamenti	gianfrancofini	pietrasanta
0.0001759634	0.9839873306	0.0007038536
futuroelibertàperlitalia	liberta	# pietrasanta
0.0105578040	0.0003519268	0.0026394510
# saràbellissima	assassinato	
0.0010557804	0.0005278902	

Gruppo 4, proporzione di mistura = 0.01207373

gianfrancofini	pietrasanta	futuroelibertàperlitalia
0.0774923814	0.0134958642	0.5067479321
tesa	convenzione	# pietrasanta
0.0004353505	0.0078363082	0.3726599913
aricò	insensibilità	# proposteconcrete
0.0043535046	0.0008707009	0.0161079669

Gruppo 5, proporzione di mistura = 0.01086151

movimentocinquestelle	# consigliomilano	mattia
0.5964912281	0.3809523810	0.0075187970
calise	ciclabili	uganda
0.0050125313	0.0075187970	0.0008354219
valico		
0.0016708438		

Gruppo 6, proporzione di mistura = 0.03337731

governotecnico	tirano	dipendesse
0.9859660574	0.0006527415	0.0003263708
consegnano	azzardano	
0.0006527415	0.0124020888	

Gruppo 7, proporzione di mistura = 0.03426962

beppegrillo	# abruzzo	eguali
0.990733591	0.003088803	0.001544402
ripristino	fininvest	
0.003088803	0.001544402	

Gruppo 8, proporzione di mistura = 0.01598787

scoperta	comportamenti	# opencamera
0.090163934	0.114754098	0.352459016
esibizione	riferisca	architetto
0.024590164	0.057377049	0.049180328
esprimersi	arresto	impossibili
0.057377049	0.245901639	0.008196721

Gruppo 9, proporzione di mistura = 0.01726959

farmacie	beppegrillo	governotecnico
0.11560694	0.17919075	0.46820809
cornuti	analisipolitica	# fininvest
0.01156069	0.01734104	0.01156069
liberta	totem	# londra
0.10404624	0.04046243	0.02312139
# licenziare		
0.02890173		

Gruppo 10, proporzione di mistura = 0.01701891

# opencamera	governotecnico	movimentocinquestelle
0.274590164	0.135245902	0.274590164
disciplinare	divertimento	inferiore
0.036885246	0.024590164	0.028688525
tirano	rivolte	colonnelli
0.053278689	0.016393443	0.020491803
marte	# ranieri	rosario
0.032786885	0.028688525	0.012295082
progress	# belpietro	comica
0.008196721	0.045081967	0.008196721

Gruppo 11, proporzione di mistura = 0.07771829

gianfrancofini	futuroelibertàperlitalia	cooperanti
0.0167529698	0.9445628998	0.0003045995
raccoglieremo	granata	virili
0.0012183978	0.0121839781	0.0003045995
# abruzzo	# rifarepalermo	briguglio
0.0021321962	0.0027413951	0.0030459945
relatori	# forzariformista	# briguglio
0.0003045995	0.0003045995	0.0027413951
aricò	# aricò	parcheggi
0.0073103868	0.0057873896	0.0003045995

Gruppo 12, proporzione di mistura = 0.01267684

# opencamera	# opensenato	granata
0.566964286	0.276785714	0.004464286
# amici	# opensud	
0.013392857	0.138392857	

Un grafico comparativo delle frequenze delle parole nei diversi gruppi è riportato in figura B.2.

4.4 Considerazioni e conclusioni

Dando un'occhiata ai cluster e al grafico, notiamo subito che le chiavi di ricerca rimangono i termini più frequenti nei diversi gruppi, e com'era prevedibile, sembrano quasi caratterizzarli. All'interno di ogni cluster, la parole con probabilità maggiori, sembrano anche quelle più coerenti tra loro. Troviamo infatti un associazione tra partito politico e rispettivi leader: beppegrillo e movimentocinquestelle nel primo gruppo, futuroelibertàperlitalia e gianfrancofini nel terzo e nel quarto, ad esempio; ma anche termini che potrebbero riferirsi ad eventi, come ad esempio scoperta, comportamenti e opencamera nel gruppo 8 e simili nei gruppi 9 e 10. Nonostante la presenza di tali associazioni, non riusciamo a definire chiaramente un topic associato ad ogni cluster, molto probabilmente a causa dell'utilizzo di un vocabolario molto ridotto; come detto precedentemente, la scelta del numero di parole del vocabolario è sicuramente influenzata dagli scopi dell'analisi. In quanto la nostra analisi si voleva proporre come esemplificativa del modello esposto nella tesi, abbiamo ritenuto valido anche un vocabolario molto ridotto. I risultati ottenuti non portano a nulla di illuminante, ma non sono nemmeno da buttare; un'analisi maggiormente approfondita e con un vocabolario più grande, potrebbe produrre dei risultati migliori.

Il processo di Dirichlet e i modelli bayesiani non parametrici, costituiscono un'area di ricerca molto attiva nella statistica, soprattutto per quanto riguardi i risvolti pratici. Abbiamo infatti proposto un esempio tratto dall'analisi testuale, ma modelli simili sono molto popolari negli ambiti informatici come ad esempio l'apprendimento automatico oppure la raccolta informazioni in generale. Possiamo trovare esempi di applicazioni anche nella genetica delle popolazioni, dove il processo di Dirichlet iniziò ad essere considerato per le prime volte. Le ricerche maggiormente di tendenza prendono in generale due direzioni. Da un lato si cerca di rendere sempre più efficienti le tecniche di inferenza per i modelli basati sul processo di Dirichlet, grazie al campionamento di Gibbs o i metodi MCMC Metropolis-Hastings ((2000)). Dall'altro invece, si vuole approfondire e migliorare il processo stesso, sia per quanto riguarda le basi teoriche, sia proponendone delle estensioni, come ad esempio il *processo Pitman-Yor* ((1995)). Concludendo, i modelli bayesiani non parametrici e il processo di Dirichlet risultano molto popolari nelle applicazioni pratiche, e per questo rappresentano un terreno fertile per

la ricerca statistica. Non resta che vedere cosa ci riserva il futuro.

Appendice A

Distribuzioni di probabilità

A.1 Distribuzione Beta

Una variabile casuale Y ha distribuzione beta con parametri di forma α ($\alpha > 0$) e β ($\beta > 0$), e si scrive sinteticamente $Y \sim Be(\alpha, \beta)$ se Y ha supporto $S_Y = [0, 1]$ e funzione di densità di probabilità

$$p_Y(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1 - y)^{\beta-1} \tag{A.1}$$

per $y \in S_Y$ e $p_Y(y; \alpha, \beta) = 0$ altrove, dove $\Gamma()$ indica la funzione gamma. Si mostra facilmente che

$$\begin{aligned} E(Y) &= \frac{\alpha}{\alpha + \beta} \\ V(Y) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{aligned} \tag{A.2}$$

A.2 Distribuzione di Dirichlet

La distribuzione di Dirichlet è una distribuzione di probabilità continua, dipendente da un vettore di numeri reali positivi $\boldsymbol{\alpha}$ che generalizza la distribuzione beta al caso multivariato.

Una distribuzione di Dirichlet di ordine $K \geq 2$ con parametri $\alpha_1, \ldots, \alpha_K > 0$ ha una funzione di densità di probabilità rispetto alla misura di Lebesgue

sul spazio euclideo R^{k-1} data da

$$p(x_1, \ldots, x_(K-1), \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{A.3}$$

per ogni $x_1, \ldots, x_{K-1} > 0$ tali che $x_1 + \cdots + x_{K-1} < 1$ e $X_K = 1 - x_1 - \cdots - x_{K-1}$, e la indicheremo con $Dir(\alpha_1, \ldots, \alpha_K)$. La costante di normalizzazione è la funzione beta multinomiale, che può essere espressa in termini di funzione gamma come:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \qquad \alpha = (\alpha_1, \ldots, \alpha_K) \tag{A.4}$$

Un caso particolare molto comune di questa distribuzione, è la distribuzione di Dirichlet simmetrica nella quale tutti gli elementi del vettore dei parametri $\boldsymbol{\alpha}$ hanno lo stesso valore. Dato che tutti gli elementi del vettore dei parametri hanno lo stesso valore, la distribuzione può essere parametrizzata da un singolo valore α scalare, chiamato parametro di concentrazione.

Sia $\mathbf{X} = (X_1, \ldots, X_K) \sim Dir(\alpha)$, ovvero i primi $K-1$ elementi hanno densità descritta sopra, e sia $\alpha_0 = \sum_{i=1}^{K} \alpha_K$; i momenti della distribuzione sono:

$$E[X_i] = \frac{\alpha_i}{\alpha_0} \tag{A.5}$$

$$V[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{A.6}$$

Inoltre una proprietà fondamentale riguarda la distribuzione delle marginali:

$$X_i \sim Beta(\alpha_i, \alpha_0 - 1) \tag{A.7}$$

Appendice B

Grafici e Tabelle

Tabella B.1: Query utilizzate e numero dei tweets

Query di ricerca	tweet
@angealfa	6793
#berlusconi	7734
#bersani	7528
#bossi	19106
#casini	1830
#dipietro	1496
@EstremoCentro	489
#fini	1051
#fli	2371
@_futuroeliberta	274
futuro e liberta per l'italia	36
@gianfranco_fini	3463
governo monti	14789
governo tecnico	3211
#grillo	459
#idv	839
@Idvstaff	11210
@ilpdl	4657
italia dei valori	629
#lega	11162
lega nord	8247
@LegaNordPadania	159
#m5s	2649
#monti	31503
movimento 5 stelle	1827
@NichiVendola	7524
#opencamera	3624
#opensenato	983
partito democratico	1544
@pbersani	15377
#pd	11861
#pdl	6654
@pdnetwork	1026
@Pierferdinando	13942
popolo della libert	4
#sel	490
sinistra ecologia libert	6
@sinistraelib	1929
udc	8767
#vendola	1494

Tabella B.2: Numero di tweet per giorno

giorno	tweet
3-1	7298
3-10	4403
3-11	4769
3-12	6721
3-13	4910
3-14	3750
3-15	5601
3-16	5553
3-17	5806
3-18	3914
3-19	4335
3-2	5421
3-20	5519
3-21	9173
3-22	6610
3-23	5113
3-24	4072
3-25	2880
3-26	6583
3-27	6945
3-28	5496
3-29	5406
3-3	4522
3-30	4975
3-31	3762
3-4	3550
3-5	6955
3-6	10189
3-7	8049
3-8	5528
3-9	4707
4-1	2694
4-2	4068
4-3	8360
4-4	11957
4-5	19143

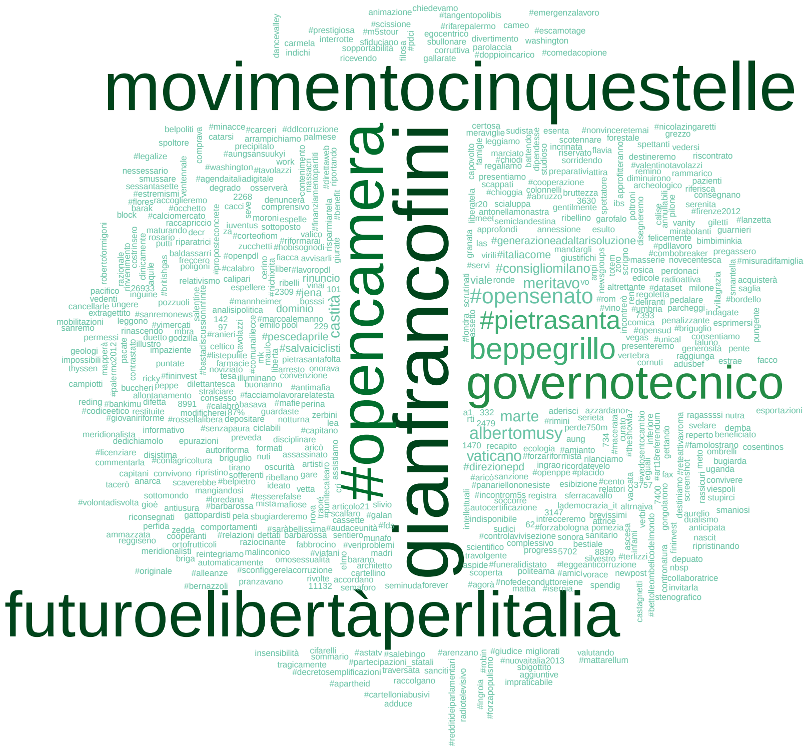


Figura B.1: Wordcloud delle frequenze

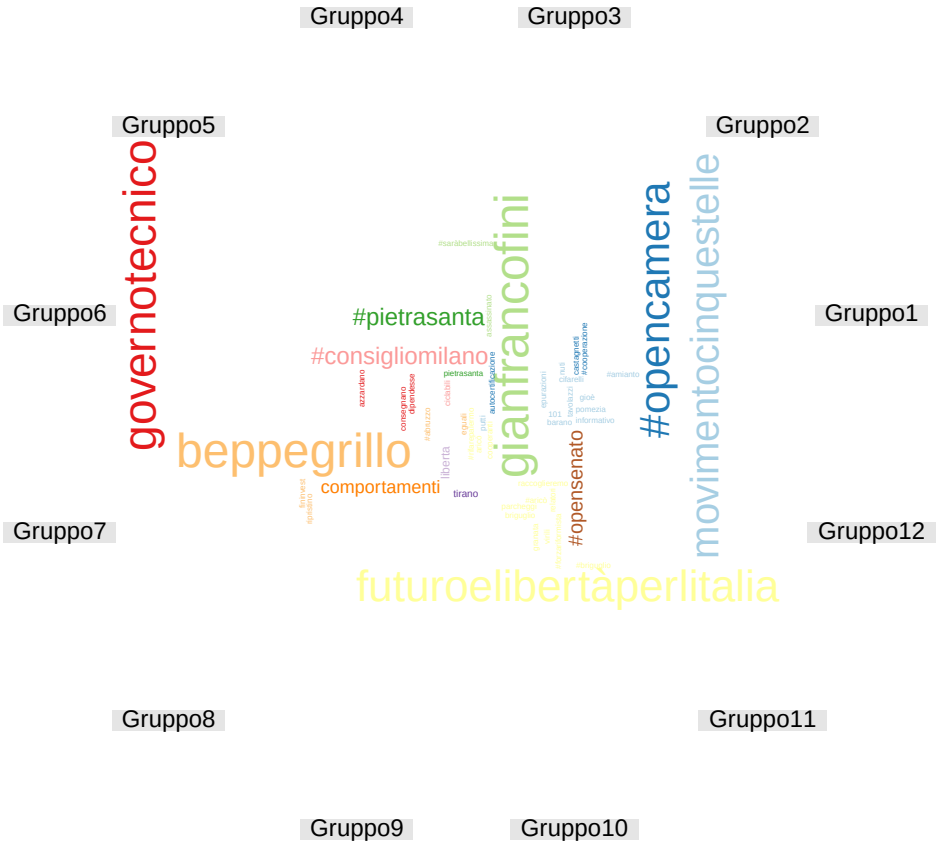


Figura B.2: Worcloud comparitvo dei gruppi