Capitolo 1

Che cos'è la Scienza dei dati?

Come ogni campo emergente, non è ancora stato completamente definito, ma lei ne sa abbastanza per essere interessato, altrimenti non starebbe leggendo questo libro.
Ritengo che la scienza dei dati si trovi all'intersezione tra informatica, statistica e domini applicativi sostanziali. Dall'informatica provengono l'apprendimento ma- chino e le tecnologie di calcolo ad alte prestazioni per gestire la scala. Dalla statistica proviene una lunga tradizione di analisi esplorativa dei dati, test di significatività e visualizzazione. Dai domini applicativi negli affari e nelle scienze provengono le sfide degne di essere combattute e gli standard di valutazione per stabilire quando sono state adeguatamente vinte.
Ma questi sono tutti campi ben consolidati. Perché la scienza dei dati e perché ora? Vedo tre ragioni per questa improvvisa esplosione di attività:

- La nuova tecnologia consente di catturare, annotare e archiviare grandi quantità di dati dei social media, dei registri e dei sensori. Dopo aver accumulato tutti questi dati, ci si chiede cosa si può fare con essi.

- I progressi informatici consentono di analizzare i dati in modi nuovi e su scale sempre maggiori. Le architetture di cloud computing consentono anche ai più piccoli di accedere a una grande potenza quando hanno bisogno. I nuovi approcci all'apprendimento automatico hanno portato a progressi sorprendenti in problemi di vecchia data, come la computer vision e l'elaborazione del linguaggio naturale.

- Aziende tecnologiche di spicco (come Google e Facebook) e fondi hedge quan- titativi (come Renaissance Technologies e TwoSigma) hanno dimostrato la potenza della moderna analisi dei dati. Le storie di successo che applicano i dati a settori diversi come la gestione dello sport e le previsioni elettorali sono servite da modello per portare la scienza dei dati a un vasto pubblico.

Questo capitolo introduttivo ha tre missioni. In primo luogo, cercherò di spiegare come pensano i bravi scienziati dei dati e come questo differisce dalla mentalità dei programmatori e degli sviluppatori di software tradizionali. In secondo luogo, esamineremo le serie di dati in termini di potenziale per cui possono essere utilizzate e impareremo a porre le domande più ampie a cui sono in grado di rispondere. Infine, introduco una raccolta di sfide di analisi dei dati che saranno utilizzate in tutto il libro come esempi motivanti.

1.1 Informatica, Scienza dei dati e Scienza reale

Gli informatici, per natura, non rispettano i dati. Tradizionalmente è stato insegnato loro che l'algoritmo era la cosa e che i dati erano solo carne passare in un tritacarne.

Quindi, per qualificarsi come data scientist efficace, deve prima imparare a pensare come un vero scienziato. I veri scienziati si sforzano di capire il mondo naturale, che è un luogo complicato e disordinato. Al contrario, gli informatici tendono a costruire i loro mondi virtuali puliti e organizzati e a vivere comodamente al loro interno. Gli scienziati sono ossessionati dalla scoperta delle cose, mentre gli informatici si dedicano a sfogare piuttosto che a scoprire.

La mentalità delle persone colora fortemente il modo in cui pensano e agiscono, causando malintesi quando cerchiamo di comunicare al di fuori delle nostre tribù. Questi pregiudizi sono così fondamentali che spesso non siamo consapevoli di averli. Esempi di differenze culturali tra l'informatica e la scienza reale sono:

- *Centrismo sui dati e centrismo sui metodi*: Gli scienziati sono guidati dai dati, mentre gli informatici sono guidati dagli algoritmi. I veri scienziati dedicano sforzi alla raccolta di dati per rispondere alla loro domanda di interesse. Inventano fantastici dispositivi di misurazione, restano svegli tutta la notte per curare gli esperimenti e dedicano la maggior parte del loro pensiero a come ottenere i dati di cui hanno bisogno.

  Al contrario, gli informatici sono ossessionati dai metodi: quale algoritmo è migliore di quale altro algoritmo, quale linguaggio di

programmazione è migliore per un lavoro, quale programma è migliore di quale altro programma. I dettagli dell'insieme di dati su cui stanno lavorando sembrano relativamente poco interessanti.

- *Preoccupazione per i risultati*: Ai veri scienziati interessano le risposte. Analizzano i dati per scoprire qualcosa sul funzionamento del mondo. I bravi scienziati si preoccupano se i risultati hanno senso, perché si preoccupano del significato delle risposte.

  Al contrario, i cattivi informatici si preoccupano di produrre numeri dall'aspetto plausibile. Non appena i numeri smettono di sembrare grossolanamente sbagliati, si presume che siano giusti. Questo perché sono personalmente meno interessati a ciò che si può imparare da un calcolo, piuttosto che a in modo rapido ed efficiente.
- *Robustezza*: Gli scienziati reali sono a proprio agio con l'idea che i dati abbiano degli errori. In generale, gli informatici non lo sono. Gli scienziati pensano molto alle possibili fonti di distorsione o di errore nei loro dati e a come questi possibili problemi possano influire sulle conclusioni che derivano. I bravi programmatori utilizzano metodologie forti di data-typing e parsing per proteggersi dagli errori di formattazione, ma le preoccupazioni in questo caso sono diverse. Diventare consapevoli del fatto che i dati possono avere degli errori è un'esperienza che dà forza. Gli informatici cantano "garbage in, garbage out" come un mantra difensivo per allontanare le critiche, un modo per dire *che non è il mio lavoro*. I veri scienziati si avvicinano abbastanza ai loro dati per annusarli, facendo la prova del fiuto per decidere se è probabile che siano spazzatura.

- *Precisione*: Niente è mai completamente vero o falso nella scienza, mentre *ogni cosa* è vera o falsa nell'informatica o nella matematica. In generale, gli informatici sono contenti di stampare i numeri in virgola mobile con il maggior numero di cifre possibile: 8/13 = 0.61538461538. Gli scienziati veri useranno solo due cifre significative: 8/13 0,62. Agli informatici interessa cosa è un numero, mentre agli scienziati reali interessa il suo significato.

Gli aspiranti data scientist devono imparare a pensare come veri scienziati. Il suo lavoro sarà quello di trasformare i numeri in intuizioni. È importante capire il perché e *il* come.

A dire il vero, anche pensare come scienziati dei dati è vantaggioso per gli scienziati veri. Le nuove tecnologie sperimentali consentono di misurare i sistemi su scala molto più ampia di quanto sia mai stato possibile in precedenza, grazie a tecnologie come il sequenziamento del genoma completo in biologia e le indagini con telescopi full-sky in astronomia. Con una nuova ampiezza di vedute, arrivano nuovi livelli di visione.
La scienza tradizionale *guidata dalle ipotesi* si basava sul porre domande specifiche sul mondo e poi generare i dati specifici necessari per confermarle o negarle. Questa è ora aumentata dalla scienza *guidata dai dati*, che si concentra invece sulla generazione di dati su una scala o una risoluzione mai viste prima, nella convinzione che le nuove scoperte arriveranno non appena si sarà in grado di guardarle.
C'è un altro modo per cogliere questa distinzione di base tra ingegneria del software e scienza dei dati. Si tratta del fatto che gli sviluppatori di software sono assunti per costruire sistemi, mentre i data scientist sono assunti per produrre intuizioni.
Questo può essere un punto di conflitto per alcuni sviluppatori. Esiste una classe importante di ingegneri che gestiscono le massicce infrastrutture distribuite necessarie per archiviare e analizzare, ad esempio, le transazioni finanziarie o i dati dei social media. su un livello di scala pari a quello di Facebook o Twitter. Infatti, dedicherò il Capitolo 12 alle sfide distintive delle infrastrutture di big data. Questi ingegneri stanno costruendo strumenti e sistemi per supportare la scienza dei dati, anche se potrebbero non estrarre personalmente i dati che gestiscono. Si qualificano come scienziati dei dati? Si tratta di una domanda giusta, che io perfezionerò un po' per massimizzare il potenziale di lettori di questo libro. Tuttavia, ritengo che quanto meglio questi ingegneri capiranno l'intera pipeline di analisi dei dati, tanto più probabilmente saranno in grado di costruire strumenti potenti in grado di fornire intuizioni importanti. Uno degli obiettivi principali di questo libro è fornire agli ingegneri dei big data gli strumenti intellettuali per pensare come gli scienziati dei big data.

1.2 Porre domande interessanti sui dati

I bravi data scientist sviluppano una curiosità intrinseca per il mondo che li circonda, in particolare per i domini e le applicazioni associate cui lavorano. Si divertono a parlare con le persone con i cui dati lavorano. Pongono loro delle domande: Qual è la cosa più bella che ha imparato in questo campo? Perché si è interessato a questo settore? Cosa spera di imparare analizzando il suo set di dati? Gli scienziati dei dati fanno sempre domande.

I bravi data scientist hanno interessi ad ampio raggio. Leggono i giornali ogni giorno per avere una prospettiva più ampia su ciò che è interessante. Capiscono che il mondo è un posto interessante. Conoscere un po' di tutto li mette in grado di giocare nel giardino degli altri. Sono abbastanza coraggiosi da uscire un po' dalla loro zona di comfort e sono spinti a saperne di più una volta arrivati.

Gli sviluppatori di software non sono incoraggiati a fare domande, ma gli scienziati dei dati sì. Gli informatici tradizionalmente non apprezzano molto i dati. Pensi al modo in cui le prestazioni degli algoritmi vengono misurate sperimentalmente. Di solito il programma viene eseguito su "dati casuali" per vedere quanto tempo impiega. Raramente si guardano i risultati del calcolo, se non per verificare che sia corretto ed efficiente. Poiché i 'dati' sono privi di significato, i risultati non possono essere importanti. Al contrario, i set di dati reali sono una risorsa scarsa, che ha richiesto un duro lavoro e l'immaginazione per essere ottenuta. Per diventare uno scienziato dei dati è necessario imparare a porre domande sui dati, quindi facciamo pratica. Ciascuna delle sottosezioni seguenti introdurrà una serie di dati interessanti. Dopo aver capito che tipo di informazioni sono disponibili, provi a proporre, ad esempio, cinque domande interessanti che potrebbe esplorare/rispondere con l'accesso a questo set di dati. La chiave è pensare in modo ampio: le risposte alle grandi domande generali spesso sono sepolte in serie di dati altamente specifici, che non sono stati assolutamente progettati per contenerle.

1.2.1 L'Enciclopedia del Baseball

Il baseball ha da tempo un'importanza fuori misura nel mondo della scienza dei dati. Questo sport è stato definito il passatempo nazionale degli Stati Uniti; infatti, lo storico francese Jacques Barzun ha osservato che "Chi vuole conoscere il cuore e la mente dell'America deve imparare il baseball". Mi rendo conto che molti lettori non sono americani, e anche quelli che lo sono

potrebbero essere completamente disinteressati allo sport. Ma seguitemi per un po'.
Ciò che rende il baseball importante per la scienza dei dati è la sua ampia documentazione statistica, che risale a ben oltre cento anni fa. Il baseball è uno sport di eventi discreti: i lanciatori lanciano le palle e i battitori cercano di colpirle - che si presta naturalmente a statistiche informative. I tifosi vengono immersi in queste statistiche da bambini, sviluppando la loro intuizione sui punti di forza e sui limiti dell'analisi quantitativa. Alcuni di questi bambini crescono e diventano scienziati dei dati. In effetti, il successo della squadra di baseball statistica di Brad Pitt nel film *Moneyball* rimane il contatto più vivo del pubblico americano con la scienza dei dati.
Otteniamo le statistiche vitali di ogni giocatore (altezza, peso, mano) e la loro durata di vita (quando/dove sono nati e morti). Otteniamo anche informazioni sul salario (quanto è stato pagato ogni stagione da ciascun giocatore) e dati sulle transazioni (come sono diventati di proprietà di ciascuna squadra per cui hanno giocato).
Ora, mi rendo conto che molti di voi non hanno la minima conoscenza o interesse per il baseball. Questo sport ricorda un po' il cricket, se questo può aiutare. Ma ricordate che come scienziato dei dati, è il vostro lavoro interessarvi mondo che vi circonda. Considerate questa occasione come un'opportunità per imparare qualcosa.
Si tratta di domande interessanti. Ma ancora più interessanti sono le domande sulle questioni demografiche e sociali. Quasi 20.000 giocatori della Major League di baseball... giocatori sono scesi in campo negli ultimi 150 anni, fornendo una coorte di uomini ampia e ampiamente documentata che può servire come proxy per popolazioni ancora più ampie e meno ben documentate. Infatti, possiamo utilizzare questi dati sui giocatori di baseball per rispondere a domande. Ci sono due temi particolari di cui essere consapevoli. In primo luogo, gli identificatori e i tag di riferimento (cioè i metadati) spesso si rivelano più interessanti in un set di dati rispetto a ciò che dovrebbe, in questo caso il record statistico del gioco. La seconda è l'idea di un *proxy statistico*, in cui si utilizza il set di dati che si ha per sostituire quello che si vuole veramente. Il set di dati dei suoi sogni probabilmente non esistere, o anche se esiste, potrebbe essere rinchiuso dietro un muro aziendale. Un buon scienziato dei dati è un pragmatico, che vede cosa può fare con ciò che ha, invece di lamentarsi di ciò su cui non può mettere le mani.

1.2.2 Il database dei film su Internet (IMDb)

Tutti amano il cinema. L'Internet Movie Database (IMDb) fornisce dati crowdsourcing e curati su tutti gli aspetti dell'industria cinematografica. Attualmente IMDb contiene dati su oltre 3,3 milioni di film e programmi TV. Per ogni film, IMDb include il titolo, la durata, i generi, la data di uscita e un elenco completo del cast e della troupe. Sono presenti dati finanziari su ogni produzione, tra cui il budget per la realizzazione del film e il suo successo al botteghino. Infine, ci sono valutazioni approfondite per ogni film da parte di spettatori e critici. Questi dati di valutazione consistono in punteggi su una scala da zero a dieci stelle, incrociati in medie per età e sesso. Spesso sono incluse recensioni scritte, che spiegano perché un determinato critico ha assegnato un determinato numero di stelle. Ci sono anche collegamenti tra i film: ad esempio, è possibile identificare quali altri film sono stati visti più spesso dagli spettatori di *La vita meravigliosa*.

Ogni attore, regista, produttore e membro della troupe associato ad un film merita una voce in IMDb, che ora contiene i dati di 6,5 milioni di persone. Si dà il caso che tra questi ci siano mio fratello, mio cugino e mia cognata. Ogni attore è collegato ad ogni film in cui è apparso, con una descrizione del suo ruolo e il suo ordine nei titoli di coda. I dati disponibili su ogni personaggio includono le date di nascita/morte, l'altezza, i premi e le relazioni familiari. Partendo dal presupposto che le persone che lavorano insieme a un film conoscono, i dati del cast e della troupe possono essere utilizzati per costruire una rete sociale del mondo del cinema. Come si presenta la rete sociale degli attori? In modo più critico, possiamo analizzare questi dati per determinare la probabilità che a qualcuno piaccia un determinato film? La tecnica del *filtraggio collaborativo* trova persone a cui sono piaciuti film che sono piaciuti anche a me, e raccomanda altri film che sono piaciuti *a loro* come buoni candidati per me. Il Premio Netflix 2007 era una petizione da 1.000.000 di dollari per produrre un motore di valutazione migliore del 10% rispetto al sistema proprietario di Netflix. Il vincitore finale di questo premio ha utilizzato una varietà di fonti di dati e di tecniche, compresa l'analisi dei link.

1.2.3 Google Ngrams

I libri stampati sono stati il principale deposito della conoscenza umana sin dall'invenzione di Gutenberg della stampa a caratteri mobili nel 1439. Gli oggetti fisici vivono un po' a disagio nel mondo digitale di oggi, ma la

tecnologia ha un modo per ridurre ogni cosa a dati. Come parte della sua missione di organizzare le informazioni del mondo, Google ha intrapreso uno sforzo per scansionare tutti i libri pubblicati nel mondo. Non ci sono ancora arrivati, ma i 30 milioni di libri digitalizzati finora rappresentano oltre il 20% di tutti i libri mai pubblicati. Google utilizza questi dati per migliorare i risultati delle ricerche e per fornire un nuovo accesso ai libri fuori catalogo. Ma forse il prodotto più interessante è *Google Ngrams*, una risorsa straordinaria per monitorare i cambiamenti nello Zeitgeist culturale. Fornisce la frequenza con cui brevi frasi ricorrono nei libri pubblicati ogni anno. Ogni frase deve comparire almeno quaranta volte nel loro corpus di libri scansionati. Questo elimina le parole e le frasi oscure, ma lascia oltre due miliardi di serie temporali disponibili per l'analisi. Questo ricco set di dati mostra come è cambiato l'uso della lingua negli ultimi 200 anni ed è stato ampiamente applicato all'analisi delle tendenze culturali. *L'elaborazione dei dati* era il termine popolare associato al settore informatico durante l'era delle schede perforate e dei nastri magnetici rotanti degli anni Cinquanta. I dati Ngrams mostrano che la rapida ascesa dell'*Informatica* ha eclissato l'*Elaborazione Dati* solo nel 1980. Ancora oggi, la *Scienza dei Dati* rimane quasi invisibile su questa scala. Consulti Google Ngrams all'indirizzo http://books.google.com/ngrams. Le assicuro che si divertirà a giocarci. Confronti *l'hot dog con il tofu*, la *scienza* con la *religione*, la *libertà* con la *giustizia* e il *sesso* con il *matrimonio*, per capire meglio questo fantastico telescopio per guardare nel passato. Ma una volta finito di giocare, pensi a cose più grandi che potrebbe fare se mettesse le mani su questi dati. Supponiamo che abbia accesso al numero annuale di riferimenti per *tutte le* parole/frasi pubblicate nei libri negli ultimi 200 anni. Google rende questi dati liberamente disponibili. Osservare le serie temporali associate a determinate parole utilizzando il Visualizzatore Ngrammi è divertente. Ma tendenze storiche più sofisticate possono essere catturate aggregando insieme più serie temporali. Può anche utilizzare questo corpus di Ngrammi per costruire un modello linguistico che catturi il significato e l'uso delle parole in una determinata lingua. Parleremo delle incorporazioni di parole che sono strumenti potenti per costruire modelli linguistici. I conteggi di frequenza rivelano quali parole sono più popolari. La frequenza delle coppie di parole che appaiono una accanto all'altra può essere utilizzata per migliorare i sistemi di riconoscimento vocale, aiutando a distinguere se l'oratore ha detto

che è troppo brutto o che *è troppo brutto*. Questi milioni di libri forniscono un'ampia serie di dati su cui costruire modelli rappresentativi.

1.2.4 Registri dei taxi di New York

Ogni transazione finanziaria oggi lascia dietro sé una traccia di dati. Seguire questi percorsi può portare a intuizioni interessanti.

I taxi costituiscono una parte importante della rete di trasporto urbano. Girano per le strade della città alla ricerca di clienti, per poi condurli a destinazione con una tariffa proporzionale alla lunghezza del viaggio. Ogni taxi contiene un dispositivo di misurazione per calcolare il costo del viaggio in funzione del tempo. Questo contatore funge da strumento di registrazione e da meccanismo per garantire che l'autista addebiti l'importo corretto per ogni viaggio. I tassametri attualmente utilizzati nei taxi di New York possono fare molte cose oltre al calcolo delle tariffe. Funzionano come terminali per le carte di credito, fornendo un modo per consentire ai clienti di pagare le corse senza contanti. Sono integrati con sistemi di posizionamento globale (GPS), che registrano la posizione esatta di ogni prelievo e consegna. E infine, essendo su una rete wireless, questi box possono comunicare tutti questi dati a un server centrale. Il risultato è un database che documenta ogni singolo viaggio di tutti i taxi di una delle più grandi città del mondo.

Poiché la New York Taxi and Limousine Commission è un'agenzia pubblica, i suoi dati non riservati sono disponibili a tutti in base al Freedom of Information Act (FOA). Ogni corsa genera due record: uno con i dati sul viaggio, l'altro con i dettagli della tariffa. Ogni corsa è associata al medaglione (licenza) di ogni vettura e all'identificatore di ogni conducente. Per ogni viaggio, otteniamo l'ora/data del ritiro e della consegna, oltre alle coordinate GPS (longitudine e latitudine) della località di partenza e della destinazione. Non otteniamo i dati GPS dell'itinerario percorso tra questi punti, ma in una certa misura questo può essere dedotto dal percorso più breve tra di essi. Per quanto riguarda i dati tariffari, otteniamo il costo misurato di ogni viaggio, comprese le tasse, i supplementi e i pedaggi. È tradizione pagare all'autista una mancia per il servizio, l'importo viene anch'esso registrato nei dati. Quindi sto parlando con lei. Questi dati sui taxi sono facilmente disponibili, con registrazioni di oltre 80 milioni di corse negli ultimi anni. Cosa intende fare con questi dati? Qualsiasi serie di dati interessanti può essere utilizzata per rispondere a domande su scale. Questi dati sulle tariffe dei taxi possono aiutarci a capire meglio il settore dei trasporti, ma anche come funziona la città e come potremmo farla funzionare ancora meglio.

Ma le domande più importanti hanno a che fare con la comprensione del trasporto in città. Possiamo utilizzare i tempi di percorrenza dei taxi come sensore per misurare il livello del traffico in città a livello fine. Quanto è più lento il traffico nelle di punta rispetto agli altri orari, e dove i ritardi sono peggiori? L'identificazione delle aree problematiche è il primo passo per proporre soluzioni, modificando i tempi dei semafori, facendo circolare più autobus o creando corsie preferenziali ad alta occupazione. Allo stesso modo, possiamo utilizzare i dati sui taxi per misurare i flussi di trasporto nella città. Dove si recano le persone, in diversi momenti della giornata? Questo ci dice molto di più della semplice congestione. Osservando i dati dei taxi, dovremmo essere in grado di vedere i turisti che si recano dagli hotel alle attrazioni, i dirigenti che si recano dai quartieri eleganti a Wall Street e gli ubriachi che tornano a casa dai locali notturni dopo una sbronza. Dati come questi sono essenziali per progettare sistemi di trasporto migliori. È uno spreco per un singolo passeggero viaggiare dal punto a al punto b, quando al punto $a+\epsilon$ c'è un altro passeggero che vuole raggiungerlo. L'analisi dei dati dei taxi consente una simulazione accurata di un sistema di ride sharing, in modo da poter valutare con precisione le richieste e le riduzioni dei costi di tale servizio.

1.3.1 Dati strutturati e non strutturati

Alcuni set di dati sono ben strutturati, come le tabelle di un database o di un programma di foglio elettronico. Altri registrano informazioni sullo stato del mondo, ma in modo più eterogeneo. Forse si tratta di un corpus di testo di grandi dimensioni con immagini e link, come Wikipedia, o del complicato mix di note e risultati di test che compaiono nelle cartelle cliniche personali. In generale, questo libro si concentrerà sulla gestione dei dati strutturati. I dati sono spesso rappresentati da una *matrice*, dove le righe della matrice rappresentano elementi o record distinti e le colonne rappresentano proprietà distinte di questi elementi. Ad esempio, un set di dati sulle città degli Stati Uniti potrebbe contenere una riga per ogni città, con colonne che rappresentano caratteristiche come lo Stato, la popolazione e l'area. Quando ci troviamo di fronte a una fonte di dati non strutturati, come una raccolta di tweet da Twitter, il nostro primo passo è generalmente quello di costruire una matrice. Un modello di *bag of words* costruirà una matrice con una riga per ogni tweet e una colonna per ogni parola del vocabolario usata di frequente. La voce della matrice $M[i, j]$ denota quindi il numero di volte in cui il tweet i

contiene la parola *j*. Tali formulazioni matriciali motiveranno la nostra discussione sull'algebra lineare.

1.3.2 Dati quantitativi vs. dati categorici

I dati quantitativi consistono in valori numerici, come l'altezza e il peso. Tali dati possono essere incorporati direttamente nelle formule algebriche e nei modelli matematici, oppure visualizzati in grafici e diagrammi convenzionali. Al contrario, i *dati categorici* sono costituiti da etichette che descrivono le proprietà degli oggetti in esame, come il sesso, il colore dei capelli e l'occupazione. Queste informazioni descrittive possono essere altrettanto precise e significative dei dati numerici, ma non possono essere trattate con le stesse tecniche. I dati categorici di solito possono essere codificati numericamente. Per esempio, il sesso può essere rappresentato come *maschio* = 0 o *femmina* = 1. Ma le cose si complicano quando ci sono più di due caratteri per caratteristica. Ma le cose diventano più complicate quando ci sono più di due caratteri per caratteristica, soprattutto quando non c'è un ordine implicito tra di essi. Potremmo essere in grado di codificare i colori dei capelli come numeri, assegnando a ciascuna tonalità un valore distinto. Tuttavia, non possiamo trattare questi valori come numeri, se non per semplici test di identità. I metodi di classificazione e clustering possono essere considerati come la generazione di etichette categoriali da dati numerici e saranno l'obiettivo principale di questo libro.

1.3.3 Big Data vs. Little Data

La scienza dei dati è stata confusa agli occhi del pubblico con i *big data*, l'analisi di enormi serie di dati derivanti dai registri dei computer e dai dispositivi sensoriali. In linea di principio, avere più dati è sempre meglio che averne meno, perché si può sempre buttarne via una parte campionando per ottenere un insieme più piccolo, se necessario. I big data sono un fenomeno entusiasmante. Ma nella pratica, ci sono difficoltà nel lavorare con grandi insiemi di dati. Nel corso di questo libro esamineremo gli algoritmi e le migliori pratiche per l'analisi dei dati. In generale, le cose diventano più difficili quando il volume diventa troppo grande. Le sfide dei big data includono: *Il tempo del ciclo di analisi rallenta con l'aumentare delle dimensioni dei dati*: Le operazioni di calcolo sulle serie di dati richiedono più tempo all'aumentare del loro volume. I fogli di calcolo di piccole dimensioni forniscono una risposta istantanea, consentendo di sperimentare e giocare a *"cosa succede se*? Ma i fogli di calcolo di grandi dimensioni possono essere lenti e goffi da utilizzare,

e gli insiemi di dati abbastanza massicci possono richiedere ore o giorni per ottenere risposte.

Gli algoritmi intelligenti possono permettere di fare cose incredibili con i grandi dati, ma rimanere piccoli in genere porta a un'analisi e a un'esplorazione più rapide. *Le serie di dati di grandi dimensioni sono complesse da visualizzare*: I grafici con milioni di punti sono impossibili da visualizzare sugli schermi dei computer o sulle immagini stampate, per non parlare della comprensione concettuale. Come possiamo sperare di capire davvero qualcosa che non possiamo vedere? *I modelli semplici non richiedono dati enormi per essere adattati o valutati*: Un compito tipico della scienza dei dati potrebbe essere quello di prendere una decisione (ad esempio, se dovrei offrire a questo collega un'assicurazione sulla vita?) sulla base di un numero ridotto di variabili: ad esempio, età, sesso, altezza, peso e presenza o assenza di condizioni mediche esistenti. Se ho questi dati su 1 milione di persone con i loro esiti di vita associati, dovrei essere in grado di costruire un buon modello generale di rischio di copertura.Probabilmente non mi aiuterebbe a costruire un modello sostanzialmente migliore se avessi questi dati su centinaia di milioni di persone. I criteri decisionali su poche variabili (come l'età e lo stato civile) non possono essere troppo complessi e devono essere robusti su un gran numero di richiedenti. Qualsiasi osservazione così sottile da richiedere dati enormi per essere individuata si rivelerà irrilevante per una grande azienda che si basa sul volume. *I big data* sono talvolta chiamati *dati cattivi*. Spesso vengono raccolti come sottoprodotto di un determinato sistema o procedura, invece di essere raccolti in modo mirato per rispondere alla domanda in . Il risultato è che potremmo dover fare eroici per dare un senso a qualcosa solo perché ce l'abbiamo. Consideriamo il problema di avere il polso delle preferenze degli elettori tra i candidati presidenziali. L'approccio basato sui big data potrebbe analizzare i feed massicci di Twitter o Facebook, interpretando gli indizi delle loro opinioni nel testo. L'approccio basato sui piccoli dati potrebbe essere quello di condurre un sondaggio, ponendo a qualche centinaio di persone questa domanda specifica e tabulando i risultati. Quale procedura pensa che si rivelerà più accurata? Il set di dati giusto è quello più direttamente rilevante per i compiti da svolgere, non necessariamente il più grande. Non aspiri ciecamente ad analizzare grandi serie di dati. Cerchi i dati *giusti* per rispondere a una determinata domanda, non necessariamente quelli più grandi su cui può mettere le mani.

1.4 Classificazione e regressione

Due tipi di problemi si presentano ripetutamente nelle applicazioni tradizionali della scienza dei dati e del riconoscimento dei modelli, le sfide

della classificazione e della regressione. Con lo sviluppo di questo libro, ho spinto le discussioni sugli approcci algoritmici alla soluzione di questi problemi verso i capitoli successivi, in modo che possano beneficiare di una solida comprensione del materiale di base della raccolta dei dati, della statistica, della visualizzazione e della modellazione matematica. Tuttavia, menzionerò i problemi legati alla classificazione e alla regressione non appena si presenteranno, quindi ha senso fermarsi qui per una rapida introduzione a questi problemi, per aiutarla a riconoscerli quando li vedrà. *Classificazione*: Spesso cerchiamo di assegnare un'etichetta a un elemento da un insieme discreto di possibilità. Problemi come quello di prevedere il vincitore di un particolare. Una gara sportiva (squadra *A* o squadra *B*?) o decidere il genere di un determinato film (commedia, dramma o animazione?) sono problemi di *classificazione*, poiché ognuno di essi comporta la selezione di un'etichetta tra le scelte possibili. *Regressione*: Un altro compito comune è quello di prevedere una determinata quantità numerica. Prevedere il peso di una persona o quanta neve ci sarà quest'anno è un problema *di regressione*, in cui si prevede il valore futuro di una funzione numerica in termini di valori precedenti e altre caratteristiche rilevanti. Forse il modo migliore per capire la distinzione che si intende fare è osservare una serie di problemi di scienza dei dati ed etichettarli (classificarli) come regressione o classificazione. Per risolvere questi due tipi di problemi si utilizzano metodi algoritmici diversi, anche se spesso le stesse domande possono essere affrontate in entrambi i modi.

1.5 Data Science Television: Il Quant Shop

Credo che l'esperienza pratica sia necessaria per interiorizzare i principi di base. Pertanto, quando insegno la scienza dei dati, mi piace dare a ciascun team di studenti una sfida di previsione interessante ma complicata, e chiedo loro di costruire e valutare un modello predittivo per il compito. Queste sfide di previsione sono associate a eventi in cui gli studenti devono fare previsioni testabili. Partono da zero: trovano i set di dati rilevanti, costruiscono i loro ambienti di valutazione e ideano il loro modello. Infine, li invito a osservare l'evento mentre si svolge, in modo da assistere alla vindi- cazione o al crollo della loro previsione. Come esperimento, nell'autunno 2014 abbiamo documentato in video l'evoluzione del progetto di ciascun gruppo. Montato professionalmente, questo è diventato *The Quant Shop*, una serie televisiva di scienza dei dati per un pubblico generale.

1.5.1 Sfide Kaggle

Un'altra fonte di ispirazione sono le sfide di Kaggle, che offre un forum competitivo per gli scienziati dei dati. Vengono pubblicate nuove sfide, che forniscono una definizione del problema, dati di addestramento e una funzione di punteggio sui dati di valutazione nascosti. Una classifica mostra i punteggi dei concorrenti più forti, in modo che lei possa vedere quanto buono il suo modello rispetto ai suoi avversari. I vincitori rivelano i loro segreti di modellazione durante le interviste post-concorso, per aiutarla a migliorare le sue capacità di modellazione. Ottenere buoni risultati nelle sfide di Kaggle è un'ottima credenziale da inserire nel curriculum per ottenere un buon lavoro come data scientist. Infatti, i potenziali datori di lavoro la rintracceranno se è una vera star di Kaggle. Ma il vero motivo per partecipare è che i problemi sono divertenti e stimolanti, e la pratica aiuta a diventare uno scienziato dei dati migliore. Attenzione: Kaggle offre una visione fuorviante e affascinante della scienza dei dati come apprendimento automatico applicato, perché presenta problemi estremamente ben definiti, con il duro lavoro di raccolta e pulizia dei dati già fatto per lei. Tuttavia, la incoraggio a visitarlo per trarre ispirazione e come fonte di dati per nuovi progetti.

1.6 Sulle storie di guerra

Genio e saggezza sono due doni intellettuali distinti. *Il genio* si manifesta nello scoprire la risposta giusta, nel fare salti mentali fantasiosi che superano gli ostacoli e le sfide. *La saggezza* si manifesta nell'evitare gli ostacoli in primo luogo, fornendo un senso di direzione o una luce guida che ci mantiene in movimento nella giusta direzione. Il genio si manifesta nella forza e nella profondità tecnica, nella capacità di vedere e fare cose che gli altri non possono fare. Al contrario, la saggezza deriva dall'esperienza e dalla conoscenza generale. Viene dall'ascolto degli altri. La saggezza deriva dall'umiltà, dall'osservazione di quanto spesso si è sbagliato in passato e dal capire perché si è sbagliato, in modo da riconoscere meglio le trappole future ed evitarle. La scienza dei dati, come la maggior parte delle cose nella vita, beneficia più della saggezza che del genio. In questo libro, cerco di trasmettere la saggezza che ho accumulato in modo difficile attraverso *storie di guerra*, raccolte da una serie di progetti diversi cui ho lavorato: *Analisi del testo su larga scala e NLP*: il mio Laboratorio di Scienza dei Dati presso la Stony Brook University lavora su una varietà di progetti nell'ambito dei big data, tra cui l'analisi dei sentimenti dai social media, l'analisi delle tendenze storiche, gli approcci di

apprendimento profondo all'elaborazione del linguaggio naturale (NLP) e l'estrazione di caratteristiche dalle reti. *Start-up*: Sono stato co-fondatore e Chief Scientist di due società di analisi dei dati: General Sentiment e Thrivemetrics. General Sentiment analizzava flussi di testo su larga scala provenienti da notizie, blog e social media per identificare le tendenze del sentimento (positivo o negativo) associato a persone, luoghi e cose. Thrivemetrics ha applicato questo tipo di analisi alle comunicazioni aziendali interne, come le email e i sistemi di messaggistica. Nessuna di queste imprese mi ha reso abbastanza ricco da rinunciare alle royalties di questo libro, ma mi hanno fornito un'esperienza sui sistemi informatici basati sul cloud e una visione del modo in cui i dati vengono utilizzati nell'industria. *Collaborare con scienziati reali*: Ho avuto diverse collaborazioni interessanti con biologi e scienziati sociali, che mi hanno aiutato a capire le complessità del lavoro con i dati reali. I dati sperimentali sono terribilmente rumorosi e pieni di errori, ma bisogna fare del proprio meglio con quello che si ha, per scoprire come funziona il mondo. *Costruire sistemi di gioco d'azzardo*: Un progetto particolarmente divertente è stato quello di costruire un sistema per prevedere i risultati delle partite di jai-alai, in modo da poter scommettere su di esse, un'esperienza raccontata nel mio libro *Scommesse calcolate: Computers, Gam- bling, and Mathematical Modeling to Win*. Il nostro sistema si è basato sullo scraping del web per la raccolta dei dati, l'analisi statistica, la simulazione/modellazione e un'attenta valutazione. Abbiamo anche sviluppato e valutato modelli predittivi per gli incassi dei film, i prezzi delle azioni e le partite di calcio utilizzando l'analisi dei social media. *Classifica delle figure storiche*: Analizzando Wikipedia per estrarre le variabili significative su oltre 800.000 personaggi storici, abbiamo sviluppato una funzione di punteggio per classificarli in base alla loro forza come memi storici. *Ognuna di queste storie di guerra è vera.* Naturalmente, le storie migliorano un po' nella rielaborazione e il dialogo è stato rafforzato per renderle più interessanti da leggere.

1.7 Storia di guerra: Rispondere alla domanda giusta

Il nostro gruppo di ricerca presso la Stony Brook University ha sviluppato un sistema basato su NLP per analizzare milioni di notizie, blog e messaggi dei social media, e ridurre questo testo alle tendenze riguardanti tutte le entità in discussione. Contare il numero di menzioni che ogni nome riceve in un flusso

di testo (volume) è facile, in linea di principio. Determinare se la connotazione di un particolare riferimento è positiva o negativa (analisi del sentimento) è difficile. Ma il nostro sistema ha fatto un buon lavoro, soprattutto se aggregato su molti riferimenti. Questa tecnologia è servita come base per un'azienda di analisi dei social media chiamata General Sentiment. È stato emozionante vivere l'avvio di una start-up, affrontare le sfide della raccolta di fondi, dell'assunzione di personale e dello sviluppo di nuovi prodotti. Ma forse il problema più grande che abbiamo affrontato è stato quello di rispondere alla domanda giusta. Il sistema General Sentiment ha registrato i trend relativi al sentimento e al volume di *ogni* persona, luogo e cosa che è stata menzionata nelle notizie, nei blog e nei social media: oltre 20 milioni di entità distinte. Abbiamo monitorato la reputazione di celebrità e politici. Ma si scopre che nessuno la paga per fare qualcosa. La pagano per fare *qualcosa*, per risolvere un problema particolare che hanno, o per eliminare un punto dolente specifico della loro attività. Essere in grado di fare qualsiasi cosa si rivela una pessima strategia di vendita, perché le impone di trovare di nuovo quell'esigenza per ogni singolo cliente.
Facebook si è aperto al mondo solo nel settembre 2006. Quindi, quando General Sentiment è partito nel 2008, eravamo all'inizio dell'era dei social media. Avevamo un grande interesse da parte di grandi marchi e agenzie pubblicitarie che *sapevano* che i social media erano pronti a esplodere. *Sapevano* che questa novità era importante e che dovevano essere presenti. *Sapevano* che un'analisi adeguata dei dati dei social media avrebbe potuto fornire loro nuove conoscenze su ciò pensavano i loro clienti. Ma non sapevano esattamente cosa volevano sapere. Il nostro sistema forniva approfondimenti diversi, provenienti da un mondo completamente diverso. Ma bisognava sapere cosa si voleva per poterli utilizzare. Siamo riusciti a ottenere contratti sostanziali da un gruppo molto eterogeneo di clienti, tra cui marchi di consumo come Toyota e Blackberry, organizzazioni governative come l'ufficio del turismo delle Hawaii e persino la campagna presidenziale del candidato repubblicano Mitt Romney nel 2012. Ogni vendita richiedeva l'ingresso in un nuovo universo, che comportava un notevole sforzo e immaginazione da parte del nostro personale di vendita e degli analisti di ricerca. Non siamo mai riusciti a ottenere due clienti nello stesso settore, il che ci avrebbe permesso di beneficiare della scala e della saggezza accumulata.
Naturalmente, il cliente ha sempre ragione. È stata colpa nostra se non siamo riusciti a spiegare loro il modo migliore per utilizzare la nostra tecnologia. La

lezione che ne deriva è che il mondo non si farà strada fino alla sua porta solo per una nuova fonte di dati. Deve essere in grado di fornire le domande giuste prima di poter trasformare i dati in denaro.

Capitolo 2

Preliminari matematici

2.1 Probabilità

La teoria della probabilità fornisce un quadro formale per ragionare sulla probabilità degli eventi. Poiché si tratta di una disciplina formale, esiste una fitta serie di definizioni associate per istanziare esattamente ciò su cui stiamo ragionando:Un *esperimento* è una procedura che produce uno dei possibili risultati. Come esempio in corso, consideriamo l'esperimento di lanciare due dadi a sei facce, una rossa e una blu, con ogni faccia che reca un interno distinto {1, . . , 6}. Uno *spazio campionario* S è l'insieme dei possibili risultati di un esperimento. Nel nostro , ci sono 36 possibili risultati. Un *evento* E è un sottoinsieme specifico dei risultati di un esperimento. L'evento che la somma dei dadi sia uguale a 7 o 11 (le condizioni per vincere ai dadi al primo lancio) è il sottoinsieme. *La probabilità di un risultato* s, indicata con $p(s)$, è un numero con le due proprietà: per ciascun risultato s nello spazio campionario S, $0 \leq p(s) \leq 1$. La somma delle probabilità di tutti i risultati è pari a uno. La *probabilità di un evento* E è la somma delle probabilità dei risultati dell'esperimento. Una formulazione alternativa è in termini di *complemento* dell'evento $\bar{E}$, il caso in cui E non si verifica. Questo è utile, perché spesso è più facile analizzare $P(\bar{E})$ che $P(E)$ in modo di- retto. Una *variabile casuale* V è una funzione numerica sui risultati di uno spazio di probabilità. La funzione "sommare i valori di due dadi" ($V((a, b)) = a + b$) produce un risultato intero compreso tra 2 e 12. Questo implica una distribuzione di probabilità dei valori della variabile casuale. La probabilità $P(V(s) = 7) = 1/6$, come mostrato in precedenza, mentre $P(V(s) = 12) = 1/36$. Il *valore atteso* di una variabile casuale V definita su uno spazio campionario S, $E(V)$ è definito. Tutto questo probabilmente l'ha già visto in precedenza. Ma fornisce il linguaggio che useremo per collegare la probabilità e la statistica. I dati che vediamo di solito derivano dalla misurazione delle proprietà degli eventi

osservati. La teoria della probabilità e della statistica fornisce gli strumenti per analizzare questi dati.

2.1.1 Probabilità vs. Statistica

La probabilità e la statistica sono aree matematiche correlate che si occupano di analizzare la frequenza relativa degli eventi. Tuttavia, ci sono differenze fondamentali nel modo in cui vedono il mondo. *La probabilità* si occupa di prevedere la probabilità di eventi futuri, mentre *La statistica* comporta l'analisi della frequenza degli eventi passati. *La probabilità* è principalmente un ramo teorico della matematica, che studia le conseguenze delle definizioni matematiche. *La statistica* è principalmente un ramo applicato della matematica, che cerca di dare un senso alle osservazioni nel mondo reale.

Entrambi gli argomenti sono importanti, rilevanti e utili. Ma sono diverse e la comprensione della distinzione è fondamentale per interpretare correttamente la rilevanza delle prove matematiche. Molti giocatori d'azzardo sono finiti in una tomba fredda e solitaria per non aver fatto la giusta distinzione tra probabilità e statistica. La teoria della probabilità ci permette di trovare le conseguenze di un determinato mondo ideale, mentre la teoria statistica ci permette di misurare la misura in cui il nostro mondo è ideale. Questa tensione costante tra teoria e pratica è il motivo per cui gli statistici si dimostrano un gruppo di persone tormentato rispetto ai probabilisti felici e contenti.

La moderna teoria delle probabilità è emersa per la prima volta dai tavoli dei dadi in Francia nel 1654. Chevalier de M'er´e, un nobile francese, si chiese se in un particolare gioco di scommesse fosse avvantaggiato il giocatore o la casa.

2.1.2 Eventi composti e indipendenza

Saremo interessati agli eventi complessi calcolati a partire da eventi più semplici A e B sullo stesso insieme di risultati. Forse l'evento A è che almeno uno dei due dadi sia un numero pari, mentre l'evento B denota il lancio di un totale di 7 o 11. Si noti che esistono alcuni esiti di A che non sono esiti di B. Questa è l'operazione di *differenza degli insiemi*. I risultati in comune tra i due eventi A e B sono chiamati l'*intersezione*, indicata con $A \cap B$. Questo può essere scritto come $A \cap B = A - (S - B)$. I risultati che appaiono sia in A che in B sono chiamati *unione*, indicata con $A\ B.U$ Con l'operazione di complemento $\bar{A} = S\ A$, otteniamo un linguaggio ricco di combinazioni di eventi. Possiamo facilmente calcolare la probabilità di qualsiasi insiemi, sommando le

probabilità dei risultati presenti negli insiemi definiti. Possiamo calcolare facilmente la probabilità di uno qualsiasi di questi insiemi sommando le probabilità dei risultati negli insiemi definiti. li eventi A e B sono *indipendenti* se e solo se $P(A \cap B) = P(A) \times P(B)$. Ciò significa che non esiste una struttura speciale di risultati condivisi tra gli eventi A e B. Supponendo che la metà degli studenti della mia classe sia di sesso femminile e che la metà degli studenti della mia classe sia superiore alla media, ci aspetteremmo che un quarto dei miei studenti sia di sesso femminile e superiore alla media, se gli eventi sono indipendenti. I teorici della probabilità amano gli eventi indipendenti, perché semplificano i loro calcoli. Ma gli scienziati dei dati generalmente non lo fanno. Quando si costruiscono modelli per prevedere la probabilità di un evento futuro B, data la conoscenza di un evento precedente A, vogliamo che la dipendenza di B da A sia il più forte possibile. Supponiamo che io usi sempre un ombrello se e solo se piove. Supponiamo che la probabilità che qui stia piovendo (evento B) sia, ad esempio, $p = 1/5$. Questo implica che la probabilità che io stia portando l'ombrello (evento A) è $q = 1/5$. Ma ancora di più, se conosce lo stato della pioggia, sa esattamente se ho l'ombrello. Questi due eventi sono perfettamente *correlati*. Al contrario, supponiamo che gli eventi siano indipendenti. E il fatto che stia piovendo non ha assolutamente alcun impatto sul fatto che io porti con me la mia attrezzatura protettiva. Le correlazioni sono la forza trainante dei modelli predittivi.

2.1.3 Probabilità condizionale

Quando due eventi sono correlati, esiste una dipendenza tra loro che rende i calcoli più difficili. La *probabilità condizionale* di A dato B, $P(A|B)$ è definita. Ricordiamo gli eventi del lancio dei dadi, ossia: l'evento A prevede che almeno uno dei due dadi sia un numero pari. L'evento B è la somma dei due dadi è un 7 o un 11. Osservi che $P(A|B) = 1$, perché *qualsiasi* lancio che somma un valore dispari deve essere composto da un numero pari e uno dispari. La probabilità condizionale sarà importante per noi, perché siamo interessati alla probabilità di un evento A (forse che una particolare e-mail sia spam) in funzione di alcune prove B (forse la distribuzione delle parole all'interno documento). I problemi di classificazione generalmente si riducono a calcolare le probabilità condizionali, in un modo o nell'altro.

Il nostro strumento principale per calcolare le probabilità condizionali sarà il *teorema di Bayes*, che inverte la direzione delle dipendenze. Spesso si dimostra più facile calcolare le probabilità in una direzione piuttosto che in un'altra, come in questo problema. Per il teorema di Bayes $P(B|A) = (1-9/36)/(25/36) = 9/25$, esattamente quello che abbiamo ottenuto prima. Il teorema di Bayes stabilirà le basi del calcolo delle probabilità di fronte alle prove.

2.1.4 Distribuzioni di probabilità

Le variabili casuali sono funzioni numeriche in cui i valori sono associati a probabilità di accadimento. Nel nostro esempio, in cui $V(s)$ è la somma di due dadi lanciati, la funzione produce un numero intero compreso tra 2 e 12. La probabilità di un particolare valore $V(s) = X$ è la somma delle probabilità di tutti i risultati che si sommano a X. Tali variabili casuali possono essere rappresentate dalla loro *funzione di densità di probabilità*, o pdf. Si tratta di un grafico in cui l'*asse delle* ascisse rappresenta la gamma di valori che la variabile casuale può e l'*asse delle* ordinate indica la probabilità di quel determinato valore. Questi grafici pdf hanno una forte relazione con gli istogrammi di frequenza dei dati, dove l'*asse x* rappresenta di nuovo l'intervallo di valori, ma y rappresenta ora la frequenza osservata di quante occorrenze di eventi si sono verificate esattamente per ogni dato valore X. La conversione di un istogramma in un pdf può essere effettuata dividendo ogni bucket per la frequenza totale su tutti i bucket. La somma delle voci diventa 1, quindi otteniamo una distribuzione di probabilità. Gli istogrammi sono statistici: riflettono le osservazioni reali dei risultati. Al contrario, i pdf sono probabilistici: rappresentano la probabilità sottostante che l'osservazione successiva abbia il valore X. Spesso, nella pratica, utilizziamo l'istogramma delle osservazioni $h(x)$ per stimare le probabilità (Una tecnica chiamata *attualizzazione* offre un modo migliore per stimare la frequenza degli eventi rari) normalizzando i conteggi per il totale delle osservazioni. numero di osservazioni. Esiste un altro modo di rappresentare le variabili casuali che spesso si rivela utile, chiamato *funzione di densità cumulativa* o cdf. La cdf è la somma delle probabilità nel pdf; come funzione di k, riflette la probabilità che $X \leq k$ invece della probabilità che $X = k$. I valori aumentano monotonicamente da sinistra a destra, perché ogni termine deriva dall'aggiunta di una probabilità positiva al totale precedente. Il valore più a destra è 1, perché tutti i risultati producono un valore non superiore al

massimo. È importante capire che il pdf $P(V)$ e il cdf $C(V)$ di una data variabile casuale V contengono *esattamente* le stesse informazioni. Possiamo muoverci avanti e indietro tra loro. Sia consapevole di quale distribuzione sta. Le distribuzioni cumulative diventano sempre più alte man mano che ci si sposta verso destra, culminando con una probabilità di $C(X) = 1$. Al contrario, l'area totale sotto la curva di un pdf è uguale a 1, quindi la probabilità in qualsiasi punto della distribuzione è generalmente molto inferiore. Un esempio divertente della differenza tra distribuzioni cumulative e incrementali. Entrambe le distribuzioni mostrano esattamente gli stessi dati sulle vendite di iPhone di Apple. Entrambe le distribuzioni mostrano esattamente gli stessi dati sulle vendite di iPhone di Apple, ma quale curva ha scelto il CEO di Apple Tim Cook per presentarla ad un importante evento per gli azionisti? La distribuzione cumulativa (in rosso) mostra che le vendite stanno esplodendo, giusto? Ma presenta una visione fuorviante del tasso di crescita, perché la variazione incrementale è la derivata di questa funzione, difficile da visualizzare. Infatti, il grafico delle vendite per trimestre (blu) mostra che il tasso di vendite di iPhone è effettivamente diminuito negli ultimi due periodi prima della presentazione.

2.2 Statistiche descrittive

Le statistiche descrittive forniscono modi per catturare le proprietà di un determinato insieme di dati o campione. Riassumono i dati osservati e forniscono un linguaggio per rappresentare un gruppo di elementi con un nuovo elemento derivato, come la media, la min, il conteggio o la somma, riduce un grande insieme di dati a una piccola statistica riassuntiva: l'aggregazione come riduzione dei dati.

Tali statistiche possono diventare caratteristiche a sé stanti quando vengono prese in considerazione gruppi o cluster naturali nell'insieme dei dati. Esistono due tipi principali di statistiche descrittive: *Misure di tendenza centrale*, che rilevano il centro attorno al quale sono distribuiti i dati. *Misure di variazione* o di *variabilità*, che descrivono la diffusione dei dati, ossia distanza delle misurazioni dal centro.Insieme, queste statistiche ci dicono molto sulla nostra distribuzione.

2.2.1 Misure di centralità

Il primo elemento della statistica a cui siamo esposti a scuola sono le misure di centralità di base: media, mediana e modalità. Sono il punto di partenza

giusto quando si pensa a un singolo numero per caratterizzare un insieme di dati. *Media*: Probabilmente si trova a suo agio con l'uso della *media aritmetica*, in cui sommiamo i valori e li dividiamo per il numero di osservazioni. Possiamo facilmente mantenere la media sotto un flusso di inserimenti e cancellazioni, mantenendo la somma dei valori separata dal conteggio della frequenza, e dividerla solo su richiesta. La media è molto significativa per caratterizzare le distribuzioni simmetriche con valori anomali, come l'altezza e il peso. Il fatto che sia simmetrica significa che il numero di articoli al di sopra della media dovrebbe essere all'incirca uguale al numero di articoli al di sopra della media. sotto. Il fatto che non ci siano valori anomali significa che la gamma di valori è molto ristretta. Si noti che un singolo MAXINT che si insinua in un insieme di osservazioni altrimenti valide, fa perdere di vista la media. La mediana è una misura di centralità che si rivela più appropriata con distribuzioni così poco educate. *Media geometrica*: La *media geometrica* è la radice n-esima del prodotto di *n* valori. La media geometrica è sempre inferiore o uguale alla media aritmetica. È molto sensibile ai valori vicini allo zero. Un singolo valore di zero distrugge la media geometrica: a prescindere dagli altri valori presenti nei dati, si finisce per avere zero. Questo è in qualche modo analogo ad avere un outlier di *infinito* in una media aritmetica. Ma le medie geometriche dimostrano il loro valore quando si fa la media dei rapporti. C'è meno 'spazio' disponibile per i rapporti inferiori a 1 rispetto ai rapporti superiori a 1, creando un'asimmetria che la media aritmetica sovrastima. La media geometrica è più significativa in questi casi, così come la media aritmetica dei *logaritmi* dei rapporti. *Mediana*: La *mediana* è l'esatto valore intermedio di un insieme di dati; tanti elementi si trovano sopra la mediana quanti sotto di essa. C'è un cavillo su cosa prendere come mediana quando si ha un numero pari di elementi. Si può prendere uno dei due candidati centrali: in qualsiasi serie di dati ragionevole, questi due valori dovrebbero essere più o meno uguali. Infatti, nell'esempio dei dadi, entrambi sono pari a 7. Una bella proprietà della mediana così definita è che deve essere un valore autentico del flusso di dati originale. Esiste effettivamente una persona di altezza mediana che si può indicare come esempio, ma presumibilmente nessuno al mondo ha un'altezza *esattamente* media. Si perde questa proprietà quando si fa la media dei due elementi centrali.

Quale misura di centralità è migliore per le applicazioni? La mediana si trova in genere abbastanza vicina alla media aritmetica nelle distribuzioni simmetriche, ma spesso è interessante vedere quanto sono distanti e su quale lato della media si trova la mediana. La mediana si rivela generalmente una statistica migliore per le distribuzioni distorte o per i dati con valori anomali, come la ricchezza e il reddito. Bill Gates aggiunge 250 dollari alla ricchezza media pro capite negli Stati Uniti, ma nulla mediana. Se la fa sentire personalmente più ricca, allora usi pure la media. Ma la mediana è la statistica più informativa in questo caso, come lo sarà per qualsiasi distribuzione a legge di potenza. *Moda*: *La moda* è l'elemento più frequente nel set di dati. Si tratta di 7 nel nostro esempio di dadi in corso, perché si verifica sei volte su trentasei elementi. Francamente, non ho mai visto la modalità fornire molte informazioni come misura di centralità, perché spesso non è vicina al centro. I campioni misurati su un ampio intervallo dovrebbero avere pochissimi elementi ripetuti o collisioni in un particolare valore. Questo rende la modalità una questione di casualità. Infatti, gli elementi più frequenti spesso rivelano artefatti o anomalie in un set di dati, come valori predefiniti o codici di errore che non rappresentano realmente elementi della distribuzione sottostante. Il concetto correlato di picco in una distribuzione di frequenza (o istogramma) è significativo, ma i picchi interessanti vengono rivelati solo attraverso un'adeguata analisi. Il picco attuale della distribuzione dei salari annuali negli Stati Uniti si trova tra i 30.000 e i 40.000 dollari all'anno, anche se la moda si colloca presumibilmente a zero.

2.2.2 Misure di variabilità

La misura più comune della variabilità è la *deviazione standard* σ, che misura la somma dei quadrati delle differenze tra i singoli elementi. Una statistica correlata, la *varianza* V , è il quadrato della deviazione standard, cioè $V = \sigma^2$. A volte è più comodo parlare di varianza che di deviazione standard, perché il termine è più corto di otto caratteri. Ma misurano esattamente la stessa cosa. A titolo di esempio, consideriamo l'umile lampadina, che in genere viene fornita con una durata di lavoro prevista, diciamo $\mu = 3000$ ore, derivata da una qualche distribuzione di base. In una lampadina convenzionale, la possibilità che duri più di μ è presumibilmente uguale a quella che si bruci prima, e questo grado di incertezza è misurato da σ. In alternativa, immaginiamo una "stampante lampadina a cartuccia", dove il produttore malvagio costruisce lampadine molto robuste, ma include un contatore per impedire che si accendano dopo 3000 ore di utilizzo. Qui $\mu = 3000$ e $\sigma = 0$. Entrambe le

distribuzioni hanno la stessa media, ma una varianza sostanzialmente diversa. La penalizzazione della somma dei quadrati nella formula di σ significa che un valore outlier $a\ d$ unità dalla media contribuisce alla varianza quanto d^2 punti ciascuno a un'unità dalla media, quindi la varianza è molto sensibile agli outlier. Una questione spesso confusa riguarda il denominatore nella formula della deviazione standard. Dobbiamo dividere per n o per n 1? La- differenza in questo caso è tecnica. La deviazione standard dell'intera *popolazione* si divide per n, mentre la deviazione standard *del campione* si divide per n 1. Il problema è che il campionamento di un solo punto non ci dice assolutamente nulla sulla deviazione standard. Il problema è che il campionamento di un solo punto non ci dice assolutamente nulla sulla varianza sottostante in qualsiasi popolazione, dove è perfettamente ragionevole dire che c'è una varianza di peso pari a zero nella popolazione di un'isola di una persona. Ma per set di dati di dimensioni ragionevoli $n\approx(n-1)$, quindi non ha davvero importanza.

2.2.3 Interpretare la varianza

Le osservazioni ripetute dello stesso fenomeno non producono sempre gli stessi risultati, a causa del rumore o dell'errore casuale. *Gli errori di campionamento* si verificano quando le nostre osservazioni catturano circostanze non rappresentative, come la misurazione del traffico dell'ora di punta nei fine settimana e durante la settimana lavorativa. *Gli errori di misurazione* riflettono i limiti di precisione insiti in qualsiasi dispositivo di rilevamento. La nozione di *rapporto segnale/rumore* cattura il grado in cui una serie di osservazioni riflette una quantità di interesse, al contrario della varianza dei dati. Come scienziati dei dati, ci interessano i cambiamenti nel segnale anziché nel rumore, e tale varianza spesso rende questo problema sorprendentemente difficile. Penso alla varianza come a una proprietà intrinseca dell'universo, simile alla velocità della luce o al valore temporale del denaro. Ogni mattina che si pesa su una bilancia, è garantito che otterrà un numero diverso, con variazioni che riflettono l'ultima volta che ha mangiato (errore di campionamento), la planarità del pavimento o l'età della bilancia (entrambi errori di misurazione), così come le variazioni della sua massa corporea (variazione reale). Quindi qual è il suo peso reale? Ogni quantità misurata è soggetta a un certo livello di varianza, ma il fenomeno è molto più profondo. Gran parte di ciò che accade nel mondo sono fluttuazioni casuali o eventi arbitrari che causano una varianza anche quando la

situazione è invariata. Gli scienziati dei dati cercano di spiegare il mondo attraverso i dati, ma spesso e volentieri non c'è un fenomeno reale da spiegare, ma solo un fantasma creato dalla varianza. Stagione buona o cattiva, o fortunata/sfortunata: è difficile distinguere il segnale dal rumore.
Prestazioni del modello: Come data scientist, in genere sviluppiamo e valutiamo diversi modelli per ogni sfida predittiva. I modelli possono variare da molto semplici a complessi e variano nelle condizioni di addestramento o nei parametri. In genere, il modello con la migliore accuratezza sul corpus di addestramento sarà portato in trionfo davanti al mondo come quello giusto. Ma le piccole differenze nelle prestazioni tra i modelli sono probabilmente spiegate da una semplice varianza piuttosto che dalla saggezza: quali coppie di formazione/valutazione sono state selezionate, come sono stati ottimizzati i parametri, ecc. Lo ricordi quando si tratta di addestrare modelli di apprendimento automatico. In effetti, quando mi viene chiesto di scegliere tra modelli con piccole differenze di prestazioni, sono più propenso a sostenere il modello più semplice che quello con il punteggio più alto. Date un centinaio di persone che cercano di prevedere il testa o croce in un flusso di lanci di monete, una di loro ha la garanzia di finire con il maggior numero di risposte giuste. Ma non c'è motivo di credere che questo individuo abbia un potere predittivo migliore di tutti noi.

2.2.4 Caratterizzare le distribuzioni

Le distribuzioni non hanno necessariamente una massa di probabilità esattamente nella media. Consideri l'aspetto sua ricchezza dopo aver preso in prestito 100 milioni di dollari e poi averli scommessi tutti su un lancio di moneta alla pari. Se vince ha 100 milioni di dollari in tasca, se perde ha 100 milioni di dollari in tasca. La sua ricchezza attesa è pari a zero, ma questa media non le dice molto sulla forma della distribuzione della sua ricchezza. Tuttavia, considerate insieme, la media e la deviazione standard fanno un lavoro decente per caratterizzare *qualsiasi* distribuzione. Anche una quantità relativamente piccola di massa posizionata lontano dalla media aggiungerebbe molto alla deviazione standard, quindi un piccolo valore di σ implica che la maggior parte della massa deve essere vicina alla media. Per essere precisi, indipendentemente dalla distribuzione dei dati, almeno $(1\ (1/k^2)$ della massa deve trovarsi entro $\pm k$ deviazioni standard della media. Ciò significa che almeno il 75% di tutti i dati deve trovarsi entro 2σ della media, e quasi l'89% entro 3σ per qualsiasi distribuzione.

Vedremo che i limiti sono ancora più stretti quando sappiamo che la distribuzione è ben educata, come la distribuzione gaussiana o normale. Ma questo è il motivo per cui è una grande pratica riportare sia μ che σ ogni volta che si parla di medie. L'altezza media delle donne adulte negli Stati Uniti è di 63,7 2,7 pollici, il che significa che μ= 63,7 e σ= 2,7. La temperatura media a Orlando, Florida, è di 60,3 gradi Fahrenheit.Tuttavia, ci sono stati molti più giorni con 100 gradi a Disney World che donne di 100 pollici in visita per goderne. Riporti sia la media che la deviazione standard per caratterizzare la sua distribuzione, scritta come $\mu \pm \sigma$.

2.3 Analisi della correlazione

Supponiamo di avere due variabili x e y, rappresentate da un campione di n punti forma (x_i, y_i), per 1 i n. Diciamo che≤≤ x e y sono *correlati* quando il valore di x ha un certo potere predittivo sul valore di y. *Il coefficiente di correlazione* $r(X, Y)$ è una statistica che misura il grado in cui Y è una funzione di X e viceversa. Il valore del coefficiente di correlazione varia da 1 a 1, dove 1 significa che è completamente correlato e 0 implica nessuna relazione, o variabili indipendenti. Le correlazioni negative implicano che le variabili sono *anticorrelate*, il che significa che quando X sale, Y scende. Le variabili perfettamente non correlate hanno una correlazione -pari a 1. Si noti che le correlazioni negative sono altrettanto valide ai fini predittivi di quelle positive. Il fatto che si abbia meno probabilità di essere disoccupati quanto maggiore è l'istruzione è un esempio di correlazione negativa, quindi il livello di istruzione può effettivamente aiutare a prevedere la situazione lavorativa. Le correlazioni intorno allo 0 sono inutili per le previsioni. Le correlazioni osservate guidano molti dei modelli predittivi che costruiamo nella scienza dei dati. Inoltre, studiamo come determinare in modo appropriato la forza e la potenza di qualsiasi correlazione osservata, per aiutarci a capire quando le connessioni tra le variabili sono reali.

2.3.1 Coefficienti di correlazione: Pearson e rango di Spearman

In effetti, esistono due statistiche principali utilizzate per misurare la correlazione. Fortunatamente, entrambe operano sulla stessa scala 1 a 1, anche se misurano cose un po' diverse. Queste statistiche diverse sono appropriate in situazioni diverse, quindi dovrebbe conoscerle entrambe.

Il coefficiente di correlazione di Pearson

La più importante delle due statistiche è la correlazione *di Pearson*, definita come Analizziamo questa equazione. Supponiamo che X e Y siano fortemente correlati. Allora ci aspetteremmo che quando x_i è maggiore della media X, allora y_i dovrebbe essere maggiore della sua media $\bar{Y}$. Quando x_i è inferiore alla sua media, y_i dovrebbe seguirla. Ora guardi il numeratore. Il segno di ciascun termine è positivo quando entrambi i valori sono superiori× (1 1) o inferiori−× −(1 1) alle rispettive medie. Il segno di ciascun termine è negativo ((1 1)× o (1 1)) se si× −muovono in direzioni opposte, suggerendo una correlazione negativa. Se X e Y non fossero correlati, i termini positivi e negativi dovrebbero verificarsi con la stessa frequenza, compensandosi a vicenda e portando il valore a zero. L'operazione del numeratore che determina il segno della correlazione è così utile che le diamo un nome, *covarianza*. Rimuovere covarianza. Il denominatore della formula di Pearson riflette la quantità di varianza delle due variabili, misurata dalle loro deviazioni standard. La covarianza tra X e Y aumenta potenzialmente con la varianza di queste variabili, e questo denominatore è la quantità magica per cui dividerla per portare la correlazione a una Scala 1 a 1.

Coefficiente di correlazione di rango Spearman

Il coefficiente di correlazione di Pearson definisce il grado in cui un predittore lineare della forma $y= m x+$ b può adattarsi- ai dati osservati. In genere, questo coefficiente fa un buon lavoro nel misurare la somiglianza tra le variabili, ma è possibile costruire esempi patologici in cui il coefficiente di correlazione tra X e Y è pari a zero, ma Y è completamente dipendente da (e quindi perfettamente prevedibile da) X. Consideriamo i punti della forma $(x, \| \| x)$, dove x è uniformemente (o simmetricamente) campionato dall'intervallo [1, 1]. La correlazione sarà pari a zero perché per ogni punto (x, x) ci sarà un punto opposto (x, x), tuttavia $y= x$ è un predittore perfetto. La correlazione di Pearson misura quanto possano funzionare i migliori predittori *lineari* , ma non dice nulla su funzioni più strane come il valore assoluto. *Il coefficiente di correlazione di rango di Spearman* conta essenzialmente il numero di coppie di punti di ingresso che sono fuori ordine. Supponiamo che il nostro set di dati contenga punti (x_1, y_1) e (x_2, y_2) dove $x_1< x_2$ e $y_1< y_2$.

Questo è un voto che i valori sono correlati positivamente, mentre il voto sarebbe per una correlazione negativa se $y_2 < y_{(1)}$.
Sommando tutte le coppie di punti e normalizzando correttamente, si ottiene la correlazione di rango di Spearman. Lasciare che $rank(x_i)$ sia la posizione di rango di x_i in ordine ordinato tra tutti gli x_i, in modo che il rango del valore più piccolo sia 1 e quello del valore più grande n. Oltre a dare punteggi elevati alle funzioni non lineari ma monotone, la correlazione di Spearman è meno sensibile agli elementi estremi di outlier rispetto a Pearson. Sia $p = (x_1, y_{max})$ il punto di dati con il valore più grande di y in un determinato set di dati. Supponiamo di sostituire p con $p' = (x_1, \infty)$. La correlazione di Pearson impazzirà, poiché il miglior adattamento ora diventa la linea verticale $x = x_1$. Ma la correlazione di Spearman rimarrà invariata, poiché tutti i punti erano sotto p, proprio come lo sono ora sotto p'.

2.3.2 Il potere e il significato della correlazione

Il coefficiente di correlazione r riflette il grado di utilizzo di x per prevedere y in un determinato campione di punti S. Man mano che $r \rightarrow 1$, queste previsioni diventano sempre migliori. Ma la vera domanda è come questa correlazione reggerà nel mondo reale, al di fuori del campione. Le correlazioni più forti hanno un r più grande, ma richiedono anche campioni di punti sufficienti per essere significativi. C'è un detto ironico che dice che se si vuole adattare i dati a una linea retta, è meglio campionarli in soli due punti. La correlazione diventa tanto più impressionante quanto più sono i punti su cui si basa.
I limiti statistici nell'interpretazione delle correlazioni, in base alla forza e alle dimensioni: *Forza della correlazione:* R : Il quadrato del coefficiente di correlazione campionaria r^2 stima la frazione della varianza di Y spiegata da X in una semplice regressione lineare. La correlazione tra altezza e peso è di circa 0,8, il che significa che spiega circa due terzi della varianza. Rapidità con cui r^2 diminuisce con r. C'è un limite prodotto a quanto dovremmo entusiasmarci nello stabilire una correlazione debole. Una correlazione di 0,5 possiede solo il 25% del potere predittivo massimo, e una correlazione di $r = 0{,}1$ solo l'1%. Pertanto, il valore predittivo delle correlazioni diminuisce rapidamente con r. Cosa intendiamo con "spiegare la varianza"? Sia $f(x) = mx + c$ il

valore predittivo di y da x, con i parametri m e c corrispondenti al miglior adattamento possibile. I valori *residui* $r_i = y_i f(x_i)$ avranno- media zero. Inoltre, la varianza dell'intero set di dati V (Y) dovrebbe essere molto più grande di V (r) se c'è un buon adattamento lineareInoltre, la varianza del set di dati completo $V(Y)$ dovrebbe essere molto più grande di $V(r)$ se esiste un buon adattamento lineare $f(x)$. Se x e y sono perfettamente correlati, non dovrebbe esserci alcun errore residuo e $V(r) = 0$. Se x e y sono totalmente non correlati, l'adattamento non dovrebbe contribuire in alcun modo e $V(y) \approx V(r)$. In generale, $1 - r^2 = V(r)/V(y)$. Una serie di punti che ammettono un buon adattamento lineare, con una correlazione $r = 0{,}94$. I residui corrispondenti $r_i = y_i f(x_i)$ sono tracciati a destra. La varianza dei valori y a sinistra $V(y) = 0{,}056$, sostanzialmente maggiore della varianza $V(r) = 0{,}0065$ a destra. Infatti, $1 - r^2 = 0{,}116 \leftrightarrow V(r)/V(y) = 0{,}116$. La significatività statistica di una correlazione dipende dalla dimensione del campione n e da r. Per tradizione, diciamo che una correlazione di n punti è *significativa* se esiste una probabilità α 1/20 = 0,05 di osservare una ≤ correlazione forte come r in qualsiasi insieme casuale di n punti. Questo non è uno standard particolarmente forte. Anche le correlazioni più piccole diventano significative al livello 0,05 con campioni sufficientemente grandi, come mostrato nella Figura 2.8 (a destra). Una correlazione di $r = 0{,}1$ diventa significativa ad = 0,05 intorno a $n = 300$, anche se tale fattore spiega solo l'1% della varianza. Le correlazioni deboli ma significative possono avere valore nei modelli di big data che coinvolgono un gran numero di caratteristiche. Ogni singola caratteristica/correlazione potrebbe spiegare/prevedere solo piccoli effetti, ma un gran numero di correlazioni deboli ma indipendenti può avere un forte potere predittivo. *Forse*. Parleremo di nuovo di significatività in modo più dettagliato nella Sezione 5.3.

2.3.3 La correlazione non implica la causalità!

La correlazione non implica la causalità: il numero di poliziotti attivi in un distretto è fortemente correlato al tasso di criminalità locale, ma la polizia non è la causa del crimine. La quantità di farmaci assunti dalle persone è correlata alla probabilità di essere malati, ma i farmaci non causano la malattia. Nel migliore dei casi, l'implicazione funziona solo in un senso. Ma molte correlazioni osservate sono completamente spurie, con nessuna delle due variabili che ha un impatto reale sull'altra. Tuttavia, la *correlazione implica la causalità* ed è un errore comune nel pensiero, anche tra coloro che comprendono il ragionamento logico. In generale, sono disponibili pochi

strumenti statistici per capire se A provoca davvero B. Possiamo condurre esperimenti controllati, se possiamo manipolare una delle variabili e osservare l'effetto su l'altro. Per esempio, il fatto che possiamo sottoporre le persone a una dieta che le fa dimagrire senza che diventino più basse è una prova convincente che il peso non *causa l'*altezza. Ma spesso è più difficile fare questi esperimenti nell'altro senso, Ad esempio, non esiste un modo ragionevole per rendere le persone più basse, se non tagliando gli arti.

2.3.4 Rilevare le periodicità attraverso l'autocorrelazione

Supponiamo che un alieno spaziale sia stato assunto per analizzare le vendite negli Stati Uniti di un'azienda di giocattoli. Invece di una bella funzione liscia che mostra una tendenza costante, rimarrebbe stupito nel vedere un gigantesco urto ogni dodicesimo mese, ogni anno. Questo alieno avrebbe scoperto il fenomeno del Natale. Le tendenze stagionali riflettono cicli di durata fissa, che salgono e scendono secondo un modello regolare. Molte attività umane procedono con un ciclo di sette giorni associato alla settimana lavorativa. Le grandi popolazioni di un tipo di insetto chiamato *cicala* emergono in un ciclo di 13 o 17 anni, nel tentativo di evitare che i predatori imparino a. Come possiamo riconoscere tali modelli ciclici in una sequenza S? Supponiamo di correlare i valori di S_i con S_{i+p}, per tutti i 1 i n p. Se i≤≤ valori sono in sincronia per un particolare periodo di lunghezza p, allora questa correlazione con se stessa sarà insolitamente alta rispetto ad altri possibili valori di ritardo. Il confronto di una sequenza con se stessa si chiama *autocorrelazione* e la serie di correlazioni per tutti i 1 k n1 si chiama *funzione di autocorrelazione*. La Figura 2.11 presenta una serie temporale di ≤≤vendite giornaliere e la funzione di autocorrelazione associata a questi dati. Il picco ad uno spostamento di sette giorni (e ad ogni multiplo di sette giorni) stabilisce che esiste una periodicità settimanale nelle vendite: si vendono più prodotti nei fine settimana. L'autocorrelazione è un concetto importante nella previsione di eventi futuri, perché significa che possiamo utilizzare le osservazioni precedenti come caratteristiche di un modello. L'euristica secondo cui il tempo di domani sarà simile a quello di oggi si basa sull'autocorrelazione, con un ritardo di p= 1 giorni. Certamente ci aspetteremmo che un modello di questo tipo sia più accurato rispetto alle previsioni fatte sui dati meteo di sei mesi fa (ritardo p = 180 giorni).

In generale la funzione di autocorrelazione per molte quantità tende ad essere più alta per ritardi molto brevi. Ecco perché le previsioni a lungo termine sono meno accurate di quelle a breve termine: le autocorrelazioni sono generalmente molto più deboli. Ma i cicli periodici a volte si estendono molto più a lungo. Infatti, una previsione meteorologica basata su un ritardo di p = 365 giorni sarà molto migliore di una di p = 180, a causa degli effetti stagionali. Il calcolo della funzione di autocorrelazione completa richiede il calcolo di n 1 correlazioni diverse sui punti della serie temporale, che può diventare costoso per n grandi. Fortunatamente, esiste un algoritmo efficiente basato sulla *trasformata veloce di Fourier* (FFT), che permette di costruire la funzione di autocorrelazione anche per sequenze molto lunghe.

2.4 Logaritmi

Il *logaritmo* è la funzione esponenziale inversa $y = b^x$, un'equazione che può essere riscritta come $x = \log_b y$. Questa definizione equivale a dire che $b^{\log_b y} = y$.

Le funzioni esponenziali crescono ad un ritmo molto veloce. Al contrario, i logaritmi crescono ad una velocità molto lenta: questi sono solo gli esponenti delle serie precedenti {1, 2, 3, 4, }. Sono associati a qualsiasi processo dove stiamo ripetutamente moltiplicando per qualche valore di b, o ripetutamente dividendo per b. Basta ricordare la definizione: $y = \log_b x \leftrightarrow b^y = x$.

I logaritmi molto utili e si presentano spesso nell'analisi dei dati. Qui illustro tre ruoli importanti dei logaritmi nella scienza dei dati. Sorprendentemente, solo uno di essi è legato alle sette applicazioni algoritmiche dei logaritmi. I logaritmi sono davvero molto utili.

2.4.1 Logaritmi e moltiplicazione delle probabilità

I logaritmi furono inventati per la prima volta come aiuto al calcolo, riducendo il problema della moltiplicazione a quello dell'addizione. In particolare, per calcolare il prodotto $p = x\,y$, potremmo- calcolare la somma dei logaritmi $s = \log_b x + \log_b y$ e poi prendere l'inverso del logaritmo (cioè alzare b alla potenza s) per ottenere p, perché: $p = x - y = b^{(\log_b x + \log_b y)}$. Questo è il trucco che alimentava i regoli calcolatori meccanici che i geek usavano nei giorni precedenti alle calcolatrici tascabili. Tuttavia, questa idea rimane importante anche oggi, soprattutto quando si moltiplicano lunghe catene di probabilità. Le probabilità sono numeri piccoli. Pertanto, moltiplicando

lunghe catene di probabilità si ottengono numeri *molto* piccoli che regolano le probabilità di eventi molto rari. Ci sono seri problemi di stabilità numerica con la moltiplicazione in virgola mobile sui computer reali. Gli errori numerici si insinueranno e finiranno per sopraffare il valore reale dei numeri abbastanza piccoli. La somma dei logaritmi delle probabilità è molto più stabile dal punto di vista numerico rispetto alla moltiplicazione, ma produce un risultato equivalente. Possiamo elevare la nostra somma a un esponenziale se abbiamo bisogno della probabilità reale, ma di solito non è necessario. Quando dobbiamo solo confrontare due probabilità per decidere quale è più grande, possiamo tranquillamente rimanere nel mondo dei logaritmi, perché i logaritmi più grandi corrispondono a probabilità più grandi. C'è una stranezza di cui bisogna essere consapevoli. Ricordiamo che il $\log_2(\frac{1}{2}) = 1$. I logaritmi delle probabilità sono tutti numeri negativi, ad eccezione di $\log(1) = 0$. Questo è il motivo per cui le equazioni con i loghi delle probabilità presentano spesso segni negativi in posti strani. Faccia attenzione a questi segni.

2.4.2 Logaritmi e rapporti

I rapporti sono quantità della forma a/b. Si trovano spesso nei set di dati come caratteristiche elementari o come valori derivati da coppie di caratteristiche. I rapporti si verificano naturalmente nella normalizzazione dei dati per le condizioni (ad esempio, il peso dopo un certo trattamento rispetto al peso iniziale) o per il tempo (ad esempio, il prezzo di oggi rispetto al prezzo di ieri). Ma i rapporti si comportano in modo diverso quando riflettono gli aumenti rispetto alle diminuzioni. Il rapporto 200/100 è del 200% superiore alla linea di base, ma 100/200 è solo del 50% inferiore, pur essendo un cambiamento di grandezza simile. Pertanto, fare cose come la media dei rapporti significa commettere un peccato statistico. Vuole davvero che un raddoppio seguito da un dimezzamento venga valutato come un aumento, invece che come un cambiamento neutro? Una soluzione in questo caso sarebbe stata quella di utilizzare la media geometrica. Ma la soluzione migliore è quella di prendere il logaritmo di questi rapporti, in modo da ottenere uno spostamento uguale, dato che $\log_2 2 = 1$ e $\log_2(1/2) = -1$. Abbiamo il bonus aggiuntivo che un rapporto unitario corrisponde a zero, quindi i numeri positivi e negativi corrispondono rispettivamente a rapporti impropri e corretti. Un errore da principianti che i miei studenti commettono spesso consiste nel tracciare il valore dei rapporti invece dei loro logaritmi.

Un grafico tratto da un elaborato di uno studente, che mostra il rapporto tra il nuovo punteggio e il vecchio punteggio sui dati di 24 ore (ogni punto rosso rappresenta la misurazione di un'ora) su quattro diverse serie di dati (ognuna delle quali ha una riga). La linea nera solida mostra il rapporto di uno, dove entrambi i punteggi danno lo stesso risultato.

2.4.3 Logaritmi e normalizzazione delle distribuzioni asimmetriche Le variabili che seguono distribuzioni simmetriche, a forma di campana, tendono ad essere buone come caratteristiche nei modelli. Mostrano una variazione sostanziale, quindi possono essere utilizzate per discriminare tra le cose, ma non su un intervallo così ampio da rendere i valori anomali schiaccianti. Ma non tutte le distribuzioni sono simmetriche. La coda a destra va molto più lontano della coda a sinistra. E siamo destinati a vedere distribuzioni molto più sbilenche quando discuteremo le leggi di potenza. La ricchezza è rappresentativa di una distribuzione di questo tipo, in cui l'uomo più povero ha una ricchezza pari a zero o forse negativa, la persona media (ottimisticamente) è nell'ordine delle migliaia di dollari, e Bill Gates sta superando i 100 miliardi di dollari al momento in cui scriviamo. Abbiamo bisogno di una normalizzazione per convertire queste distribuzioni in qualcosa di più facile da gestire. Per suonare la campana di una distribuzione a legge di potenza, abbiamo bisogno di qualcosa di non lineare, che riduca i valori grandi in modo sproporzionato rispetto ai valori più modesti. Il logaritmo è la trasformazione preferita per le variabili a legge di potenza. Se si colpisce la distribuzione a coda lunga con un log, spesso accadono cose positive. La distribuzione era la distribuzione *log-normale*, quindi il logaritmo ha prodotto una perfetta curva a campana sulla destra. L'assunzione del logaritmo delle variabili con una distribuzione a legge di potenza le rende più in linea con le distribuzioni tradizionali. Ad esempio, in qualità di professionista di classe medio-alta, la mia ricchezza dista all'incirca lo stesso numero di log dai miei studenti affamati che da Bill Gates! A volte prendere il logaritmo si rivela un colpo troppo drastico, e una trasformazione non lineare meno drammatica come la radice quadrata funziona meglio per normalizzare una distribuzione. Il test acido consiste nel tracciare una distribuzione di frequenza dei valori trasformati e vedere se appare a forma di campana: grossolanamente simmetrica, con un rigonfiamento al centro. A quel punto si sa di avere la funzione giusta.

2.5 Storia di guerra: Adattare i geni del designer

La parola *bioinformatico* è la parola di scienze della vita per "scienziato dei dati", il praticante di una disciplina emergente che studia collezioni massicce di dati di sequenze di DNA alla ricerca di modelli. I dati di sequenza sono molto interessanti da lavorare e ho partecipato come bioinformatico a progetti di ricerca sin dagli inizi del progetto sul genoma umano. Le sequenze di DNA sono stringhe dell'alfabeto di quattro lettere *{A,C,G,T}*. Le proteine costituiscono il materiale con cui siamo fisicamente costruiti e sono composte da stringhe di 20 tipi diversi di unità molecolari, chiamate aminoacidi. *I geni* sono le sequenze di DNA che descrivono esattamente come creare proteine specifiche, con le unità descritte ciascuna da una tripletta di *A, C, G, T* chiamata *codoni*. Per i nostri scopi, è sufficiente sapere che esiste *{* un numero*}* enorme di possibili sequenze di DNA che descrivono i geni che *potrebbero* codificare una particolare sequenza proteica desiderata. Ma solo una di esse *viene* utilizzata. Io e i miei collaboratori biologi volevamo sapere perché. In origine si pensava che tutte queste diverse codifiche sinonime fossero essenzialmente identiche, ma le statistiche eseguite sui dati di sequenza hanno chiarito che alcuni codoni sono utilizzati più spesso di altri. La convinzione biologica è che "i codoni contano", e ci sono buone ragioni biologiche per cui ciò dovrebbe avvenire. Ci siamo interessati se "le coppie di codoni vicine contano". Forse alcune coppie di triple sono come l'olio e l'acqua e non si mescolano. Alcune coppie di lettere in inglese hanno preferenze di ordine: si vede il bigramma *gh* molto più spesso di *hg*. Forse questo vale anche per il DNA? Se così fosse, ci sarebbero coppie di triple che dovrebbero essere sottorappresentate nei dati di sequenza del DNA. Per verificare questo, avevamo bisogno di un punteggio che confrontasse il numero di volte in cui vediamo effettivamente una particolare tripla (ad esempio *x= CAT*) accanto ad un'altra particolare tripla (ad esempio *y= GAG*) con quello che ci aspetteremmo per caso. Lasciate che *F* (*xy*) sia la frequenza di *xy*, il numero di volte in cui vediamo effettivamente il codone *x* seguito dal codone *y* nel database delle sequenze di DNA. Questi codoni codificano per aminoacidi specifici, ad esempio *a* e *b* rispettivamente. Per l'aminoacido *a*, la probabilità che sia codificato da *x* è *P* (*x*) = *F* (*x*)*/F* (*a*), e allo stesso modo *P* (*y*) = *F* (*y*)*/F* (*b*). Quindi il numero atteso di volte in cui si vede *xy.*

In base a questo, possiamo calcolare un punteggio di coppia di codoni per ogni dato esamerale *xy*

La cosa più importante è che il segno del punteggio distingueva le coppie sovra-rappresentate da quelle sotto-rappresentate. Poiché le grandezze erano simmetriche (+1 era altrettanto impressionante di 1), potevamo sommare o calcolare la media di questi punteggi in modo sensato per ottenere un punteggio per ogni gene. Abbiamo utilizzato questi punteggi per progettare geni che dovrebbero essere negativi per i virus, il che ha fornito una nuova ed entusiasmante tecnologia per la produzione di vaccini. Sapere che certe coppie di codoni erano negative non spiegava *perché* lo fossero. Ma calcolando due punteggi correlati (dettagli non importanti) e ordinando le triplette in base ad essi, sono alcuni schemi. Nota gli schemi? Tutte le sequenze negative a sinistra contengono *TAG*, che risulta essere un codone speciale che indica al gene di fermarsi. E tutte le sequenze negative a destra sono costituite da *C* e *G* in sequenze ripetitive molto semplici. Questo spiega biologicamente perché i modelli sono evitati dall'evoluzione, il che significa che abbiamo scoperto qualcosa di molto significativo sulla vita. Ci sono due lezioni da trarre da questa storia. In primo luogo, lo sviluppo di funzioni di punteggio numerico che evidenziano aspetti specifici degli articoli può essere molto utile per rivelare i modelli.. In secondo luogo, colpire tali quantità con un logaritmo può renderle ancora più utili, permettendoci di vedere la foresta per gli alberi.

Capitolo 3

Munging dei dati

La maggior parte dei data scientist passa gran parte del tempo a pulire e formattare i dati. Gli altri passano la maggior parte del tempo a lamentarsi che non ci sono dati disponibili per fare ciò che vogliono fare.
In questo capitolo, esamineremo alcuni dei meccanismi di base del lavoro con i dati. Non si tratta di cose altisonanti come la statistica o l'apprendimento automatico, ma del lavoro di ricerca dei dati e della loro pulizia, che va sotto il nome di "*data munging"*. Mentre le domande pratiche come "Qual è la migliore libreria o linguaggio di programmazione disponibile?" sono chiaramente importanti, le risposte cambiano così rapidamente che un libro come questo è il posto sbagliato per affrontarle. Pertanto, mi atterrò al livello dei principi

generali, invece di modellare questo libro su un particolare insieme di strumenti software. Tuttavia, in questo capitolo discuteremo il panorama delle risorse disponibili: perché esistono, cosa fanno e come al meglio.
Il primo passo in qualsiasi progetto di scienza dei dati è mettere le mani sui dati giusti. Ma questo è spesso difficile da gestire. Questo capitolo esaminerà le zone di caccia più ricche di risorse di dati, e poi introdurrà le tecniche per pulire ciò che si uccide. La gestione dei dati in modo da poterli analizzare in modo sicuro è fondamentale per ottenere risultati significativi. Come Babbage stesso avrebbe detto in modo più conciso, "spazzatura dentro, spazzatura fuori".

3.1 Linguaggi per la scienza dei dati

In teoria, ogni linguaggio di programmazione sufficientemente potente è in grado di eseguire qualsiasi algoritmo che valga la pena di essere calcolato.
Ma in pratica, alcuni linguaggi di programmazione
I linguaggi si dimostrano molto migliori di altri per compiti specifici. Meglio in questo caso potrebbe significare *più facile per il programmatore* o forse *più efficiente dal punto di vista computazionale*, a seconda della missione da svolgere.
I principali linguaggi di programmazione per la scienza dei dati da conoscere sono: *Python*: È il linguaggio di programmazione più diffuso per la scienza dei dati. Python contiene una serie di funzioni linguistiche che facilitano la raccolta dei dati di base, come le espressioni regolari. Si tratta di un linguaggio interpretato, che rende il processo di sviluppo più rapido e piacevole. Python è supportato da un'enorme varietà di librerie, che fanno di tutto, dallo scraping alla visualizzazione, all'algebra lineare e all'apprendimento automatico. Forse il colpo più grande contro Python è l'efficienza: le lingue interpretate non possono competere con quelle compilate per la velocità. Ma i compilatori di Python esistono, e supportano il collegamento di efficienti librerie in linguaggio C/assemblaggio per i compiti ad alta intensità di calcolo. In conclusione, Python dovrebbe essere il suo strumento principale per lavorare sul materiale che presentiamo in questo libro. *Perl*: Questo era il linguaggio preferito per la raccolta di dati sul web, prima che Python se lo mangiasse per pranzo. Nell'indice di popolarità dei linguaggi di programmazione TIOBE, Python ha superato Perl in popolarità nel 2008 e non si è più guardato indietro. Le ragioni sono diverse, tra cui il supporto più forte per la programmazione orientata agli oggetti e le migliori

librerie disponibili, ma la conclusione è che attualmente ci sono poche buone ragioni per iniziare progetti in Perl. Tuttavia, non si sorprenda se lo incontra in qualche progetto legacy. *R*: è il linguaggio di programmazione degli statistici, con le librerie più profonde disponibili per l'analisi e la visualizzazione dei dati. Il mondo della scienza dei dati è diviso tra i campi R e Python, con R forse più adatto all'esplorazione e Python più adatto all'uso in produzione. Lo stile di interazione con R è un po' un gusto acquisito, quindi la invito a giocarci un po' per vedere se le viene naturale. Esistono collegamenti tra R e Python, per cui può chiamare comodamente le funzioni della libreria R nel codice Python. Ciò consente di accedere a metodi statistici avanzati, che potrebbero non essere supportati dalle librerie native di Python. *Matlab*: Mat qui sta per *matrice*, in quanto Matlab è un linguaggio de- firmato per la manipolazione veloce ed efficiente delle matrici. Come vedremo, molti algoritmi di apprendimento automatico si riducono a operazioni sulle matrici, rendendo Matlab una scelta naturale per gli ingegneri che programmano ad un alto livello di astrazione. Matlab è un sistema proprietario. Tuttavia, molte delle sue funzionalità sono disponibili in GNU Octave, un'alternativa open-source. *Java e C/C++*: Questi linguaggi di programmazione mainstream per sviluppo di grandi sistemi sono importanti nelle applicazioni di big data. Sistemi di elaborazione particolare come Hadoop e Spark si basano rispettivamente su Java e C++. Se vive nel mondo dell'informatica distribuita, allora vive in un mondo di Java e C++, invece che negli altri linguaggi qui elencati. *Mathematica/Wolfram Alpha*: Mathematica è un sistema proprietario che fornisce supporto computazionale per tutti gli aspetti della matematica numerica e simbolica, costruito sul linguaggio di programmazione Wolfram, meno proprietario. È la base del motore di conoscenza computazionale Wolfram Alpha, che elabora query in linguaggio naturale attraverso un mix di algoritmi e fonti di dati predigeriti. Confesso di avere un debole per Mathematica È quello che tendo a usare quando faccio una piccola analisi dei dati o una simulazione, ma il costo lo ha tradizionalmente messo fuori dalla portata di molti utenti. Il rilascio del linguaggio Wolfram forse ora lo apre a una comunità più ampia. *Excel*: I programmi di foglio elettronico come Excel sono strumenti potenti per l'analisi esplorativa dei dati, come ad esempio giocare con una determinata serie di dati per vedere cosa contiene. Meritano il nostro rispetto per queste applicazioni. I programmi di fogli di calcolo completi contengono una quantità sorprendente di funzionalità nascoste per gli utenti esperti. Un mio studente che è diventato un dirigente di Microsoft mi ha detto che il 25% di tutte le nuove richieste di funzionalità per Excel proponevano funzionalità già presenti. Le

funzioni speciali e le caratteristiche di manipolazione dei dati che desidera sono probabilmente presenti in Excel, se cerca bene, allo stesso modo in cui una libreria Python per ciò che le serve sarà probabilmente trovata se la cerca.

3.1.1 L'importanza degli ambienti notebook

Il risultato principale di un progetto di scienza dei dati non dovrebbe essere un programma. Non dovrebbe essere un set di dati. Non dovrebbe essere il risultato dell'esecuzione del programma sui dati. Non dovrebbe essere solo un rapporto scritto. Il risultato di ogni progetto di scienza dei dati dovrebbe essere un taccuino computabile che riunisca il codice, i dati, i risultati computazionali e l'analisi scritta di ciò che si è appreso durante il processo. Un taccuino Jupyter/IPython, integra codice, grafica e documentazione in un documento descrittivo che può essere eseguito come un programma.
Il motivo per cui questo è così importante è che i risultati computazionali sono il prodotto di lunghe catene di selezioni di parametri e decisioni di progettazione. Questo crea diversi problemi che vengono risolti dagli ambienti di calcolo dei notebook: i calcoli devono essere *riproducibili*. Dobbiamo essere in grado di eseguire nuovamente gli stessi programmi da zero e ottenere esattamente lo stesso risultato. Ciò significa che le pipeline di dati devono essere *complete*: prendere l'input grezzo e produrre l'output finale. È un karma terribile iniziare con una serie di dati grezzi, eseguire un'elaborazione, modificare/formattare i file di dati a mano, e poi eseguire un*'altra* elaborazione - perché ciò che si è fatto a mano non può essere prontamente rifatto su un'altra serie di dati, o annullato dopo essersi resi conto di aver commesso un errore. i calcoli devono essere *modificabili*. Spesso la riconsiderazione o la valutazione richiederà una modifica di uno o più parametri o algoritmi. Ciò richiede una nuova esecuzione del notebook per produrre il nuovo calcolo. Non c'è niente di più scoraggiante che ricevere un prodotto di big data senza provenienza e sentirsi dire che *questo* è il risultato finale e che non si può cambiare nulla. Un notebook non è mai finito se non dopo che l'intero progetto è stato completato. Le pipeline di dati devono essere *documentate*. Il fatto che i notebook le permettano di integrare testo e visualizzazioni con il suo codice fornisce un modo potente per comunicare ciò che sta facendo e perché, in modi che gli ambienti di programmazione tradizionali non possono eguagliare. *Utilizzare* un ambiente notebook come IPython o Mathematica per costruire e riportare i risultati di qualsiasi progetto di scienza dei dati.

3.1.2 Formati dati standard

I dati provengono da ogni tipo di luogo e in ogni tipo di formato. La rappresentazione migliore dipende da chi il consumatore finale. Grafici e diagrammi sono modi meravigliosi per trasmettere il significato dei dati numerici alle persone. Infatti, il Capitolo 6 si concentrerà sulle tecniche di visualizzazione dei dati. Ma queste immagini sono essenzialmente inutili come fonte di dati cui fare calcoli. C'è molta strada da fare dalle mappe stampate a Google Maps. I migliori formati di dati computazionali hanno diverse proprietà utili: *Sono facili da analizzare per i computer*: I dati scritti in un formato utile sono destinati ad essere utilizzati di nuovo, altrove. I formati di dati sofisticati sono spesso supportati da API che regolano i dettagli tecnici che garantiscono un formato corretto. *Sono facili da leggere per le persone*: Guardare i dati è un'operazione essenziale in molti contesti. Quale dei file di dati in questa directory è quello giusto da utilizzare? Cosa sappiamo dei campi dati di questo file? Qual è l'intervallo di valori lordi per ogni campo particolare? Questi casi d'uso parlano dell'enorme valore di poter aprire un file di dati in un editor di testo per guardarlo. In genere, ciò significa presentare i dati in un formato di testo codificato leggibile dall'uomo, con i record delimitati da linee separate e i campi separati da simboli di delimitazione. *Sono ampiamente utilizzati da altri strumenti e sistemi*: L'impulso a inventare standard di dati proprietari batte forte nel cuore delle aziende, e la maggior parte degli sviluppatori di software preferirebbe condividere uno spazzolino da denti piuttosto che un formato di file. Ma questi sono impulsi da evitare. La potenza dei dati deriva dalla combinazione e dall'abbinamento con altre risorse di dati, il che è meglio facilitato dall'uso di formati standard popolari. Una proprietà che ho omesso da questo elenco è la *concisione*, poiché in genere non è una preoccupazione primaria per la maggior parte delle applicazioni in esecuzione sui sistemi informatici moderni. La ricerca di minimizzare i costi di archiviazione dei dati spesso va contro altri obiettivi. Impacchettare abilmente più campi nei bit di ordine superiore degli interi consente di risparmiare spazio, ma al costo di renderli incompatibili e illeggibili. Le utility di compressione generale come gzip si dimostrano incredibilmente brave a rimuovere la ridondanza della formattazione a misura d'uomo. I prezzi dei dischi sono incredibilmente bassi: nel momento in cui scrivo è possibile acquistare un'unità da 4 TB per circa 100 dollari, il che significa meno del costo di un'ora di tempo sprecata dallo sviluppatore per programmare un formato più stretto. A meno che non

si operi alla scala di Facebook o Google, la concisione non ha l'importanza che si pensa abbia. I formati/rappresentazioni di dati più importanti da conoscere sono descritti di seguito: *File CSV (comma separated value)*: Questi file rappresentano il formato più semplice e popolare per lo scambio di dati tra programmi. Il fatto che ogni riga rappresenti un singolo record, con i campi separati da virgole, è ovvio a prima vista. Ma le sottigliezze riguardano i caratteri speciali e le stringhe di testo: cosa succede se i dati sui nomi contengono una virgola, come "Thurston Howell, Jr.". Il formato csv offre un modo per sfuggire al codice di tali caratteri, in modo che non vengano trattati come delimitatori, ma è complicato. Un'alternativa migliore è quella di utilizzare un carattere delimitatore più raro, come nei file tsv o *tab separated value.* Il miglior test per verificare se il suo file csv è formattato correttamente è che Microsoft Excel o un altro programma di foglio elettronico possa leggerlo senza problemi. Si assicuri che i risultati di ogni progetto superino questo test non appena il primo file csv è stato scritto, per evitare problemi in seguito. *XML (eXtensible Markup Language)*: I dati strutturati ma non tabellari sono spesso scritti come testo con annotazioni. L'output naturale di un tagger di nomi di entità per il testo avvolge le sottostringhe rilevanti di un testo in parentesi che indicano una persona, un luogo o una cosa. Sto scrivendo questo libro in La- Tex, un linguaggio di formattazione con comandi di parentesi posizionati intorno alle espressioni matematiche e al *testo in corsivo*. Tutte le pagine web sono scritte in HTML, il linguaggio di markup per ipertesti che organizza i documenti utilizzando comandi di parentesi come e per racchiudere il **testo in grassetto**. XML è un linguaggio per scrivere le specifiche di tali linguaggi di markup. Una corretta specifica XML consente all'utente di analizzare qualsiasi documento conforme alla specifica. Progettare tali specifiche e aderire completamente richiede disciplina, ma ne vale la pena. Nella prima versione del nostro sistema di analisi testuale Lydia, abbiamo scritto i nostri markup in uno "pseudoXML", letto da parser ad hoc che gestivano correttamente il 99% dei documenti, ma si rompevano ogni volta che cercavamo di estenderli. Dopo un doloroso passaggio all'XML, tutto ha funzionato in modo più affidabile *ed* efficiente, perché abbiamo potuto impiegare parser XML veloci e open-source per gestire tutto il lavoro sporco di enforcing delle nostre specifiche. *Database SQL (structured query language)*: I fogli di calcolo sono naturalmente strutturati su tabelle singole di dati. Al contrario, i database relazionali si rivelano eccellenti per la manipolazione di tabelle

multiple distinte ma correlate, utilizzando l'SQL per fornire un linguaggio di query goffo ma potente. Qualsiasi sistema di database ragionevole importa ed esporta i record come file csv o XML, oltre a un dump del contenuto interno. La rappresentazione interna dei database è opaca, quindi non è corretto descriverli come un formato di dati. Tuttavia, li sottolineo qui perché i database SQL si dimostrano generalmente una soluzione migliore e più potente rispetto alla manipolazione di più file di dati in modo ad hoc. *JSON (JavaScript Object Notation)*: Si tratta di un formato per la trasmissione di oggetti di dati tra programmi. È un modo naturale per comunicare lo stato delle variabili/strutture di dati da un sistema all'altro. Questa rappresentazione è fondamentalmente un elenco di coppie attributo-valore corrispondenti a nomi di variabili/campi e ai valori associati. Poiché le funzioni di libreria che supportano la lettura e la scrittura di oggetti JSON sono facilmente disponibili in tutti i moderni linguaggi di programmazione, è diventato un modo molto comodo per archiviare strutture di dati da utilizzare successivamente. Gli oggetti JSON sono leggibili per l'uomo, ma hanno un aspetto piuttosto disordinato, in quanto rappresentano array di record rispetto ai file CSV. Li utilizzi per oggetti strutturati complessi, ma non per semplici tabelle di dati. *Buffer di protocollo*: Si tratta di un modo neutro dal punto di vista del linguaggio/piattaforma di serializzare i dati strutturati per le comunicazioni e l'archiviazione tra le applicazioni. Si tratta essenzialmente di versioni più leggere di XML (dove si definisce il formato dei dati strutturati), progettate per comunicare piccole quantità di dati attraverso programmi come JSON. Questo formato di dati è utilizzato per gran parte delle comunicazioni tra macchine di Google. Apache Thrift è uno standard correlato, utilizzato da Facebook.

3.2 Raccolta dei dati

La questione più critica in qualsiasi progetto di data science o di modellazione è trovare il set di dati giusto. L'identificazione di fonti di dati valide è un'arte, che ruota attorno a tre domande fondamentali:

- Chi potrebbe avere i dati di cui ho bisogno?
- Perché potrebbero decidere di metterlo a mia disposizione?
- Come posso metterci mani sopra?

In questa sezione, esploreremo le risposte a queste domande. Esaminiamo le fonti comuni di dati e ciò che è probabile che sia in grado di trovare e perché. Passiamo poi in rassegna i meccanismi principali per ottenere l'accesso, tra cui le API, lo scraping e il logging.

3.2.1 Caccia

Chi ha i dati e come può ottenerli?

de e fonti di dati proprietari

Grandi aziende come Facebook, Google, Amazon, American Express e Blue Cross dispongono di quantità incredibili di dati interessanti sugli utenti e sulle transazioni, dati che potrebbero essere utilizzati per migliorare il funzionamento del mondo. Il problema è che ottenere un accesso esterno è solitamente impossibile. Le aziende sono riluttanti a condividere i dati per due buoni motivi:

- Problemi di business e la paura di aiutare la concorrenza.
- Problemi di privacy e paura di offendere i clienti.

Una storia commovente di ciò che può accadere con il rilascio di dati aziendali si è verificata quando AOL ha fornito agli accademici una serie di dati di milioni di query al suo motore di ricerca, accuratamente privati delle informazioni di identificazione. La prima cosa che gli studiosi hanno scoperto è che le query inserite più frequentemente erano tentativi disperati di fuga verso altri motori di ricerca come Google. Questo non ha aumentato la fiducia del pubblico nella qualità della ricerca di AOL.

La seconda scoperta è stata che l'anonimizzazione delle query di ricerca si è rivelata molto più difficile di quanto si pensasse in precedenza. Certo, si possono sostituire i nomi degli utenti con i numeri id, ma non è così difficile capire chi l'uomo di Long Island che interroga ripetutamente *Steven Skiena*, *Stony Brook* . In effetti, non appena è stato reso noto che l'identità delle persone è stata rivelata da questo rilascio di dati, il responsabile è stato licenziato e il set di dati è scomparso. La privacy degli utenti è importante e le questioni etiche relative alla scienza dei dati. Quindi non pensi di convincere le aziende a rilasciare i dati confidenziali degli utenti. Tuttavia, molte aziende responsabili come *il New York Times*, Twitter, Facebook e Google rilasciano alcuni dati, in genere tramite interfacce di programmi applicativi (API) a tariffa limitata. In genere hanno due motivazioni:

- Fornire ai clienti e alle terze parti dati che possono aumentare le vendite. Ad esempio, rilasciare dati sulla frequenza delle query e sui prezzi degli annunci può incoraggiare un maggior numero di persone a pubblicare annunci su una determinata piattaforma.

- In genere è meglio per l'azienda fornire API ben educate, piuttosto che avere cowboy che ripetutamente martellano e raschiano il loro sito.

Altre organizzazioni forniscono download di massa di dati interessanti per l'analisi off- line, come nel caso dei set di dati di Google Ngrams, IMDb e delle tariffe dei taxi di cui si è parlato nel Capitolo 1. I set di dati di grandi dimensioni sono spesso accompagnati da metadati preziosi, come i titoli dei libri, le didascalie delle immagini e la cronologia delle modifiche, che possono essere riutilizzati con la giusta immaginazione.
Infine, la maggior parte delle organizzazioni dispone di set di dati interni rilevanti per la loro attività. In qualità di dipendente, dovrebbe essere in grado di ottenere un accesso privilegiato durante il suo lavoro. Tenga presente che le aziende hanno politiche di accesso ai dati interni, per cui sarà comunque soggetto a determinate restrizioni. Violare i termini di queste politiche è un modo eccellente per diventare un ex dipendente.

Fonti di dati governativi

La raccolta di dati è una delle cose importanti che i governi fanno. In effetti, il requisito che gli Stati Uniti conducano un censimento della popolazione è previsto dalla nostra Costituzione, e viene eseguito puntualmente ogni dieci anni dal 1790. I governi delle città, degli Stati e dei Paesi federali si sono impegnati sempre di più nell'apertura dei dati, per facilitare nuove applicazioni e migliorare il modo in cui il governo può adempiere alla sua missione. Il sito web http://Data.gov è un'iniziativa del Governo federale per raccogliere in modo centralizzato le sue fonti di dati e, all'ultimo conteggio, conta oltre 100.000 set di dati! I dati governativi si distinguono da quelli industriali , in linea di principio, appartengono al popolo. *La Legge sulla Libertà di Informazione* (FOI) consente a qualsiasi cittadino di fare una richiesta formale per qualsiasi documento o serie di dati governativi. Tale richiesta avvia un processo per determinare ciò che può essere rilasciato senza compromettere l'interesse nazionale o violare la privacy.

I governi statali operano in base a cinquanta legislazioni diverse, per cui i dati che sono strettamente conservati in una giurisdizione possono essere liberamente disponibili in altre. Le grandi città come New York hanno operazioni di elaborazione dati più grandi di molti Stati, anche in questo caso con restrizioni che variano a seconda della località. Consiglio il seguente modo di pensare ai documenti governativi. Se non riesce a trovare quello che le serve online dopo aver curiosato un po', cerchi di capire quale agenzia probabilmente lo possiede. Faccia una telefonata amichevole per vedere se possono aiutarla a trovare ciò che desidera. Ma se le fanno ostruzionismo, si senta libero di fare una richiesta di legge FOI. La tutela della privacy è in genere la questione principale nel decidere se un particolare set di dati governativi può essere rilasciato.

Set di dati accademici

Esiste un vasto mondo di studi accademici, che copre tutto ciò che l'umanità ha ritenuto degno di essere conosciuto. Una parte crescente della ricerca accademica comporta la creazione di grandi serie di dati. Molte riviste ora richiedono di mettere i dati di partenza a disposizione di altri ricercatori prima della pubblicazione. Se si cerca bene, si possono trovare grandi quantità di dati economici, medici, demografici, storici e scientifici. La chiave per trovare questi set di dati è rintracciare i documenti pertinenti. Esiste una letteratura accademica su qualsiasi argomento di interesse. Google Scholar è la fonte più accessibile di pubblicazioni di ricerca. Effettui una ricerca per argomento, e magari per "Scienza aperta" o "dati". Le pubblicazioni di ricerca in genere forniscono indicazioni su dove si possono trovare i dati associati. In caso contrario, contattando direttamente l'autore con una richiesta, si otterrà rapidamente il risultato desiderato. L'ostacolo maggiore nell'utilizzo di serie di dati pubblicati è che qualcun altro ha lavorato duramente per analizzarli prima che lei li raggiungesse, quindi queste fonti precedentemente estratte potrebbero essere state prosciugate di nuovi risultati interessanti. Tuttavia, se si pongono nuove domande ai vecchi dati, in genere si aprono nuove possibilità. Spesso i progetti di scienza dei dati interessanti prevedono la collaborazione tra ricercatori di discipline diverse, come le scienze sociali e naturali. Queste persone parlano lingue diverse dalle sue e possono sembrare intimidatorie all'inizio. Ma spesso accolgono con favore la collaborazione e, una volta superato il gergo, di solito è possibile comprendere i loro problemi a un livello ragionevole senza

studi specialistici. Sia certo che le persone di altre discipline non sono generalmente più intelligenti di lei.

Equità del sudore

A volte dovrà lavorare per i suoi dati, invece di prenderli da altri. Molti dati storici esistono ancora solo nei libri o in altri documenti cartacei, il che richiede l'inserimento manuale e la cura. Un grafico o una tabella potrebbero contenere le informazioni di cui abbiamo bisogno, ma può essere difficile ottenere i numeri da un grafico bloccato in un file PDF (portable document format). Ho osservato che le persone orientate al calcolo sopravvalutano enormemente la quantità di sforzi necessari per l'inserimento manuale dei dati. Con un record al minuto, si possono facilmente inserire 1.000 record in soli due giorni lavorativi. Invece, le persone orientate alla computazione tendono a dedicare sforzi enormi per evitare questo lavoro pesante, come la ricerca vana di sistemi di riconoscimento ottico dei caratteri (OCR) che non facciano un pasticcio del file, o spendere più tempo per ripulire una scansione rumorosa di quanto ne occorrerebbe per digitarla di nuovo. Una via di mezzo è rappresentata dal pagamento di qualcun altro che faccia il lavoro sporco per lei. Le piattaforme di crowdsourcing come Amazon Turk e CrowdFlower le consentono di pagare eserciti di persone che la aiutino ad estrarre i dati, o addirittura a raccoglierli. I compiti che richiedono un'annotazione umana, come l'etichettatura delle immagini o la risposta ai sondaggi , sono particolarmente utili per i lavoratori a distanza.Molte incredibili risorse di dati aperti sono state costruite da team di collaboratori, come Wikipedia, Freebase e IMDb. Ma c'è un concetto importante da ricordare: in genere le persone lavorano meglio quando vengono.

3.2.2 Raschiamento

Le pagine web spesso contengono testi e dati numerici preziosi, sui quali vorremmo mettere le mani. Ad esempio, nel nostro progetto di costruire un sistema di gioco d'azzardo per lo sport del jai-alai, avevamo bisogno di fornire al nostro sistema i risultati delle partite di ieri e il programma delle partite di oggi. La nostra soluzione è stata quella di scrutare i siti web delle agenzie di scommesse di jai-alai, che hanno pubblicato queste informazioni per i loro fan. Ci sono due fasi distinte per far sì che ciò avvenga, lo spidering e lo scraping:

- Lo *spidering* è il processo di download del giusto set di pagine da analizzare.

- Lo *scraping* è l'arte raffinata di togliere questo contenuto da ogni pagina per prepararlo all'analisi computazionale.

La prima cosa da capire è che le pagine web sono generalmente scritte in linguaggi di formattazione semplici da capire, come HTML e/o JavaScript. Il suo browser conosce questi linguaggi e interpreta il testo della pagina web come un programma per specificare cosa visualizzare. Chiamando una funzione che emula/finge di essere un browser web, il suo programma può scaricare qualsiasi pagina web e interpretarne il contenuto per analizzarlo. Tradizionalmente, i programmi di scraping erano script specifici per il sito, modificati per cercare particolari modelli HTML che affiancano il contenuto di interesse. In questo modo si sfruttava il fatto che un gran numero di pagine di siti web specifici sono generate dai programmi stessi, e quindi altamente prevedibili nel loro formato. Ma tali script tendono ad essere brutti e fragili, e si rompono ogni volta che il sito web di destinazione modifica la struttura interna delle sue pagine. Oggi le librerie in linguaggi come Python (veda BeautifulSoup) rendono più facile scrivere spider e scraper robusti. In effetti, probabilmente qualcun altro ha *già* scritto uno spider/scraper per ogni sito web popolare e lo ha reso disponibile su SourceForge o Github, quindi cerchi prima di scrivere. Alcune missioni di spidering possono essere banali, ad, colpire un singolo URL (localizzatore uniforme di risorse) a intervalli di tempo regolari. Tali schemi si verificano nel monitoraggio, ad esempio, della posizione di vendita di questo libro dalla sua pagina Amazon. Approcci un po' più sofisticati allo spidering si basano sulla regolarità del nome del sito. URL sottostanti. Se tutte le pagine di un sito sono specificate dalla data o dal numero ID del prodotto, l'iterazione dell'intera gamma di valori interessanti diventa solo una questione di conteggio. La forma più avanzata di spidering è il *web crawling*, in cui si attraversano sistematicamente tutti i link in uscita da una determinata pagina principale, continuando in modo ricorsivo fino a visitare ogni pagina del sito web di destinazione. Questo è ciò che fa Google nell'indicizzazione del web. Può farlo anche lei, con sufficiente pazienza e con librerie di web crawling in Python facili da trovare. La preghiamo di comprendere che la cortesia limita la velocità di spider/crawling di un determinato sito web. È considerata una cattiva forma colpire un sito più di una volta al secondo, e in effetti le migliori pratiche impongono ai provider di bloccare l'accesso alle persone che li stanno martellando.

Ogni sito web importante contiene un documento *sui termini di servizio* che limita ciò che lei può fare legalmente con i dati associati. In linea di massima, la maggior parte dei siti la lascerà in pace a condizione che non li martelli e che non ridistribuisca i dati che ha raccolto. Si tratta di un'osservazione, non di un parere legale. In effetti, legga il caso di Aaron Schwartz, dove una nota figura di Internet è stata perseguita con gravi accuse penali per aver violato i termini di servizio nello spider/scraping di articoli di riviste, e letteralmente perseguitata a morte. Se sta tentando un progetto di web-scraping a livello professionale, si assicuri che il management comprenda i termini di servizio prima di essere troppo creativo con la proprietà di qualcun altro.

3.2.3 Registrazione

Se possiede una potenziale fonte di dati, la tratti come se di sua proprietà. L'accesso interno a un servizio web, a un dispositivo di comunicazione o a uno strumento di laboratorio le garantisce il diritto e la responsabilità di registrare tutte le attività per l'analisi a valle. Si possono fare cose incredibili con la raccolta di dati ambientali dai weblog e dai dispositivi di rilevamento, presto destinati ad esplodere con l'imminente "Internet delle cose". Gli accelerometri dei telefoni cellulari possono essere utilizzati per misurare la forza dei terremoti, con la correlazione degli eventi all'interno di una regione sufficiente a filtrare le persone che guidano su strade dissestate o che lasciano il telefono nell'asciugatrice. Il monitoraggio dei dati GPS di una flotta di taxi traccia la congestione del traffico nelle strade cittadine. L'analisi computazionale dei flussi di immagini e video apre le porte a innumerevoli applicazioni. Un'altra idea interessante è quella di utilizzare le fotocamere come strumenti meteorologici, osservando il colore del cielo sullo sfondo dei milioni di fotografie caricate ogni giorno sui siti fotografici. Il motivo principale per strumentare il suo sistema per raccogliere dati è che può farlo. Forse non sa esattamente cosa farne ora, ma qualsiasi serie di dati ben costruita diventerà probabilmente di valore una volta raggiunta una certa massa critica di dimensioni.

I costi attuali dell'archiviazione chiariscono quanto sia bassa la barriera per la strumentazione di un sistema. Il mio locale Costco sta vendendo tre unità disco da un terabyte a un prezzo inferiore a 100 dollari, ovvero Big O di niente. Se ogni record di transazione richiede 1 kilobyte (uno

migliaia di caratteri), questo dispositivo in linea di principio ha spazio per 3 miliardi di record, circa uno ogni due persone sulla terra. Le considerazioni importanti nella progettazione di qualsiasi sistema di registrazione sono: lo costruisca per durare nel tempo con una manutenzione limitata. Lo imposti e lo dimentichi, dotandolo di spazio di archiviazione sufficiente per un'espansione illimitata e di un backup. Memorizzi tutti i campi di possibile valore, senza impazzire. Utilizzi un formato leggibile dall'uomo o un database di transazioni, in modo da poter capire esattamente cosa contiene quando arriva il momento, mesi o anni dopo, di sedersi e analizzare i suoi dati.

3.3 Dati di pulizia

"Garbage in, garbage out" è il principio fondamentale dell'analisi dei dati. Il percorso che porta dai dati grezzi a una serie di dati puliti e analizzabili può essere lungo. Nella pulizia dei dati per l'analisi possono sorgere molti problemi potenziali. In questa sezione, discutiamo l'identificazione degli artefatti di elaborazione e l'integrazione di serie di dati diverse. La nostra attenzione si concentra sull'elaborazione *prima dell*'analisi vera e propria, per assicurarci che la spazzatura non arrivi mai. La pulizia dei dati viene sempre eseguita su una copia dei dati originali, idealmente da una pipeline che apporta modifiche in modo sistematico e ripetibile.

3.3.1 Errori vs. artefatti

Secondo l'antica legge ebraica, se un sospetto sotto processo veniva giudicato colpevole all'unanimità da tutti i giudici, questo sospetto veniva *assolto*. I giudici avevano notato che il consenso unanime spesso indica la presenza di un errore sistemico nel processo giudiziario. Hanno ragionato sul fatto che quando qualcosa sembra troppo bello per essere vero, è probabile che sia stato commesso un errore da qualche parte. Se consideriamo i dati come misurazioni su qualche aspetto del mondo, gli *errori* dei dati rappresentano informazioni che sono fondamentalmente perse durante l'acquisizione. Il rumore gaussiano che offusca la risoluzione dei nostri sensori rappresenta un errore, una precisione che è stata persa in modo permanente. Le due ore di registri mancanti a causa del crash del server rappresentano un errore di dati: si tratta di informazioni che non possono essere ricostruite. Al contrario, gli *artefatti* sono generalmente problemi sistematici derivanti dall'elaborazione effettuata sulle informazioni grezze da cui sono state costruite. La buona notizia è che gli artefatti di elaborazione possono essere corretti, a condizione

che il set di dati grezzi originali rimanga disponibile. La cattiva notizia è che questi artefatti devono essere rilevati prima di poter essere corretti.
La chiave per individuare gli artefatti di elaborazione è la "prova del fiuto", esaminando il prodotto abbastanza da vicino da percepire qualcosa di negativo. Qualcosa di negativo è solitamente qualcosa di inaspettato o sorprendente, perché le persone sono naturalmente ottimiste. Le osservazioni sorprendenti sono ciò per cui gli scienziati dei dati vivono. In effetti, tali intuizioni sono la ragione principale per cui facciamo quello che facciamo. Ma nella mia esperienza, la maggior parte delle sorprese si rivelano artefatte, quindi dobbiamo guardarle con scetticismo. La chiave per trovare gli artefatti è cercare le anomalie nei dati, che contraddicono ciò che ci si aspetta di vedere. Come *dovrebbe essere* la distribuzione del numero di autori vergini e come dovrebbe cambiare nel tempo? Per prima cosa, costruisca una distribuzione preventiva di ciò che si aspetta di vedere, in modo da poter valutare correttamente le potenziali anomalie rispetto ad essa. Secondo la mia intuizione, la distribuzione dei nuovi scienziati di punta dovrebbe essere piuttosto piatta, perché ad ogni classe successiva di nascono nuove stelle. Immagino anche che ci possa essere una deriva graduale verso l'alto con l'espansione della popolazione e l'ingresso di più persone nella comunità scientifica. Ma questo non è ciò che vedo nella Figura 3.2. Quindi cerchi di enumerare quali sono le anomalie/potenziali artefatti. . . 1965, e un picco che esplode nel 2002. Riflettendoci, il picco più a sinistra ha senso. Questo picco a sinistra si verifica nell'anno in cui Pubmed ha iniziato a raccogliere sistematicamente i record bibliografici. Sebbene esistano alcuni dati completi molto interiori dal 1960 al 1964, la maggior parte degli scienziati più anziani che avevano pubblicato articoli per diversi anni sarebbero "emersi" solo con l'inizio della raccolta sistematica di dati nel 1965. Questo spiega il picco a sinistra, che poi si stabilizza entro il 1970 in quella che sembra la distribuzione piatta che ci aspettavamo. Ma che dire del gigantesco picco del 2002? E il declino dei nuovi autori, quasi nullo negli anni precedenti? Un calo simile è visibile anche a destra del grande picco. Tutti i principali scienziati del mondo erano destinati a nascere nel 2002? Un'attenta ispezione dei record nel grande picco ha rivelato la fonte dell'anomalia: i nomi. Agli albori di Pubmed, gli autori erano identificati dalle loro iniziali e dai loro cognomi. Ma alla fine del 2001, *SS Skiena* è diventato *Steven S. Skiena*, quindi *sembrava* che un nuovo autore stesse emergendo dal cielo.

Ma perché il declino verso il nulla a sinistra e a destra di questo picco? Ricordiamo che abbiamo limitato questo studio ai 100.000 scienziati più prolifici. È improbabile che una rock star della scienza, nata nel 1998, appaia in questa classifica, perché il suo nome era destinato a cambiare pochi anni dopo, non lasciando abbastanza tempo per accumulare una carriera completa di articoli. Cose simili accadono all'estrema destra della distribuzione: gli scienziati appena nati nel 2010 non sarebbero mai in grado di realizzare un'intera carriera di lavori in un paio d'anni. Entrambi i fenomeni sono perfettamente spiegati da questa base di nomi. La pulizia di questi dati per unificare i riferimenti ai nomi ci ha richiesto alcune iterazioni per essere corretta. Anche dopo aver eliminato il picco del 2002, abbiamo riscontrato un calo sostanziale di scienziati di spicco che hanno iniziato la loro carriera a metà degli anni Novanta. Questo perché molte persone che hanno avuto un'ottima mezza carriera prima dei nomi e una seconda ottima mezza carriera dopo i nomi non hanno raggiunto la soglia di un'ottima carriera completa. in entrambi i periodi. Pertanto, abbiamo dovuto abbinare tutti i nomi nell'intero *periodo precedente.* identificando chi fossero i 100.000 scienziati migliori. Siate sempre sospettosi se i vostri dati sono abbastanza puliti da potersi fidare.

3.3.2 Compatibilità dei dati

Diciamo che un confronto tra due elementi è "mele a mele" quando si tratta di un confronto equo, cioè quando gli elementi coinvolti sono abbastanza simili da poter essere messi a confronto in modo significativo. Al contrario, i confronti "mele contro arance" sono in definitiva privi di significato. Questi tipi di problemi di comparabilità dei dati sorgono ogni volta che i set di dati vengono uniti. Qui spero di mostrarle quanto possano essere insidiosi questi problemi di comparabilità, per sensibilizzarla sul perché deve esserne consapevole. Inoltre, per alcune importanti classi di conversioni, indico i modi per. Esamini il significato di ogni campo di qualsiasi serie di dati con cui lavora. Se non capisce cosa c'è dentro, fino alle unità di misura, non c'è un modo sensato per utilizzarlo.

Conversioni di unità

La quantificazione delle osservazioni nei sistemi fisici richiede unità di misura standard. Purtroppo esistono molti sistemi di misura funzionalmente

equivalenti ma incompatibili. Io e mia figlia di 12 anni pesiamo entrambe circa 70 chili, ma una di noi è in chili e l'altra in chilogrammi.

Quando le misurazioni vengono inserite nei sistemi informatici utilizzando le unità di misura sbagliate, accadono cose disastrose come le esplosioni di razzi. In particolare, la NASA ha perso la missione spaziale Mars Climate Orbiter, costata 125 milioni di dollari, il 23 settembre 1999, a causa di un problema di conversione metrica-inglese. Questi problemi si affrontano meglio selezionando un unico sistema di misura e attenendosi ad esso. Il sistema metrico offre diversi vantaggi rispetto al sistema tradizionale inglese. In particolare, le singole misure sono naturalmente espresse come singole quantità decimali (come 3,28 metri) invece di coppie di quantità incomparabili (5 piedi, 8 pollici). Lo stesso problema si presenta nella misurazione degli angoli (radianti vs. gradi/secondi) e del peso (chilogrammi vs. libbre/oz). Attenersi al sistema metrico non risolve di per sé tutti i problemi di comparabilità, poiché non c'è nulla che impedisca di mescolare altezze in metri e centimetri. Ma è un buon inizio. Come può difendersi dalle unità incompatibili quando unisce i set di dati? La vigilanza deve essere la sua arma principale. Si assicuri di conoscere le unità previste per ogni colonna numerica del suo set di dati e verifichi la compatibilità al momento della fusione. Qualsiasi colonna che non abbia un'unità o un tipo di oggetto associato deve essere immediatamente sospettata. Quando si uniscono record provenienti da fonti diverse, è una pratica eccellente creare un nuovo campo "origine" o "fonte" per identificare la provenienza di ciascun record. Questo fornisce almeno la speranza che gli errori di conversione delle unità possano essere corretti in seguito, operando sistematicamente sui record provenienti dalla fonte problematica. Una procedura parzialmente automatizzata per rilevare tali problemi può essere ideata a partire dal test di significatività statistica. Supponiamo di tracciare le frequenze delle altezze umane in un insieme di dati fusi di misure inglesi (piedi) e metriche (metri). Vedremmo un picco nella distribuzione intorno a 1,8 e un secondo intorno a 5,5. L'esistenza di più picchi in una distribuzione dovrebbe renderci sospettosi. Il *valore* p risultante dal test di significatività sulle due popolazioni di input fornisce una misura rigorosa del grado di convalida dei nostri sospetti.

Conversioni di rappresentazione numerica

Le caratteristiche numeriche sono le più facili da incorporare nei modelli matematici. Infatti, alcuni algoritmi di apprendimento automatico, come la

regressione lineare e le macchine vettoriali supporto, funzionano solo con dati numerici. Ma anche trasformare i numeri in numeri può essere un problema sottile. I campi numerici possono essere rappresentati in modi diversi: come numeri interi (123), come decimali (123,5) o addirittura come frazioni (123 1/2). I numeri possono anche essere rappresentati come testo, richiedendo la conversione da "dieci milioni" a 10000000 per l'elaborazione numerica. I problemi di rappresentazione numerica possono avere il merito di aver distrutto un altro razzo. Un razzo Ariane 5 lanciato al costo di 500 milioni di dollari il 4 giugno 1996 è esploso quaranta secondi dopo il decollo, con la causa che alla fine è stata attribuita a una conversione non riuscita di un numero in virgola mobile a 64 bit in un numero intero a 16 bit.
La distinzione tra numeri interi e numeri in virgola mobile (reali) è importante da mantenere. I numeri interi sono numeri di conteggio: quantità che sono realmente discreti devono essere rappresentati come numeri interi. Le quantità misurate fisicamente non sono mai quantificate con precisione, perché viviamo in un mondo continuo. Pertanto, tutte le misurazioni devono essere riportate come numeri reali. Le approssimazioni integrali dei numeri reali sono talvolta utilizzate nel tentativo malriuscito di risparmiare spazio. Non lo faccia: gli effetti di quantificazione dell'arrotondamento o del troncamento introducono artefatti. In un set di dati particolarmente maldestro che abbiamo incontrato, i pesi dei bambini erano rappresentati come due campi interi (libbre e once rimanenti). Sarebbe stato molto meglio combinarli in un'unica quantità decimale.

Unificazione del nome

L'integrazione di record provenienti da due set di dati distinti richiede che essi condividano un campo chiave comune. I nomi sono spesso utilizzati come campi chiave, ma spesso sono riportati in modo incoerente. *Jos'e* è la stessa persona di *Jose*? Tali segni diacritici sono banditi dai registri ufficiali delle nascite di diversi Stati americani, nel tentativo aggressivo di costringerli ad essere coerenti. Per fare un altro esempio, i database mostrano le mie pubblicazioni come autori del prodotto cartesiano del mio nome (*Steve*, *Steven*, o *S.*), del secondo nome (*Sol*, *S.*, o vuoto) e del cognome (*Skiena*), consentendo nove diverse varianti. E le cose peggiorano se includiamo gli errori di ortografia. Posso trovarmi su Google con un nome *Stephen* e cognomi *Skienna* e *Skeina*. L'unificazione dei record in base alla chiave è un problema molto brutto, che non ha una soluzione magica. Questo è

esattamente il motivo per cui sono stati inventati i numeri identificativi, quindi li utilizzi come chiavi, se possibile. La migliore tecnica generale è l'unificazione: eseguire semplici trasformazioni del testo per ridurre ogni nome ad un'unica versione canonica. La conversione di tutte le stringhe in minuscole aumenta il numero di collisioni (solitamente corrette). L'eliminazione dei secondi nomi o almeno la loro riduzione a un'abbreviazione crea un numero ancora maggiore di corrispondenze/collisioni, così come la mappatura dei nomi in versioni canoniche (come trasformare tutti gli *Steves* in *Stevens*). Qualsiasi trasformazione di questo tipo corre il rischio di creare persone Frankenstein, singoli record assemblati da corpi multipli. Le applicazioni si differenziano a seconda che il pericolo maggiore risieda in una fusione troppo aggressiva o troppo timida. Scopra dove si colloca il suo compito in questo spettro e agisca di conseguenza.Una preoccupazione importante nella fusione di set di dati è l'unificazione *del codice dei caratteri*. Ai caratteri nelle stringhe di testo vengono assegnate rappresentazioni numeriche, con la mappatura tra simboli e numeri regolata dallo standard del codice dei caratteri. Sfortunatamente, esistono diversi standard di codice carattere nell'uso comune, il che significa che ciò che si estrae da una pagina web potrebbe non essere nello stesso codice carattere assunto dal sistema che lo elaborerà. Storicamente, il buon vecchio standard di codice *ASCII* a 7 bit è stato ampliato al codice dell'alfabeto *latino IS0 8859-1* a 8 bit, che aggiunge caratteri e segni di punteggiatura di diverse lingue europee. *UTF-8* è una codifica di tutti i caratteri Unicode che utilizza un numero variabile di blocchi a 8 bit, che è retrocompatibile con ASCII. È la codifica dominante per le pagine web, anche se altri sistemi rimangono in uso.
Unificare correttamente i codici dei caratteri dopo la fusione è praticamente impossibile. Deve avere la disciplina di scegliere un unico codice come standard, e controllare la codifica di ogni file di input durante la preelaborazione, convertendolo nell'obiettivo prima di proseguire il lavoro.

Unificazione di data e ora

I dati/timbri temporali vengono utilizzati per dedurre l'ordine relativo degli eventi e per raggruppare gli eventi in base alla relativa simultaneità. L'integrazione dei dati degli eventi da più fonti richiede un'attenta pulizia per garantire risultati significativi.

Innanzitutto, consideriamo i problemi di misurazione del tempo. Gli orologi di due computer non coincidono mai esattamente, per cui allineare con precisione i registri di sistemi diversi richiede un mix di lavoro e di congetture. Ci sono anche problemi di fuso orario quando si ha a che fare con dati provenienti da regioni diverse, nonché diversità nelle regole locali che disciplinano i cambiamenti dell'ora legale. La risposta giusta è allineare tutte le misurazioni del tempo al *Tempo Uni- versale Coordinato* (UTC), uno standard moderno che sostituisce il tradizionale *Tempo Medio di Greenwich* (GMT). Uno standard correlato è l'*ora UNIX*, che riporta l'ora precisa di un evento in termini di numero di secondi trascorsi dalle 00:00:00 UTC di giovedì 1 gennaio 1970. Il calendario gregoriano è comune in tutto il mondo tecnologico, anche se molti altri sistemi di calendario sono in uso in diversi Paesi. Per convertire i sistemi di calendario è necessario utilizzare algoritmi sottili. Un problema più grande per l'allineamento delle date riguarda la corretta interpretazione dei fusi orari e della linea di data internazionale. L'unificazione delle serie temporali è spesso complicata dalla natura del calendario degli affari. I mercati finanziari sono chiusi nei fine settimana e nei giorni festivi, il che problemi di interpretazione quando si correla, ad esempio, i prezzi delle azioni alla temperatura locale. Qual è il momento giusto durante il fine settimana per misurare la temperatura, in modo da essere coerenti con gli altri giorni della settimana? Linguaggi come Python contengono ampie librerie per trattare i dati delle serie temporali finanziarie, per correggere problemi come questo. Problemi simili si presentano con i dati mensili, perché i mesi (e persino gli anni) hanno lunghezze diverse.

Unificazione finanziaria

Il denaro fa girare il mondo, ed è per questo che molti progetti di scienza dei dati ruotano intorno alle serie temporali finanziarie. Ma il denaro può essere sporco, quindi questi dati richiedono una pulizia. Un problema è la *conversione di valuta*, che rappresenta i prezzi internazionali utilizzando un'unità finanziaria standardizzata. I tassi di cambio delle valute possono variare di qualche punto percentuale nell'arco di una giornata, per cui alcune applicazioni richiedono conversioni sensibili al tempo. I tassi di conversione non sono veramente standardizzati. Mercati diversi avranno tassi e *spread* diversi, il divario tra i prezzi di acquisto e di vendita che coprono il costo della conversione.

L'altra correzione importante riguarda l'inflazione. *Il valore temporale del denaro* implica che un dollaro oggi ha (generalmente) più valore di un dollaro tra anno, con i tassi di interesse che rappresentano il modo giusto per scontare i dollari futuri. I tassi di inflazione sono stimati seguendo le variazioni di prezzo di un paniere di articoli e forniscono un modo per standardizzare il potere d'acquisto di un dollaro nel tempo. L'utilizzo di prezzi non aggiustati in un modello per periodi di tempo non banali è solo un'occasione per creare problemi. Una volta, un gruppo di miei studenti si è entusiasmato per forte correlazione osservata tra i prezzi delle azioni e i prezzi del petrolio in un periodo di trent'anni, e ha cercato di utilizzare i prezzi delle azioni in un modello di previsione delle materie prime. Ma entrambi i beni erano prezzati in dollari, senza alcun aggiustamento in seguito all'inflazione. Le serie temporali dei prezzi di *qualsiasi* coppia di articoli saranno fortemente correlate nel tempo quando non si corregge l'inflazione. In effetti, il modo più significativo per rappresentare le variazioni di prezzo
nel tempo probabilmente non sono le differenze ma i *rendimenti*, che normalizzano la differenza rispetto al prezzo iniziale. Questo è più analogo a una variazione percentuale, con il vantaggio il logaritmo di questo rapporto diventa simmetrico ai guadagni e alle perdite. Le serie temporali finanziarie contengono molte altre sottigliezze che richiedono una pulizia. Molte azioni distribuiscono *dividendi* programmati all'azionista in una data particolare ogni anno. Diciamo, ad esempio, che Microsoft pagherà un dividendo di 2,50 dollari il 16 gennaio. Se lei possiede un'azione di Microsoft all'inizio dell'attività di quel giorno, riceve questo assegno, quindi il valore dell'azione scende immediatamente di 2,50 dollari nel momento successivo all'emissione del dividendo. Questo calo di prezzo non riflette alcuna perdita reale per l'azionista, ma i dati adeguatamente puliti devono tener conto del dividendo nel prezzo dell'azione. È facile immaginare un modello addestrato sui dati dei prezzi non corretti che impara a vendere le azioni appena prima dell'emissione dei dividendi e si sente ingiustamente orgoglioso di se stesso per averlo fatto.

3.3.3 Gestire i valori mancanti

Non tutti i set di dati sono completi. Un aspetto importante della pulizia dei dati è identificare i campi per i quali i dati non ci sono. I set di dati numerici si aspettano un valore per ogni elemento di una matrice. Impostare i valori mancanti a zero è allettante, ma generalmente sbagliato, perché c'è sempre un'ambiguità sul fatto che questi valori debbano essere interpretati come dati o meno.

Lo stipendio di qualcuno è pari a zero perché è disoccupato, oppure non ha risposto alla domanda? Il pericolo di utilizzare valori senza senso come simboli di non-dati è che possono essere interpretati erroneamente come dati quando arriva il momento di costruire i modelli. Un modello di regressione lineare addestrato a prevedere gli stipendi in base all'età, all'istruzione e al sesso problemi con le persone che si sono rifiutate di rispondere alla domanda. L'utilizzo di un valore come 1 come simbolo di assenza di dati ha esattamente le stesse carenze dello zero. Anzi, faccia come il matematico che ha paura dei numeri negativi: non si fermi davanti a nulla per evitarli. *C*onservare separatamente sia i dati grezzi che la loro versione ripulita. I dati grezzi sono la verità di base e devono essere conservati intatti per le analisi future. I dati puliti possono essere migliorati utilizzando l'imputazione per riempire i valori mancanti. Ma manteniamo i dati grezzi distinti da quelli puliti, in modo da poter studiare diversi approcci all'imputazione. Come gestire i valori mancanti? L'approccio più semplice è quello di eliminare tutti i record contenenti valori mancanti. Questo funziona bene quando ci sono abbastanza dati di formazione, a condizione che i valori mancanti siano assenti per motivi non sistematici. Se le persone che si rifiutano di dichiarare il proprio stipendio sono generalmente quelle che si trovano al di sopra della media, l'eliminazione di questi record porterà a risultati falsati. Ma in genere vogliamo utilizzare i record con campi mancanti. Può meglio stimare o *imputare* i valori mancanti, invece di lasciarli vuoti. Abbiamo bisogno di metodi generali per riempire i valori mancanti. I candidati includono: *imputazione basata su euristiche*: Data una conoscenza sufficiente del dominio sottostante, dovremmo essere in grado di fare un'ipotesi ragionevole per il valore di alcuni campi. Se devo inserire un valore per l'anno in cui lei morirà, indovinare l'*anno di nascita+80* si rivelerà in media corretto, e molto più veloce che aspettare la risposta finale. *Imputazione del valore medio*: Utilizzare il valore medio di una variabile come proxy per i valori mancanti è generalmente ragionevole. In primo luogo, l'aggiunta di più valori con la media lascia la media invariata, quindi non alteriamo le nostre statistiche con questa imputazione. In secondo luogo, i campi con i valori medi aggiungono un sapore di vaniglia alla maggior parte dei modelli, quindi hanno un impatto attenuato su qualsiasi previsione fatta utilizzando i dati. Ma la media potrebbe non essere appropriata se c'è una ragione sistematica per i dati mancanti. Supponiamo di utilizzare l'anno medio di morte in Wikipedia per imputare il valore mancante per tutte le persone viventi. Questo si rivelerebbe

disastroso, con molte persone registrate come morte prima di essere effettivamente nate. *Imputazione di valori casuali*: Un altro approccio consiste nel selezionare un valore casuale dalla colonna per sostituire il valore mancante. Questo sembrerebbe a fare delle ipotesi potenzialmente sbagliate, ma in realtà è proprio questo il punto. La selezione ripetuta di valori casuali permette di valutare statisticamente l'impatto dell'imputazione. Se eseguiamo il modello dieci volte con dieci diversi valori imputati, il modello viene sostituito da un valore casuale. e otteniamo risultati molto diversi, allora probabilmente non dovremmo avere molta fiducia nel modello. Questo controllo di accuratezza è particolarmente prezioso quando c'è una frazione sostanziale di valori mancanti nel set di dati. *Imputazione tramite il vicino più prossimo*: E se identificassimo il record completo che corrisponde maggiormente a tutti i campi presenti e utilizzassimo questo vicino più prossimo per dedurre i valori di ciò che manca? Tali previsioni dovrebbero essere più accurate della media, quando ci sono ragioni sistematiche per spiegare la varianza tra i record.Questo approccio richiede una funzione di distanza per identificare i record più simili. I metodi di *prossimità* sono una tecnica importante nella scienza dei dati. *Imputazione per interpolazione*: Più in generale, possiamo utilizzare un metodo come la regressione lineare per prevedere i valori della colonna target, dati gli altri campi del record. Tali modelli possono essere addestrati su record completi e poi applicati a quelli con valori mancanti. L'utilizzo della regressione lineare per prevedere i valori mancanti funziona meglio quando c'è un solo campo mancante per record. Il pericolo potenziale in questo caso è la creazione di outlier significativi attraverso previsioni errate. I modelli di regressione possono facilmente trasformare un record incompleto in un outlier, riempiendo i campi mancanti con valori insolitamente alti o bassi. Questo porterebbe l'analisi a valle a concentrarsi maggiormente sui record con valori mancanti, esattamente il contrario di ciò che vogliamo fare. Tali preoccupazioni enfatizzano l'importanza del rilevamento degli outlier, la fase finale del processo di pulizia che verrà presa in considerazione qui.

3.3.4 Rilevamento dei valori anomali

Gli errori nella raccolta dei dati possono facilmente produrre degli outlier che possono interferire con un'analisi corretta. Un esempio interessante riguarda la più grande vertebra di dinosauro mai scoperta. Misurata a 1500 millimetri, implica che un individuo era lungo 188 piedi. Si tratta di un dato

sorprendente, soprattutto perché il *secondo* esemplare più grande mai scoperto arriva a soli 122 piedi. La spiegazione più probabile in questo caso è che questo fossile gigante non sia mai esistito: manca dal Museo Americano di Storia Naturale da oltre cento anni. Forse la misurazione originale è stata effettuata su un osso di dimensioni convenzionali e le due cifre centrali sono state accidentalmente trasposte, riducendo la vertebra a 1050 millimetri.

Gli elementi outlier sono spesso creati da errori di inserimento dei dati, come sembra essere il caso in questione. Possono anche derivare da errori di scraping, ad esempio un'irregolarità nella formattazione che causa l'interpretazione di un numero di nota come un valore numerico. Il fatto che qualcosa sia scritto non lo rende corretto. Come nell'esempio del dinosauro, un singolo elemento anomalo può portare a interpretazioni errate.

Il controllo generale della correttezza richiede di esaminare i valori più grandi e più piccoli di ogni variabile/colonna per vedere se sono troppo fuori linea. Il modo migliore per farlo è tracciare l'istogramma di frequenza e osservare la posizione degli elementi estremi. L'ispezione visiva può anche confermare che la distribuzione ha l'aspetto che dovrebbe avere, tipicamente a forma di campana. Nei dati distribuiti normalmente, la probabilità che un valore si trovi a k devizioni standard dalla media diminuisce in modo esponenziale con k. Questo spiega perché non esistono giocatori di basket di 3 metri e fornisce una soglia valida per identificare gli outlier. Le distribuzioni a legge di potenza sono meno facili da individuare: *esiste* davvero un Bill Gates che vale oltre 10.000 volte di più dell'individuo medio.

È troppo semplice eliminare le righe contenenti i campi anomali e andare avanti. I valori anomali spesso indicano problemi più sistematici che devono essere affrontati. Consideriamo un insieme di dati di personaggi storici in base alla durata della vita. È facile individuare il biblico Matusalemme (con 969 anni) come un outlier. Ma è meglio capire se è indicativo di altre figure che dovremmo considerare di rimuovere. Osserviamo che Matusalemme non aveva date di nascita e di morte ben stabilite. Forse le età pubblicate di chiunque non abbia date dovrebbero essere considerate abbastanza sospette da essere eliminate. Al contrario, la persona con la durata di vita più breve in Wikipedia (Giovanni I, Re di Francia) visse solo cinque giorni. Ma le sue date di nascita (15 novembre) e di morte (20 novembre) nel 1316 mi convincono che la sua durata di vita fosse accurata.

3.4 Storia di guerra: Battere il mercato

Ogni volta che ci incontravamo, il mio studente laureato Wenbin mi diceva che stavamo soldi. Ma sembrava sempre meno fiducioso ogni volta che glielo chiedevo. Il nostro sistema di analisi del sentimento Lydia ha accolto enormi flussi di testo di notizie e social media, riducendoli a serie temporali giornaliere di frequenza e sentimento per i milioni di persone, luoghi e organizzazioni diversi citati all'interno. Quando qualcuno vince un campionato sportivo, vengono scritti molti articoli che descrivono la grandezza dell'atleta. Ma quando questo giocatore viene poi arrestato per droga, il tono degli articoli che lo riguardano cambia immediatamente. Tenendo il conto della frequenza relativa di associazione con parole positive ("vittorioso") rispetto a parole negative ("arrestato") nel flusso di testo, possiamo costruire segnali di sentimento per qualsiasi entità degna di nota. Wenbin ha studiato come i segnali di sentiment possano essere utilizzati per prevedere eventi futuri, come l'incasso di un determinato film, in risposta alla qualità delle recensioni pubblicate o al buzz. Ma in particolare voleva utilizzare questi dati per giocare con il mercato azionario. Le azioni salgono e scendono in base alle notizie. Un rapporto sui guadagni non raggiunto è una cattiva notizia per un'azienda, quindi il prezzo scende. L'approvazione di un nuovo farmaco da parte della Food and Drug Administration (FDA) è un'*ottima* notizia per l'azienda che lo possiede, quindi il prezzo sale. Se Wenbin potesse utilizzare il nostro segnale di sentiment per prevedere i prezzi futuri delle azioni, beh, diciamo che non dovrei più pagarlo come assistente di ricerca. Così ha simulato una strategia di acquisto delle azioni che mostravano i valori più alti. sentimento nelle notizie di quel giorno, e poi shortando quelli con il sentimento più basso. Ha ottenuto ottimi risultati. "Vedete", ha detto. "Stiamo facendo soldi". I numeri sembravano fantastici, ma avevo un cavillo. Usare i risultati delle notizie di oggi per prevedere i movimenti attuali dei prezzi non era davvero corretto, perché l'evento descritto nell'articolo potrebbe aver già spostato il prezzo prima che avessimo la possibilità di leggerlo. I prezzi delle azioni dovrebbero reagire molto rapidamente alle notizie importanti. Wenbin ha quindi simulato la strategia di acquisto di azioni in base al sentiment delle notizie del giorno precedente, per creare un divario tra le notizie osservate e le variazioni di prezzo. Il tasso di rendimento è diminuito notevolmente, ma era ancora positivo. "Vede", ha detto. "Stiamo ancora guadagnando". Ma sono rimasto un po' a disagio per questo. Molti economisti ritengono che i mercati finanziari siano *efficienti*, il che significa che tutte le

notizie pubbliche si riflettono istantaneamente nella variazione dei prezzi. I prezzi cambiano certamente in risposta alle notizie, ma non sareste in grado di entrare abbastanza velocemente per sfruttare le informazioni. Abbiamo dovuto rimanere abbastanza scettici per assicurarci che non ci fossero problemi di dati/tempi che potessero spiegare i nostri risultati. Così ho chiesto a Wenbin come avesse eseguito esattamente la sua simulazione. La sua strategia comprava e vendeva al prezzo di chiusura ogni giorno. Ma questo lasciava sedici ore prima dell'apertura del giorno successivo, un tempo sufficiente perché il mondo reagisse agli eventi accaduti mentre dormivo. Ha cambiato l'acquisto simulato al prezzo di apertura. Anche in questo caso, il tasso di ritorno si è ridotto notevolmente, ma era ancora positivo. "Vede", ha detto. "Stiamo ancora facendo soldi". Ma potrebbero esserci altri artefatti nel modo in cui abbiamo cronometrato i nostri dati, sostanzialmente il giornale di domani oggi? In buona fede, abbiamo cercato tutte le altre possibilità che ci venivano in mente, come ad esempio che le date degli articoli pubblicati riflettessero il momento in cui sono apparsi invece di quello in cui sono stati scritti. Dopo aver fatto del nostro meglio per essere scettici, le sue strategie sembravano ancora mostrare rendimenti positivi dal sentimento delle notizie. Il nostro articolo su questa analisi è stato ben accolto, e Wenbin ha continuato a diventare un esperto di successo, utilizzando il sentiment tra gli altri segnali per fare trading sui mercati finanziari. Ma resto un po' inquieto riguardo a questo risultato. Pulire i nostri dati per datare con precisione ogni articolo di notizie è stato molto difficile da fare correttamente. Il nostro sistema è stato originariamente progettato per produrre serie temporali giornaliere in modalità batch, quindi è difficile essere sicuri di aver fatto tutto correttamente nei milioni di articoli scaricati nel corso di diversi anni per eseguire ora un'analisi su scala più fine. La lezione da trarre è che la pulizia è importante quando c'è ballo del denaro. Inoltre, è meglio progettare un ambiente pulito all'inizio dell'analisi, invece di lavarsi furiosamente alla fine.

3.5 Crowdsourcing

Nessuna persona ha tutte le risposte. Nemmeno io. Gran parte di ciò che passa per saggezza è il modo in cui aggreghiamo le competenze, mettendo insieme le opinioni derivanti dalla conoscenza e dall'esperienza degli altri. *Il crowdsourcing* sfrutta le intuizioni e il lavoro di un gran numero di persone per raggiungere un obiettivo comune. Sfrutta la *saggezza delle folle*, secondo cui la conoscenza collettiva di un gruppo di persone potrebbe essere

superiore a quella dell'individuo più intelligente tra loro. Questa idea è iniziata con un bue. Francis Galton, fondatore della scienza statistica e parente di Charles Darwin, partecipò a una fiera del bestiame locale nel 1906. Nell'ambito dei festeggiamenti, gli abitanti del villaggio furono invitati a indovinare il peso di questo particolare bue; la persona che si fosse avvicinata di più al segno avrebbe vinto un premio. Quasi 800 partecipanti si sono cimentata nell'impresa. Nessuno ha indovinato il peso effettivo di 1.178 libbre, ma Galton ha osservato che la media delle ipotesi era incredibilmente vicina: 1.179 libbre! L'esperimento di Galton suggerisce che per alcuni compiti si possono ottenere risultati migliori coinvolgendo un gruppo eterogeneo di persone, invece di chiedere solo agli esperti.

Il crowdsourcing è un'importante fonte di dati per la costruzione di modelli, soprattutto per i compiti associati alla percezione umana. Gli esseri umani rimangono il sistema all'avanguardia nell'elaborazione del linguaggio naturale e nella visione computerizzata, raggiungendo il più alto livello di prestazioni. Il modo migliore per raccogliere dati di formazione spesso richiede di chiedere alle persone di assegnare un punteggio a un determinato testo o immagine. Per fare questo su una scala sufficientemente ampia da costruire dati di addestramento sostanziali, è necessario un gran numero di annotatori, in pratica una folla. I social media e altre nuove tecnologie hanno reso più facile la raccolta e l'aggregazione di opinioni su vasta scala.

3.5.1 Il Penny Demo

Cominciamo con un piccolo esperimento di saggezza delle folle. La Figura 3.4 contiene le foto di un barattolo di penny che ho accumulato nel mio ufficio nel corso di molti anni. Quanti penny ho in questo barattolo? Provi a indovinare ora, perché le dirò la risposta nella prossima pagina. Per ottenere la risposta giusta, ho chiesto al mio collaboratore biologo Justin Garden di pesare i penny su una bilancia da laboratorio di precisione. Dividendo per il peso di un singolo penny si ottiene il conteggio. Quindi le chiedo di nuovo: quanti penny pensa che ci siano in questo barattolo? Ho eseguito questo esperimento sugli studenti del mio corso di scienze dei dati. Come sarà la sua risposta rispetto alla loro? Per prima cosa ho chiesto a undici dei miei studenti di scrivere le loro opinioni su dei cartoncini e di passarmeli silenziosamente davanti alla sala. In questo modo le ipotesi erano completamente indipendenti l'una dall'altra. I risultati, ordinati per comodità, sono stati: 537, 556, 600, 636, 1200, 1250, 2350, 3000, 5000, 11,000, 15,000. Poi ho scritto questi numeri sulla lavagna e ho calcolato alcune

statistiche. La mediana di queste ipotesi era di 1250, con una media di 3739. In effetti, c'erano esattamente 1879 penny nel barattolo. Il punteggio mediano tra i miei studenti era più vicino alla somma giusta di qualsiasi singola ipotesi. Ma prima di rivelare il totale effettivo, ho chiesto ad un'altra dozzina di studenti di indovinare. L'unica differenza era che questa coorte aveva visto le ipotesi del primo gruppo di studenti scritte alla lavagna. Le loro scelte sono state: 750, 750, 1000, 1000, 1000, 1250, 1400, 1770, 1800, 3500, 4000, 5000. L'esposizione della coorte alle ipotesi di altre persone ha condizionato fortemente la distribuzione, eliminando tutti gli outlier: il minimo tra il secondo gruppo era superiore a quattro delle ipotesi precedenti e il massimo inferiore o uguale a tre del turno precedente. All'interno di questa coorte, la mediana era di 1325 e la media di 1935. Entrambi si avvicinano un po' di più alla risposta reale, ma è chiaro che il pensiero di gruppo si è instaurato. *L'ancoraggio* è il noto pregiudizio cognitivo secondo cui i giudizi delle persone si fissano in modo irrazionale sul primo numero che sentono. I concessionari di auto sfruttano questo aspetto continuazione, indicando inizialmente un costo gonfiato per il veicolo, in modo che i prezzi successivi sembrino un affare. Poi ho fatto un ultimo test prima di rivelare la risposta. Ho permesso ai miei studenti di fare un'offerta sul vaso, il che significa che dovevano essere abbastanza sicuri da rischiare del denaro sul risultato. Questo ha prodotto esattamente due offerte da parte di studenti coraggiosi, rispettivamente di 1500 e 2000 centesimi. Ho intascato 1,21 dollari dal babbeo con l'offerta più alta, ma entrambi si sono dimostrati abbastanza vicini. Non è una sorpresa: le persone disposte a scommettere il proprio denaro su un evento sono, per definizione, fiduciose nella loro selezione.

3.5.2 Quando la folla è saggia?

Secondo James Surowiecki nel suo libro *La saggezza delle folle*, le folle sono sagge quando sono soddisfatte quattro condizioni: *Quando le opinioni sono indipendenti*: Il nostro esperimento ha evidenziato quanto sia facile per un gruppo cadere nel pensiero di gruppo. Le persone vengono naturalmente influenzate dagli altri. Se vuole avere la vera opinione di una persona, deve chiederla in modo isolato. *Quando le folle sono persone con conoscenze e metodi diversi*: Le folle aggiungono informazioni solo quando c'è disaccordo. Un comitato composto da esperti perfettamente correlati non apporta nulla di più di quanto si potrebbe imparare da ognuno di loro. Nel problema dell'indovinello, alcune persone hanno stimato il volume del contenitore,

mentre altre hanno misurato la flessione del mio braccio mentre sollevavo la massa pesante. Altri approcci avrebbero potuto stimare quanti centesimi avrei potuto accumulare in vent'anni di svuotamento occasionale delle tasche, o ricordare le proprie esperienze di accumulo. *Quando il problema riguarda un ambito che non richiede conoscenze specialistiche*: Mi fido del consenso della folla in alcune decisioni importanti, come il tipo di auto da acquistare o chi dovrebbe essere il Presidente del mio Paese (gulp). Ma quando si tratta di decidere se il mio campione di tumore è canceroso o benigno, mi fiderò della parola di un medico piuttosto che di un cast di 1.000 nomi estratti a caso dall'elenco telefonico. Perché? Perché la domanda in questione trae grande beneficio dalla conoscenza e dall'esperienza specialistica. C'è una vera ragione per cui il medico *dovrebbe* sapere più di tutti gli altri. Per i compiti percettivi più semplici, la folla la fa da padrona, ma bisogna fare attenzione a non chiedere alla folla qualcosa che non ha modo di sapere. *Le opinioni possono essere abbastanza aggregate*: La parte meno utile di qualsiasi modulo di sondaggio di massa è il campo di risposta aperto "Ci dica cosa ne pensa!". Il problema è che non c'è modo di combinare queste opinioni per formare un consenso, perché persone diverse hanno problemi e preoccupazioni diverse. Forse questi testi potrebbero essere suddivisi in gruppi in base alla somiglianza, ma è difficile farlo in modo efficace. L'uso più comune di queste risposte libere è l'aneddotica. Le persone scelgono più positive e le inseriscono in una diapositiva per impressionare il capo. Essere un elemento incomparabile nell'ordine parziale della vita. Il pensiero diversificato e indipendente apporta la maggior saggezza alla folla.

3.5.3 Meccanismi di aggregazione

Raccogliere la saggezza da un insieme di risposte richiede l'utilizzo del giusto meccanismo di aggregazione. Per stimare le quantità numeriche, sono adatte le tecniche standard come il grafico della distribuzione di frequenza e il calcolo delle statistiche riassuntive. Sia la media che la mediana presuppongono implicitamente che gli errori sono distribuita in modo simmetrico. Una rapida occhiata alla forma della distribuzione può generalmente confermare o respingere questa ipotesi. La mediana è, in generale, una scelta più appropriata della media in questi problemi di aggregazione. Riduce l'influenza degli outlier, problema particolare nel caso di esperimenti di massa, in cui è probabile che una certa frazione dei partecipanti siano degli idioti. Sui nostri dati relativi all'indovinello del penny, la media ha prodotto una terribile sovrastima di 3739, che si è ridotta a 2843 dopo aver eliminato l'indovinello più grande e quello più

piccolo, per poi scendere a 2005 una volta tagliati i due outlier su ciascuna estremità (ricordiamo che la risposta corretta era 1879). Eliminare gli outlier è un'ottima strategia, ma potremmo avere altri motivi per giudicare l'affidabilità dei nostri soggetti, come le loro prestazioni in altri test di cui conosciamo la risposta. L'adozione di una *media ponderata*, in cui diamo maggior peso ai punteggi ritenuti più affidabili, offre un modo per prendere in considerazione tali misure di fiducia.

Per i problemi di classificazione, il voto è il meccanismo di aggregazione di base. *Il teorema della giuria di Condorcet* giustifica la nostra fiducia nella democrazia. Afferma che se la probabilità che ciascun elettore abbia ragione su una determinata questione è $p > 0{,}5$, la probabilità che la maggioranza degli elettori abbia ragione ($P(n)$) è maggiore di p. I grandi conteggi degli elettori conferiscono validità statistica anche alle elezioni altamente contestate. Supponiamo che $p = 0{,}51$, il che significa che le forze di destra sono una maggioranza risicata. Una giuria di 101 membri raggiungerebbe la decisione corretta il 57% delle volte, mentre $P(1001) = 0{,}73$ e $P(10001) = 0{,}9999$. La probabilità di una decisione corretta si avvicina a 1 con l'aumentare di n. Tuttavia, esistono dei limiti naturali al potere dei sistemi elettorali. *Il teorema di impossibilità di Arrow* afferma che nessun sistema elettorale per sommare le per- mutazioni delle preferenze come voti soddisfa quattro condizioni naturali per l'equità di un'elezione. Questo aspetto sarà discusso nella Sezione 4.6, nel contesto dei punteggi e delle classifiche.

3.5.4 Servizi di crowdsourcing

I servizi di crowdsourcing come Amazon Turk e CrowdFlower le offrono l'opportunità di assumere un gran numero di persone per svolgere piccole quantità di lavoro. L'aiutano a gestire le persone, al fine di creare dati da gestire. Questi servizi di crowdsourcing mantengono una grande scuderia di lavoratori freelance, fungendo da intermediari tra loro e i potenziali datori di lavoro. A questi lavoratori, generalmente chiamati *Turker*, vengono forniti elenchi di lavori disponibili e la relativa retribuzione. I datori di lavoro hanno generalmente una certa capacità di controllare la posizione e le credenziali di chi assumono, e il potere di rifiutare gli sforzi di un lavoratore senza retribuzione, se lo ritengono inadeguato. Ma le statistiche sui tassi di accettazione dei datori di lavoro sono pubblicate, ed è improbabile che buoni lavoratori lavorino per cattivi attori. I compiti assegnati ai Turker in genere comportano sforzi cognitivi semplici che attualmente non possono essere

eseguiti bene dai computer. Buone applicazioni dei Turker includono: *misurare gli aspetti della percezione umana*: I sistemi di crowdsourcing offrono modi efficienti per raccogliere opinioni rappresentative su compiti semplici. Una bella applicazione è stata quella di stabilire collegamenti tra i colori nello spazio rosso-verdeblu e i nomi con cui le persone li identificano tipicamente in una lingua. Questo è importante da sapere quando si scrivono descrizioni di prodotti e immagini. Quindi, dov'è il confine nello spazio dei colori tra "blu" e "azzurro", o "blu uovo di pettirosso" e "verde acqua"? I nomi giusti sono una funzione della cultura e delle convenzioni, non della fisica. Per , deve chiedere alle persone, e il crowdsourcing le permette di interrogare facilmente centinaia o migliaia di persone diverse. *Ottenere dati di formazione per i classificatori di apprendimento automatico*: Il nostro interesse principale nel crowdsourcing sarà quello di produrre annotazioni umane che servano come dati di addestramento. Molti problemi di apprendimento automatico cercano di svolgere un particolare compito "come fanno le persone". Per farlo, è necessario un gran numero di istanze di addestramento per stabilire cosa hanno fatto le persone, quando ne hanno avuto la possibilità. Supponiamo di voler costruire un sistema di analisi del sentimento in grado di leggere una recensione scritta e decidere se la sua opinione su un prodotto è favorevole o sfavorevole. Avremo bisogno di un gran numero di recensioni etichettate da annotatori, che serviranno come dati di test/addestramento. Inoltre, abbiamo bisogno delle stesse recensioni etichettate ripetutamente da annotatori diversi, in modo da identificare eventuali disaccordi tra gli annotatori sul significato esatto di un testo. *Ottenere dati di valutazione per i sistemi informatici*: *Il test A/B* è un metodo standard per ottimizzare le interfacce utente: si mostra a metà dei giudici la versione *A* di un determinato sistema e all'altra metà la versione *B*. Poi si verifica quale gruppo ha fatto meglio secondo una certa metrica. I turker possono fornire un feedback sull'interesse di una determinata app, o sulle prestazioni di un nuovo classificatore. Uno dei miei studenti laureati (Yanqing Chen) ha utilizzato CrowdFlower per valutare un sistema da lui costruito per identificare la categoria di Wikipedia più rilevante per una determinata entità. Quale categoria descrive meglio Barack Obama: *Presidenti degli Stati Uniti* o *Autori afroamericani*? Per 200 dollari, ha chiesto alle persone di rispondere a un totale di 10.000 domande a scelta multipla, sufficienti per valutare correttamente il suo sistema. *Mettere gli esseri umani nella macchina*: Esistono ancora molti compiti cognitivi che le

persone svolgono molto meglio delle macchine. Un'interfaccia progettata in modo intelligente può fornire le domande dell'utente a persone sedute all'interno del computer, in attesa di servire chi ne ha bisogno. Supponiamo che lei voglia creare un'app per aiutare gli ipovedenti, consentendo all'utente di scattare una foto e chiedere aiuto a qualcuno. Magari sono in cucina e hanno bisogno di qualcuno che legga loro l'etichetta di una lattina. Questa applicazione potrebbe chiamare un Turker come subroutine, per svolgere tale compito quando è necessario. Naturalmente, queste coppie immagine-annotazione dovrebbero essere conservate per future. Potrebbero servire come dati di addestramento per un programma di apprendimento automatico, per eliminare il più possibile le persone dal ciclo. *Sforzi creativi indipendenti*: Il crowdsourcing può essere utilizzato per commissionare un gran numero di lavori creativi su richiesta. Può ordinare post o articoli di blog su richiesta, o recensioni scritte di prodotti sia buoni che cattivi. Tutto ciò che può immaginare può essere creato, se solo specifica ciò che desidera. Ecco due esempi sciocchi che in qualche modo trovo ispiranti: The Sheep Market ha commissionato 10.000 disegni di pecore per pochi centesimi l'uno. Come opera d'arte concettuale, cerca di venderle al miglior offerente. Quali imprese creative le vengono in mente che la gente farebbe per lei a 0,25 dollari l'? Emoji Dick è stato un progetto di crowdsourcing per tradurre completamente il grande romanzo americano *Moby Dick* in immagini emoji. I suoi creatori hanno suddiviso il libro in circa 10.000 parti e hanno affidato la traduzione di ogni parte a tre Turker distinti. Altri Turker sono stati assunti per selezionare la migliore tra queste da incorporare nel libro finale. Sono stati coinvolti oltre 800 Turker, con un costo totale di 3.676 dollari raccolti dal sito di crowd-funding Kickstarter. *Esperimenti economici/psicologici*: Il crowdsourcing si è rivelato una manna per gli scienziati sociali che conducono esperimenti di economia e psicologia comportamentale. Invece di corrompere i laureandi locali per partecipare ai loro studi, questi ricercatori possono ora espandere il loro pool di soggetti a tutto il mondo. Hanno la possibilità di sfruttare popolazioni più ampie, di eseguire repliche indipendenti in Paesi diversi e, quindi, di verificare se ci sono pregiudizi culturali nelle loro ipotesi.

Ci sono molti compiti interessanti che possono essere portati a termine con profitto utilizzando il crowd sourcing. Tuttavia, è destinato alla delusione se impiega i Turker per il compito sbagliato, nel modo sbagliato. I cattivi usi del crowdsourcing includono: *Qualsiasi compito che richieda una formazione avanzata*: Anche se ogni persona possiede abilità e competenze uniche, i

lavoratori del crowdsourcing non hanno una formazione specifica. Sono progettati per essere trattati come parti intercambiabili. Non si instaura un rapporto personale con questi lavoratori, e qualsiasi lavoro sensato sarà troppo breve per consentire più di qualche minuto di formazione.

I compiti che richiedono competenze tecniche specifiche non possono essere ragionevolmente affidati in crowdsourcing. Tuttavia, potrebbero essere ragionevolmente subappaltati, in accordi tradizionali a lungo termine. *Qualsiasi compito che non può specificare chiaramente*: Non ha un meccanismo di comunicazione con i Turker. In generale, non hanno modo di farle domande. Pertanto, il sistema funziona solo se può specificare i suoi compiti in modo chiaro, conciso e non ambiguo. È molto più difficile di quanto sembri. Si renda conto che sta cercando di programmare delle persone invece che dei computer, con tutti i relativi bug associati al "fai come dico io" che prevale sul "fai come intendo io". Provi le sue specifiche su persone locali prima di aprire il suo lavoro alle masse, e poi faccia una piccola prova sulla sua piattaforma di crowdsourcing per valutare come va, prima di liberare la maggior parte del suo budget. Potrebbe delle sorprese culturali. Le cose che a lei sembrano ovvie potrebbero avere un significato molto diverso per un lavoratore dall'altra parte del mondo. *Qualsiasi compito in cui non può verificare se stanno facendo un buon lavoro*: I turchi hanno un'unica motivazione per accettare il suo lavoro a cottimo: cercano di convertire il loro tempo in denaro nel modo più efficiente possibile. Cercano lavori che offrano il miglior guadagno, e i più intelligenti cercheranno di portare a termine il suo compito nel modo più rapido e sconsiderato possibile. Le piattaforme di crowdsourcing consentono ai datori di lavoro di trattenere il pagamento se il lavoro commissionato non è accettabile. Per trarne vantaggio è necessario un modo efficiente per verificare la qualità del prodotto. Forse dovrebbe chiedere loro di completare alcuni compiti di cui conosce già la risposta corretta. Forse può confrontare le loro risposte con quelle di altri lavoratori indipendenti e scartare il loro lavoro se si discosta troppo spesso dal consenso. È molto importante impiegare un meccanismo di controllo della qualità. Una parte dei lavoratori disponibili su qualsiasi piattaforma sono bot, alla ricerca di compiti a scelta multipla da attaccare attraverso la casualità. Altri possono essere persone con competenze linguistiche del tutto inadeguate al compito da svolgere. È necessario controllare e rifiutare per evitare di essere un babbeo. Tuttavia, non può lamentarsi in modo equo dei risultati ottenuti da compiti

mal specificati. Rifiutare una frazione troppo alta di lavoro ridurrà la sua reputazione, con i lavoratori e con la piattaforma. È un karma particolarmente negativo rifiutare di pagare le persone ma utilizzare comunque il loro prodotto di lavoro. *Qualsiasi compito illegale o troppo disumano per sottoporlo alle persone*: Non le è consentito chiedere a un Turker di fare qualcosa di illegale o non etico. Un esempio clas- sico è quello di assumere qualcuno che scriva recensioni negative sui prodotti del suo concorrente. Assumere un sicario la rende colpevole di omicidio tanto quanto la persona che ha sparato. Sia consapevole che esistono tracce elettroniche che possono essere seguite dalla pubblicazione del suo annuncio direttamente fino a lei. Le persone che lavorano presso le istituzioni educative e di ricerca sono tenute a rispettare uno standard più elevato rispetto alla legge, attraverso il loro *comitato di revisione istituzionale* o IRB. L'IRB è un comitato di ricercatori e funzionari amministrativi che devono approvare qualsiasi ricerca su soggetti umani prima di intraprenderla. Le applicazioni di crowdsourcing benigne, come quelle di cui abbiamo parlato, sono approvate di routine, dopo che i ricercatori hanno seguito un breve corso di formazione online per assicurarsi di comprendere le regole. Si renda sempre conto che c'è una persona all'altro capo della macchina. Non assegni loro compiti offensivi, degradanti, che violano la privacy o troppo stressanti. Probabilmente otterrà risultati migliori dai suoi dipendenti se li tratterà come esseri umani. Per convincere le persone a fare i suoi ordini occorrono incentivi adeguati, non solo istruzioni chiare. Nella vita, in genere si ottiene ciò che si paga. Si informi sulla retribuzione oraria minima attualmente prevalente nel suo Paese e fissi il prezzo dei suoi incarichi di conseguenza. Non si tratta di un requisito legale, ma in genere è una buona pratica commerciale. Il fascino sinistro che deriva dall'assunzione di lavoratori a 0,50 dollari l'ora si esaurisce rapidamente una volta constatata la bassa qualità dei lavoratori che i suoi compiti attirano. Può facilmente consumare tutti i suoi risparmi per la necessità di correggere rigorosamente il loro prodotto di lavoro, magari pagando più lavoratori per farlo ripetutamente. Le attività più pagate trovano lavoratori molto più rapidamente, quindi si prepari ad aspettare se non paga la tariffa prevalente. I bot e i loro equivalenti funzionali sono più felici di accettare salari da schiavi che non i lavoratori che lei vuole realmente assumere.

3.5.5 Gamification

Esiste un'alternativa al pagare persone per annotare o trascrivere i suoi dati. Invece, renda le cose così divertenti che le persone lavoreranno per lei gratuitamente! *I giochi con uno scopo* (GWAP) sono sistemi che mascherano la raccolta di dati come un gioco che le persone vogliono fare, o un compito che le persone stesse vogliono svolgere. Con con la giusta combinazione di gioco, motivazione e immaginazione, si possono fare cose incredibili. Gli esempi di successo includono: *CAPTCHA per il riconoscimento ottico dei caratteri (OCR)*: I CAPTCHA sono quelle immagini di testo distorte che si incontrano spesso quando si crea un account sul web. Richiedono che lei digiti il contenuto delle stringhe di testo mostrate nell'immagine per dimostrare che lei è un essere umano, consentendo così di negare l'accesso ai bot e ad altri sistemi programmati. I ReCAPTCHA sono stati inventati per ottenere dati utili dagli oltre 100 milioni di CAPTCHA visualizzati ogni giorno. In ognuno di essi vengono visualizzate due stringhe di testo, una delle quali viene controllata dal sistema per consentire l'ingresso. L'altra rappresenta un caso difficile per un sistema OCR che sta digitalizzando vecchi libri e giornali. Le risposte vengono mappate per migliorare la digitalizzazione dei documenti d'archivio, trascrivendo oltre 40 milioni di parole al giorno. *Test psicologici/Quali nei giochi/app*: Gli psicologi hanno stabilito cinque tratti fondamentali della personalità come aspetti importanti e riproducibili della per- sonalità. Gli psicologi accademici utilizzano test di personalità a scelta multipla per misurare la posizione degli individui lungo le scale di personalità per ciascuno dei cinque grandi tratti: apertura, coscienziosità, estroversione, gradevolezza e nevroticismo. Trasformando questi sondaggi in app di gioco ("Quali sono *i tuoi* tratti di personalità?"), gli psicologi hanno raccolto misurazioni della personalità su oltre 75.000 persone diverse, insieme ad altri dati sulle preferenze e sul comportamento. Questo ha creato un'enorme serie di dati per studiare molti temi interessanti della psicologia della personalità. *Il gioco FoldIt per la previsione delle strutture proteiche*: Prevedere le strutture formate dalle molecole proteiche è una delle grandi sfide computazionali della scienza. Nonostante molti anni di lavoro, ciò che fa sì che una proteina si ripieghi in una forma particolare non è ancora ben compreso. FoldIt è un gioco che sfida i non biologi a progettare molecole proteiche che si ripiegano in una particolare. I giocatori ricevono un punteggio in base a quanto il loro progetto si avvicina all'obiettivo dato, e i giocatori che ottengono i punteggi più alti vengono classificati in una

classifica. Sono stati pubblicati diversi articoli scientifici sulla forza dei progetti vincenti. La chiave del successo è creare un gioco che sia abbastanza giocabile da popolare. Questo è molto più difficile di quanto possa sembrare. Ci sono milioni di applicazioni gratuite nell'App Store, soprattutto giochi. Pochissime vengono provate da più di qualche centinaio di persone, il che non è affatto sufficiente per essere interessante dal punto di vista della raccolta dati. Se si aggiunge il vincolo supplementare che il gioco generi dati scientifici interessanti e allo stesso tempo sia giocabile, questo compito è ancora più difficile. Le tecniche motivazionali dovrebbero essere utilizzate per migliorare la giocabilità. Tenere il punteggio è una parte importante di qualsiasi gioco, e il gioco dovrebbe essere progettato in modo che le prestazioni aumentino rapidamente all'inizio, per agganciare il giocatore. Barre di progresso incoraggiare a raggiungere il livello successivo. L'assegnazione di distintivi e la creazione di classifiche viste da altri incoraggiano un maggiore impegno. Napoleone istituì una vasta gamma di nastri e decorazioni per i suoi soldati, osservando che "è sorprendente ciò che gli uomini fanno per una striscia di stoffa". Il principio di progettazione principale di giochi come FoldIt è quello di astrarre la tecnicità del dominio, nella funzione di punteggio. Il gioco è configurato in modo che i giocatori non debbano capire realmente le questioni della dinamica molecolare, ma solo che certe modifiche fanno salire i punteggi, mentre altre li fanno scendere. Il giocatore costruirà la propria intuizione sul dominio mentre gioca, dando vita a progetti che potrebbero non venire in mente agli esperti del settore.

Capitolo 4

Punteggi e classifiche

Le funzioni di punteggio sono misure che riducono i record multidimensionali a un singolo valore, evidenziando una particolare proprietà dei dati. Un esempio familiare di funzioni di punteggio sono quelle utilizzate per assegnare i voti agli studenti in corsi come il mio. Gli studenti possono poi essere classificati (ordinati) in base a questi punteggi numerici e successivamente assegnare voti in lettere in base a questo ordine.

I voti sono in genere calcolati da funzioni su caratteristiche numeriche che riflettono le prestazioni degli studenti, come i punti assegnati a ciascun compito ed esame. Ogni studente riceve un singolo punteggio combinato, spesso con una scala da 0 a 100. Questi punteggi derivano in genere da una combinazione lineare delle variabili di input, magari attribuendo un peso dell'8% a ciascuno dei cinque compiti a casa e un peso del 20% a ciascuno dei tre esami.
Ci sono diverse cose da osservare su queste rubriche di valutazione, che useremo come modello per funzioni di punteggio e di classificazione più generali:

- *Grado di arbitrarietà*: Ogni insegnante/professore utilizza un diverso compromesso tra i punteggi dei compiti e gli esami quando giudica i suoi studenti. Alcuni pesano l'esame finale più di tutte le altre variabili. Alcuni normalizzano ogni valore a 100 prima di fare la media, mentre altri convertono ogni punteggio in un punteggio Z. Tutti differiscono nella filosofia, ma ogni insegnante/professore è certo che il suo sistema di valutazione sia il modo migliore di fare le cose.

- *Mancanza di dati di convalida*: Non esiste un gold standard che informi gli istruttori sul voto "giusto" che i loro studenti *avrebbero* dovuto ricevere nel corso. Gli studenti spesso si lamentano che dovrei dare loro un voto migliore, ma dietro queste richieste sembra nascondersi più l'interesse personale che 'obiettività. Infatti, raramente sento gli studenti raccomandarmi di abbassare il loro voto.
  Senza un feedback oggettivo o degli standard con cui confrontarsi, non c'è un modo rigoroso per valutare il mio sistema di valutazione e migliorarlo.

- *Robustezza generale*: Eppure, nonostante l'utilizzo di approcci molto diversi e totalmente non convalidati, i diversi sistemi di valutazione producono generalmente risultati simili. In ogni scuola c'è una coorte di studenti con la sufficienza piena che monopolizzano una parte considerevole dei voti migliori in ogni corso. Questo non potrebbe accadere se tutti questi diversi sistemi di valutazione ordinassero arbitrariamente le prestazioni degli studenti. Gli studenti con la C in genere si muovono nelle fasce medio-basse della maggior parte dei loro corsi, invece di alternare A e F nel percorso verso la media finale. Tutti i sistemi di valutazione sono diversi, ma quasi tutti sono difendibili.

Le funzioni di punteggio spesso sembrano arbitrarie e ad hoc, e nelle mani sbagliate possono produrre numeri dall'aspetto impressionante che sono essenzialmente privi di significato. Poiché la loro efficacia in genere non può essere convalidata, queste tecniche non sono scientificamente valide come i metodi statistici e di apprendimento automatico che presenteremo nei capitoli successivi.

Ma credo che sia importante apprezzare le funzioni di punteggio per quello che sono: modi utili ed euristici per trarre comprensione da grandi insiemi di dati. Una funzione di punteggio viene talvolta chiamata *statistica*, il che le conferisce maggiore dignità e rispetto. Introdurremo diversi metodi per ottenere punteggi significativi dai dati.

4.1 L'indice di massa corporea (BMI)

Tutti amano mangiare e il nostro moderno mondo di abbondanza offre numerose opportunità per farlo. Il risultato è che una percentuale considerevole della popolazione supera il proprio peso corporeo ottimale. Ma come si fa a capire se è uno di loro? L'*indice di massa corporea* (BMI) è un punteggio o una statistica che serve a capire se il suo peso è sotto controllo. Si definisce come massa/ altezza^2 dove la massa è misurata in chilogrammi e l'altezza in metri. Nel momento in cui, sono alto 68 pollici (1,727 metri) e mi sento leggermente grassoccio con 150 libbre (68,0 kg). Pertanto, il mio IMC è $68{,}0/(1{,}727^2) = 22{,}8$. Questo non è così terribile, tuttavia, perché gli intervalli di IMC comunemente accettati negli Stati Uniti definiscono:

- *Sottopeso*: inferiore a 18,5.
- *Peso normale*: da 18,5 a 25.
- *Sovrappeso*: da 25 a 30.
- *Obeso*: oltre i 30 anni.

Quindi sono considerata nella fascia normale, con un'altra dozzina di chili da prendere prima di diventare ufficialmente in sovrappeso. Ogni punto di questo grafico a dispersione è una persona, colorata in base alla classificazione del suo peso secondo il BMI. Le regioni di colore apparentemente solido sono così dense di persone che i punti si

sovrappongono. I punti anomali a destra corrispondono agli individui più pesanti. L'IMC è un esempio di statistica/funzione di punteggio di grande successo. È ampiamente utilizzato e generalmente accettato, anche se alcuni nel campo della salute pubblica sostengono che sono disponibili statistiche migliori.

La logica dell'IMC è quasi sana. Il quadrato dell'altezza dovrebbe essere proporzionale all'area. Ma la massa dovrebbe crescere proporzionalmente al *volume*, non all'area, quindi perché non è *massa/altezza*? Storicamente, l'IMC è stato progettato per correlarsi alla percentuale di grasso corporeo di un individuo, che è una misura molto più difficile da realizzare rispetto all'altezza e al peso. Gli esperimenti con diverse funzioni di punteggio semplici, tra cui m/l e m/l^3, hanno rivelato che l'IMC funziona meglio. È molto interessante osservare le distribuzioni dell'IMC per le popolazioni estreme. Consideriamo gli atleti professionisti del football americano (NFL) e del basket (NBA): I giocatori di basket sono notoriamente persone alte. Inoltre, devono correre su e giù per il campo tutto il giorno, favorendo una forma fisica superiore

I giocatori di football americano sono notoriamente individui pesanti. In particolare, gli uomini di linea esistono solo per bloccare o spostare altri uomini di linea, il che comporta un premio per la massa.

In effetti, quasi tutti i giocatori di basket hanno un IMC normale, nonostante le loro altezze molto anormali. E i giocatori di calcio sono quasi uniformemente animali, con la maggior parte classificata come obesa, nonostante siano anche atleti ben allenati. Questi giocatori di calcio sono generalmente ottimizzati per la forza, invece che per la forma cardiovascolare. Utilizziamo i *grafici a dispersione* per mostrare ogni individuo come un punto nello spazio altezza-peso, con le etichette (classe di peso o posizione del giocatore) visualizzate come colori. Anche la ripartizione dell'IMC in base alla posizione è rivelatrice. Nel basket, le guardie sono veloci e slanciate, mentre i centri sono alti e intimidatori. Quindi tutte queste posizioni si dividono nettamente in base alle dimensioni. Nel calcio, i giocatori di abilità (i quarterback, i kicker e i punter) si dimostrano notevolmente più piccoli dei fianchi di manzo sulla linea.

4.2 Sviluppare sistemi di punteggio

I punteggi sono funzioni che mappano le caratteristiche di ogni entità in un valore numerico di merito. Questa sezione esaminerà gli approcci di base per costruire sistemi di punteggio efficaci e per valutarli.

4.2.1 Standard d'oro e deleghe

Storicamente, le valute cartacee erano sostenute dall'oro, il che significa che un dollaro cartaceo poteva sempre essere scambiato con un dollaro d'oro. Per questo sapevamo che il nostro denaro valeva più della carta su cui era stampato. Nella scienza dei dati, un *gold standard* è un insieme di etichette o risposte che riteniamo corrette. Nella formulazione originale del BMI, il gold standard era costituito dalle percentuali di grasso corporeo misurate accuratamente su un piccolo numero di soggetti. Naturalmente, tali misurazioni sono soggette a qualche errore, ma definendo questi valori come il gold standard per la forma fisica, accettiamo che siano la misura giusta. Confidiamo nell'oro. La presenza di un gold standard fornisce un modo rigoroso per sviluppare un buon sistema di punteggio. Possiamo utilizzare una tecnica di adattamento della curva, come la regressione lineare, per pesare le caratteristiche in ingresso in modo da approssimare al meglio le "risposte giuste" sulle istanze del gold standard.

Ma può essere difficile trovare dei veri e propri gold standard. *I proxy* sono dati più facili da trovare, che *dovrebbero essere* ben correlati con la verità di base desiderata ma introvabile. L'IMC è stato progettato per essere un proxy delle percentuali di grasso corporeo. È facilmente calcolabile a partire da altezza e peso e fa un ottimo lavoro di correlazione con il grasso corporeo. Ciò significa che raramente è necessario effettuare test di galleggiamento in vasche d'acqua o "pizzicare un pollice" con i calibri, misure più invasive che quantificano direttamente l'entità della ciccia di un individuo. Supponiamo che io voglia migliorare il sistema di valutazione che utilizzo per il corso di scienza dei dati del prossimo anno. Ho i dati degli studenti dell'anno precedente, vale a dire i loro punteggi nei compiti e nei test, ma non ho un gold standard su quali voti *meritassero* questi studenti. Ho solo il voto che ho dato loro, che non ha senso se sto cercando di migliorare il sistema. Ho bisogno di una proxy per il loro merito sconosciuto del corso "reale". Un buon candidato potrebbe essere il GPA cumulativo di ogni studente negli *altri* corsi. In generale, le prestazioni degli studenti dovrebbero essere conservate tra i vari corsi. Se il mio sistema di punteggio danneggia il GPA degli studenti migliori e aiuta quelli di livello inferiore, probabilmente sto facendo qualcosa di sbagliato. I proxy sono particolarmente utili quando si valutano i sistemi di punteggio/classifica. Nel nostro libro *Chi è più grande?* abbiamo utilizzato Wikipedia per classificare i personaggi storici in base alla loro "importanza". Non disponevamo di dati di importanza gold standard che misurassero la *reale* importanza queste persone. Ma abbiamo utilizzato diversi proxy per valutare il nostro operato, per mantenerci onesti:I prezzi che i collezionisti pagheranno per gli autografi delle celebrità *dovrebbero* generalmente *essere* correlati all'importanza della celebrità. Più alto è il prezzo che le persone sono disposte a pagare, più grande è la star. Le statistiche sulla bravura di un giocatore di baseball *dovrebbero* generalmente *essere* correlate all'importanza del giocatore. Più l'atleta è bravo, più è probabile che sia importante. Le classifiche pubblicate nei libri e nelle riviste elencano i migliori presidenti, le star del cinema, i cantanti, gli autori, ecc. I presidenti classificati più in alto dagli storici dovrebbero generalmente essere classificati più in alto da noi. Tali opinioni, in generale, dovrebbero essere correlate all'importanza di queste figure storiche.

4.2.2 Punteggi e classifiche

Le classifiche sono permutazioni che ordinano n entità per merito, generalmente costruite ordinando i risultati di un sistema di punteggio.

Esempi popolari di classifiche/sistemi di valutazione includono:

Top 20 di calcio/basket: Le agenzie di stampa generalmente classificano le migliori squadre sportive universitarie aggregando i voti degli allenatori o degli scrittori sportivi. In genere, ogni votante fornisce la propria classifica personale delle venti squadre migliori, e ogni squadra riceve più punti quanto più in alto compare nella lista dei votanti. Sommando i punti di ogni votante si ottiene un punteggio totale per ogni squadra, e l'ordinamento di questi punteggi definisce la classifica. *Classifiche accademiche universitarie*: La rivista *U.S. News and World Report* pubblica una classifica annuale dei migliori college e università americani. La loro metodologia è proprietaria e cambia ogni anno, presumibilmente per motivare le persone ad acquistare le nuove classifiche. Ma in genere si tratta di un punteggio prodotto da statistiche come il rapporto facoltà/studenti, il rapporto di accettazione, i punteggi dei test standardizzati degli studenti e dei candidati e forse le prestazioni delle squadre di calcio/basket. Anche i sondaggi degli esperti accademici fanno parte del mix. *Google PageRank/risultati di ricerca*: Ogni richiesta a un motore di ricerca attiva una quantità sostanziale di calcoli, classificando implicitamente la pertinenza di ogni documento sul web rispetto alla richiesta. I documenti vengono classificati in base alla corrispondenza con il testo della query, insieme alla valutazione della qualità intrinseca di ogni pagina. La metrica di qualità della pagina più famosa è il *PageRank*, l'algoritmo di centralità della rete. *Classifica di classe*: la maggior parte delle scuole superiori classifica gli studenti in base ai loro voti, con lo studente che si classifica al primo posto e che viene premiato come valedictorian della classe. La funzione di punteggio alla base di queste classifiche è in genere la media dei voti (GPA), in cui il contributo di ogni corso è ponderato in base al numero di crediti, e ogni possibile voto in lettere è mappato in un numero (in genere $A = 4.0$). Ma ci sono delle varianti naturali: molte scuole scelgono di pesare maggiormente i corsi di specializzazione rispetto ai corsi leggeri come la ginnastica, per riflettere la maggiore difficoltà di ottenere buoni voti. In generale, l'ordinamento dei risultati di un sistema di punteggio produce una classifica numerica. Ma pensando al contrario, la posizione di classifica di ogni articolo (ad esempio, 493° su 2196) produce anche un punteggio numerico per l'articolo.

Poiché i punteggi e le classifiche sono duali l'uno all'altro, quale dei due fornisce una rappresentazione più significativa dei dati? Come in ogni

confronto, la risposta migliore è che dipende da aspetti quali: *I numeri saranno presentati in modo isolato?* Le classifiche sono ottime per fornire un contesto per interpretare i punteggi. Nel momento in cui scrivo, la squadra di pallacanestro di Stony Brook è al 111° posto tra le 351 squadre universitarie della nazione, grazie al nostro RPI (indice percentuale di valutazione) di 39,18. Quale numero le dà un'idea migliore del fatto che abbiamo una squadra buona o cattiva, il 111° posto o il 39,18? *Qual è la distribuzione sottostante dei punteggi?* Per definizione, l'entità prima classificata ha un punteggio migliore di quella seconda classificata, ma questo non ci dice nulla sull'entità della differenza tra loro. Sono praticamente alla pari, o il numero 1 è in ? Le differenze nelle classifiche *sembrano* essere lineari: la differenza tra 1 e 2 sembra uguale a quella tra 111 e 112. Ma questo non è generalmente vero nei sistemi di punteggio. Infatti, piccole differenze di punteggio assoluto possono spesso produrre grandi differenze di classifica. *Le interessano gli estremi o il centro?* I sistemi di punteggio ben progettati hanno spesso una distribuzione a campana. Con i punteggi concentrati intorno alla media, piccole differenze nel punteggio possono significare grandi differenze nel ranking. In una distribuzione normale, un aumento del punteggio rispetto alla media di una deviazione standard (σ) la sposta dal 50° percentile all'84° percentile. Ma la stessa variazione di dimensioni da 1σ a 2σ la porta solo dall'84° al 92,5° percentile. Quindi, quando un'organizzazione scivola dal primo al decimo posto, le teste dovrebbero rotolare. Ma quando la squadra di Stony Brook scivola dalla 111esima alla 120esima posizione, probabilmente rappresenta una differenza insignificante nel punteggio e dovrebbe essere ignorata. Le classifiche sono ottime per evidenziare le entità migliori e peggiori del gruppo, ma meno le differenze vicino alla mediana.

4.2.3 Riconoscere le buone funzioni di punteggio

Le buone funzioni di punteggio sono valide perché sono facilmente interpretabili e geneticamente credibili. Qui rivediamo le proprietà delle statistiche che puntano in queste direzioni: *Facilmente calcolabile*: Una buona statistica può essere facilmente descritta e presentata. L'IMC è un esempio eccellente: contiene solo due parametri e si valuta con una semplice algebra. È stato trovato come risultato di una ricerca di tutte le forme funzionali semplici su un piccolo numero di variabili rilevanti

facilmente ottenibili. È un esercizio eccellente per fare un brainstorming di possibili statistiche da una serie data di caratteristiche su una serie di dati che si conoscono bene, per fare pratica. *Facilmente comprensibile*: Dalla descrizione delle statistiche dovrebbe essere chiaro che la classifica è rilevante per la domanda in questione. "La massa aggiustata per l'altezza" spiega perché l'IMC è associato all'obesità. Spiegare chiaramente le idee alla base della sua statistica è necessario affinché le altre persone si fidino abbastanza da utilizzarla. *Interpretazioni monotone delle variabili*: Dovrebbe avere un'idea di come ciascuna delle caratteristiche utilizzate nella sua funzione di punteggio sia correlata all'obiettivo. La massa *dovrebbe essere* correlata positivamente con l'IMC, perché essere pesanti richiede un peso elevato. L'altezza *dovrebbe essere* correlata negativamente, perché le persone alte pesano naturalmente di più di quelle basse. In generale, si sta producendo una funzione di punteggio senza un gold standard effettivo cui confrontarsi. Ciò richiede la comprensione del significato delle sue variabili, in modo che la sua funzione di punteggio si correli correttamente con questo obiettivo moscetto. *Produce risultati generalmente soddisfacenti sui valori anomali*: Idealmente, si conosce abbastanza di certi punti individuali per avere un'idea del loro posto in qualsiasi sistema di punteggio ragionevole. Se sono veramente sorpreso dall'identità delle entità top rivelate dal sistema di punteggio, probabilmente si tratta di un bug, non di una caratteristica. Quando calcolo i voti degli studenti dei miei corsi, conosco già i nomi di diverse stelle e di diversi idioti grazie alle loro domande in classe. Se i miei voti calcolati non corrispondono grossolanamente a queste impressioni, c'è un potenziale bug che deve essere. Se i dati sono davvero completamente anonimi per lei, probabilmente dovrebbe dedicare un po' di tempo a conoscere meglio il suo dominio. Come minimo, costruisca degli esempi artificiali ("Superstar" e "Superdork") con valori di caratteristica tali da essere vicini alla parte superiore e inferiore della classifica, e poi veda come si adattano ai dati reali. *Utilizza variabili sistematicamente normalizzate*: Le variabili tratte da distribuzioni a campana si comportano in modo sensato nelle funzioni di punteggio. Ci saranno degli outlier alle code di entrambe le estremità, che corrispondono agli elementi migliori/peggiori, oltre a un picco al centro degli elementi i cui punteggi dovrebbero essere tutti relativamente simili. Queste variabili distribuite normalmente devono essere trasformate in punteggi Z prima di, in modo

che tutte le caratteristiche abbiano medie e varianze comparabili. Questo riduce la dipendenza della funzione di punteggio dalle costanti magiche per regolare i pesi, in modo che nessuna singola caratteristica abbia un impatto troppo dominante sui risultati. In generale, la somma dei punteggi Z utilizzando i segni corretti (più per le variabili correlate positivamente e meno per le correlazioni negative) con pesi uniformi farà più o meno la cosa giusta. Una funzione migliore potrebbe pesare queste variabili per importanza, in base alla forza della correlazione con l'obiettivo. Ma è improbabile che non faccia molta differenza. *Rompe i legami in modo significativo*: Le funzioni di classificazione hanno un valore molto limitato quando ci sono molti legami. Classificare la maneggevolezza delle persone in base a quante dita hanno non sarà molto rivelatore. Ci sarà un gruppo molto selezionato con dodici, una grande maggioranza con dieci, e poi piccoli gruppi di vittime di incidenti sempre più disabili, fino ad arrivare a zero. In generale, i punteggi dovrebbero essere numeri reali in un intervallo sano, al fine di ridurre al minimo la probabilità di creare dei legami. L'introduzione di caratteristiche secondarie per rompere i legami è preziosa e ha senso a condizione che queste caratteristiche siano anche correlate alla proprietà che le interessa.

4.3 Punteggi Z e normalizzazione

Un principio importante della scienza dei dati è che dobbiamo cercare di rendere il più semplice possibile per i nostri modelli fare la cosa giusta. Le tecniche di apprendimento automatico, come la regressione lineare, pretendono di trovare la linea che si adatta in modo ottimale a un dato insieme di dati. Ma è fondamentale normalizzare tutte le diverse variabili per rendere la loro gamma/distribuzione comparabile, prima di cercare di usarle per adattarsi a qualcosa.

I punteggi Z saranno il nostro metodo principale di normalizzazione. La trasformazione del punteggio Z viene calcolata: $Z_i = (a_i - \mu)/\sigma$ dove μ è la media della distribuzione e σ la deviazione standard associata. I punteggi Z trasformano insiemi arbitrari di variabili in un intervallo uniforme. Il punteggio Z dell'altezza misurata in pollici sarà esattamente uguale a quello dell'altezza misurata in miglia. Il valore medio di un punteggio Z su tutti i punti è zero. La Figura 4.4 mostra una serie di numeri interi ridotti a punteggi Z. I valori superiori media diventano positivi, mentre quelli inferiori alla media diventano negativi. La deviazione standard dei punteggi Z è pari a 1, quindi tutte le distribuzioni di punteggi Z hanno proprietà simili. La trasformazione dei valori in punteggi Z raggiunge due obiettivi. In primo luogo, aiutano a visualizzare i modelli e le correlazioni, assicurando che tutti i campi abbiano un valore identico. media (zero) e operano su un intervallo simile. Comprendiamo che un punteggio Z di 3,87 deve rappresentare l'altezza a livello di giocatore di basket in un modo che 79,8 non rappresenta, senza la familiarità con l'unità di misura (ad esempio i pollici). In secondo luogo, l'uso dei punteggi Z facilita i nostri algoritmi di apprendimento automatico, rendendo tutte le diverse caratteristiche di una scala comparabile. In teoria, l'esecuzione di una trasformazione lineare come lo Z-score non fa nulla che la maggior parte degli algoritmi di apprendimento non possa capire da sola. Questi algoritmi in genere trovano il miglior coefficiente con cui moltiplicare ogni variabile, che è libero di essere vicino a σ se l'algoritmo lo desidera veramente. Tuttavia, qui entrano in gioco le realtà del calcolo numerico. Supponiamo di voler costruire un modello lineare su due variabili associate alle città degli Stati Uniti, ad esempio l'area in miglia quadrate e la popolazione. La prima ha una media di circa 5 e un massimo di circa 100. La seconda ha una media di circa 25.000 abitanti. La seconda ha una media di circa 25.000 e un massimo di 8.000.000. Affinché le due variabili abbiano un effetto simile sul nostro modello, dobbiamo dividere la seconda variabile per un fattore di 100.000 circa. Questo causa problemi di precisione numerica, perché un cambiamento molto piccolo nel valore del coefficiente causa un cambiamento molto grande nella misura in cui la variabile della popolazione domina il modello. Sarebbe molto meglio che le variabili avessero la stessa scala e lo stesso intervallo di distribuzione, in modo che il problema sia se una caratteristica viene ponderata, ad esempio, due volte più fortemente di un'altra. I punteggi Z sono utilizzati al meglio su variabili distribuite normalmente, che, dopo tutto, sono completamente descritte dalla media μ e dalla deviazione standard σ. Ma funzionano meno bene quando la distribuzione è una legge di potenza. Consideriamo la distribuzione della ricchezza negli Stati Uniti, che può avere una media di (diciamo) 200.000 dollari, con una σ= di 200.000 dollari. Lo Z-score di Bill Gates, che ha un valore di 80 miliardi di dollari, sarebbe quindi di

4999, ancora un incredibile outlier data la media di zero. Il suo più grande peccato nell'analisi dei dati sarà l'utilizzo di variabili non correttamente normalizzate nella sua analisi. Cosa possiamo fare per Bill Gates? Possiamo colpirlo con un tronco.

4.4 Tecniche di classificazione avanzate

La maggior parte dei compiti di classificazione sono risolti calcolando i punteggi come combinazioni lineari di caratteristiche e poi ordinandoli. In assenza di un gold standard, questi metodi producono statistiche spesso rivelatrici e informative. Detto questo, sono state sviluppate diverse tecniche potenti per calcolare le classifiche da tipi specifici di input: i risultati di confronti accoppiati, reti di relazioni e persino assemblaggi di altre classifiche. Passiamo qui in rassegna questi metodi, a titolo di ispirazione.

4.4.1 Classifiche Elo

Le classifiche sono spesso formate analizzando sequenze di confronti binari, che sorgono naturalmente nelle competizioni tra entità:

- *Risultati di gare sportive*: I tipici eventi sportivi, che siano partite di calcio o di scacchi, mettono le squadre A e B una contro l'altra. Solo una di esse vincerà. Quindi ogni partita è essenzialmente un confronto binario di merito.

- *Voti e sondaggi*: Alle persone competenti viene spesso chiesto di confrontare le opzioni e di decidere quale scelta ritengono migliore. In un'elezione, questi confronti sono chiamati voti. Una componente importante di alcune classifiche universitarie deriva dal chiedere ai professori: quale scuola è migliore, A o B?

Nel film *The Social Network*, Mark Zuckerberg di Facebook viene mostrato mentre inizia a lavorare con FaceMash, un sito web che mostrava agli spettatori due volti e chiedeva loro di scegliere quale fosse il più attraente. Il suo sito classificava poi tutti i volti dal più al meno attraente, in base a questi confronti accoppiati.

- *Confronti impliciti*: Dal giusto punto di vista, i dati sulle caratteristiche possono interpretati in modo significativo come confronti a coppie. Supponiamo che uno studente sia stato accettato da entrambe le università A e B, ma che opti per A. Questo può essere interpretato come un voto implicito che A è migliore di B.

Qual è il modo giusto di interpretare le raccolte di questi voti, soprattutto quando ci sono molti candidati e non tutte le coppie di giocatori affrontano? Non è ragionevole dire che chi ha ottenuto il maggior numero di vittorie, perché (a) potrebbe aver partecipato a più confronti rispetto ad altri giocatori, e (b) potrebbe aver evitato avversari forti e aver battuto solo concorrenti inferiori. Il *sistema Elo* inizia valutando tutti i giocatori, presumibilmente in modo uguale, e poi aggiusta in modo incrementale il punteggio di ogni giocatore in risposta al risultato di ogni partita, secondo la formula: $r'(A) = r(A) + k(S_A - \mu_A)$, dove

- $r(A)$ e $r'(A)$ rappresentano i punteggi precedenti e aggiornati del giocatore A.

- k è un parametro fisso che riflette il massimo aggiustamento possibile del punteggio in risposta ad una singola partita. Un valore piccolo di k dà luogo a classifiche abbastanza statiche, mentre l'utilizzo di un k troppo grande causerà oscillazioni selvagge nella classifica in base all'ultima partita.

- S_A è il risultato di punteggio ottenuto dal giocatore A nella partita in esame. In genere, $S_A = 1$ se A ha vinto, e $S_A = -1$ se A *ha* perso.

- μ_A è il risultato atteso per A quando compete contro B. Se A ha esattamente lo stesso livello di abilità di B, allora presumibilmente $\mu_A = 0$. Ma supponiamo che A sia un campione e che B sia un principiante o un incapace. La nostra aspettativa è che A vinca quasi certamente in un incontro testa a testa, quindi $\mu_A > 0$ e probabilmente sarà molto vicino a 1. Tutto è chiaro qui, tranne come determinare μ_A. Data una stima della probabilità che A batta B ($P_{A>B}$), allora $\mu_A = 1 \cdot P_{A>B} + (-1) \cdot (1 - P_{A>B})$.

Questa probabilità di vittoria dipende chiaramente dall'entità della differenza di abilità tra i giocatori A e B, che è esattamente ciò che si suppone venga misurato dal sistema di classificazione. Quindi $x = r(A) \quad r(B)$ rappresenta questa differenza di abilità. Per completare il sistema di classifica Elo, abbiamo bisogno di un modo per prendere questa variabile reale x e convertirla in una probabilità significativa. Si tratta di un problema importante che incontreremo ripetutamente in questo libro e che viene risolto da una piccola matematica chiamata *funzione logit*.

La funzione Logit

Supponiamo di voler prendere una variabile reale $-\infty < x < \infty$ e di convertirla in una probabilità $0 \leq p \leq 1$. Ci sono molti modi in cui si può immaginare di farlo, ma una trasformazione particolarmente semplice è $p = f(x)$. Si notino in particolare i casi speciali ai punti medi e finali:

Quando due giocatori hanno la stessa abilità, $x = 0$, e $f(0) = 1/2$, riflette che entrambi i giocatori hanno la stessa probabilità di vincere.

- Quando il giocatore A ha un grande vantaggio, $x \to \infty$ e $f(\ ) = 1, \infty$ definendo che A è in grado di vincere la partita.

- Quando il giocatore B ha un grande vantaggio, $-x \to \infty$ $0,$ e $f($ $-\infty$ $)=$ denotando che B è in grado di vincere la partita.

Questi sono esattamente i valori che vogliamo se x misura la differenza di abilità tra i giocatori. La funzione logit interpola in modo uniforme e simmetrico tra questi poli. Il parametro c della funzione logit regola la pendenza della transizione. Piccole differenze di abilità si traducono in grandi differenze nella probabilità di vincere? Per $c = 0$, il paesaggio è piatto come una frittella: $f(x) = 1/2$ per tutti x. Più grande è c, più netta è la transizione. Infatti, $c = \infty$ produce una funzione passo-passo da 0 a 1. Impostare $c = 1$ è un inizio ragionevole, ma la scelta giusta è specifica del dominio. L'osservazione della frequenza con cui una determinata grandezza di differenza di abilità si traduce in una sconfitta (vittoria della parte più debole) aiuta a specificare il parametro. Il sistema di classificazione Elo Chess è stato progettato in modo che $r(A)$ $r(B) = 400$ significa che A ha una probabilità di vittoria dieci volte superiore a B. Qui $k = 40$, il che implica un'oscillazione di punteggio massima possibile di 80 punti come conseguenza di una singola partita. La funzione logit standard dava a Kasparov una probabilità di 0,999886 di battere Skiena al primo turno, ma per un miracolo simile alla resurrezione di Lazzaro, la partita è andata nella direzione opposta. Di conseguenza, 80 punti sono passati dalla classifica di Kasparov alla mia. Dall'altra parte della staffa si sono affrontati due veri campioni di scacchi, con il più immaginabile sconvolgimento da parte di Polgar che ha spostato solo 55 punti. Ha spazzato via il pavimento con me all'ultimo turno, un risultato così chiaramente atteso che ha guadagnato essenzialmente zero punti di rating. Il metodo Elo è molto efficace nell'aggiornare le valutazioni in risposta alla sorpresa, non solo alla vittoria.

4.4.2 Unire le classifiche

Qualsiasi caratteristica numerica singola f, come l'altezza, può seminare 2^n confronti a coppie tra n elementi, verificando se $f(A) > f(B)$ per ogni coppia di elementi A e B. Potremmo inserire queste coppie nel metodo Elo per ottenere una graduatoria, ma questo sarebbe un modo stupido di pensare alle cose. Dopo tutto, il risultato di una tale analisi rifletterebbe semplicemente l'ordine ordinato di f. Tuttavia, l'integrazione di una raccolta di classifiche di diverse caratteristiche rende problema più interessante. Qui interpretiamo l'ordine ordinato della caratteristica *iesima* come se definisse una permutazione P_i sugli elementi di interesse. Cerchiamo la permutazione di consenso P, che in qualche modo riflette meglio tutte le permutazioni componenti $P_1, \ldots, P_k$. Ciò richiede la definizione di una funzione di distanza per misurare la somiglianza tra due permutazioni. Un problema simile si è presentato nella definizione del coefficiente di correlazione di rango di Spearman (vedere la Sezione 2.3.1), dove abbiamo confrontato due variabili con misura dell'accordo nell'ordine relativo degli elementi. *Il metodo di Borda* crea una classifica di consenso da molteplici altre classifiche, utilizzando un semplice sistema di punteggio. In particolare, assegniamo un costo o un peso a ciascuna delle n posizioni della permutazione. Poi, per ciascuno degli n elementi, sommiamo i pesi delle sue posizioni su tutte le k classifiche di ingresso. L'ordinamento di questi n punteggi determina la classifica di consenso finale. Ora tutto è chiaro, tranne la mappatura tra posizioni e costi. La funzione di costo più semplice assegna i punti per apparire nella posizione iesima in ogni permutazione, cioè sommiamo i gradi dell'elemento su tutte le permutazioni. L'elemento A ottiene 3 1+ 1 2= 5 punti su grazie al fatto di apparire primo in tre classifiche e secondo in una. L'ar˙ticolo C ottiene 12 punti grazie ai piazzamenti 2, 3, 3 e 4. La classifica finale di consenso di A, B, C, D, E integra tutti i voti di tutte le classifiche di ingresso, anche se il consenso è in disaccordo almeno in parte con tutte e quattro le classifiche di ingresso. Ma non è chiaro se l'utilizzo di pesi lineari rappresenti la scelta migliore, perché presuppone una fiducia uniforme nella nostra precisione nel posizionare gli elementi. in tutta la permutazione. In genere, conosceremo la maggior parte mer- cato delle nostre scelte principali, ma saremo piuttosto confusi sul modo in cui quelli che si trovano vicino 'ordine intermedio si posizionano tra di loro. In questo caso, un approccio migliore potrebbe essere quello di assegnare più punti per la distinzione tra il 1° e il 2° che tra il 110° e il 111°. Questo tipo di ponderazione è implicitamente eseguito da una curva a campana. Supponiamo di campionare n articoli a intervalli uguali da una distribuzione normale. Assegnando questi valori x come pesi posizionali, si ottiene una maggiore diffusione ai ranghi più alti e più bassi rispetto al centro. Le regioni di coda sono davvero ampie come

sembrano per questi 50 punti equidistanti: ricordiamo che il 95% della massa di probabilità si trova entro 2σ dal centro. In alternativa, se la nostra fiducia non è simmetrica, potremmo campionare dalla distribuzione seminormale, in modo che la coda dei nostri ranghi sia ponderata dal picco della distribuzione normale. In questo modo, c'è la massima separazione tra gli elementi classificati più alti, ma poca distinzione tra gli elementi della coda. La scelta della funzione di ponderazione dipende dal dominio, quindi ne scelga una che sembra fare un buon lavoro per il suo problema. L'identificazione della *migliore* funzione di costo risulta essere un problema mal posto. E succedono cose strane quando cerchiamo di progettare il sistema elettorale perfetto.

4.4.3 Classifiche basate sui digrammi

Le reti offrono un modo alternativo di pensare a un insieme di voti della forma "*A* è più avanti di *B*". Possiamo costruire un grafo/rete diretto in cui c'è un vertice corrispondente a ciascuna entità, e un bordo diretto (*A, B*) per ogni voto che *A* è superiore a *B*. La classifica ottimale sarebbe quindi una permutazione *P* dei vertici che viola il minor numero di bordi, dove il bordo (*A, B*) è violato se *B* viene prima di *A* nella permutazione finale della classifica *P*. Se i voti fossero totalmente coerenti, allora questa permutazione ottimale violerebbe esattamente zero bordi. In effetti, questo è il caso quando non ci sono cicli diretti nel grafico. Un ciclo diretto come (*A, C*), (*C, E*), (*E, A*) rappresenta una contraddizione intrinseca a qualsiasi ordine di classifica, perché ci sarà sempre un bordo infelice, indipendentemente dall'ordine scelto. Un grafo diretto senza cicli si chiama *grafo aciclico diretto* o DAG. Un lettore attento con un po' di background sugli algoritmi ricorderà che la ricerca di questo ordine ottimale dei vertici si chiama *ordinamento topologico* del DAG, che può essere eseguito in modo efficiente in tempo lineare. La Figura 4.9 (a sinistra) è un DAG e presenta esattamente due ordini distinti coerenti con i bordi diretti: *{A, B, C, D, E}* e *{A, C, B, D, E}*. Tuttavia, è estremamente improbabile che un insieme reale di caratteristiche o di elettori siano tutti coerenti tra loro. Il problema del *massimo sottografo aciclico* cerca di trovare il minor numero di bordi da eliminare per lasciare un DAG. La rimozione del bordo (*E, A*) è sufficiente. Sfortunatamente, il problema di trovare il miglior ranking in questo caso è NP-completo, il che significa che non esiste un algoritmo efficiente per trovare la soluzione ottimale. Ma esistono delle euristiche naturali. Un buon indizio sull'appartenenza di un vertice *v* è la differenza d_v tra il suo grado in e il suo grado out. Quando d_v è altamente negativo, probabilmente appartiene alla parte anteriore della permutazione, poiché

domina molti elementi ma è dominato solo da alcuni. Si può costruire una permutazione di classifica decente ordinando i vertici in base a queste differenze. Ancora meglio è inserire in modo incrementale il vertice più negativo (o più positivo) v nella sua posizione logica, eliminando i bordi incidenti su v, e poi aggiustare i conteggi prima di posizionare il prossimo vertice migliore.

4.4.4 PageRank

Esiste un metodo diverso e più famoso per ordinare i vertici di una rete in base all'importanza: l'algoritmo PageRank alla base del motore di ricerca di Google. Il web è costituito da pagine web, la maggior parte delle quali contiene link ad altre pagine web. La sua pagina web che rimanda alla mia è un'approvazione implicita del fatto che lei pensa che la mia pagina sia abbastanza buona. Se viene interpretato come un voto che "pensa che la mia pagina sia migliore della sua", possiamo costruire la rete di link e come un problema di massimo sottografo aciclico, discusso nella sottosezione precedente. Ma la dominanza non è l'interpretazione giusta per i link sul web. Il PageRank invece premia i vertici che hanno il maggior numero di in-link: se tutte le strade portano a Roma, Roma deve essere un luogo abbastanza importante. Inoltre, valuta questi in-link in base alla forza della fonte: un link a me da una pagina importante dovrebbe contare di più di uno da un sito di spam. I dettagli sono interessanti. Tuttavia, spero che questa breve introduzione al PageRank la aiuti ad apprezzare il racconto seguente.

4.5 Storia di guerra: La vendetta di Clyde

Durante il secondo anno di liceo, ho avuto l'idea di scrivere un programma per prevedere l'esito delle partite di calcio professionistico. Non ero molto interessato al calcio come sport, ma osservavo diversi miei compagni di classe che scommettevano i soldi del pranzo sull'esito delle partite di calcio del fine settimana. Mi è sembrato chiaro che scrivere un programma che prevedesse con precisione l'esito delle partite di calcio avrebbe potuto avere un valore significativo ed essere una cosa molto bella da fare.

Con il senno di poi, il programma che ho ideato mi sembra irrimediabilmente rozzo. Il mio programma avrebbe fatto una media dei punti segnati dalla squadra x e dei punti concessi dalla squadra y per prevedere il numero di punti che x segnerà contro y. Poi aggiustavo questi numeri verso l'alto o verso il basso in base ad altri fattori, in particolare il vantaggio del campo di casa, arrotondavo i numeri in modo appropriato e chiamavo ciò che rimaneva il mio punteggio previsto per la partita.

Questo programma per computer, *Clyde*, è stato il mio primo tentativo di costruire una funzione di punteggio per qualche aspetto del mondo reale. Aveva una certa logica. Le buone squadre segnano più punti di quelli che concedono, mentre le cattive squadre concedono più punti di quelli che segnano. Se la squadra *x* gioca contro una squadra *y* che ha molti punti, allora *x* dovrebbe segnare più punti contro *y* che contro squadre con difese migliori. Allo stesso modo, più punti ha segnato la squadra *x* contro il resto del campionato, più punti è probabile che segni contro *y*. Naturalmente, questo modello grezzo non può catturare tutti gli aspetti della realtà calcistica. Supponiamo che la squadra *x* abbia giocato contro tutte le squadre di scarso valore fino a questo momento della stagione, mentre la squadra *y* abbia giocato contro le migliori squadre del campionato. La squadra *y* potrebbe essere una squadra molto migliore di *x*, anche se il suo record finora è scarso. Questo modello ignora anche gli infortuni di cui soffre una squadra, il fatto che il clima sia caldo o freddo e che la squadra sia calda o fredda. Non tiene conto di tutti i fattori che rendono lo sport intrinsecamente imprevedibile. Eppure, anche un modello così semplice può fare un lavoro ragionevole di previsione dell'esito delle partite di calcio. Se si calcolano le medie dei punti come sopra, e si assegnano alla squadra di casa tre punti in più come bonus, si sceglie il vincitore in quasi due terzi di tutte le partite di calcio. Confrontatelo con il modello ancora più rozzo del lancio di una moneta, che predice correttamente solo la metà delle partite. Questa è stata la prima grande lezione che *Clyde* mi ha insegnato: *Anche i modelli matematici più grezzi possono avere un reale potere predittivo.* Da audace sedicenne, scrissi al nostro giornale locale, *The New Brunswick Home News*, spiegando che avevo un programma informatico per prevedere i risultati delle partite di calcio ed ero pronto ad offrire loro l'opportunità esclusiva di pubblicare i miei pronostici ogni settimana. Ricordiamo che questo accadeva nel 1977, ben prima che i personal computer si affacciassero alla coscienza pubblica. A quei tempi, l'idea che un liceale *usasse* davvero un computer aveva un notevole valore di novità. Ho ottenuto il lavoro. *Clyde* predisse il risultato di ogni partita della National Football League del 1977. Se non ricordo male, io e *Clyde* finimmo la stagione con un record apparentemente impressionante di 135-70. Ogni settimana, si confrontavano le mie previsioni con quelle dei giornalisti sportivi del giornale. Se non ricordo male, ci siamo ritrovati tutti a poche partite di distanza l'uno dall'altro, anche se la maggior parte dei redattori sportivi ha concluso con record migliori di quelli del computer. L'*Home News* fu così impressionato dal mio lavoro che non mi rinnovò la stagione successiva. Tuttavia, le scelte di *Clyde* per la stagione 1978 furono pubblicate

sul *Philadelphia Inquirer*, un giornale molto più grande. Tuttavia, non avevo la colonna tutta per me. Invece, l'*Inquirer* mi ha incluso tra dieci pronosticatori dilettanti e professionisti, o tout. Ogni settimana dovevamo prevedere l'esito di quattro partite contro lo spread di punti. Il divario di punti nel calcio è un modo per handicappare le squadre più forti ai fini delle scommesse. Lo spread di punti è progettato per rendere ogni partita una proposta 50/50, e quindi rende molto più difficile prevedere l'esito delle partite. Io e *Clyde* non siamo andati molto bene contro lo spread durante la stagione
1978 della National Football League, e nemmeno la maggior parte degli altri touts *del Philadelphia Inquirer*. Abbiamo pronosticato solo il 46% delle partite correttamente contro lo spread, una performance abbastanza buona (o cattiva) da classificarci al 7° posto tra i dieci pronosticatori pubblicati. Scegliere contro lo spread mi ha insegnato una seconda importante lezione di vita: *I modelli matematici grezzi non hanno un reale potere predittivo quando c'è in gioco del denaro vero.* Quindi Clyde non era destinato a rivoluzionare il mondo dei pronostici sul calcio. Me ne sono praticamente dimenticato fino a quando non ho assegnato la sfida di prevedere il Super Bowl come progetto nel mio corso di scienza dei dati. La squadra che ha ottenuto il lavoro era composta da studenti indiani, il che significa che conoscevano molto di più il cricket che il football americano quando hanno iniziato. Tuttavia, hanno raccolto la sfida, diventando fan mentre costruivano un'ampia serie di dati sui risultati di ogni partita professionale e universitaria giocata negli ultimi dieci anni. Hanno eseguito un'analisi di regressione logistica su 142 caratteristiche diverse, tra cui la corsa, il passaggio e il guadagno dei calci, il tempo di possesso e il numero di punizioni. Poi mi hanno riferito con orgoglio l'accuratezza del loro modello: previsioni corrette sul 51,52% delle partite NFL. "*Cosa!*" Ho urlato: "*È terribile!*". "Il cinquanta per cento è quello che si ottiene lanciando una moneta. Provi a fare una media dei punti segnati e subiti dalle due squadre e dia tre punti alla squadra di casa. Come va questo semplice modello?". Sul loro set di dati, questo modello Clyde-light ha scelto il 59,02% di tutte le partite in modo corretto, molto meglio del loro modello di apprendimento automatico dall'aspetto sofisticato. Si erano persi nella nebbia di troppe caratteristiche, che non erano normalizzate correttamente, e costruite utilizzando statistiche raccolte in storia troppo lunga per essere rappresentative della composizione attuale della squadra. Alla fine gli studenti sono riusciti a trovare un modello basato su PageRank che ha fatto un po' meglio (60,61%), ma Clyde ha fatto quasi altrettanto bene come modello di base. Ci sono diverse lezioni importanti qui. In primo luogo, l'immondizia

entra, l'immondizia esce. Se non si prepara un set di dati pulito e normalizzato correttamente, gli algoritmi di apprendimento automatico più avanzati di non possono. In secondo luogo, i punteggi semplici basati su una modesta quantità di conoscenze specifiche del dominio possono dare risultati sorprendenti. Inoltre, aiutano a mantenere l'onestà. Costruisca e valuti linee di base semplici e comprensibili prima di investire in approcci più potenti. Il fatto che Clyde abbia scelto la linea di base ha lasciato il suo modello di apprendimento automatico senza difese.

4.6 Teorema dell'impossibilità di Arrow

Abbiamo visto diversi approcci per costruire classifiche o funzioni di punteggio dai dati. Se disponiamo di un gold standard che riporta l'ordine relativo "giusto" per almeno alcune delle entità, allora questo potrebbe essere utilizzato per addestrare o valutare la nostra funzione di punteggio, in modo da concordare con queste classifiche nella misura più ampia possibile. Ma senza un gold standard, si può dimostrare che non esiste un sistema di classificazione migliore. Questa è una conseguenza *del teorema di impossibilità di Arrow*, che dimostra che nessun sistema elettorale per aggregare permutazioni di preferenze soddisfa le seguenti proprietà desiderabili e dall'aspetto innocente:

- Il sistema deve essere completo, nel senso che quando si chiede di scegliere tra le alternative A e B, deve dire (1) A è preferito a B, (2) B è preferito ad A, oppure (3) c'è una preferenza uguale tra loro.
- I risultati devono essere transitivi, ossia se A è preferito a B e B è preferito a C, allora A deve essere preferito a C.

• Se ogni individuo preferisce A a B, allora il sistema deve preferire A a B.

- Il sistema non deve dipendere solo dalle preferenze di un individuo, un dittatore.

- La preferenza di A rispetto a B dovrebbe essere indipendente dalle preferenze per qualsiasi altra alternativa, come C.

Teorema di Arrow e la natura non transitiva dell'ordinamento di tipo "sassocarta-forbice". Tre elettori (x, y e z) che classificano le loro preferenze tra i colori. Per stabilire la preferenza tra due colori a e b, un sistema logico potrebbe confrontare quante permutazioni classificano a prima di b rispetto a b prima di a. Secondo questo sistema, il rosso è preferito al verde da x e y,

quindi il rosso vince. Allo stesso modo, il verde è preferito al blu da x e z, quindi vince il verde. Per transitività, il rosso dovrebbe essere preferito al blu per implicazione su questi risultati. Tuttavia, y e z preferiscono il blu al rosso, violando una proprietà intrinseca che vogliamo che il nostro sistema elettorale preservi. Il teorema di Arrow è molto sorprendente, ma significa che dovremmo alle classifiche come strumento di analisi dei dati? Ovviamente no, così come il teorema di Arrow significa che dovremmo rinunciare alla democrazia. I sistemi di voto tradizionali, basati sull'idea che la *maggioranza governa*, in genere fanno un buon lavoro nel riflettere le preferenze popolari, una volta generalizzati in modo appropriato per trattare un gran numero di candidati Arrow. E le tecniche di questo capitolo fanno generalmente un buon lavoro nel classificare gli elementi in modo interessante e significativo. *Non* cerchiamo classifiche corrette, perché questo è un obiettivo mal definito. Cerchiamo invece classifiche utili e interessanti.

4.7 Storia di guerra: Chi è più grande?

I miei studenti a volte mi dicono che io sono la storia. Spero che non sia ancora vero, ma sono molto interessata alla storia, come il mio ex postdoc Charles Ward. Charles e io abbiamo chiacchierato su chi sono le figure più significative della storia... e come si può misurare. Come la maggior parte delle persone, abbiamo trovato le nostre risposte in Wikipedia.
Wikipedia è una cosa incredibile, un prodotto di lavoro distribuito costruito da oltre 100.000 autori che in qualche modo mantiene uno standard generalmente valido di accuratezza e profondità. Wikipedia cattura una quantità sorprendente di conoscenza umana in una forma aperta e leggibile dalle macchine. Abbiamo deciso di utilizzare la Wikipedia inglese come fonte di dati su cui basare le classifiche storiche. Il nostro primo passo è stato quello di estrarre dalla pagina Wikipedia di ogni persona le variabili di caratteristica che dovrebbero essere chiaramente correlate all'importanza storica. Questo includeva caratteristiche come:

- *Lunghezza*: I personaggi storici più significativi dovrebbero avere pagine di Wikipedia più lunghe rispetto ai comuni mortali. Pertanto, la lunghezza dell'articolo in parole fornisce una caratteristica naturale che riflette la potenza storica, almeno in una certa misura.

- *Hits*: Le figure più significative hanno le loro pagine Wikipedia lette più spesso di altre, perché sono di maggiore interesse per un numero maggiore di persone. La mia pagina di Wikipedia viene visitata in media venti volte al giorno, il che è piuttosto interessante. Ma la pagina di Issac

Newton viene visitata in media 7700 volte al giorno, il che è molto meglio.

- *PageRank*: Le figure storiche significative interagiscono con altre figure storiche significative, che si riflettono come riferimenti ipertestuali negli articoli di Wikipedia. Questo definisce un grafo diretto in cui i vertici sono articoli e i bordi diretti sono collegamenti ipertestuali. Il calcolo del PageRank di questo grafico misurerà la centralità di ogni figura storica, che si correla bene con l'importanza.

totale, abbiamo estratto sei caratteristiche per ogni figura storica. Successivamente, abbiamo normalizzato queste variabili prima di aggregarle, essenzialmente combinando le classifiche sottostanti con pesi distribuiti in modo normale. Abbiamo utilizzato una tecnica chiamata *analisi statistica dei fattori*, collegata all'analisi delle componenti principali, per isolare due fattori che spiegavano la maggior parte della varianza dei nostri dati. Una semplice combinazione lineare di queste variabili ci ha fornito una funzione di punteggio e abbiamo ordinato i punteggi per determinare la nostra classifica iniziale, che abbiamo chiamato *fama*. Le prime venti figure in base al nostro punteggio di. Abbiamo studiato queste classifiche e abbiamo deciso che non catturavano realmente ciò volevamo. La top 20 della fama comprendeva musicisti pop come Madonna e Michael Jackson, e tre presidenti degli Stati Uniti contemporanei. Era chiaro che le figure contemporanee si posizionavano molto più in alto di quanto pensassimo: la nostra funzione di punteggio catturava la fama attuale molto più dell'importanza storica.

La nostra soluzione è stata quella di decadere i punteggi dei personaggi contemporanei per tenere conto del passare del tempo. Il fatto che una celebrità attuale ottenga molte visite su Wikipedia è impressionante, ma il fatto che ci interessiamo ancora a qualcuno che è morto 300 anni fa è molto più impressionante. Ecco cosa stavamo cercando! Abbiamo convalidato le classifiche utilizzando tutti i proxy di importanza storica che siamo riusciti a trovare: altre classifiche pubblicate, prezzi degli autografi, statistiche sportive, libri di testo di storia e risultati delle elezioni della Hall of Fame. Le nostre classifiche hanno mostrato una forte correlazione con tutti questi indicatori. In effetti, credo che queste classifiche siano meravigliosamente rivelatrici. Abbiamo scritto un libro che descrive tutti i tipi di cose che si possono imparare da esse. La incoraggio a leggerlo se è interessato alla storia e alla cultura. Più studiavamo queste classifiche, più rimanevo colpito dalla loro solidità generale. Detto questo, le nostre classifiche pubblicate non

hanno incontrato un consenso universale. Tutt'altro. Sono stati pubblicati decine di articoli di giornali e riviste sulle nostre classifiche, molti dei quali piuttosto ostili. Perché la gente non le rispettava, nonostante la nostra ampia convalida? In retrospettiva, la maggior parte delle critiche che abbiamo ricevuto erano dovute a tre motivi diversi:

- *Differenti nozioni implicite di significato*: I nostri metodi sono stati concepiti per misurare *la forza dei meme*, ovvero il successo di queste figure storiche nel propagare i loro nomi nella storia. Ma molti lettori pensavano che i nostri metodi dovessero catturare le nozioni di *grandezza* storica. Chi è stato più importante, in termini di cambiamento del mondo? E intendiamo il mondo o solo il mondo di lingua inglese? Come è possibile che non ci siano cinesi o indiani figure nella lista quando rappresentano oltre il 30% della popolazione mondiale?

  Dobbiamo essere d'accordo su ciò che stiamo cercando di misurare prima di misurarlo. L'altezza è un'ottima misura delle dimensioni, ma non è in grado di catturare l'obesità. Tuttavia, l'altezza è molto utile per selezionare i giocatori di una squadra di basket.

- *Gli outlier*: I test di sniff sono importanti per valutare i risultati di un'analisi. Per quanto riguarda le nostre classifiche, questo significava controllare il posizionamento delle persone che conoscevamo, per confermare che rientravano in posizioni ragionevoli.

Mi sono sentito bene per quanto riguarda la classificazione della stragrande maggioranza dei personaggi storici da parte del nostro metodo. Ma ci sono state alcune persone che il nostro metodo ha classificato più in alto di qualsiasi persona ragionevole, in particolare il Presidente George W. Bush (36) e la star televisiva adolescente Hilary Duff (1626). Si potrebbe guardare a questi bugiardi e liquidare l'intera faccenda. Ma si deve comprendere che abbiamo classificato quasi 850.000 personaggi storici, all'incirca la popolazione di San Francisco. Alcuni cattivi esempi selezionati devono essere inseriti nel giusto contesto.

- *Vincoli di piccionaia*: La maggior parte dei recensori ha visto solo le classifiche dei nostri primi 100 personaggi, e si sono lamentati di dove abbiamo collocato le persone e di chi non ha fatto il taglio. Il programma televisivo femminile *The View* ha commentato che non avevamo abbastanza donne. Ricordo articoli britannici che si lamentavano del fatto che avevamo Winston Churchill (37) classificato troppo in basso, articoli sudafricani che ritenevano che non avessimo Nelson Mandela (356),

articoli cinesi che dicevano che non avevamo abbastanza cinesi e persino una rivista cilena che si lamentava dell'assenza di cileni.
In parte questo riflette le differenze culturali. Questi critici avevano una nozione implicita di importanza diversa da quella riflessa dalla Wikipedia inglese. Ma in gran parte riflette il fatto che ci sono esattamente cento posizioni nella top 100. Molte delle figure che consideravano mancanti erano solo leggermente al di fuori dell'orizzonte visibile. Per ogni nuova persona che abbiamo inserito nella top 100, abbiamo dovuto eliminare qualcun altro. Ma i lettori non hanno quasi mai suggerito nomi che dovevano essere omessi, ma solo quelli che dovevano essere aggiunti.

Qual è la morale? Cerchi di anticipare le preoccupazioni del pubblico per le sue classifiche. Siamo stati incoraggiati a chiamare esplicitamente la nostra misura "*forza del meme"* invece di "*significato"*. In retrospettiva, l'utilizzo di questo nome meno carico avrebbe permesso ai nostri lettori di apprezzare meglio ciò che stavamo facendo. Probabilmente avremmo anche dovuto scoraggiare i lettori dall'agganciarsi alle nostre classifiche top 100 e concentrarci invece sugli ordini relativi all'interno dei gruppi di interesse: chi erano i migliori musicisti, scienziati e artisti? Questo avrebbe potuto risultare meno controverso, aiutando meglio le persone a creare fiducia in ciò che stavamo facendo.

Capitolo 5

Analisi statistica

Confesso che non ho mai avuto una conversazione veramente soddisfacente con uno statistico. Questo non è del tutto dovuto alla mancanza di tentativi. Diverse volte, nel corso degli anni, ho portato problemi di interesse agli statistici, ma ho sempre ricevuto risposte del tipo: "Non si può fare così" o "Ma non è indipendente", invece di sentirmi dire: "Ecco il modo in cui può gestirlo". Ad essere onesti, anche questi statistici in genere non hanno gradito parlare con me. Gli statistici pensano seriamente ai dati da molto più tempo degli e hanno molti metodi e idee potenti da mostrare. In questo capitolo,

introdurrò alcuni di questi strumenti importanti, come le definizioni di alcune distribuzioni fondamentali e i test di significatività statistica. Questo capitolo introdurrà anche l'analisi baysiana, un modo per valutare in modo rigoroso come i nuovi dati dovrebbero influenzare le nostre precedenti stime di eventi futuri. Il processo di ragionamento statistico. C'è una popolazione non ancora definita di cose possibili che possiamo potenzialmente osservare. Solo un sottoinsieme relativamente piccolo di essi viene effettivamente campionato, idealmente in modo casuale, il che significa che possiamo osservare le proprietà degli elementi campionati. La teoria della probabilità descrive quali proprietà dovrebbe avere il nostro campione, date le proprietà della popolazione sottostante. Ma l'*inferenza statistica* funziona in modo opposto: cerchiamo di dedurre le caratteristiche dell'intera popolazione in base all'analisi del campione. Idealmente, impareremo a pensare come un esperto di statistica: abbastanza da rimanere vigili e da evitare la sovrainterpretazione e l'errore, pur mantenendo la fiducia di poter giocare con i dati e portarli dove ci portano.

5.1 Distribuzioni statistiche

Ogni variabile che osserviamo definisce una particolare distribuzione di frequenza, che riflette la frequenza con cui si presenta un determinato valore. Le proprietà uniche di variabili come l'altezza, il peso e il QI sono catturate dalle loro distribuzioni. Ma le forme di queste distribuzioni non sono uniche: in gran , la ricca varietà di dati del mondo appare solo in un piccolo numero di forme classiche. Queste distribuzioni classiche hanno due belle proprietà: (1) descrivono le forme delle distribuzioni di frequenza che si presentano spesso nella pratica e (2) spesso possono essere descritte matematicamente utilizzando espressioni in forma chiusa con pochissimi parametri. Una volta astratte dalle osservazioni di dati specifici, diventano *distribuzioni di probabilità*, degne di uno studio indipendente. La familiarità con le distribuzioni di probabilità classiche è importante. Si presentano spesso nella pratica, per cui bisogna stare attenti a queste distribuzioni. forniscono un vocabolario per parlare dell'aspetto dei nostri dati. Rivedremo le distribuzioni statistiche più importanti (binomiale, normale, di Poisson e legge di potenza) nelle sezioni seguenti, sottolineando le proprietà che definiscono il loro carattere essenziale. Si noti che i dati osservati non derivano necessariamente da una particolare distribuzione teorica solo perché la sua forma è simile. I test statistici possono essere utilizzati per dimostrare in modo rigoroso se i dati osservati sperimentalmente riflettono campioni tratti da una particolare distribuzione.

Ma le risparmierò la fatica di eseguire effettivamente uno di questi test. Affermerò con grande sicurezza che i suoi dati reali *non* si adattano precisamente a nessuna delle famose distribuzioni teoriche.
Perché? Comprenda che il mondo è un luogo complicato, il che rende la misurazione un processo disordinato. Le sue osservazioni saranno probabilmente tratte da più popolazioni campione, ognuna delle quali ha una distribuzione di fondo un po' diversa. In genere, alle code di qualsiasi distribuzione osservata accade qualcosa di strano: un'improvvisa esplosione di valori insolitamente alti o bassi. Le misurazioni avranno degli errori associati, a volte in modo strano e sistematico. Ma detto questo, la comprensione delle distribuzioni di base è davvero molto importante. Ogni distribuzione classica è classica per un motivo. La comprensione di questi motivi ci dice molto sui dati osservati, per cui saranno esaminati in questa sede.

5.1.1 La distribuzione binomiale

Consideri un esperimento costituito da prove identiche e indipendenti che hanno due possibili esiti P_1 e P_2, con le rispettive probabilità di p e q= (1 p). Forse il suo esperimento consiste nel lanciare delle monete giuste, dove la probabilità di avere testa (p= 0,5) è uguale a quella di avere croce (q= 0,5). Forse si tratta di accendere ripetutamente un interruttore della luce, dove la probabilità di scoprire improvvisamente che deve cambiare la lampadina (p= 0,001) è molto inferiore a quella di vedere la luce (q = 0,999). La *distribuzione binomiale* riporta la probabilità di ottenere esattamente x eventi P_1 nel corso di n prove indipendenti, in nessun ordine particolare. L'indipendenza è importante in questo caso: stiamo assumendo che la probabilità di guasto di una lampadina non abbia alcuna relazione con il numero di volte in cui è stata utilizzata in precedenza. Il pdf della distribuzione binomiale. Ci sono diverse cose da osservare sulla distribuzione binomiale: *È discreta*: entrambi gli argomenti della distribuzione binomiale (n e x) devono essere numeri interi. La scorrevolezza della Figura 5.2 (a sinistra) è un'illusione, perché n= 200 è abbastanza grande. Non è possibile ottenere 101,25 teste in 200 lanci di moneta. *Probabilmente è in grado di spiegare la teoria che ne sta alla base*: Ha incontrato per la prima volta la distribuzione binomiale superiori. Ricorda il triangolo di Pascal? La conclusione con esattamente x teste in n lanci in una particolare sequenza si verifica con la probabilità $p^x(1-p)^{(n-x)}$,

per ciascuna delle ^{n} sequenze di lanci distinte. *È una specie di campana*: Per una moneta giusta (p= 0,5), la distribuzione binomiale è perfettamente simmetrica, con la media al centro. Questo non è vero nel caso della lampadina: se accendiamo la lampadina n= 1000 volte, il numero più probabile di guasti sarà zero. Detto questo, con l'aumentare di n otterremo una distribuzione simmetrica con un picco nella media. *Viene definito utilizzando solo due parametri*: Tutto ciò di cui abbiamo bisogno sono i valori di p e di n per definire completamente una determinata distribuzione binomiale. Molte cose possono essere ragionevolmente modellate dalla distribuzione binomiale. Ricordiamo la varianza nelle prestazioni di ttito. un bare p= 0,300. In quel caso, la proIn quel caso, la probabilità di ottenere un colpo ad ogni prova era p= 0,3, con n= 500 prove per stagione. Pertanto, il numero di battitori per stagione viene estratto da una distribuzione binomiale. Rendersi conto che si trattava di una distribuzione binomiale significava che non era necessario ricorrere alla simulazione per costruire la distribuzione. Proprietà come il valore atteso numero o$\underline{\text{f hits } \mu = np}$= 500× 0,3= 150 e la sua deviazione standard semplicemente si tratta di formule chiuse che che può consultare in caso di necessità.

5.1.2 La distribuzione normale

Molti fenomeni naturali sono modellati da curve a campana. Le caratteristiche misurate come l'altezza, il peso, la durata della vita e il QI si adattano tutte allo stesso schema di base: la maggior parte dei valori si trova abbastanza vicino alla media, la distribuzione è simmetrica e nessun valore è troppo estremo. Nell'intera storia del mondo, non c'è mai stato un uomo alto ^{3} metro e 80 o una donna di 140 anni. La madre di tutte le curve a campana è la *distribuzione gaussiana* o *normale*, che è completamente parametrizzata dalla sua media e dalla deviazione standard *È continua*: gli argomenti della distribuzione normale (media μ e deviazione standard σ) sono liberi di essere numeri reali arbitrari, con il solo vincolo che $\sigma >$ 0. *Probabilmente non sa spiegare da dove deriva*: la distribuzione normale è una generalizzazione della distribuzione binomiale, dove $n \to \infty$ e il grado di concentrazione intorno alla media è specificato dal parametro σ. Prenda spunto dalla distribuzione binomiale e confidi che Gauss abbia fatto bene i suoi calcoli: il grande matematico elaborò la distribuzione normale per la sua tesi di dottorato. Oppure consulti un qualsiasi libro di statistica decente, se è davvero curioso di vedere da dove proviene. *È davvero a forma di campana*: La distribuzione gaussiana è l'esempio platonico di una curva a campana. Poiché opera su una

variabile continua (come l'altezza) invece che su un conteggio discreto (ad esempio, il numero di eventi), è perfettamente regolare. Poiché va all'infinito in entrambe le direzioni, non c'è troncamento delle code alle due estremità. La distribuzione normale è un costrutto teorico che aiuta a spiegare questa perfezione. *Inoltre, viene definita utilizzando solo due parametri*: Tuttavia, si tratta di parametri diversi rispetto alla distribuzione binomiale! La distribuzione normale è completamente definita dal suo punto centrale (dato dalla media μ) e dalla sua diffusione (data dalla deviazione standard σ). Sono le uniche manopole che possiamo utilizzare per modificare la distribuzione.

Cosa è normale?

Un numero incredibile di fenomeni che si verificano in natura sono modellati dalla distribuzione normale. Forse il più importante è l'errore di misurazione. Ogni volta che misura il suo peso su una bilancia da bagno, otterrà una risposta in qualche modo diversa, anche se il suo peso non è cambiato. A volte la bilancia leggerà un valore alto e altre volte un valore basso, a seconda della temperatura della stanza e deformazione del pavimento. Gli errori piccoli sono più probabili di quelli grandi, e un valore leggermente alto è altrettanto probabile di un valore leggermente basso. L'errore sperimentale è generalmente distribuito in modo normale come un *rumore gaussiano*. Fenomeni fisici come l'altezza, il peso e la durata della vita hanno tutti distribuzioni a campana, con argomenti simili. Tuttavia, l'affermazione che tali distribuzioni sono *normali* viene solitamente fatta con troppa disinvoltura, senza specificare con precisione la popolazione sottostante. L'altezza umana è distribuita in modo normale? Certamente no: uomini e donne hanno altezze medie diverse e distribuzioni associate. L'altezza maschile è distribuita normalmente? Certamente no: includendo i bambini nel mix e riducendo gli anziani, si ottiene nuovamente la somma di diverse distribuzioni sottostanti. L'altezza dei maschi adulti negli Stati Uniti è normale? No, probabilmente nemmeno in questo caso. Esistono popolazioni non banali con disturbi della crescita. come il nanismo e l'acromegalia, che lasciano gruppi di persone sostanzialmente più basse e più alte di quanto possa essere spiegato dalla distribuzione normale. Forse la più famosa distribuzione a campana ma non normale è quella dei rendimenti giornalieri (movimenti percentuali dei prezzi) nei mercati finanziari. Un grande crollo del mercato è definito da un grande calo percentuale dei prezzi: il 10 ottobre 1987, la media Dow Jones perse il 22,61% del suo valore. I grandi crolli del mercato azionario si verificano con una frequenza molto maggiore di quella che può essere modellata in modo accurato dalla distribuzione normale. In effetti, ogni crollo sostanziale del

mercato spazza via un certo numero quants che avevano ipotizzato la normalità e si erano assicurati in modo inadeguato contro questi eventi estremi. Si scopre che il *logaritmo* dei rendimenti azionari si rivela normalmente distribuito, risultando in una distribuzione con code molto più grasse della norma. Anche se dobbiamo ricordare che distribuzioni a campana non sono sempre normali, fare questa ipotesi è un modo ragionevole per iniziare a pensare in assenza di conoscenze migliori.

5.1.3 Implicazioni della Distribuzione Normale

Ricordiamo che la media e la deviazione standard insieme caratterizzano sempre in modo approssimativo qualsiasi distribuzione di frequenza, come discusso nella Sezione 2.2.4. Ma fanno un lavoro spettacolare nel caratterizzare la distribuzione normale, perché *definiscono* la distribuzione normale. Regola del 68%-95%-99% della distribuzione normale. Il 68% della massa di probabilità deve trovarsi nella regione 1σ della media. Inoltre, il ±95% della probabilità si trova entro 2σ e il 99,7% entro 3σ. Ciò significa che i valori lontani dalla media (in termini di σ) sono estremamente rari in qualsiasi variabile distribuita normalmente. Infatti, il termine *sei sigma* viene utilizzato per Coniuga standard di qualità così elevati che i difetti sono eventi incredibilmente rari. Vogliamo che gli incidenti aerei siano eventi Six Sigma. La probabilità di un evento 6σ distribuzione normale è di circa 2 parti per miliardo.

L'intelligenza misurata dal QI è distribuita in modo normale, con una media di 100 e una deviazione standard $\sigma = 15$. Quindi il 95% della popolazione si trova all'interno di 2σ della media, da 70 a 130. Pertanto, il 95% della popolazione si trova entro 2σ della media, da 70 a 130. Rimane solo il 2,5% delle persone con un QI superiore a 130 e un altro 2,5% inferiore a 70. Un totale del 99,7% della massa si trova entro 3σ media, ossia le persone con un QI compreso tra 55 e 145. Quanto è intelligente la persona più intelligente del mondo? Se ipotizziamo una popolazione di 7 miliardi di persone, la probabilità che una persona selezionata a caso sia la più intelligente è di circa $1{,}43\ 10^{-10}$. Questa è circa la stessa probabilità che un singolo campione si × discosti di più di $6{,}5\sigma$ dalla media. Quindi la persona più intelligente del mondo dovrebbe avere un QI di circa 197,5, secondo questo calcolo.

Il grado di accettazione dipende dalla convinzione che il QI sia davvero distribuito normalmente. Tali modelli sono di solito in grave pericolo di rottura agli estremi. In effetti, secondo questo modello c'è quasi la stessa

probabilità che ci sia qualcuno abbastanza stupido da ottenere un punteggio negativo in un test del QI.

5.1.4 Distribuzione di Poisson

La *distribuzione di Poisson* misura la frequenza degli intervalli tra eventi rari. Supponiamo di modellare la durata della vita umana con una sequenza di eventi quotidiani, dove esiste una probabilità piccola ma costante 1 p che oggi si smetta di respirare. Una durata di vita di esattamente n giorni significa respirare con successo per ciascuno dei primi n-1 giorni e poi interrompere per sempre lo schema all'*ennesimo* giorno. La probabilità di vivere esattamente n giorni è data -da $Pr(n)= p^{n-1}(1 \quad p)$, ottenendo una durata di vita attesa. La distribuzione di Poisson deriva fondamentalmente da questa analisi, ma adotta un argomento più conveniente di p. Si basa invece su μ, il valore medio della distribuzione. Poiché ogni p definisce un particolare valore di μ, questi parametri sono in un certo senso equivalenti, ma la media è molto più facile da stimare o misurare. La distribuzione di Poisson produce una forma chiusa molto semplice. Una volta che si inizia a pensare nel modo giusto, molte distribuzioni iniziano a sembrare di Poisson, perché rappresentano intervalli tra eventi rari. Ricordiamo il modello di lampadina a distribuzione binomiale della sezione precedente. In questo modo è stato facile calcolare il numero previsto di cambiamenti, ma non la distribuzione della durata della vita, che è di Poisson. Quasi tutte le lampadine hanno una durata compresa tra 900 e 1100 ore prima di spegnersi. In alternativa, supponiamo di modellare il numero di figli con un processo in cui la famiglia continua ad avere figli fino a quando, dopo un numero eccessivo di capricci, vendite di torte o carichi di biancheria, un genitore finalmente cede. *"! ho abbastanza. Basta!"*.

In base a tale modello, le dimensioni della famiglia dovrebbero essere modellate come una distribuzione di Poisson, dove ogni giorno c'è una probabilità piccola ma non nulla di un guasto che comporta la chiusura della fabbrica. Quanto funziona il modello "Ho avuto" per prevedere le dimensioni della famiglia? La linea poligonale rappresenta la distribuzione di Poisson con il parametro λ= 2,2, il che significa che le famiglie hanno una media di 2,2 figli. I punti rappresentano la frazione di famiglie con k figli, ricavata dalla classifica U.S. General Social 2010. Sondaggio (GSS). C'è un eccellente accordo su tutte le dimensioni della famiglia, tranne k= 1, e francamente, la mia esperienza personale suggerisce che ci sono più bambini singleton di quanto questa serie di dati rappresenti. Insieme, conoscere solo

la media e la formula della distribuzione di Poisson ci permette di costruire una stima ragionevole della reale distribuzione delle dimensioni della famiglia.

5.1.5 Distribuzioni a legge di potenza

Molte distribuzioni di dati presentano code molto più lunghe di quelle che potrebbero essere possibili con le distribuzioni normali o di Poisson. Consideriamo, ad esempio, la popolazione delle città. Nel 2014 c'erano esattamente 297 città statunitensi con una popolazione superiore a 100.000 persone, secondo Wikipedia. La popolazione *della* kesima città più grande, per 1 k 297. Mostra che un numero≤ relativamente ≤ piccolo di città ha una popolazione che domina selvaggiamente il resto. In effetti, le diciassette città più grandi hanno popolazioni così grandi che sono state ritagliate da questo grafico per poter vedere il resto. Queste città hanno una popolazione media di 304.689, con una deviazione standard spaventosa di 599.816. C'è qualcosa che non va quando la deviazione standard è così grande rispetto alla media. In una distribuzione normale, il 99,7% della massa si trova entro 3σ dalla media, rendendo così improbabile che una qualsiasi di queste città abbia una popolazione superiore a 2,1 milioni di persone. Eppure Houston ha una popolazione di 2,2 milioni di persone, e New York (con 8,4 milioni di persone) è più di 13σ sopra la media! Le popolazioni delle città *non sono* chiaramente distribuite in modo normale. Infatti, osservano una distribuzione diversa, chiamata *legge di potenza*. Per una data variabile X definita da una distribuzione a legge di potenza. Questo è parametrizzato da due costanti: l'esponente α e la costante di normalizzazione c.

Le distribuzioni a legge di potenza richiedono una certa riflessione per essere analizzate correttamente. La probabilità totale definita da questa distribuzione è l'area sotto la curva. Il valore particolare di A è definito dai parametri α e c. La costante di normalizzazione c è scelta specificamente per un dato α, per assicurarsi che $A = 1$, come richiesto dalle leggi della probabilità. A parte questo, c non è di particolare importanza per noi. L'azione reale avviene con α. Si noti che quando raddoppiamo il valore dell'ingresso (da x a $2x$), diminuiamo la probabilità di un fattore di $f = 2^{-\alpha}$. Questo sembra negativo, ma per qualsiasi dato α è solo una costante. Quindi, ciò che la legge di potenza dice in realtà è che la probabilità di un evento *di dimensioni* 2x è 2^{α} volte meno frequente di un evento *di dimensioni* x, per tutte le x.La ricchezza personale è ben modellata da una legge di potenza, in≈ cui f 0,2= $1/5$. Ciò significa che in un ampio intervallo, se Z persone hanno x dollari, Z/5 persone hanno $2x$ dollari. Un quinto delle persone

ha 200.000 dollari rispetto a 100.000 dollari. Se ci sono 625 persone al mondo che valgono 5 miliardi di dollarici dovrebbero essere circa 125 multimiliardari che valgono ciascuno 10 miliardi di dollari. Inoltre, dovrebbero esserci 25 super-miliardari da 20 miliardi di dollari ciascuno, cinque iper-miliardari da 40 miliardi di dollari e infine un solo Bill Gates da 80 miliardi di dollari. Le leggi di potere definiscono le regole dell'"80/20" che spiegano tutta l'ineguaglianza del nostro mondo: l'osservazione che il 20% superiore dell'A ottiene l'80% dei guadagni. Le leggi di potere tendono a sorgere quando i ricchi diventano più ricchi, dove c'è una probabilità crescente di ottenere di più in base a ciò che si possiede già. Le grandi città crescono in modo sproporzionato perché più persone sono attratte dalle città quando sono grandi. Grazie alla sua ricchezza, Bill Gates ha accesso a opportunità di investimento molto migliori rispetto me, quindi il suo denaro cresce più velocemente del mio. Molte distribuzioni sono definite da tali modelli di crescita preferenziale o di attaccamento, tra cui:

- *Siti Internet con x utenti*: I siti web diventano più popolari perché hanno più utenti. Y Lei è più probabile che si iscriva a Instagram o Facebook perché i suoi amici si sono già iscritti a Instagram o Facebook. L'adesione preferenziale porta ad una distribuzione a legge di potenza.

- *Parole utilizzate con una frequenza relativa di x*: Esiste una lunga coda di milioni di parole come *algorist* o *defenestrate*[4] che sono raramente utilizzate nella lingua inglese. linguaggio. D'altra parte, un piccolo gruppo di parole come *"the"* viene utilizzato con una frequenza molto maggiore rispetto al resto.

La legge di Zipf regola la distribuzione dell'uso delle parole nelle lingue naturali e afferma che la $kesima$ parola più popolare (misurata in base al rango di frequenza) è usata solo 1/kesimo della parola più popolare. Per valutare l'efficacia questa legge, consideri i ranghi delle parole basati sulla frequenza, tratti Wikipedia inglese. Dovrebbe essere convincente che la frequenza d'uso diminuisce rapidamente con il rango: ricordiamo che *grandmom* è solo una forma gergale di *grandma*, non il vero McCoy. Perché si tratta di una legge di potenza? Una parola di rango $2x$ ha una frequenza di $F_{2x} \sim F(1)/2x$, rispetto a $F_x \sim F_1/x$. Quindi, dimezzando il rango si raddoppia la frequenza, e ciò corrisponde alla legge di potenza con $\alpha = 1$. Qual è il meccanismo alla base dell'evoluzione delle lingue che porta a questa distribuzione? Una spiegazione plausibile è che le persone imparano e usano le parole perché le sentono usare da altre persone.

Qualsiasi meccanismo che favorisca le parole già popolari porta a una legge di potenza.

- *Frequenza dei terremoti di magnitudo* x: La scala Richter per misurare la forza dei terremoti è logaritmica, il che significa che un terremoto di 5,3 è dieci volte più forte di un evento di scala 4,3. Aggiungendo uno alla magnitudo, la forza viene moltiplicata per un fattore di dieci.

Con una scala in rapida crescita, è logico che gli eventi più grandi siano più rari di più piccoli. Io provoco un terremoto di magnitudo 0,02 ogni volta che tiro lo sciacquone del bagno. In effetti ci sono miliardi di eventi di questo tipo ogni giorno, ma i terremoti più grandi diventano sempre più rari con le dimensioni. Quando una quantità cresce in modo potenzialmente illimitato, ma la probabilità che si verifichi diminuisce in modo esponenziale, si ottiene una legge di potenza. I dati dimostrano che questo è vero per l'energia rilasciata dai terremoti, così come lo è per le vittime delle guerre: fortunatamente il numero di conflitti che uccidono x persone diminuisce come una legge di potenza. Impari a tenere gli occhi aperti per le distribuzioni della legge di potenza. Le troverà ovunque nel nostro mondo ingiusto. Sono rivelate dalle seguenti proprietà:

- *Le leggi di potenza si presentano come linee rette sui grafici del valore logico e della frequenza logici*: Guardi il grafico della popolazione delle città. Sebbene ci siano degli spazi vuoti ai margini, dove i dati scarseggiano, in linea di massima i punti si trovano ordinatamente su una linea. Questa è *la* caratteristica principale di una legge di potenza. A proposito, la pendenza di questa linea è determinata da α, la costante che definisce la forma della distribuzione della legge di potenza.

- *La media non ha senso*: Bill Gates da solo aggiunge circa 250 dollari ricchezza della persona media negli Stati Uniti. Questo è strano. In una distribuzione a legge di potenza, c'è una probabilità molto piccola ma non nulla che qualcuno abbia una ricchezza infinita, quindi cosa comporta questo per la media? La mediana fa un lavoro molto migliore per catturare la maggior parte di queste distribuzioni rispetto alla media osservata.

- *La deviazione standard non ha senso*: in una distribuzione a legge di potenza, la deviazione standard è tipicamente grande o più grande della media. Ciò significa che la distribuzione è molto poco caratterizzata da μ e σ, mentre la legge di potenza fornisce un'ottima descrizione in termini di α e c.
- *La distribuzione è invariante di scala*: Supponiamo di tracciare le popolazioni delle città statunitensi dalla 300esima alla 600esima più grande, invece delle prime 300 come nella Figura 5.7 (a sinistra). La forma sarebbe molto simile, con la popolazione della 300esima città più grande che sovrasta la coda. Qualsiasi funzione esponenziale è *invariante di scala*, perché ha lo stesso aspetto a qualsiasi risoluzione. Questa è una conseguenza del fatto che è una linea retta su un grafico log-log: qualsiasi sottogamma è un segmento di linea retta, che ha gli stessi parametri nella sua finestra della distribuzione completa. Attenzione alle distribuzioni a legge di potenza. Esse riflettono le disuguaglianze del mondo, il che significa che sono ovunque.

5.2 Campionamento da distribuzioni

Il campionamento di punti da una determinata distribuzione di probabilità è un'operazione comune, che è utile sapere come fare. Forse ha bisogno di dati di prova da una distribuzione a legge di potenza per eseguire una simulazione, o per verificare che il suo programma funzioni in condizioni estreme. Per verificare se i dati si adattano effettivamente a una particolare distribuzione, è necessario un confronto con i dati, che in genere dovrebbero essere dati sintetici generati in modo appropriato e tratti dalla distribuzione canonica. Esiste una tecnica generale per il campionamento da qualsiasi distribuzione di probabilità, chiamata *campionamento con trasformazione inversa*.

Ricordiamo che possiamo spostarci tra la funzione di densità di probabilità P e la funzione di densità cumulativa C mediante integrazione e differenziazione. Possiamo muoverci avanti e indietro tra di esse perché: Supponiamo di voler campionare un punto da questa distribuzione forse molto complicata. Posso utilizzare un generatore di numeri casuali uniformi per selezionare un valore p nell'intervallo $[0, \ldots, 1]$. Possiamo interpretare p come una probabilità e utilizzarlo come indice sulla distribuzione cumulativa C. Precisamente, riportiamo il valore esatto di x tale che $C(X \leq x) = p$. La Figura 5.8 illustra l'approccio, in questo caso campionando dalla distribuzione normale. Supponiamo che $p = 0{,}729$ sia il numero casuale selezionato dal

nostro generatore uniforme. Restituiamo il valore x tale che $y = 0{,}729$, quindi $x = 0{,}62$ come da questa cdf.

Se sta lavorando con una distribuzione di probabilità popolare in un linguaggio ben supportato come Python, quasi certamente è già disponibile una funzione di libreria per generare campioni casuali. Quindi cerchi la libreria giusta prima di scrivere la sua.

5.2.1 Campionamento casuale oltre una dimensione

Il campionamento corretto da una determinata distribuzione diventa un problema molto delicato quando si aumenta il numero di dimensioni. Consideriamo il compito di campionare i punti in modo uniforme all'interno di un cerchio. Rifletta un attimo su come potrebbe farlo prima di procedere.

I più intelligenti potrebbero avere l'idea di campionare l'angolo e la distanza dal centro in modo indipendente. L'angolo che qualsiasi punto campionato deve formare rispetto all'origine e all'*asse x* positivo varia tra 0 e 2π. La distanza dall'origine deve essere un valore compreso tra 0 e r. Selezionando queste coordinate in modo uniforme e casuale, si ottiene un punto casuale nel cerchio. Questo metodo è intelligente, ma sbagliato. Certo, qualsiasi punto così creato deve trovarsi all'interno del cerchio. Ma i punti non vengono selezionati con una frequenza uniforme. Questo metodo genererà punti in cui la metà di essi si troverà a una distanza massima di $r/2$ dal centro. Ma la maggior parte dell'area del cerchio è molto più lontana dal centro! Pertanto, sovracampioneremo vicino all'origine, a scapito della massa vicino al confine. Questo è dimostrato da un grafico di 10.000 punti generato con questo metodo. Una tecnica stupida che si rivela corretta è il *campionamento Monte Carlo*. Le coordinate x e y di ogni punto del cerchio vanno da r a r, così- come molti punti al di fuori del cerchio. Quindi, campionando questi valori in modo uniforme e casuale, si ottiene un punto che si trova in una casella di delimitazione del cerchio, ma non sempre all'interno del cerchio stesso. Questo può essere facilmente verificato: la distanza da (x, y) all'origine è al massimo r, cioè $\sqrt{x^2+y^2} \leq r$? Se sì, abbiamo trovato un punto casuale nel cerchio. In caso, lo scartiamo e riproviamo. La Figura 5.9 (a destra) illustra 10.000 punti costruiti con questo metodo: si veda come coprono uniformemente il cerchio, senza evidenti punti di sovracampionamento o

sottocampionamento. L'efficienza in questo caso dipende interamente dal rapporto tra il volume della regione desiderata (l'area del cerchio) e il volume del rettangolo di selezione l'area di un quadrato). Poiché il 78,5% di questo riquadro delimitato è occupato dal cerchio, in media sono sufficienti meno di due tentativi per trovare ogni nuovo punto del cerchio.

5.3 Significatività statistica

Gli statistici si preoccupano soprattutto di stabilire se le osservazioni sui dati sono significative. L'analisi computazionale troverà facilmente una serie di modelli e correlazioni in qualsiasi serie di dati interessanti. Ma una particolare correlazione riflette un fenomeno reale, piuttosto che un semplice caso? In altre parole, quando un'osservazione è davvero *significativa*?

Le correlazioni sufficientemente forti su grandi serie di dati possono sembrare "ovviamente" significative, ma i problemi sono spesso molto sottili. Per prima, la *correlazione non implica la causalità.* La Figura 5.10 dimostra in modo convincente che il volume studi avanzati in informatica è correlato alla quantità di videogiochi. Mi piace pensare di aver spinto più persone verso gli algoritmi rispetto a Nintendo, ma forse si tratta della stessa cosa? I grafici di queste correlazioni spurie riempiono letteralmente un libro, per di più molto divertente. La disciplina della statistica si rivela utile per fare sottili distinzioni sulla significatività o meno di un'osservazione. L'esempio classico viene dalla statistica medica, nel determinare l'efficacia dei trattamenti farmacologici. Un'azienda farmaceutica conduce un esperimento di confronto tra due farmaci. Il farmaco A ha curato 19 pazienti su 34. Il farmaco B ha curato 14 pazienti su 21. Il farmaco B è davvero migliore del farmaco A? L'approvazione di nuovi farmaci da parte della FDA può aggiungere o sottrarre miliardi al valore delle aziende farmaceutiche. Ma si può essere sicuri che un nuovo farmaco rappresenti un reale miglioramento? Come si fa a capirlo?

5.3.1 Il significato del significato

La significatività statistica misura la nostra fiducia che esista una differenza reale tra due distribuzioni date. Questo è importante. Ma la significatività statistica non misura l'importanza o la grandezza di questa differenza. Per le grandi con dimensioni del campione sufficienti, differenze estremamente piccole possono essere registrate come altamente significative nei test statistici.

Per esempio, supponiamo che io venga convinto a scommettere croce su una moneta che esce testa il 51% delle volte, invece del 50% che associamo a una moneta giusta. Dopo 100 lanci di una moneta equa, mi aspetterei di vedere il

51% o più di teste il 46,02% delle volte, quindi non ho assolutamente motivo di lamentarmi quando succede. Dopo 1.000 lanci, la probabilità di vedere almeno 510 teste scende a 0,274. Dopo 10.000 lanci, la probabilità di vedere così tante teste è solo 0,0233, e dovrei iniziare a sospettare che la moneta sia equa. Dopo 100.000 lanci, la probabilità di correttezza scenderà a $1{,}29 \times 10^{-10}$, così piccola che dovrò presentare un reclamo formale, anche se ritengo che il mio avversario sia un gentiluomo.

Ma ecco il punto. Anche se ora è chiarissimo che× sono stato ingannato con una moneta di parte, le conseguenze di questo atto non sono sostanziali. Per quasi tutte le questioni della vita che vale la pena di girare, sarei disposto a prendere il lato corto della moneta, perché la posta in gioco non è abbastanza alta. Con una scommessa di 1 dollaro a lancio, la mia perdita attesa, anche dopo 100.000 lanci, sarebbe di soli 1.000 dollari.

La significatività indica quanto è improbabile che qualcosa sia dovuto al caso, ma non se è importante. A noi interessa davvero la *dimensione dell'effetto*, l'entità della differenza tra i due gruppi. In via informale, classifichiamo una dimensione dell'effetto *di livello medio* come visibile a occhio nudo da un osservatore attento. Su questa scala, gli effetti *grandi* spiccano e gli effetti *piccoli* non sono del tutto banali. Esistono diverse statistiche che cercano di misurare la dimensione dell'effetto, tra cui:

- *D di Cohen*: L'importanza della differenza tra due medie μ e μ' dipende dalla grandezza assoluta della variazione, ma anche dalla variazione naturale delle distribuzioni, misurata da σ o σ'. Questa dimensione dell'effetto può essere misurata da: $d = (|\mu - \mu'|)/\sigma$.

  Una soglia ragionevole per una dimensione d'effetto piccola è> 0,2, un effetto medio> 0,5, e una dimensione d'effetto grande > 0,8.

- *Coefficiente di correlazione di Pearson r*: Misura il grado di relazione lineare tra due variabili, su una scala che va da− 1 a 1. Le soglie per

Le dimensioni dell'effetto sono paragonabili allo spostamento medio: i piccoli effetti iniziano a± 0,2, effetti medi circa± 0,5, e grandi dimensioni di effetto richiedono correlazioni di ±0.8.

- *Il coefficiente di variazione r*: il ^{2}quadrato del coefficiente di correlazione riflette la proporzione della varianza di una variabile che viene spiegata dall'altra. Le soglie derivano dalla quadratura di quelle precedenti. Gli

effetti piccoli spiegano almeno il 4% della varianza, gli effetti medi il 25% e gli effetti grandi almeno il 64%.

≥

- *Percentuale di sovrapposizione*: L'area sotto qualsiasi singola distribuzione di probabilità è, per definizione, 1. L'area d'intersezione tra due distribuzioni date è una buona misura della loro somiglianza, come illustrato nella Figura 5.11. Le distribuzioni identiche si sovrappongono al 100%, mentre gli intervalli disgiunti si sovrappongono allo 0%. Le soglie ragionevoli sono: per gli effetti piccoli 53% di sovrapposizione, per gli effetti medi 67% di sovrapposizione e per gli effetti grandi 85% di sovrapposizione.

Naturalmente, qualsiasi effetto considerevole che non sia statisticamente significativo è intrinsecamente sospetto. La correlazione studio CS vs. gioco al videogioco nella Figura 5.10 era così alta (r= 0,985) che la dimensione dell'effetto sarebbe stata enorme, se il numero di punti campione e la metodologia fossero stati sufficientemente solidi da sostenere la conclusione. *La* significatività statistica dipende dal numero di , mentre la dimensione dell'effetto no.

5.3.2 Il test T: Confronto delle medie della popolazione

Abbiamo visto che grandi spostamenti medi tra due popolazioni suggeriscono grandi dimensioni di effetto. Ma di quante misurazioni abbiamo bisogno prima di poter credere con sicurezza che il fenomeno sia reale. Supponiamo di misurare il QI di venti uomini e venti donne. I dati mostrano che un gruppo è più intelligente, in media? Certamente le medie dei campioni differiranno, almeno un po', ma questa differenza è significativa?
Il test t valuta se le medie della popolazione di due campioni sono diverse. Questo problema si presenta comunemente nei *test AB*, associati alla valutazione di se un cambiamento di prodotto fa la differenza nelle prestazioni. Supponiamo di mostrare a un gruppo di utenti la versione A e a un altro gruppo la versione B. Inoltre, supponiamo di misurare un valore di performance del sistema per ogni utente, come il numero di volte in cui clicca sugli annunci o il numero di stelle che gli assegna quando gli viene chiesto dell'esperienza. Il test t misura se la differenza osservata tra i due gruppi è significativa. Due mezzi differiscono significativamente se:

- *La differenza media è relativamente grande*: Questo ha senso. Si può concludere che gli uomini pesano di più delle donne in media abbastanza facilmente, perché la dimensione dell'effetto è così grande. Secondo il Center for Disease Control[2], l'uomo medio americano pesava 195,5 chili nel 2010, mentre la donna media americana pesava 166,2 chili. Si tratta di un dato enorme. Dimostrare che una differenza molto più sottile, come il QI, è reale, richiede molte più prove per essere altrettanto convincente.

- *Le deviazioni standard sono abbastanza piccole*: Anche questo ha senso. È facile convincersi che uomini e donne hanno, in media, lo stesso numero di dita, perché i conteggi che osserviamo sono molto stretti intorno alla media: { 10, 10, 10, 10, 9, 10 } L'ipotesi del conteggio delle dita uguali...
  L'analisi richiederebbe molte più prove se i numeri saltassero molto. Sarei riluttante a impegnarmi per una vera media distributiva di $\mu = 10$ se quello che ho osservato fosse *{3*, 15, 6, 14, 17, *5}*.

- *Il numero di campioni è sufficientemente grande*: Anche questo ha senso. Più dati vedo, più mi convinco che il campione rappresenterà accuratamente la distribuzione sottostante. Ad esempio, è indubbio che gli uomini abbiano in media *meno* dita delle donne, come conseguenza di un maggior numero di avventure con utensili elettrici.[3] Ma sarebbe necessario un numero molto elevato di campioni per osservare e convalidare questo fenomeno relativamente raro.

Il test t inizia calcolando una *statistica di test* sui due gruppi di osservazioni. Il t-statistico di Welch è definito come dove $\bar{x}_i$, σ_i e n_i sono la media, la deviazione standard e la dimensione della popolazione del campione i, rispettivamente. Analizziamo attentamente questa equazione. Il numeratore è la differenza tra le medie, quindi maggiore è questa differenza, maggiore è il valore del tstatistico. Le deviazioni standard si trovano nel denominatore, quindi più piccolo è σ_i, più grande è il valore del tstatistico. Se questo è confuso, ricordi cosa succede quando si divide x per un numero che si avvicina a zero. Aumentando le dimensioni del campione n_i rende anche il denominatore più piccolo, quindi più grande è n_i, più grande è il valore del statistico. In tutti i casi, i

fattori che ci rendono più sicuri dell'esistenza di una differenza reale tra le due distribuzioni aumentano il valore del t-statistico. L'interpretazione del significato di un particolare valore della statistica t

deriva dalla ricerca di un numero in una tabella appropriata. Per un *livello di significatività* α e un numero di *gradi di libertà* desiderati (essenzialmente le dimensioni del campione), la voce della tabella specifica il valore v che la statistica t deve superare. Se $t > v$, l'osservazione è significativa al livello α.

Perché funziona?

I test statistici come il t-test mi sembrano spesso un voodoo, in quanto si cerca un numero da una tabella magica e lo si tratta come un vangelo. *L'oracolo ha parlato: la differenza è significativa!* Naturalmente c'è una matematica reale dietro i test di significatività, ma la derivazione coinvolge il calcolo e strane funzioni (come la funzione gamma $\Gamma(n)$, una generalizzazione dei numeri reali dei fattoriali). Questi calcoli complessi sono il motivo per cui è nata la convenzione di cercare le cose in una tabella precompilata, invece di calcolarle da soli.Può trovare le derivazioni delle formule pertinenti in qualsiasi buon libro di statistica, se è interessato. Questi test si basano su idee come il campionamento casuale. Abbiamo visto come la media e la deviazione standard vincolino la forma di qualsiasi distribuzione di probabilità sottostante. Ottenere una media del campione molto lontana media implica sfortuna. Scegliere casualmente valori che si discostano di diverse deviazioni standard dalla media della popolazione è molto improbabile, secondo la teoria. Questo rende più probabile che l'osservazione di una differenza così grande sia il risultato di un'estrazione da una distribuzione diversa.
Gran parte della tecnicità in questo caso è una conseguenza del fatto che si ha a che fare con un fenomeno sottile e con piccole serie di dati. Storicamente, i dati osservati erano una risorsa molto scarsa, e lo sono ancora in molte situazioni. Ricordiamo la nostra discussione sui test di efficacia dei farmaci, dove per ogni singolo punto raccolto deve morire una persona nuova. Il mondo dei big data che probabilmente abiterà è caratterizzato da un maggior numero di osservazioni (tutti coloro che visitano la nostra pagina web), da una posta in gioco più bassa (i clienti acquistano di più quando si mostra loro uno sfondo verde anziché blu?) e da dimensioni dell'effetto più piccole (di quanto miglioramento abbiamo bisogno per giustificare la modifica del colore dello sfondo?).

5.3.3 Il test di Kolmogorov-Smirnov

Il test t confronta due campioni tratti da distribuzioni presumibilmente normali in base alla distanza tra le rispettive medie. Invece, il test di *Kolmogorov-Smirnov* (KS) confronta le funzioni di distribuzione cumulativa (cdf) delle due distribuzioni campionarie e valuta quanto siano simili.

Le cdf dei due diversi campioni sono tracciate sullo stesso grafico. Se i due campioni sono tratti dalla stessa distribuzione, gli intervalli dei valori di x dovrebbero sovrapporsi ampiamente. Inoltre, poiché entrambi Le distribuzioni sono rappresentate come cdf, l'*asse y* rappresenta la probabilità cumulativa da 0 a 1. Entrambe le funzioni aumentano monotonamente da sinistra a destra, dove $C(x)$ è la frazione del campione x. Cerchiamo di identificare il valore di x per il quale i valori y associati delle due cdf differiscono il più possibile. La distanza $D(C_1, C_2)$ tra le distribuzioni C_1 e C_2 è la differenza dei valori y a questo valore critico x. Quanto più sostanzialmente due distribuzioni di campioni differiscono in qualche valore, tanto più è probabile che siano stati estratti da distribuzioni diverse. Due campioni indipendenti della stessa distribuzione normale. Si noti il piccolo divario risultante tra di loro. Al contrario, la Figura 5.13 (a destra) confronta un campione estratto da una distribuzione normale con uno estratto dalla distribuzione uniforme. Il test KS non si lascia ingannare: osservi i grandi divari vicino alle code, dove ci aspetteremmo di vederli.

Il test KS confronta il valore di $D(C_1, C_2)$ con un obiettivo particolare, dichiarando che due distribuzioni differiscono al livello di significatività di α quando dove $c(\alpha)$ è una costante da cercare in una tabella. La funzione delle dimensioni del campione ha una certa intuizione base. Supponiamo per semplicità che entrambi i campioni abbiano la stessa dimensione, n. Allora La quantità n si presenta naturalmente nei problemi di campionamento, come lo standard deviazione della distribuzione binomiale. La differenza prevista tra il numero di teste e di code in n lanci di moneta è dell'ordine di n. Nel contesto del test KS, riflette in modo simile la deviazione attesa quando due campioni devono essere considerati uguali. Il test KS riflette ciò che accade nel cuore della distribuzione, dove è possibile effettuare una determinazione robusta. Mi piace il test di Kolmogorov-Smirnov. Fornisce immagini delle distribuzioni comprensibili, che identificano il punto più debole dell'ipotesi che siano identiche. Questo test ha meno presupposti e varianti tecniche rispetto al test t, il che significa che è meno probabile che si commetta un errore

utilizzandolo. Inoltre, il test KS può essere applicato a molti problemi, tra cui verificare se i punti sono tratti da una distribuzione normale.

Test di normalità

Quando viene tracciata, la distribuzione normale produce una curva a campana. Ma non tutte le distribuzioni a forma di campana sono normali, e a volte è importante conoscere la differenza.

Esistono test statistici specializzati per verificare la normalità di un dato campione distributivo f_1. Ma possiamo usare il test KS generale per fare il lavoro, a condizione che possiamo identificare una f_2 significativa con cui confrontare f_1. Utilizzando il metodo della distribuzione cumulativa che abbiamo descritto, si possono estrarre campioni di n punti statisticamente validi da qualsiasi distribuzione di cui si conosce la cdf, per qualsiasi n. Per f_2, dovremmo scegliere un numero significativo di punti da . Possiamo utilizzare $n_2 = n_1$, o forse un campione un po' più grande se n_1 è molto piccolo. Vogliamo essere sicuri di catturare la forma della distribuzione desiderata con il nostro campione. Quindi, se costruiamo il nostro campione casuale per f_2 dalla distribuzione normale, il test KS non dovrebbe essere in grado di distinguere f_1 da f_2 se anche f_1 proviene da una distribuzione normale sugli stessi μ e σ. Una parola di cautela. Un test statistico sufficientemente sensibile probabilmente rifiuterà la normalità di *qualsiasi* distribuzione osservata. La distribuzione normale è un'astrazione e il mondo è un luogo complicato. Ma osservando il grafico del test KS si vede esattamente dove si verificano le deviazioni. Le code sono troppo grasse o troppo magre? La distribuzione è distorta? Con questa consapevolezza, può decidere se le differenze sono abbastanza grandi da essere importanti per lei.

5.3.4 La correzione di Bonferroni

La convenzione scientifica prevede da tempo l'utilizzo di $\alpha = 0{,}05$ come limite tra la significatività statistica e l'irrilevanza. Una significatività statistica di 0,05 significa che c'è una probabilità di 1/20 che questo risultato sarebbe arrivato per puro caso. Non si tratta di uno standard irragionevole quando si raccolgono dati per testare un'ipotesi difficile. Scommettere su un cavallo con una quota di 20 a 1 e vincere la scommessa è un risultato degno di nota. A meno che non abbia scommesso contemporaneamente su milioni di altri cavalli. Allora vantarsi del piccolo numero di scommesse 20 a 1 in cui ha effettivamente vinto sarebbe quantomeno fuorviante. Pertanto, le spedizioni di pesca che testano milioni di ipotesi devono essere sottoposte a standard più elevati. *Questa* è la

fallacia che ha creato la forte ma spuria correlazione tra i dottorati in informatica e l'attività videoludica. È stato scoperto nel corso della comparazione di migliaia di serie temporali tra loro, e conservando solo le coppie più divertenti che mostravano un elevato punteggio di correlazione. La *correzione di Bonferroni* fornisce un equilibrio importante per valutare quanto ci fidiamo di un risultato statistico apparentemente significativo. Si riferisce al fatto che il modo in cui si è trovata la correlazione può essere importante quanto la forza della correlazione stessa. Chi compra un milione di biglietti della lotteria e vince una volta, ha una fortuna molto meno impressionante di chi compra un solo biglietto e vince. La correzione di Bonferroni stabilisce che quando si testano contemporaneamente n ipotesi diverse, il *valore p* risultante deve salire a un livello α/n, per essere considerato significativo al livello α. Come per ogni test statistico, si nascondono molte sottigliezze nell'applicazione corretta della correzione. Ma il principio fondamentale è importante da cogliere. Le persone che si occupano di informatica sono particolarmente inclini a eseguire confronti su larga scala di tutto contro tutto, o a cercare anomalie e modelli insoliti. Dopo tutto, una volta scritto il programma di analisi, perché non eseguirlo su tutti i dati? Presentando solo i risultati migliori e selezionati, è facile ingannare gli altri. La correzione di Bonferroni è il modo per evitare di ingannare se stessi.

5.3.5 Tasso di scoperta dei falsi

La correzione di Bonferroni ci protegge dall'accettare troppo velocemente la significatività di un'unica ipotesi di successo tra molte prove. Ma spesso, quando si lavora con dati grandi e ad alta dimensionalità, ci troviamo di fronte a un problema diverso. Forse tutte le m variabili sono correlate (forse debolmente) con la variabile target. Se n è abbastanza grande, molte di queste correlazioni saranno statisticamente significative. Abbiamo davvero fatto tante scoperte importanti? La procedura *Benjamini-Hochberg* per la minimizzazione *del tasso di scoperta falsa* (FDR) fornisce un modo molto semplice per tracciare la linea di demarcazione tra variabili interessanti e non interessanti in base alla significatività. Ordini le variabili in base alla forza del loro *p-value*, in modo che le variabili più estreme si trovino a sinistra e quelle meno significative a destra. Ora consideriamo la variabile *iesima* classificata in questo ordinamento. Accettiamo la significatività di questa variabile al livello α se

I *valori p* sono ordinati in ordine crescente da sinistra a destra, come indicato dalla curva blu irregolare. Se accettassimo tutti i *valori p* inferiori ad α, ne accetteremmo troppi. Questo è il motivo per cui Bonferroni ha sviluppato la sua correzione. Ma richiedere che tutti i *valori p* soddisfino lo standard della correzione di Bonferroni (dove la curva attraversa α/m) è troppo severo. La procedura Benjamini-Hochberg riconosce che se molti valori sono davvero significativi per un certo standard, una certa frazione di essi dovrebbe essere significativa per uno standard molto più alto.

5.4 Storia di guerra: Scoprire la Fonte della Giovinezza?

È stato un matrimonio bellissimo. Eravamo molto felici per Rachel e David, gli sposi. Avevo mangiato come un re, avevo ballato con la mia adorabile moglie e stavo godendo di un caldo bagliore post-prima colazione, quando mi sono reso conto che qualcosa non andava. Mi sono guardato intorno nella stanza e ho fatto un doppio sguardo. In qualche modo, per la prima volta dopo molti anni, ero diventato più giovane della maggior parte delle persone presenti. Questo potrebbe non sembrarle un problema, ma è perché lei, il lettore, probabilmente *è* più giovane della maggior parte delle persone in molti ambienti. Ma mi creda, arriverà il momento in cui noterà queste cose. Ricordo quando mi sono resa conto di aver frequentato l'università nel periodo in cui nasceva la maggior parte dei miei studenti. Poi hanno iniziato a nascere quando ero alla scuola di specializzazione. Gli studenti universitari di oggi non solo sono nati dopo che sono diventato professore, ma anche dopo che ho ottenuto la cattedra qui. Quindi, come potrei essere più giovane della maggior parte delle persone presenti a questo matrimonio? Ci sono due possibilità. O è stato un caso che così tanti anziani. Questo è il motivo per cui sono stati inventati i test di significatività statistica e i valori *p,* per aiutare a distinguere qualcosa dal nulla. Quindi, qual era la probabilità che io, allora all'età di 54 anni, fossi più giovane della maggior parte delle 251 persone presenti al matrimonio di Rachel? Secondo Wolfram Al- pha (più precisamente, gli esiti quinquennali dell'American Community Survey 2008-2012), negli Stati Uniti c'erano 309,1 milioni di persone, di cui 77,1 milioni avevano 55 anni o . Quasi esattamente il 25% della popolazione è più anziana di me mentre scrivo queste parole. La probabilità che la maggioranza di 251 americani selezionati a caso abbia più di 55 anni. Questa probabilità è incredibilmente piccola, paragonabile all'estrazione di una moneta giusta

dalla tasca e al fatto che esca testa per 56 volte di seguito. Questo non poteva essere il risultato di un evento casuale. Doveva esserci un motivo per cui ero più giovane della maggior parte di questa folla, e la risposta non era che stavo diventando più giovane. Quando ho chiesto a Rachel di parlarne, mi ha detto che, per motivi di budget, hanno deciso di non invitare bambini al matrimonio. Questa sembrava essere una spiegazione ragionevole. Dopo tutto, questa regola ha escluso 73,9 milioni di persone di età inferiore ai diciotto anni dalla partecipazione al matrimonio, risparmiando così miliardi di dollari rispetto a quanto sarebbe costato invitarli tutti. La frazione f di persone più giovani di me che non sono bambini è pari a f= 1(77,*1/*(309,1 73,9))= 0,672. Tuttavia, si tratta di un valore sostanzialmente superiore a− 0,5. La probabilità− che io sia più giovane della mediana in un campione casuale estratto da questa coorte è: Sebbene sia molto più grande del precedente *valore p*, è ancora incredibilmente piccolo: è come lanciare 27 teste di fila sulla sua moneta giusta. Il solo divieto di avere figli non è stato abbastanza potente da farmi tornare giovane. Sono tornata da Rachel e l'ho costretta a . È emerso che sua madre ha avuto un numero insolitamente elevato di cugini crescendo, ed è stata eccezionalmente brava a mantenere i contatti con *tutti* loro. Ricordiamo la Teoria della Relatività di Einstein, dove $E= mc^2$ indica che tutti sono cugini di mia madre, due volte separati. *Tutti* questi cugini sono stati invitati al matrimonio. Con la famiglia di Rachel che ha superato il clan insolitamente piccolo dello sposo, questa coorte di cugini anziani ha dominato la pista da ballo. Infatti, possiamo calcolare il numero di cugini più anziani (c) che devono essere invitati per ottenere una probabilità del 50/50 che io sia più giovane dell'invitato mediano, supponendo che il resto dei 251 invitati sia stato selezionato a caso. Si scopre che c = 65 cugini singoli (o 32,5 coppie sposate) sono sufficienti, una volta esclusi i bambini (f = 0,672). La morale è che è importante calcolare la probabilità di qualsiasi osservazione interessante prima di dichiararla un miracolo. Non si fermi mai a una spiegazione parziale, se non riduce la sorpresa a livelli plausibili. Probabilmente c'è un fenomeno genuino alla base di qualsiasi evento sufficientemente raro, e scoprire di cosa si tratta rende la scienza dei dati entusiasmante.

5.5 Test di permutazione e valori P

I tradizionali test di significatività statistica si dimostrano abbastanza efficaci nel decidere se due campioni sono effettivamente tratti dalla stessa distribuzione. Tuttavia, questi test devono essere eseguiti correttamente per

svolgere il loro lavoro. Molti test standard presentano delle sottigliezze, come la questione dei test a un lato o a due lati, le ipotesi distributive e altro ancora. L'esecuzione corretta di questi test richiede attenzione e formazione.

I test di permutazione consentono un modo più generale e computazionalmente a prova di idiota per stabilire la significatività. Se la sua ipotesi è supportata dai dati, allora Le serie di dati mescolati in modo casuale dovrebbero avere meno probabilità di sostenerlo. Conducendo molte prove su dati randomizzati, possiamo stabilire esattamente quanto sia insolito il fenomeno che state testando. Consideri la Figura 5.15, dove denotiamo la variabile indipendente (sesso: maschile o femminile) e la variabile dipendente (ad esempio, l'altezza) usando i colori. La distribuzione originale dei colori dei risultati (a sinistra) appare nettamente diversa tra uomini e donne, riflettendo le reali differenze di altezza. Ma quanto è insolita questa differenza? Possiamo costruire un nuovo set di dati assegnando casualmente il genere alle variabili di risultato originali (centro). L'ordinamento all'interno di ciascun gruppo rende evidente che la distribuzione pseudo-maschile/femminile dei risultati è ora molto più equilibrata rispetto ai dati originali (destra). Questo dimostra che il sesso *era* effettivamente un fattore significativo nel determinare l'altezza, una conclusione a cui crederemo ancora di più dopo che si ripeterà più volte nel corso di 1.000 o 1.000.000 di prove.

Il rango della statistica del test sui dati reali tra la distribuzione dei valori statistici delle permutazioni casuali determina il livello di significatività o il *valore p*. La Figura 5.16 (a sinistra) mostra ciò che stiamo cercando. Il valore reale si trova all'estrema destra della distribuzione, a testimonianza della significatività. Nella figura a destra, il valore reale si trova nella parte centrale della distribuzione, suggerendo che non c'è alcun effetto.

I test di permutazione richiedono lo sviluppo di una statistica che rifletta la sua ipotesi sui dati. Il coefficiente di correlazione è una scelta ragionevole se si vuole stabilire una relazione importante tra una coppia specifica di variabili. Idealmente, la correlazione osservata nei dati reali sarà più forte di qualsiasi permutazione casuale. Per convalidare la connessione tra sesso e altezza, forse la nostra statistica potrebbe essere la differenza tra le altezze medie di uomini e donne. Anche in questo caso, speriamo che si dimostri più grande nei dati reali che nella maggior parte delle permutazioni casuali.

Siate creativi nella scelta della statistica: la potenza dei test di permutazione è che possono funzionare praticamente con qualsiasi cosa lei possa proporre per dimostrare il suo caso. È meglio che la sua statistica riduca al minimo la possibilità di pareggi, dal momento che è tenuto a contare tutti i pareggi contro la sua ipotesi. I test di permutazione forniscono la probabilità dei dati

in base alla sua ipotesi, ossia che la statistica sia un outlier rispetto alla distribuzione casuale del campione. Questo non equivale a dimostrare la sua ipotesi in base ai dati, che è l'obiettivo tradizionale dei test di significatività statistica. Ma è molto meglio di niente. Il punteggio di significatività o il *valore* p di un test di permutazione dipende da quanti tentativi casuali vengono eseguiti. Cerchi sempre di eseguire almeno 1.000 prove casuali, e anche di più se è possibile. Più permutazioni si provano, più impressionante può essere il *valore p* di significatività, almeno fino a un certo punto. Se l'input dato è in effetti il migliore di tutte le $k!$ permutazioni, il *valore* p più estremo che può ottenere è $1/k!$, indipendentemente dal numero di permutazioni casuali provate. Il sovracampionamento gonfierà il denominatore senza aumentare di una virgola la fiducia reale. I valori P sono calcolati per aumentare *la* fiducia che un'osservazione sia reale e interessante. Questo funziona solo se si esegue il test di permutazione in modo onesto, eseguendo esperimenti che possano fornire una giusta misura di sorpresa.

5.5.1 Generazione di permutazioni casuali

La generazione di permutazioni casuali è un altro importante problema di campionamento che le persone spesso sbagliano. I due algoritmi che seguono utilizzano entrambi sequenze di scambi casuali per rimescolare la permutazione iniziale $\{1, 2, \ldots, n\}$. Ma garantire che tutte le $n!$ permutazioni siano generate in modo uniforme e casuale è un'impresa difficile. Infatti, solo uno di questi algoritmi ci riesce. Ci pensi attentamente: la differenza qui è molto sottile. È così sottile che potrebbe non notarla nemmeno nel codice. La differenza critica è l'1 o l'i chiamata a Random. Uno di questi algoritmi è giusto e uno di questi algoritmi è sbagliato. Se pensa di poterlo dire, spieghi in modo convincente perché uno funziona e l'altro no. Se proprio deve saperlo, il primo algoritmo è corretto. Sceglie un elemento casuale da 1 a n per la prima posizione, poi lo lascia in pace e ricorre sul resto. Genera permutazioni uniformemente a caso. Il secondo algoritmo offre a determinati elementi una maggiore possibilità di finire per primi, dimostrando che la distribuzione non è uniforme. Ma se non è possibile dimostrarlo teoricamente, si può utilizzare l'idea di un test di permutazione. Implementa entrambi gli algoritmi ed esegui

1.000.000 di esecuzioni di ciascuno, costruendo permutazioni casuali di, ad esempio, n= 4 elementi. Conta la frequenza con cui ciascun algoritmo genera ognuna delle 4! = 24 permutazioni distinte. I risultati di questo esperimento sono mostrati nella Figura 5.17. L'algoritmo 1 si dimostra incredibilmente

costante, con una deviazione standard di 166,1 occorrenze. Al contrario, c'è una differenza di otto volte tra le permutazioni più e meno frequenti con l'algoritmo 2, con $\sigma = 20,923,9$.

La morale è che la generazione casuale può essere molto sottile. E che gli esperimenti di tipo Monte Carlo, come i test di permutazione, possono eliminare la necessità di un ragionamento sottile. Verificate, poi fidatevi.

5.5.2 La serie di battute di DiMaggio

Uno dei record più sorprendenti del baseball è la striscia di battuta di 56 partite di Joe DiMaggio. Il compito di un battitore è quello di ottenere battute, e ogni partita ha forse quattro possibilità di ottenerne una. Anche i battitori più bravi spesso falliscono. Ma nel 1941, Joe DiMaggio riuscì a ottenere successi in 56 partite consecutive, un risultato davvero incredibile. Nessun giocatore nei settantacinque anni successivi si è avvicinato a questo record, né nessuno prima di lui. Ma quanto è stata insolita una striscia così lunga nel contesto della sua carriera? DiMaggio ha giocato 1736 partite, con 2214 successi in 6821 battute. Quindi dovrebbe ottenere successi in circa− 1 (1 $(-2214/6821))^4 = 79,2\%$ delle sue partite con quattro battute. Qual è la probabilità che qualcuno con il suo livello di abilità possa gestire una tale striscia di partite consecutive nel corso della sua carriera? Per coloro che sono stanchi delle mie analogie con il baseball, mettiamo questo in un altro contesto. Supponiamo che lei sia uno studente che ha una media di 90 nei test. È sicuramente un ottimo studente, ma non perfetto. Quante possibilità ci sono che lei abbia una striscia positiva in cui ottiene un punteggio superiore a 90 in dieci esami consecutivi? E se ne avesse venti di fila? Potrebbe forse superare 56 esami di fila? Se si verificasse una striscia così lunga, significherebbe che ha portato i suoi studi ad un altro livello, o che è stato solo fortunato? Quindi, quando DiMaggio ha avuto la sua striscia di battitori, era solo una conseguenza attesa delle sue indiscusse capacità e della sua costanza, o è stato semplicemente fortunato? È stato uno dei migliori battitori della sua epoca o di qualsiasi altra epoca, un All-Star in ogni stagione dei suoi tredici anni di carriera. Ma sappiamo anche che DiMaggio è stato fortunato di tanto in . Dopo tutto, *era* sposato con la star del cinema Marilyn Monroe. Per risolvere questa domanda, abbiamo utilizzato dei numeri casuali per simulare quando ha ottenuto dei successi nel corso di una "carriera" sintetica di 1736 partite. Ad ogni partita, il Joe simulato ha ricevuto quattro possibilità di colpire e ci è riuscito con una probabilità di $p = (2214/6821) = 0,325$.

Possiamo quindi identificare la striscia di battute più lunga nel corso di di questa carriera simulata. Simulando 100.000 carriere di DiMaggio, otteniamo una distribuzione di frequenza delle strisce che può contestualizzare la rarità del suo successo, ottenendo un *valore p* nel processo. I risultati sono mostrati nella Figura 5.18. Solo in 44 delle 100.000 carriere simulate (p= 0,00044) DiMaggio ha gestito una striscia di almeno 56 partite. Quindi la lunghezza è abbastanza fuori linea rispetto a quanto ci si aspetterebbe da lui. Il secondo La striscia più lunga di qualsiasi battitore della Major League è di sole 44 partite, quindi è anche fuori linea rispetto a tutti gli altri. Ma una volta ha anche colpito in 61 partite consecutive a un livello inferiore di competizione, quindi sembra che abbia avuto una straordinaria capacità di coerenza. Le strisce di battute possono essere considerate come corse tra partite senza battute, e quindi possono essere modellate utilizzando una distribuzione di Poisson. Ma le simulazioni Monte Carlo forniscono risposte senza una matematica dettagliata. I test di permutazione ci danno una visione con una conoscenza e uno sforzo intellettuale minimi.

5.6 Ragionamento bayesiano

La probabilità condizionale $P(A \mid B)$ misura la probabilità che si verifichi l'evento A, dato che si sa che l'evento B si è verificato. Faremo affidamento sulla probabilità condizionale in tutto questo libro, perché ci permette di aggiornare la nostra fiducia in un evento in risposta a nuove prove, come i dati osservati. *Il Teorema di Bayes* è uno strumento importante per lavorare con le probabilità condizionali. Con il teorema di Bayes, possiamo convertire la domanda di $P(\textit{esito dei dati} \mid)$ in $P(\textit{esito dei dati})$, che spesso è molto più facile da calcolare. In un certo senso, il teorema di Bayes è solo una conseguenza dell'algebra, ma porta a un modo diverso di pensare alla probabilità. Il teorema di Bayes in azione. Lo spazio degli eventi consiste nello scegliere uno dei quattro blocchi. Gli eventi complessi A e B rappresentano dei sottogruppi di blocchi, dove $P(A)$= 3/4 e $P(B)$= 2/4= 1/2. Contando i blocchi da la figura, si può vedere che $P(A \mid$ B)= 1/2 e $P(B \mid$ A)= 1/3. Il ragionamento bayesiano riflette il modo in cui una probabilità anteriore $P(A)$ viene aggiornata per dare la probabilità posteriore $P(A \mid B)$ di fronte a una nuova osservazione B, in base al rapporto tra la probabilità $P(B \mid A)$ e la probabilità marginale $P(B)$. La probabilità *anteriore* $P(A)$ riflette

la nostra ipotesi iniziale sul mondo, da rivedere in base all'evidenza aggiuntiva B. Il ragionamento bayesiano è un modo importante per guardare il mondo. Entrando al matrimonio di Rachel e David, la mia ipotesi preliminare era che la distribuzione dell'età avrebbe rispecchiato quella del mondo in generale. Ma la mia fiducia si è indebolita con ogni cugino anziano che ho incontrato, fino a crollare. Utilizzeremo il ragionamento bayesiano per costruire i classificatori. Ma tenga presente questa filosofia mentre analizza i dati. Dovrebbe arrivare a ogni compito con una concezione precedente di quali dovrebbero essere le risposte, e poi rivederle in base alle prove statistiche.

Capitolo 6

Visualizzazione dei dati

La visualizzazione efficace dei dati è un aspetto importante della scienza dei dati, per almeno tre motivi distinti:

- *Analisi esplorativa dei dati*: Che aspetto hanno realmente i suoi dati? Capire con cosa si ha a che fare è il primo passo di qualsiasi analisi seria. I grafici e le visualizzazioni sono il modo migliore che conosco per .
- *Rilevamento degli errori*: Ha fatto qualcosa di stupido nella sua analisi? Dare in pasto dati non visualizzati a qualsiasi algoritmo di apprendimento automatico significa andare incontro a problemi. I problemi con i punti di anomalia, la pulizia insufficiente e le ipotesi errate si rivelano immediatamente quando si visualizzano correttamente i dati. Troppo spesso una statistica riassuntiva (77,8% di precisione!) nasconde ciò che il suo modello sta realmente facendo. Guardare con attenzione a ciò che sta ottenendo di giusto o di sbagliato è il primo passo per migliorare le prestazioni.
- *Comunicazione*: È in grado di presentare in modo efficace agli altri ciò che ha imparato? I risultati significativi diventano azionabili solo dopo essere stati condivisi. Il suo successo come scienziato dei dati si basa sul convincere gli altri che lei sa di cosa sta parlando. Un'immagine vale più

di 1.000 parole, soprattutto quando deve fare una presentazione ad un pubblico scettico.

Probabilmente crea grafici e diagrammi fin dalle elementari. Un software universale rende facile la creazione di immagini dall'aspetto professionale. Allora, cosa c'è di così difficile nella visualizzazione dei dati?

Per rispondere, offro una parabola. Un terribile incidente avvenuto durante la mia giovinezza riguardava un'aggressione a una campionessa di pattinaggio sul ghiaccio. Un teppista la colpì al ginocchio con un bastone, sperando di metterla fuori gioco alle imminenti Olimpiadi. Fortunatamente, mancò ginocchio e la pattinatrice vinse la medaglia d'argento.

Ma dopo aver conosciuto il suo cliente, l'avvocato del delinquente ha proposto una difesa interessante. Questo crimine, disse, era chiaramente troppo complesso perché il suo cliente potesse concepirlo da solo. Questo mi ha colpito, perché significava che avevo sottovalutato la capacità cognitiva necessaria per colpire qualcuno alla gamba con un bastone.

La mia morale è che molte cose sono più complicate di quanto sembrino. In particolare, parlo del problema di tracciare i dati su un grafico per cogliere ciò che . Una percentuale incredibilmente alta di grafici che ho visto nelle presentazioni sono *terribili*, o non trasmettono alcun messaggio, o non rivelano ciò che i dati effettivamente mostrano. I cattivi grafici possono avere un valore negativo, in quanto portano nella direzione sbagliata.

In questa sezione, arriveremo a comprendere i principi che fanno funzionare i disegni dei grafici standard e mostreremo come possono essere fuorvianti se non vengono utilizzati correttamente. A partire da questa esperienza, cercheremo di sviluppare il suo senso di quando i grafici possono mentire e come può costruirne di migliori.

6.1 Analisi dei dati esplorativi

L'avvento di enormi serie di dati sta cambiando il modo di fare scienza. Il metodo scientifico tradizionale è *basato sulle ipotesi*. Il ricercatore formula una teoria sul funzionamento del mondo e poi cerca di sostenere o respingere questa ipotesi sulla base dei dati. Al contrario, la scienza *guidata dai dati* inizia con l'assemblaggio di una serie di dati sostanziali, e poi va a caccia di modelli che idealmente svolgeranno il ruolo di ipotesi per le analisi future. *L'analisi esplorativa dei dati* è la ricerca di modelli e tendenze in un dato

insieme di dati. Le tecniche di visualizzazione svolgono un ruolo importante in questa ricerca. Osservare attentamente i dati è importante per diverse ragioni, tra cui identificare gli errori nella raccolta/elaborazione, trovare violazioni dei presupposti statistici e suggerire ipotesi interessanti.
In questa sezione, discuteremo di come procedere all'analisi esplorativa dei dati e di cosa apporta la visualizzazione come parte del processo.

6.1.1 Confrontarsi con un nuovo set di dati

Cosa fare quando si incontra una nuova serie di dati? Questo dipende in qualche modo dal motivo per cui è interessato, ma le fasi iniziali dell'esplorazione sono quasi indipendenti dall'applicazione.
Ecco alcuni passi fondamentali che raccomando di fare per familiarizzare con qualsiasi nuovo set di dati, che illustro esplorando il set di dati sulla misurazione del corpo NHANES. Si tratta di dati tabellari, ma i principi generali qui esposti sono applicabili a una classe più ampia di risorse: *Rispondere alle domande di base*: Ci sono diverse cose che dovrebbe sapere sul suo set di dati prima ancora di aprire il file. Faccia domande come: *Chi ha costruito questo set di dati, quando e perché?* Capire come sono stati ottenuti i suoi dati fornisce indizi sulla loro probabile rilevanza e sull'opportunità di fidarsi di essi. Inoltre, ci indica le persone giuste se abbiamo bisogno di saperne di più sull'origine o sulla provenienza dei dati. Con un po' di ricerche, ho scoperto che provenivano dal National Health and Nutrition Examination Survey 2009-2010, e chi era il responsabile della pubblicazione. *Quanto è grande?* Quanto è ricco il set di dati in termini di numero di campi o colonne? Quanto è grande in termini di numero di record o righe? Se è troppo grande per essere esplorato facilmente con gli strumenti interattivi, estragga un piccolo campione e faccia le sue esplorazioni iniziali su quello. Questo set di dati ha 4978 record (2452 uomini e 2526 donne), ciascuno con sette campi dati più il sesso. *Cosa significano i campi?* Esamini ogni colonna del suo set di dati e si assicuri di capire il loro significato. Quali campi sono numerici o categorici? quali unità sono state misurate le quantità? Quali campi sono ID o descrizioni, invece di dati con cui fare calcoli? Un rapido esame mostra che le lunghezze e i pesi qui sono stati misurati utilizzando il sistema metrico, rispettivamente in centimetri e chilogrammi. *Cerchi record familiari o interpretabili*: Trovo estremamente prezioso familiarizzare con alcuni record, al punto da conoscerne i nomi. In genere, i record sono associati a una persona, un luogo o una cosa di cui si ha già una certa conoscenza, in modo da poterli contestualizzare e valutare la solidità dei dati possesso. Ma se così non

fosse, trovi alcuni record di particolare interesse da conoscere, magari quelli con i valori massimi o minimi del campo più importante. Se le cartelle cliniche familiari non, a volte conviene. Un ingegnoso sviluppatore di un database di cartelle cliniche mi ha detto di aver utilizzato i primi 5000 nomi storici di *Who's Bigger* come nomi di pazienti durante lo sviluppo del prodotto. Si trattava di un'idea molto più ispirata rispetto alla creazione di nomi artificiali come "Paziente F1253". Erano abbastanza divertenti da incoraggiare il gioco con il sistema, e abbastanza memorabili da poter segnalare e segnalare i casi anomali: ad esempio, "C'è qualcosa di gravemente sbagliato in Franz Kafka". *Statistiche di riepilogo*: Esamini le statistiche di base di ogni colonna. Il *riepilogo dei cinque numeri* di Tukey è un ottimo inizio per i valori numerici, composto dai valori estremi (max e min), oltre agli elementi mediani e quartili. Applicato ai componenti del nostro set di dati di altezza/peso, otteniamo: Questo è molto istruttivo. Innanzitutto, come mai l'età mediana è di 584 anni? Tornando ai dati, apprendiamo che l'età è misurata in mesi, il che significa che la mediana è di 48,67 anni. La lunghezza delle braccia e delle gambe sembra avere circa la stessa mediana, ma la lunghezza delle gambe ha una variabilità molto maggiore. *Non lavevo mai saputo.* Ma improvvisamente mi rendo conto che le persone sono più spesso descritte come con gambe lunghe/corte che con braccia lunghe/corte, quindi forse è questo il motivo. Per i campi categoriali, come l'occupazione, il riepilogo analogo sarebbe un rapporto su quanti tipi di etichette diverse appaiono nella colonna e quali le tre categorie più popolari, con le frequenze associate. *Correlazioni a coppie*: Una matrice di coefficienti di correlazione tra tutte le coppie di colonne (o almeno le colonne rispetto alle variabili dipendenti di interesse) dà un'idea di quanto sarà facile costruire un modello di successo. Idealmente, avremo diverse caratteristiche che si correlano fortemente con il risultato, mentre non si correlano fortemente tra loro. Solo una colonna di un insieme di caratteristiche perfettamente correlate ha un valore, perché tutte le altre caratteristiche sono completamente definite da ogni singola colonna. Queste correlazioni a coppie sono piuttosto interessanti. Perché l'altezza è correlata *negativamente* con l'età? Le persone qui presenti sono tutte adulte (241 mesi = 20,1 anni), quindi sono tutte completamente cresciute. Ma la generazione precedente era più bassa delle persone di oggi. Inoltre, le persone si rimpiccioliscono quando invecchiano, quindi questo insieme probabilmente spiega il fenomeno. La forte correlazione tra peso e circonferenza vita (0,89) riflette una sfortunata verità sulla natura. *Suddivisioni per classi*: Ci sono modi interessanti per suddividere le cose in base alle principali variabili categoriali, come il sesso o la località? Attraverso il riepilogo statistiche, può valutare se c'è

una differenza tra le distribuzioni quando sono condizionate dalla categoria. Guardi soprattutto dove pensa che ci *dovrebbero* essere delle differenze, in base alla sua comprensione dei dati e dell'applicazione. Le correlazioni sono generalmente simili per sesso, ma ci sono alcune differenze interessanti. Ad esempio, la correlazione tra altezza e peso è più forte per gli uomini (0,443) rispetto alle donne (0,297). *Grafici delle distribuzioni*: Questo capitolo si concentra sulle tecniche di visualizzazione dei dati. Utilizzi i tipi di grafici di cui parleremo nella Sezione 6.3 per osservare le distribuzioni, alla ricerca di schemi e valori anomali. Qual è la forma generale di ogni distribuzione? I dati devono essere puliti o trasformati per renderli più a campana? La potenza di una griglia di diagrammi a punti di diverse variabili. A colpo d'occhio vediamo che non ci sono outlier selvaggi, quali coppie sono correlate e la natura di eventuali linee di tendenza. Armati di questo singolo grafico, siamo ora pronti ad applicare questa serie di dati a qualsiasi sfida.

6.1.2 Statistiche di sintesi e quartetto di Anscombe

Ci sono limiti profondi alla comprensione dei dati senza tecniche di visualizzazione. Questo è rappresentato al meglio dal quartetto di Anscombe: quattro serie di dati bidimensionali, ognuna con undici punti. Tutti e quattro Gli insiemi di dati hanno medie identiche per i valori x e y, varianze identiche per i valori x e y. valori x e y, e la stessa identica correlazione tra i valori x e y. Queste serie di dati devono essere tutte piuttosto simili, giusto? Studi i numeri per un po', in modo da avere un'idea di come sono. Ora guardi i diagrammi a punti di queste serie di dati. Sono tutti diversi e raccontano storie sostanzialmente diverse. Hanno tutti un aspetto diverso e raccontano storie sostanzialmente diverse. Uno ha un andamento lineare, mentre un secondo sembra quasi parabolico. Altri due hanno un andamento quasi perfettamente lineare, con i valori anomali, ma con pendenze molto diverse. Il punto è che si possono apprezzare immediatamente queste differenze con uno sguardo al grafico di dispersione. Anche le visualizzazioni semplici sono strumenti potenti per capire cosa succede in un insieme di dati. Qualsiasi scienziato dei dati ragionevole si sforza di trarre il massimo vantaggio dalle tecniche di visualizzazione.

6.1.3 Strumenti di visualizzazione

Per supportare la visualizzazione è disponibile una vasta collezione di strumenti software. In generale, i compiti di visualizzazione rientrano in tre categorie e la scelta giusta degli strumenti dipende da quale sia la sua vera missione:

- *Analisi dei dati esplorativi*: qui cerchiamo di eseguire esplorazioni rapide e interattive di una serie di dati. I programmi di foglio elettronico come Excel e gli ambienti di programmazione basati su notebook come iPython, R e Mathe- matica sono efficaci per creare i tipi di diagrammi standard. La chiave qui è nascondere la complessità, in modo che le routine di plottaggio facciano di default qualcosa di ragionevole, ma possano essere personalizzate se necessario.

- *Grafici di qualità da pubblicazione/presentazione*: Solo perché Excel è molto popolare, non significa che produca i migliori grafici/trame possibili. Le migliori visualizzazioni sono un'interazione tra scienziato e software, sfruttando appieno la flessibilità di uno strumento per massimizzare il contenuto informativo del grafico.

Le librerie di plottaggio come MatPlotLib o Gnuplot supportano una serie di opzioni che consentono al suo grafico di apparire esattamente come lo desidera. Il linguaggio statistico R ha una libreria molto ampia di visualizzazioni di dati. Consulti i cataloghi dei tipi di grafici supportati dalla sua libreria preferita, per aiutarla a trovare la rappresentazione migliore per i suoi dati. *Visualizzazione interattiva per applicazioni esterne*: La costruzione di *dashboard* che facilitano l'interazione dell'utente con set di dati proprietari è un compito tipico degli ingegneri del software orientati alla scienza dei dati. La missione tipica in questo caso è costruire strumenti che supportino l'analisi esplorativa dei dati per il personale meno esperto tecnicamente e più orientato alle applicazioni. Tali sistemi possono essere facilmente costruiti in linguaggi di programmazione come Python, utilizzando librerie di plottaggio standard. Esiste anche una classe di sistemi di terze parti per la creazione di cruscotti, come Tableau. Questi sistemi sono pro- grammabili a un livello superiore rispetto ad altri strumenti, supportando particolari paradigmi di inter-azione e viste collegate tra viste distinte dei dati. Quale dipinto le piace di più? La formazione di preferenze intelligenti nell'arte o nelle visualizzazioni dipende dall'avere un'estetica visiva distintiva.

6.2 Sviluppare un'estetica di visualizzazione

Un apprezzamento sensato dell'arte o del vino richiede lo sviluppo di un gusto o di un'estetica particolare. Non si tratta tanto di capire se una cosa le piace, ma di capire perché le piace. Gli esperti d'arte parlano della gamma del palato di un pittore, dell'uso della luce o dell'energia/tensione di una composizione.

Gli intenditori di vino testimoniano fragranza, il corpo, l'acidità e la limpidezza del loro vino preferito, e la quantità di rovere o tannino che contiene. Hanno sempre qualcosa di meglio da dire che "ha un buon sapore". Per distinguere le visualizzazioni buone o cattive, è necessario sviluppare un'analisi del design e un vocabolario per parlare delle rappresentazioni dei dati. La Figura 6.4 presenta due famosi punti di riferimento della pittura occidentale. Quale dei due è migliore? Questa domanda non ha senso senza un senso estetico e un vocabolario per descriverlo.
La mia estetica e il mio vocabolario visivo derivano in gran parte dai libri di Edward Tufte. È un artista: infatti una volta ho avuto modo di incontrarlo nella sua ex galleria d'arte di fronte a Chelsea Piers a Manhattan. Ha riflettuto a lungo su ciò che rende un grafico o un diagramma informativo e bello, basando l'estetica del design sui seguenti principi:

- *Massimizzare il rapporto dati-inchiostro*: La sua visualizzazione dovrebbe mettere in risalto i suoi dati. Allora perché la maggior parte di ciò che si vede nei grafici sono le griglie di sfondo, le ombreggiature e i segni di spunta?
- *Ridurre al minimo il fattore menzogna*: Come scienziato, i suoi dati dovrebbero rivelare la verità, idealmente la verità che lei vuole vedere rivelata. Ma è onesto con il suo pubblico o utilizza dispositivi grafici che lo ingannano nel vedere qualcosa che in realtà non c'è?

- *Ridurre al minimo la spazzatura dei grafici*: I moderni software di visualizzazione spesso aggiungono effetti visivi interessanti che hanno poco a che fare con i suoi dati. Il suo grafico è interessante grazie ai suoi dati, o nonostante essi?

- *Utilizzi scale adeguate e un'etichettatura chiara*: L'interpretazione accurata dei dati dipende da elementi diversi dai dati, come la scala e l'etichettatura. I suoi dati descrittivi sono materiali ottimizzati per la chiarezza e la precisione?

- *Utilizzi efficacemente il colore*: l'occhio umano ha il potere di discriminare tra piccole gradazioni di tonalità e saturazione del colore. Utilizza il colore per evidenziare proprietà importanti dei suoi dati o solo per fare una dichiarazione artistica?

- *Sfrutti il potere della ripetizione*: Gli array di grafici simili con elementi di dati diversi ma correlati forniscono un modo conciso e potente per

consentire confronti visivi. I multipli dei suoi grafici facilitano i confronti o sono semplicemente ridondanti?

6.2.1 Massimizzare il rapporto dati-inchiostro

In qualsiasi grafica, una parte dell'inchiostro viene utilizzata per rappresentare i dati reali sottostanti, mentre il resto viene impiegato per gli effetti grafici. In generale, le visualizzazioni dovrebbero concentrarsi sulla rappresentazione dei dati stessi. Definiamo il *rapporto datiinchiostro* : Rapporto dati-inchiostro= dati inchiostro/inchiostro totale. Il salario medio in base al sesso (Bureau of Labor Statistics, 2015) e aiuta a chiarire questa nozione. Quale rappresentazione dei dati preferisce? La massimizzazione del rapporto dati-inchiostro lascia parlare i dati, che è lo scopo principale dell'esercizio di visualizzazione. La prospettiva piatta sulla destra consente un confronto più equo delle altezze delle barre, per cui i maschi non sembrano come i pipsqueak per le donne. I colori fanno un buon lavoro, permettendoci di fare un confronto tra le due cose. Esistono modi più estremi per aumentare il rapporto dati-inchiostro. Perché abbiamo bisogno di barre? La stessa informazione potrebbe essere trasmessa tracciando un punto dell'altezza appropriata, e sarebbe chiaramente un miglioramento se tracciassimo molto più degli otto punti mostrati qui. Sia consapevole che meno può essere più nella visualizzazione dei dati.

6.2.2 Ridurre al minimo il fattore menzogna

Una visualizzazione cerca di raccontare una storia vera su ciò che i dati dicono. La forma più semplice di menzogna è quella di falsificare i dati, ma è possibile riportare i dati in modo accurato, ma fuorviare deliberatamente il pubblico su ciò che dicono. Tufte definisce il *fattore di menzogna* di un grafico come: - (dimensione di un effetto nel grafico) menzogna / (dimensione dell'effetto dei dati). L'integrità grafica richiede di minimizzare questo fattore di menzogna, evitando le tecniche che tendono ad ingannare. Le cattive pratiche includono:

- *Presentare le medie senza varianza*: I valori dei dati 100, 100, 100*{* , 100, 100 *}* e 200*{* , 0, 100, 200, 0 raccontano*}* storie diverse, anche se entrambe le medie sono 100. Se non può tracciare i punti effettivi con la media, mostri almeno la varianza, per rendere chiaro il grado in cui la media riflette la distribuzione.

- *Presentare interpolazioni senza i dati reali*: Le linee di regressione e le curve adattate sono efficaci per comunicare le tendenze e semplificare

grandi serie di dati. Ma senza mostrare i punti di dati su cui si basano, è impossibile accertare la qualità dell'adattamento.

- *Distorsioni di scala*: Il rapporto di aspetto di una figura può avere un effetto enorme sul modo in cui interpretiamo ciò che vediamo. La Figura 6.6 presenta tre rappresentazioni di una determinata serie temporale finanziaria, identiche tranne che per il rapporto di aspetto del grafico.

  Nel rendering inferiore, la serie appare piatta: non c'è nulla di cui preoccuparsi. A destra, i profitti sono scesi da un precipizio: il cielo sta cadendo! Il grafico dell'angolo sinistro presenta un grave declino, ma con segnali di un rimbalzo autunnale.

  Qual è il grafico giusto? Le persone sono generalmente abituate a vedere i grafici presentati secondo il rapporto aureo, che implica che la larghezza dovrebbe essere circa 1,6 volte l'altezza. Dia loro questa forma, a meno che non abbia dei motivi ben sviluppati per cui è inappropriata. Gli psicologi ci informano che le linee a 45 gradi sono le più facilmente interpretabili, quindi eviti le forme che amplificano o smorzano sostanzialmente le linee di questo obiettivo.
- *Eliminare le etichette di spunta dagli assi numerici*: Anche le peggiori disfunzioni di scala possono essere completamente nascoste non stampando le etichette di riferimento numerico sugli assi. Solo con le etichette di scala numerica si possono ricostruire i valori reali dei dati dal grafico.

• *Nascondere il punto di origine dal grafico*: Il presupposto implicito nella maggior parte dei grafici è che l'intervallo di valori sull'asse y vada da zero a y_{max}. Perdiamo la capacità di confrontare visivamente le grandezze se l'*intervallo delle* y va invece da y_{min} ϵ a y_{max}. Il valore più grande appare improvvisamente molte volte più grande del valore più piccolo, invece di essere scalato nella proporzione corretta.

Se la Figura 6.5 (a destra) fosse disegnata con un intervallo y stretto [900, 2500], il messaggio sarebbe che i consulenti stanno morendo di fame, invece di guadagnare stipendi vicini a quelli degli insegnanti, come gli sviluppatori di software a quelli dei farmacisti. Questi inganni possono essere riconosciuti se le scale sono segnate sull'asse, ma sono difficili da cogliere. Nonostante la formula di Tufte, il fattore di menzogna non può essere calcolato meccanicamente, perché richiede la comprensione dell'agenda che sta dietro alla distorsione. Nel leggere qualsiasi grafico, è importante sapere chi lo ha

prodotto e perché. Capire il loro programma dovrebbe sensibilizzarla ai messaggi potenzialmente fuorvianti codificati nel grafico.

6.2.3 Ridurre al minimo la spazzatura dei grafici

Gli elementi visivi estranei distraggono dal messaggio che i dati stanno cercando di trasmettere. In un grafico emozionante, sono i dati a raccontare la storia, non la spazzatura del grafico. Una serie temporale mensile delle vendite di un'azienda che sta iniziando a vivere momenti difficili. Il grafico in questione è un *diagramma a barre*, un modo perfettamente valido per rappresentare i dati delle serie temporali, e viene disegnato utilizzando le opzioni convenzionali, forse predefinite, di un pacchetto di plottaggio ragionevole. Le operazioni critiche sono:

- *Prigionieri dei suoi dati (in alto a sinistra)*: Le griglie pesanti imprigionano i suoi dati, dominando visivamente i contenuti. Spesso i grafici possono essere migliorati eliminando la griglia, o almeno alleggerendola.

  Il valore potenziale della griglia di dati è che facilita un'interpretazione più precisa delle quantità numeriche. Pertanto, le griglie tendono ad essere più utili nei grafici con un gran numero di valori che potrebbero dover essere citati con precisione. Le griglie leggere possono gestire adeguatamente tali compiti.

- *Smetta di fare ombra (in alto a destra)*: Lo sfondo colorato qui non apporta nulla all'interpretazione del grafico. Rimuovendolo, aumenta il rapporto dati-inchiostro e lo rende meno invadente.

- *Pensare fuori dagli schemi (in basso a sinistra)*: Il riquadro di delimitazione non contribuisce realmente all'informazione, in particolare i confini superiori e quelli di destra, che non definiscono gli assi. Li elimini e faccia entrare più aria nelle sue trame.

- *Faccia in modo che l'inchiostro mancante lavori per lei (in basso a destra)*: L'effetto della griglia di riferimento può essere recuperato rimuovendo le linee dalle barre invece di aggiungere elementi. In questo modo è più facile confrontare l'entità dei numeri più grandi, concentrando l'attenzione sui grandi cambiamenti nel pezzo superiore relativamente piccolo, invece che sui piccoli cambiamenti nella barra lunga.

L'architetto Mies van der Rohe disse notoriamente che "meno è meglio". La rimozione di elementi dalle trame spesso le migliora molto di più dell'aggiunta

di elementi. Lo faccia diventare parte della sua filosofia di progettazione grafica.

6.2.4 Scala ed etichettatura adeguate

Le carenze nella scala e nell'etichettatura sono la fonte principale di disinformazione intenzionale o accidentale nei grafici. Le etichette devono riportare la giusta grandezza dei numeri e la scala deve mostrare questi numeri alla giusta risoluzione, in modo da facilitare il confronto. In generalei dati devono essere scalati in modo da riempire lo spazio assegnato loro sul grafico. Le persone ragionevoli possono divergere sul fatto di scalare gli assi sull'intera gamma teorica della variabile, o di ridurli per riflettere solo i valori osservati. Ma alcune decisioni sono chiaramente irragionevoli. Poiché la correlazione è compresa tra [1, 1], ha forzato il grafico per rispettare questo intervallo. L'immenso mare di bianco in questo grafico cattura solo l'idea che avremmo potuto fare meglio, avvicinando la correlazione a 1,0. Ma il grafico è altrimenti illeggibile. La Figura 6.9 (a destra) presenta esattamente gli stessi dati, ma con una scala troncata. *Ora* possiamo vedere dove ci sono aumenti di prestazioni man mano che ci spostiamo da sinistra a destra e leggiamo il punteggio per qualsiasi lingua. In precedenza, le barre erano così distanti dalle etichette che era difficile stabilire la corrispondenza nome-barra.

Il peccato più grande delle scale troncate si verifica quando non si mostra l'intera barra, per cui la lunghezza della barra non riflette più il valore relativo variabile. Qui mostriamo la linea $y = 0$, aiutando il lettore a capire che ogni barra deve intera. Anche far uscire i dati dalla griglia della prigione sarebbe stato utile.

6.2.5 Uso efficace del colore e delle ombreggiature

I colori sono sempre più considerati parte integrante di qualsiasi comunicazione grafica. In effetti, ho appreso con piacere che i costi di stampa della mia casa editrice sono ora identici per il colore e il bianco e nero, quindi lei, il lettore, non pagherà di più per vedere la mia grafica a colori qui. I colori svolgono due ruoli principali nei grafici, ovvero marcare le distinzioni di classe e codificare i valori numerici. Rappresentare i punti di diversi tipi, cluster o classi con colori diversi codifica un altro livello di informazioni su un convenzionale grafico a punti. Questa è un'ottima idea quando cerchiamo di stabilire l'entità delle differenze nella distribuzione dei dati tra le classi. La

cosa più importante è che le classi siano facilmente distinguibili l'una dall'altra, utilizzando colori primari in grassetto. La cosa migliore è che i colori siano selezionati per avere valori mnemonici da collegare naturalmente alla classe in questione. Le perdite dovrebbero essere stampate con inchiostro rosso, le cause ambientali associate con il verde, le nazioni con i colori della loro bandiera e le squadre sportive con i colori della loro maglia. Colorare i punti per rappresentare i maschi in blu e le femmine in rosso offre un indizio sottile per aiutare l'osservatore a interpretare un grafico a dispersione.
La selezione dei colori per rappresentare una scala numerica è un problema più difficile. Le mappe dei colori dell'arcobaleno sono percettivamente non lineari, il che significa che non è ovvio per nessuno se il viola si trova prima o dopo il verde. Pertanto, mentre i numeri tracciati nei colori dell'arcobaleno raggruppano numeri simili in colori simili, le grandezze relative sono impercettibili senza fare esplicitamente riferimento alla scala di colori.
Molto meglio sono le scale di colore basate sulla variazione della luminosità o della saturazione. La *luminosità* di un colore viene modulata mescolando la tinta con una tonalità di grigio, a metà tra il bianco e il nero. *La saturazione* viene controllata mescolando una frazione di grigio, dove 0 produce la tinta pura e 1 elimina tutto il colore. Un'altra scala di colori popolare presenta colori positivi/negativi distinti (ad esempio, blu e rosso) riflessi intorno a un centro bianco o grigio a zero. In questo modo, la tonalità indica all'osservatore la polarità del numero, mentre la luminosità/saturazione riflette la magnitudine. Alcune scale di colori sono molto più adatte alle persone daltoniche, in particolare quelle che evitano l'uso del rosso e del verde.
Come regola generale, le aree grandi sui grafici dovrebbero essere rappresentate con colori non saturi. Il contrario è vero per le regioni piccole, che risaltano meglio con i colori saturi. I sistemi di colore sono una questione sorprendentemente tecnica e complicata, il che significa che dovrebbe sempre utilizzare scale di colori ben consolidate, invece di inventarne di proprie.

6.2.6 Il potere della ripetizione

I piccoli grafici multipli e le tabelle sono modi eccellenti per rappresentare i dati multivariati. Ricordiamo la potenza delle griglie che mostrano tutte le distribuzioni bivariate.Ci sono molte applicazioni dei piccoli grafici multipli. Possiamo usarli per suddividere una distribuzione per classi, magari tracciando grafici separati ma comparabili per regione, sesso o tempo. Gli array di grafici facilitano i confronti: cosa è cambiato tra distribuzioni

diverse. I grafici delle serie temporali ci permettono di confrontare le stesse quantità in diversi punti caldici. Ancora meglio è confrontare più serie temporali, sia come linee sullo stesso grafico, sia come grafici multipli in un array logico che riflette la loro relazione.

6.3 Grafico Tipi

In questa sezione, esamineremo le motivazioni alla base dei principali tipi di visualizzazione dei dati. Per ogni grafico, presento le migliori pratiche per il loro utilizzo e delineo i gradi di libertà di cui dispone per rendere la sua presentazione il efficace possibile.
Non c'è niente che dica "Ecco un grafico di alcuni dati" come un grafico prodotto in modo sconsiderato, creato utilizzando le impostazioni predefinite di uno strumento software. I miei studenti mi presentano troppo spesso questi prodotti di dati non digeriti, e questa sezione è in qualche modo una reazione personale contro di essi. Lei ha il potere e la responsabilità di produrre presentazioni significative e interpretabili del suo lavoro. Una visualizzazione efficace implica un processo iterativo che consiste nell'osservare i dati, decidere quale storia si sta cercando di raccontare e poi migliorare la visualizzazione per raccontare meglio la storia. Pratico albero decisionale per aiutare a selezionare la giusta rappresentazione dei dati, tratto da Abela [Abe13]. I grafici più importanti saranno esaminati in questa sezione, ma usi questo albero per capire meglio *perché* certe visualizzazioni sono più appropriate in determinati contesti. Dobbiamo produrre il grafico giusto per una determinata serie di dati, non solo la prima cosa che ci viene in .

6.3.1 Dati tabellari

Le tabelle di numeri possono essere belle e sono un modo molto efficace per presentare i dati. Sebbene possano *sembrare* prive del fascino visivo delle presentazioni grafiche, le tabelle presentano diversi vantaggi rispetto ad altre rappresentazioni, tra cui:

- *Rappresentazione della precisione*: La risoluzione di un numero ci dice qualcosa sul processo con cui è stato ottenuto: un salario medio di 79.815 dollari dice qualcosa di diverso da 80.000 dollari. Queste sottigliezze si perdono in genere nei diagrammi, ma sono chiaramente visibili nelle tabelle numeriche.
- *Rappresentazione della scala*: Le lunghezze delle cifre dei numeri in una tabella possono essere paragonate a grafici a barre, su una scala logaritmica. La giustificazione dei numeri a destra comunica meglio le

differenze di ordine di grandezza, così come (in misura minore) la scansione delle cifre iniziali dei numeri in una colonna.

La giustificazione a sinistra dei numeri impedisce tali confronti, quindi si assicuri sempre di giustificarli a destra.

- *Visualizzazione multivariata*: La geometria diventa complicata da capire quando si va oltre le due dimensioni. Ma le tabelle possono rimanere gestibili anche per un gran numero di variabili. Ricordiamo le statistiche sul baseball di Babe Ruth, una tabella di ventotto colonne che è facilmente interpretabile da qualsiasi tifoso esperto.
- *Dati eterogenei*: Le tabelle sono generalmente il modo migliore per presentare un mix di attributi numerici e categorici, come il testo e le etichette. I glifi, come le emoji, possono anche essere utilizzati per rappresentare i valori di alcuni campi.
- *Compattezza*: Le tabelle sono particolarmente utili per rappresentare piccoli numeri di punti. Due punti in due dimensioni possono essere disegnati come una linea, ma perché preoccuparsi? Una tabella piccola è generalmente migliore di una visualizzazione rada.

La presentazione di dati tabellari sembra semplice da fare ("basta metterli in una tabella"), battere una gamba con un bastone. Ma ci sono delle sottigliezze nel produrre le tabelle più informative. Le migliori pratiche includono:

- *Ordinare le righe per invitare al confronto*: Ha la libertà di ordinare le righe di una tabella nel modo che desidera, quindi ne approfitti. Ordinare le righe in base ai valori di una colonna importante è generalmente una buona idea. Quindi, raggruppare le righe è utile per facilitare il confronto, mettendo i "mi piace" con i "mi piace".

  L'ordinamento per dimensione o data può essere più rivelatore del nome in molti contesti. L'utilizzo di un ordine canonico delle righe (ad esempio, lessicografico per nome) può essere utile per cercare gli elementi per nome, ma in genere questo non è un problema a meno che la tabella non abbia molte righe.
- *Ordini le colonne per evidenziare l'importanza o le relazioni di coppia*: Gli occhi che attraversano la pagina da sinistra a destra non possono fare confronti visivi efficaci, ma i campi vicini sono facili da contrastare. In generale, le colonne dovrebbero essere organizzate in modo da raggruppare i campi simili, nascondendo quelli meno importanti a destra.

- *Giustificare a destra i numeri a precisione uniforme*: Confronto visivo tra 3,1415 e 39,2 in una tabella è un compito senza speranza: il numero più grande deve sembrare più grande. La cosa migliore è giustificarli a destra e impostare tutti la stessa precisione: 3,14 contro 39,20.
- *Utilizzi l'***enfasi***, il* carattere *o il colore per evidenziare le voci importanti*: Evidenziando i valori estremi di ogni colonna in modo che risaltino, rivela a colpo d'occhio le informazioni importanti. Tuttavia, è facile **esagerare**, quindi cerchi di essere delicato.

- *Eviti descrittori di colonna di lunghezza eccessiva*: I nastri bianchi nelle tabelle distraggono e di solito derivano da etichette di colonna più lunghe dei valori che rappresentano. Utilizzi abbreviazioni o accatastamenti di parole su più righe per minimizzare il problema, e chiarisca qualsiasi ambiguità nella didascalia allegata alla tabella. Per aiutare a illustrare questi possibili peccati, ecco una tabella che registra sei proprietà di quindici nazioni diverse, con gli ordini delle righe e delle colonne dati a caso. Ci sono molti ordinamenti possibili delle righe (Paesi). L'ordinamento per singola colonna è un miglioramento rispetto a quello casuale, anche se potremmo anche raggrupparle per regione/continente. L'ordine delle colonne può essere reso più incomprensibile mettendo i like accanto ai like. Infine, trucchi come la giustificazione dei numeri, la rimozione delle cifre non informative, l'aggiunta di virgole e l'evidenziazione valore più grande in ogni colonna rendono i dati più facili da leggere:

6.3.2 Grafici a punti e a linee

I grafici a punti e a linee sono le forme più diffuse di grafici di dati, che forniscono una rappresentazione visiva di una funzione $y = f(x)$ definita da un insieme di punti (x, y). I grafici a punti mostrano solo i punti dei dati, mentre i grafici a linee li collegano o li interpolano per definire una funzione continua $f(x)$. I vantaggi dei grafici a linee includono:

- *Interpolazione e adattamento*: La curva di interpolazione derivata dai punti fornisce una previsione per $f(x)$ sull'intera gamma di x possibili. Questo ci permette di controllare la correttezza o di fare riferimento ad altri valori e di rendere esplicite le tendenze mostrate nei dati.

Sovrapporre una curva adattata o smussata allo stesso grafico dei dati di partenza è una combinazione molto potente. L'adattamento fornisce un modello che spiega cosa dicono i dati, mentre i punti effettivi ci permettono di esprimere un giudizio educato su quanto ci fidiamo del modello.

- *Grafici a punti*: L'aspetto positivo dei tracciati a linee è che non è mostrare la linea, con il risultato di un *tracciato a punti*. Collegare i punti con segmenti di linea (polilinee) si rivela fuorviante in molte situazioni. Se la funzione è definita solo in punti interi, o i *valori* x rappresentano condizioni distinte, allora non ha alcun senso interpolare tra loro. Inoltre, le polilinee oscillano drammaticamente per catturare gli outlier, incoraggiandoci visivamente a concentrarci proprio sui punti che dovremmo. I movimenti su e giù ad alta frequenza ci distraggono da vedere la tendenza più ampia, che è la ragione principale per fissare un grafico.

Le migliori pratiche con i grafici a linee includono: *Mostrare i punti dati, non solo i fits*: In genere, è importante mostrare i dati effettivi, anziché solo le linee adattate o interpolate. La chiave è assicurarsi che l'uno non sovrasti l'altro. Per rappresentare un gran numero di punti in modo discreto, possiamo (a) ridurre dimensioni dei punti, possibilmente a puntini, e/o (b) schiarire la tonalità dei punti in modo che si posizionino sullo sfondo. Si ricordi che esistono cinquanta sfumature di grigio e che la chiave è la delicatezza. *Se possibile, mostri l'intera gamma di variabili*: Per impostazione predefinita, la maggior parte dei software grafici traccia da x_{min} a x_{max} e da y_{min} a y_{max}, dove i minimi e i massimi sono definiti sui valori dei dati di input. Ma il minimo e il massimo logici sono specifici del contesto e possono ridurre il fattore di menzogna per mostrare l'intera gamma. I conteggi dovrebbero logicamente partire da zero, non da y_{min}. Ma a volte mostrare l'intera gamma è poco informativo, in quanto appiattisce completamente l'effetto che si sta cercando di illustrare. Una possibile soluzione è quella di utilizzare una scala logica per l'asse, in modo da incorporare una gamma più ampia di numeri in un modo efficiente dal punto di vista dello spazio. Ma se deve troncare l'intervallo, chiarisca cosa sta facendo, utilizzando etichette dell'asse con segni di spunta e chiarendo qualsiasi ambiguità nella didascalia associata.*Ammettere l'incertezza quando si tracciano le medie*: I punti che appaiono su un grafico a linee o a punti sono spesso ottenuti dalla media di più osservazioni. La media risultante cattura meglio la distribuzione rispetto a qualsiasi singola . Ma le medie hanno interpretazioni diverse in base alla varianza. Sia la media di { 8.5, 11.0, 13.5,

7.0, 10.0} che quella di { 9.9, 9.6, 10.3, 10.1, 10.1} sono pari a 10.0, ma il grado di fiducia nella loro precisione è sostanzialmente diverso.
Ci sono diversi modi per confessare il livello di incertezza della misurazione nei nostri grafici. Il mio preferito consiste nel tracciare tutti i valori dei dati sottostanti sullo stesso grafico delle medie, utilizzando lo stesso *valore* x come media associata. Questi punti saranno visivamente poco appariscenti rispetto alla linea di tendenza più pesante, a condizione che siano disegnati come piccoli punti e leggermente ombreggiati, ma ora sono disponibili per l'ispezione e l'analisi. Un secondo approccio traccia la deviazione standard σ intorno a y come un baffo, mostrando l'intervallo $[y\ \sigma,- y+\ \sigma]$. Questa rappresentazione dell'intervallo è onesta e denota l'intervallo con il 68% dei valori in una distribuzione normale. Un baffo più lungo significa che dovrebbe essere più sospettoso dell'accuratezza delle medie, mentre un baffo corto implica una maggiore precisione. *I diagrammi a riquadri* registrano in modo conciso l'intervallo e la distribuzione dei valori in un punto con un riquadro. Questo riquadro mostra l'intervallo di valori dei quartili (25% e 75%) e viene tagliato in corrispondenza della mediana (50° percentile). In genere i baffi (capelli) vengono aggiunti per mostrare l'intervallo dei valori più alti e più bassi. La Figura 6.13 mostra un grafico box-and-whisker del peso in funzione dell'altezza in un campione di popolazione. Il peso mediano aumenta con l'altezza, ma non il massimo, perché un minor numero di punti nella fascia più alta riduce la possibilità di un valore massimo anomalo. I veri scienziati sembrano amare i grafici box-and-whisker, ma personalmente li trovo eccessivi. Se proprio non è possibile rappresentare i punti dati effettivi, forse basta mostrare le linee di contorno che affiancano la media/mediana al 25° e 75° percentile. Questo trasmette esattamente le stesse informazioni del riquadro nel grafico a riquadri, con una minore quantità di rifiuti grafici.

- *Non collegare mai i punti per i dati categorici*: Supponiamo di misurare una variabile (forse, il reddito mediano) per diverse classi (ad esempio, i cinquanta Stati, dall'Alabama al Wyoming). Potrebbe avere senso visualizzare questo dato come un grafico a punti, con 1 x 50, ma sarebbe sciocco e fuorviante collegare i punti. Perché? Perché non c'è un'adiacenza≤≤ significativa tra lo Stato i e lo Stato $i+ 1$. In effetti, tali grafici sono meglio considerati come grafici a barre, di cui si parla nella Sezione 6.3.4.
  In effetti, collegare i punti tramite polilinee è molto spesso una schifezza di grafico. Le linee di tendenza o di adattamento sono spesso

più rivelatrici e informative. Cerchi di mostrare i punti di dati grezzi stessi, anche se in modo leggero e discreto.

- *Utilizzi il colore e il tratteggio per distinguere le linee/le classi*: Spesso ci troviamo di fronte alla rappresentazione stessa funzione $f(x)$ disegnata su due o più classi, ad esempio il reddito come funzione della scolarità, separatamente per uomini e donne.
  Il modo migliore per gestirli è assegnare colori distinti alle linee/punti per ogni classe. Si possono utilizzare anche i tratteggi delle linee (punteggiate, tratteggiate, solide e in grassetto), ma spesso sono più difficili da distinguere rispetto ai colori, a meno che il supporto di output non sia in e nero. In pratica, si possono distinguere da due a quattro linee di questo tipo su un singolo grafico, prima che la visualizzazione collassi in un pasticcio. Per visualizzare un numero elevato di gruppi, li suddivida in cluster logici e utilizzi tracciati multipli, ciascuno con un numero di linee sufficiente a non creare confusione.

6.3.3 Piani di dispersione

Gli insiemi di dati massicci sono una vera sfida da presentare in efficace, perché un gran numero di punti facilmente sovrasta le rappresentazioni grafiche, dando luogo 'immagine della palla nera della morte. Ma se disegnati correttamente, i grafici a dispersione sono in grado di mostrare migliaia di punti bivariati (bidimensionali) in modo chiaro e comprensibile. I diagrammi di dispersione mostrano i valori di ogni punto (x, y) in una determinata serie di dati. Abbiamo utilizzato i grafici a dispersione nella Sezione 4.1 per rappresentare lo stato di massa corporea degli individui, rappresentandoli come punti nello spazio altezza-peso. Il colore di ogni punto rifletteva la sua classificazione come normale, sovrappeso o obeso. Le migliori pratiche associate ai grafici a dispersione includono:

- *Sparga i punti della giusta dimensione*: Nel film *Oh G-d*, George Burns nei panni del creatore guarda all'avocado come al suo più grande errore. Perché? Perché aveva fatto la buca troppo grande. L'errore più grande della maggior parte delle persone con i grafici a dispersione è quello di rendere i punti troppo grandi.

Ora vediamo una struttura fine di un nucleo denso, pur riuscendo a rilevare l'alone chiaro dei valori anomali. La dimensione predefinita dei punti per la maggior parte dei programmi di plottaggio è appropriato per una cinquantina di punti. Ma per set di dati più grandi, utilizzi punti più piccoli.

- *Colorare o scuotere i punti interi prima di eseguire il grafico a dispersione*: I grafici a dispersione rivelano schemi a griglia quando i valori *x* e *y* hanno valori interi, perché non ci sono gradazioni uniformi tra di loro. Questi grafici a dispersione hanno un aspetto innaturale, ma peggio ancora tendono a oscurare i dati, perché spesso più punti condividono esattamente le stesse coordinate.
  Ci sono due soluzioni ragionevoli. La prima consiste nel colorare ogni punto in base alla sua frequenza di occorrenza. Questi grafici sono chiamati *mappe di calore* e i centri di concentrazione diventano facilmente visibili, a condizione che si utilizzi una scala di colori ragionevole. La Figura 6.15 mostra una mappa di calore dei dati di altezzapeso, che fa un lavoro molto migliore per rivelare le concentrazioni dei punti rispetto al semplice grafico a punti associato. Un'idea correlata è quella di ridurre l'*opacità* (equivalentemente, di aumentare la *trans-parenzialità*) dei punti che tracciamo a dispersione. Per impostazione predefinita, i punti sono generalmente disegnati in modo da essere opachi, producendo una massa quando ci sono punti sovrapposti. Ma ora supponiamo di permettere a questi punti di essere leggermente ombreggiati e transgenere. Ora i punti che si sovrappongono appaiono più scuri dei singoli, ottenendo una mappa di calore in più tonalità di grigio.
  Il secondo approccio consiste nell'aggiungere una piccola quantità di rumore casuale a ciascun punto, per farlo oscillare all'interno di un cerchio di raggio subunitario intorno alla sua posizione originale. Ora vedremo l'intera molteplicità dei punti e romperemo la regolarità distraente della griglia.
- *Proiettare i dati multivariati fino a due dimensioni, oppure utilizzare matrici di grafici a coppie*: Gli esseri del nostro universo hanno difficoltà a visualizzare le serie di dati in quattro o più dimensioni. Le serie di dati a più alta dimensione possono spesso essere proiettate a due dimensioni prima di renderle su grafici a dispersione, utilizzando tecniche come l'analisi delle componenti principali e l'auto-organizzazione. mappe. Un bell'esempio dove proiettiamo un centinaio di dimensioni fino a due, rivelando una visione molto coerente di questa serie di dati ad alta dimensionalità. L'aspetto positivo di queste trame è che possono fornire una visione efficace di cose che altrimenti non potremmo vedere. L'aspetto negativo è che le due dimensioni non hanno più alcun significato. Più precisamente, le nuove dimensioni non hanno nomi di variabili che possano trasmettere un significato, perché ciascuna delle due 'nuove' dimensioni codifica le proprietà di tutte le dimensioni originali. Una rappresentazione alternativa consiste nel tracciare una griglia o un reticolo di tutte le proiezioni a coppie, ognuna delle quali mostra solo due delle dimensioni originali. Questo è un modo

meraviglioso per avere un'idea di quali coppie di dimensioni sono correlate tra loro.

- *I diagrammi di dispersione tridimensionali sono utili solo quando c'è una struttura reale da mostrare*: I notiziari televisivi sulla scienza dei dati presentano sempre qualche ricercatore che afferra una nuvola di punti tridimensionale e la fa ruotare nello spazio, alla ricerca di qualche importante intuizione scientifica. Non la trovano mai, perché la vista di una nuvola da qualsiasi direzione appare praticamente uguale a quella di qualsiasi altra direzione. In genere non c'è un punto di osservazione in cui diventa improvvisamente chiaro il modo in cui le dimensioni interagiscono.

  L'eccezione è quando i dati sono stati effettivamente ricavati da oggetti tridimensionali strutturati, come le scansioni laser di una determinata scena. La maggior parte dati che incontriamo nella scienza dei dati non corrisponde a questa descrizione, quindi le aspettative per la visualizzazione interattiva sono basse. Utilizzi la tecnica della griglia di tutte le proiezioni bidimensionali, che essenzialmente visualizza la nuvola da tutte le direzioni ortogonali.
- *I diagrammi a bolle variano il colore e la dimensione per rappresentare dimensioni aggiuntive*: Modificando il colore, la forma, la dimensione e l'ombreggiatura dei punti, i tracciati a punti possono rappresentare dimensioni aggiuntive, sui tracciati a bolle. In genere, questo funziona meglio che tracciare i punti in tre dimensioni.

6.3.4 Grafici a barre e a torta

I diagrammi a barre e i grafici a torta sono strumenti per presentare le proporzioni relative delle variabili categoriche. Entrambi funzionano suddividendo un insieme geometrico, sia esso una una barra o un cerchio, in aree proporzionali alla frequenza di ciascun gruppo. Entrambi gli elementi sono efficaci in multipli, per consentire i confronti. Infatti, suddividendo ogni barra in pezzi si ottiene il *grafico a barre sovrapposte*. Dati dei votanti di tre anni di elezioni presidenziali statunitensi, presentati come grafici a torta e a barre. Il blu rappresenta i voti democratici, il rosso i voti repubblicani. Le torte mostrano più chiaramente quale schieramento ha vinto ogni elezione, ma le barre mostrano che il totale dei voti repubblicani è rimasto abbastanza costante, mentre i democratici sono stati generalmente in crescita. Si noti che queste barre possono essere facilmente confrontate perché sono giustificate a sinistra. Alcuni critici si scatenano in una febbre quasi religiosa contro i grafici a torta, perché occupano più spazio del necessario e sono generalmente più difficili da leggere e confrontare. Ma i grafici a torta sono

probabilmente migliori per mostrare le percentuali della totalità. A molte persone sembrano piacere, quindi probabilmente sono innocui in piccole quantità. Le migliori pratiche per i grafici a barre e i grafici a torta includono: *Etichetta direttamente le fette della torta*: I grafici a torta sono spesso accompagnati da chiavi di legenda che indicano cosa corrisponde ogni fetta di colore. Questo distrae molto, perché gli occhi devono spostarsi avanti e indietro tra la chiave e la torta per interpretarla.

Molto meglio è etichettare ogni fetta direttamente, all'interno della fetta o appena oltre il bordo. Questo ha il vantaggio secondario di scoraggiare l'uso di fette, perché le fette diventano generalmente non interpretabili. È utile raggruppare le fette in un'unica fetta chiamata "*altro"*, per poi presentare magari una Il secondo grafico a torta è stato scomposto in *altri* componenti principali. *Utilizzi i grafici a barre per consentire confronti precisi*: Se ancorati a una linea fissa, gli array di barre consentono di identificare facilmente i valori minimi e massimi di una serie e se una tendenza è in aumento o in diminuzione. I grafici a barre impilati sono concisi, ma più difficili da usare per questi scopi. Predisporre una serie di piccoli grafici a barre, in questo caso uno per ogni gruppo di sesso/etnia, ci permette di fare confronti così fini. *Riduca la scala in modo appropriato, a seconda che cerchi di evidenziare la grandezza assoluta o la proporzione*: I grafici a torta esistono per rappresentare frazioni dell'intero. Nel presentare una serie di grafici a torta o a barre, la decisione più critica è se vuole mostrare le dimensioni dell'insieme o invece le frazioni di ogni sottogruppo. Due grafici a barre impilati che presentano le statistiche di sopravvivenza. sulla nave condannata *Titanic*, riportati per classe di biglietto. L'istogramma registra con precisione le dimensioni di ciascuna classe e gli risultanti. Il grafico con barre di uguale lunghezza cattura meglio come il tasso di mortalità sia aumentato per le classi inferiori. I grafici a torta possono anche essere utilizzati per mostrare i cambiamenti di grandezza, variando l'area del cerchio che definisce la torta. Ma per l'occhio è più difficile calcolare l'area che la lunghezza, rendendo difficili i confronti. Modulare il raggio per riflettere la magnitudine invece dell'area è ancora più ingannevole, perché raddoppiando il raggio di un cerchio si moltiplica l'area per quattro.

Grafici a <u>torta</u> sbagliati

Due grafici a torta che riportano la distribuzione dei delegati alla convention repubblicana del 2016, per candidato. La torta di sinistra è bidimensionale, mentre il grafico di destra presenta fette spesse e ben separate per evidenziare la profondità. Quale dei due è migliore nel trasmettere la distribuzione dei voti?

Dovrebbe essere chiaro che gli effetti tridimensionali e la separazione sono pura spazzatura grafica, che oscura solo la relazione tra le dimensioni delle fette. Anche i

valori reali dei dati sono scomparsi, forse perché non c'era abbastanza spazio per loro dopo tutte quelle ombre. Ma perché bisogno di un grafico a torta? Una piccola tabella di etichette/colori con una colonna aggiuntiva di percentuali sarebbe più concisa e informativa.

6.3.5 Istogrammi

Le proprietà interessanti delle variabili o delle caratteristiche sono definite dalla loro distribuzione di frequenza sottostante. Dove si trova il picco della distribuzione e la modalità è vicina alla media? La distribuzione è simmetrica o obliqua? Dove sono le code? Potrebbe essere bimodale, suggerendo che la distribuzione proviene da un mix di due o più popolazioni sottostanti?

Spesso ci troviamo di fronte a un gran numero di osservazioni di una particolare variante e cerchiamo di tracciarne una rappresentazione. *Gli istogrammi* sono grafici delle distribuzioni di frequenza osservate. Quando la variabile è definita su un'ampia gamma di valori possibili rispetto alle n osservazioni, è improbabile che si veda una duplicazione esatta. Tuttavia, suddividendo l'intervallo di valori in un numero adeguato di bins di uguale larghezza, possiamo accumulare diversi conteggi per bins e approssimare la distribuzione di probabilità sottostante.

Il problema principale nella costruzione di un istogramma è decidere il numero giusto di bins da utilizzare. Troppi bins, e ci saranno solo pochi punti anche nel bucket più popolare. Abbiamo scelto il binning risolvere proprio questo problema. Ma se si utilizzano troppo pochi bins, non si vedranno abbastanza dettagli per capire la forma della distribuzione.

Le conseguenze della dimensione dei bin sull'aspetto di un istogramma. I grafici nella riga superiore raggruppano 100.000 punti di una distribuzione normale in dieci, venti e cinquanta secchi rispettivamente. Ci sono abbastanza punti per riempire cinquanta secchi, e la distribuzione sulla destra appare bellissima. I tracciati nella fila inferiore hanno solo 100 punti, quindi il grafico a trenta bucket sulla destra è rado e scabroso. In questo caso, sette bins (mostrati a sinistra) sembrano produrre il grafico più rappresentativo.

È impossibile fornire regole rigide e rapide per selezionare il miglior numero di bin b per mostrare i suoi dati. Si renda conto che non sarà mai in grado di discriminare tra più di un centinaio di bins a occhio, quindi questo fornisce un limite superiore logico. In generale, mi piace vedere una media di 10-50 punti per binario per rendere le cose più omogenee, quindi b= n/25 fornisce una prima ipotesi ragionevole. Ma sperimenti diversi valori di b, perché il giusto numero di binari funzionerà molto meglio degli altri. Lo riconoscerà quando lo vedrà.

Una serie di dati di grandi dimensioni beneficia di molti bins (in alto). Ma la struttura di un insieme di dati più piccolo è meglio mostrata quando ogni bin ha un numero non banale di elementi.

Le migliori pratiche per gli istogrammi includono: *Trasformi il suo istogramma in un pdf*: In genere, interpretiamo i nostri dati come osservazioni che approssimano la funzione di densità di probabilità (pdf) di una variabile casuale sottostante. In questo caso, diventa più interpretabile etichettare l'*asse delle y* con la frazione di elementi in ogni bucket, invece del conteggio totale. Questo è particolarmente vero nei set di dati di grandi dimensioni, dove i bucket sono abbastanza pieni da non preoccuparci del livello esatto di supporto. La Figura 6.23 mostra gli stessi dati, tracciati a destra come pdf invece che come istogramma. La forma è esattamente la stessa in entrambi i grafici: tutto ciò che cambia è l'etichetta sull'asse y. Tuttavia, il risultato è più facile da interpretare, perché è in termini di probabilità invece che di conteggi. *Consideri la cdf*: La funzione di densità cumulativa (cdf) è l'integrale del pdf e le due funzioni contengono esattamente le stesse informazioni. Quindi, consideri l'utilizzo di una cdf invece di un istogramma per rappresentare la sua distribuzione. Un aspetto positivo del tracciare la cdf è che non si basa su un parametro di conteggio dei bin, quindi presenta una visione reale e non adulterata dei suoi dati. Ricordiamo come apparivano grandi i cdf nel test di KolmogorovSmirnov. Le distribuzioni cumulative richiedono una lettura leggermente più sofisticata rispetto agli istogrammi. La cdf è monotonamente crescente, quindi non ci sono picchi nella distribuzione. Invece, la modalità è contrassegnata dal segmento di linea verticale più lungo. Ma i cdf evidenziano molto meglio le code di una distribuzione. Il motivo è chiaro: i piccoli conteggi nelle code sono oscurati dall'asse di un istogramma, ma si accumulano in elementi visibili nella cdf.

6.3.6 Mappe di dati

Le mappe utilizzano la disposizione spaziale delle regioni per rappresentare luoghi, concetti o cose. Tutti noi abbiamo imparato a navigare nel mondo attraverso le mappe, abilità che si traducono nella comprensione delle visualizzazioni correlate.

Le mappe di dati tradizionali utilizzano il colore o l'ombreggiatura per evidenziare le proprietà delle regioni nella mappa. Le mappe non si limitano alle regioni geografiche. La mappa più potente nella storia della visualizzazione scientifica è la tavola periodica degli elementi della chimica. Le regioni collegate mostrano dove risiedono i metalli e i gas Nobel, oltre ai punti in cui dovevano trovarsi gli elementi non ancora scoperti. La tavola periodica è una mappa sufficientemente dettagliata da essere citata ripetutamente dai chimici che lavorano, ma è facilmente comprensibile per i bambini delle scuole.

Cosa conferisce alla tavola periodica un tale potere di visualizzazione?

- *La mappa ha una storia da raccontare*: Le mappe sono preziose quando codificano informazioni che vale la pena consultare o assimilare. La tavola periodica è la *giusta* visualizzazione degli elementi a causa della struttura dei gusci di elettroni e della loro importanza sulle proprietà chimiche e sul legame. La mappa deve essere esaminata ripetutamente perché ha cose importanti da dirci.

  Le mappe di dati sono affascinanti, perché la suddivisione delle variabili per regione porta spesso a storie interessanti. Le regioni sulle mappe in genere riflettono le continuità culturali, storiche, economiche e linguistiche, per cui i fenomeni che derivano da uno di questi fattori in genere si mostrano chiaramente sulle mappe di dati.

- *Le regioni sono contigue e la contiguità ha un significato*: la continuità delle regioni nella Figura 6.26 si riflette nello schema di colori, che raggruppa elementi che condividono proprietà simili, come i metalli alcalini e i gas Nobel. Due elementi che si trovano uno accanto all'altro di solito significano che hanno qualcosa di importante in comune.
- *I quadrati sono abbastanza grandi da vedere*: Una decisione critica nella canonizzazione della tavola periodica è stata la collocazione dei metalli lantanidi (elementi 57-71) e attinidi (elementi 89-103). Sono presentati convenzionalmente come le due file inferiori, ma logicamente appartengono al corpo della tavola quadrati verdi non etichettati.

  Tuttavia, il rendering convenzionale evita due problemi. Per farlo "bene", questi elementi dovrebbero essere compressi in schegge irrimediabilmente sottili, oppure il tavolo dovrebbe diventare due volte più largo per .

- *Non è troppo fedele alla realtà*: Migliorare la realtà per ottenere mappe migliori è una lunga e onorevole tradizione. Ricordiamo che il progetto Mercator distorce le dimensioni delle terre vicine ai poli (sì, la Groenlandia, sto guardando te) per preservarne la forma.

  I cartogrammi sono mappe distorte in modo che le regioni riflettano una variabile sottostante, come la popolazione. La Figura 6.25 (a destra) riporta i risultati elettorali del 2012 su una mappa in cui l'area di ogni Stato è proporzionale alla sua popolazione/voti elettorali. Solo ora diventa chiara l'entità della vittoria di Obama: quei giganteschi Stati rossi del Midwest si riducono a una dimensione adeguata, producendo una mappa con più blu che rosso.

6.4 Grandi visualizzazioni

Sviluppare la propria estetica di visualizzazione le dà un linguaggio per parlare di ciò che le piace e di ciò che non le piace. Ora la incoraggio ad applicare il suo giudizio per valutare i meriti e i demeriti di alcuni grafici e diagrammi.In questa sezione, esamineremo un gruppo selezionato di visualizzazioni classiche che considero grandiose. Esiste un gran numero di grafiche terribili che mi piacerebbe contrapporre a queste, ma mi è stato insegnato che non è bello prendere in giro le persone. Questo è particolarmente vero quando ci sono restrizioni di copyright sull'uso di tali immagini in un libro come questo.

6.4.1 Orario del treno di Marey

Tufte indica l'orario ferroviario di E.J. Marey come un punto di riferimento nel design grafico. Le ore del giorno sono rappresentate sull'*asse delle ascisse*. In effetti, questo grafico rettangolare è in realtà un cilindro tagliato alle 6 del mattino e appoggiato in piano. L'*asse y* rappresenta tutte le stazioni della linea Parigi-Lione. Ogni linea rappresenta il percorso di un particolare treno, riportando dove dovrebbe trovarsi in ogni momento. I normali orari dei treni sono tabelle, con una colonna per ogni treno, una riga per ogni stazione e la voce (*i, j*) che riporta l'orario di arrivo del treno *j* alla stazione *i*. Tali tabelle sono utili per dirci a che ora arrivare per prendere il nostro treno. Ma la tabella di Marey fornisce molte più informazioni. Cos'altro può vedere qui che non può vedere con le tabelle orarie convenzionali?

- *A che velocità si muove il treno?* La pendenza di una linea misura quanto ripida. Più veloce è il treno, maggiore è la pendenza assoluta. I treni più lenti sono contrassegnati da linee più piatte, perché impiegano più tempo per coprire il terreno dato.

  Un caso particolare è l'identificazione dei periodi in cui il treno è fermo in stazione. In questi momenti, la linea è orizzontale e indica che non c'è movimento lungo la linea.

- *Quando i treni si ?* La direzione di un treno è data pendenza della linea associata. I treni diretti a nord hanno una pendenza positiva, mentre quelli diretti a sud hanno una pendenza negativa. Due treni si incrociano in corrispondenza dei punti di intersezione di due linee, per cui i passeggeri sanno quando guardare fuori dal finestrino e salutare.

- *Quando è l'ora di punta?* C'è una concentrazione di treni in partenza sia da Parigi che da Lione intorno alle 19.00, il che mi dice che quello doveva essere l'orario più popolare per viaggiare. Il viaggio durava in genere circa undici ore, quindi

doveva trattarsi di un treno con vagone letto, dove i viaggiatori arrivavano a destinazione il giorno successivo di buon'ora.

Naturalmente sono presenti anche gli orari di partenza dei treni in una stazione. Ogni stazione è contrassegnata da una linea orizzontale, quindi cerchi l'orario in cui i treni attraversano la sua stazione nella direzione corretta.
Il mio unico dubbio è che sarebbe stato ancora meglio con una griglia di dati più leggera. Non imprigioni mai i suoi dati!

6.4.2 Mappa del colera di Snow

Una mappa di dati particolarmente famosa ha cambiato il corso della storia della medicina. Il colera era una terribile malattia che uccideva un gran numero di persone nelle città del XIX secolo. La peste arrivava all'improvviso e colpiva a morte le persone, la cui causa era un mistero per la scienza dell'epoca.

John Snow ha tracciato i casi di colera di un'epidemia del 1854 su una mappa stradale di Londra, sperando di vedere uno schema. Ogni punto nella Figura 6.28 rappresenta una famiglia colpita dalla malattia. Cosa vede? Snow notò un gruppo di casi centrati su Broad Street. Inoltre, al centro del gruppo c'era una croce, che indicava un pozzo dove i residenti prendevano l'acqua potabile. La fonte dell'epidemia fu rintracciata nella maniglia di una singola pompa dell'acqua. Hanno cambiato la maniglia e improvvisamente le persone hanno smesso di ammalarsi. Questo dimostrò che il colera era una malattia infettiva causata dall'acqua contaminata e indicò la strada per prevenirla.

6.4.3 L'anno meteorologico di New York

Vale quasi la pena sopportare un inverno newyorkese per vedere un particolare grafico che appare sul *New York Times* ogni gennaio, che riassume il clima dell'anno precedente. La Figura 6.29 presenta una rappresentazione indipendente degli stessi dati, che cattura il motivo per cui questo grafico è emozionante. Per ogni giorno dell'anno, vediamo le temperature alte e basse tracciate su un grafico, insieme a dati storici che le contestualizzano: la media giornaliera delle temperature alte e basse, nonché le temperature più alte e più basse mai registrate per quella data.
Che cosa c'è di così bello in questo? Innanzitutto, mostra× 6 365 = 2190 numeri in coerente, il che facilita i confronti sulla curva sinusoidale delle stagioni: Possiamo dire quando ci sono stati periodi di caldo e di freddo e quanto sono durati.

- Possiamo dire quali giorni hanno avuto grandi sbalzi di temperatura e quando il termometro non si è mosso affatto.
- Possiamo dire se il clima è stato insolito quest'anno. È stato un anno insolitamente caldo o freddo, o entrambi? Quando sono stati stabiliti dei record di alte/basse temperature e quando si è arrivati vicini a stabilirli?

Questo singolo grafico è ricco, chiaro e informativo. Si lasci ispirare da esso.

6.5 Leggere i grafici

Quello che si vede non è sempre quello che si ottiene. Ho visto molti grafici portati dai miei studenti nel corso degli anni. Alcuni sono stati straordinari, mentre molti altri sono stati sufficienti per svolgere il lavoro.
Ma vedo anche ripetutamente trame con gli stessi problemi di base. In questa sezione, per alcuni dei miei problemi, presento il diagramma originale e il modo per rimediare al problema. Con l'esperienza, dovrebbe essere in grado di identificare questo tipo di problemi semplicemente osservando il grafico iniziale.

6.5.1 La distribuzione oscurata

La frequenza di 10.000 parole inglesi, ordinate per frequenza. Non sembra molto entusiasmante: tutto ciò che si vede è un singolo punto a (1, 2.5). Cosa è successo e come si può risolvere? Se fissa a lungo la figura, vedrà che in realtà ci sono molti più punti. Ma tutti si trovano sulla linea $y = 0$ e si sovrappongono l'uno all'altro, tanto da formare una massa indifferenziata. Il lettore attento si renderà conto che questo singolo punto è in realtà un outlier, con una magnitudine così grande da ridurre tutti gli altri totali verso lo zero. Una reazione naturale sarebbe quella di eliminare il punto più grande, ma curiosamente i punti rimanenti avranno lo stesso aspetto.
Il problema è che questa distribuzione è una legge di potenza, e tracciare una legge di potenza su una scala lineare non mostra nulla. La chiave qui è tracciare su scala logica. Ora può vedere i punti e la mappatura dai ranghi alla frequenza. Ancora meglio è tracciare su una scala log-log. La linea retta qui conferma che abbiamo a che fare con una legge di potenza.

6.5.2 Interpretare in modo eccessivo la varianza

Nella bioinformatica, si cerca di scoprire come funziona la vita osservando i dati. Confondere la varianza con il segnale: i valori estremi nella figura a sinistra sono artefatti dovuti alla media di un numero ridotto di campioni. abbiamo appena scoperto che l'energia varia inversamente alla lunghezza del gene?No. Abbiamo solo sovra interpretato la varianza. Il primo indizio è che ciò che sembrava molto stabile inizia ad andare in tilt con l'aumento della lunghezza. La maggior parte dei geni ha una lunghezza molto ridotta. Pertanto, i punti sul lato destro del grafico di sinistra si basano su pochi dati. La media di pochi punti non è robusta come la media di molti punti. Infatti, un grafico di frequenza del numero di geni in base alla lunghezza mostra che i conteggi si riducono a zero proprio nel punto in cui inizia il salto.
Come possiamo risolvere il problema? La cosa giusta da fare in questo caso è settare il grafico mostrando solo i valori con un supporto di dati sufficiente, forse con una lunghezza di 500 o poco più. Al di là di questo, potremmo suddividerli in base alla lunghezza e prendere la media, in per dimostrare che l'effetto di salto scompare.

6.6 Visualizzazione interattiva

Le tabelle e i grafici di cui abbiamo parlato finora erano tutte immagini statiche, progettate per essere studiate dall'osservatore, ma non manipolate. Le tecniche di visualizzazione interattiva stanno diventando sempre più importanti per l'analisi esplorativa dei dati. Le app mobili, i notebook e le pagine web con widget di visualizzazione interattiva possono essere particolarmente efficaci nel presentare i dati e nel diffonderli a un pubblico più vasto, che li può esplorare. Offrire agli spettatori la possibilità di giocare con i dati reali aiuta a garantire che la storia presentata da un grafico sia quella vera e completa. Se intende visualizzare i suoi dati online, ha senso farlo utilizzando widget interattivi. In genere si tratta di estensioni dei grafici di base che abbiamo descritto in questo capitolo, con caratteristiche come l'offerta di pop-up con ulteriori informazioni quando l'utente scorre sui punti, oppure incoraggiando l'utente a modificare gli intervalli di scala con i cursori. Ci sono alcuni potenziali svantaggi della visualizzazione interattiva. In primo luogo, è più difficile comunicare esattamente ciò che sta vedendo agli altri, rispetto alle immagini statiche. Potrebbero non vedere esattamente la stessa cosa che vede lei. Le schermate dei sistemi interattivi in genere non sono paragonabili ai grafici di qualità editoriale ottimizzati sui sistemi tradizionali. Il problema dei sistemi "what you see is what you get" (WYSIWYG) è che, in genere, ciò che si vede è tutto ciò che si ottiene. I sistemi interattivi sono migliori

per l'esplorazione, non per la presentazione. Ci sono anche degli eccessi che tendono a verificarsi nella visualizzazione interattiva. Manopole e funzioni vengono aggiunte perché possono farlo, ma gli effetti visivi che aggiungono possono distrarre piuttosto che aggiungere al messaggio. La rotazione delle nuvole di punti tridimensionali è sempre bella, ma trovo che siano difficili da interpretare e molto raramente trovo queste visualizzazioni di utili. Per far sì che i dati raccontino una storia, è necessario capire come si raccontano le storie. I film e la televisione rappresentano lo stato dell'arte della presentazione narrativa. Come loro, le migliori presentazioni interattive mostrano una narrazione, come lo spostamento nel tempo o la ricerca di ipotesi alternative. Per trarre ispirazione, la invito a guardare il discorso TED del compianto Han Rosling che ha utilizzato una bolla animata. Le recenti tendenze nella visualizzazione basata sul cloud la incoraggiano a caricare i suoi dati su un sito come Google Charts, in modo da poter sfruttare gli strumenti di visualizzazione interattiva e i widget che essi forniscono. Questi strumenti producono grafici interattivi molto belli e sono facili da usare. Il possibile punto critico è la sicurezza, dal momento che si stanno fornendo i propri dati a una terza parte. Spero e confido che gli analisti della CIA abbiano accesso alle loro soluzioni interne che mantengono la riservatezza dei loro dati. Ma si senta libero di sperimentare questi strumenti in situazioni meno sensibili.

6.7 Storia di guerra: Mappare il mondo con il testo

La mia più grande esperienza nella visualizzazione dei dati è nata dal nostro sistema di analisi delle notizie/sentimenti su larga scala. TextMap presentava un cruscotto di analisi delle notizie, per ogni entità il cui nome compariva negli articoli di cronaca dell'epoca. La pagina di Barack Obama. Il nostro cruscotto era composto da una serie di sottocomponenti:

- *Serie temporale di riferimento*: Qui abbiamo riportato la frequenza con cui la data entità è apparsa nelle notizie, in funzione del tempo. I picchi corrispondevano a eventi di cronaca più importanti. Inoltre, abbiamo suddiviso questi conteggi in base a alle apparizioni in ogni sezione del giornale: notizie, sport, intrattenimento o affari.
- *Distribuzione del settore delle notizie*: Questo grafico contiene *esattamente* gli stessi dati della serie temporale dei riferimenti, ma presentati come un grafico ad area sovrapposta per mostrare la frazione di riferimenti alle entità in ogni sezione del giornale. Obama si presenta chiaramente come un personaggio di notizie. Ma altre persone hanno mostrato transizioni interessanti, come Arnold Schwarzenegger quando è passato dalla recitazione alla politica.

- *Analisi del sentimento*: qui presentiamo una serie temporale che misura il sentimento dell'entità, presentando la differenza normalizzata del numero di menzioni di notizie positive rispetto al totale delle referenze. In questo modo, lo zero rappresenta una reputazione neutra e abbiamo fornito una linea di riferimento centrale per mettere in prospettiva il posizionamento. In questo caso, il sentimento di Obama oscilla con gli eventi, ma in genere rimane sul lato destro della linea.

  È stato bello vedere le cose brutte accadere alle persone brutte, osservando il calo del sentimento delle notizie. Se ricordo bene, la notizia più bassa mai raggiunta è stata quella di una mamma che ha raggiunto la notorietà grazie al cyberbullismo che ha spinto una delle rivali sociali di sua figlia a suicidarsi.

- *Heatmap*: Qui abbiamo presentato una mappa di dati che mostra la frequenza di riferimento relativa dell'entità. La grande macchia rossa di Obama intorno all'Illinois è dovuta al fatto che era in carica come senatore di quella regione al momento della sua prima candidatura alla presidenza. Molte classi di entità hanno mostrato forti pregiudizi regionali, come i personaggi dello sport e i politici, ma meno i personaggi dell'intrattenimento come le star del cinema e i cantanti.

- *Giustapposizioni e rete relazionale*: Il nostro sistema ha costruito una rete sulle entità delle notizie, collegando due entità ogni volta che venivano menzionate nello stesso articolo. La forza di questa relazione può essere misurata dal numero di articoli che le collegano. Le associazioni più forti sono riportate come giustapposizioni, con la loro frequenza e forza mostrata da un grafico a barre su una scala logaritmica.

  Finora non abbiamo parlato molto della visualizzazione della rete, ma l'idea chiave è quella di posizionare i vertici in modo che i vicini si trovino vicini, il che significa che i bordi sono corti. Le amicizie più forti sono indicate da linee più spesse.

- *Articoli correlati*: Abbiamo fornito dei link ad articoli di notizie rappresentativi che menzionano l'entità in questione.

Una carenza del nostro cruscotto era che non era interattivo. In effetti, era anti interattivo: l'unica animazione era una gif del mondo che girava in modo forzato nell'angolo in alto a sinistra. Molti dei grafici che abbiamo renderizzato richiedevano grandi quantità di accesso ai dati dal nostro goffo database, in particolare la mappa di calore. Poiché non potevano essere renderizzate in modo interattivo, abbiamo precompilato queste mappe offline e le abbiamo mostrate su richiesta. Dopo che Mikhail ha aggiornato la nostra infrastruttura, è diventato possibile supportare alcuni

grafici interattivi, in particolare le serie temporali. Abbiamo sviluppato una nuova interfaccia utente chiamata TextMap Access, che consente agli utenti di giocare con i nostri dati. Ma quando General Sentiment ha concesso in licenza la tecnologia, la prima cosa che ha disconosciuto è stata la nostra interfaccia utente. Non aveva senso per gli analisti aziendali che erano i nostri clienti. Era troppo complicata. C'è una differenza sostanziale tra il fascino superficiale e la reale efficacia di un'interfaccia per gli utenti. Guardi attentamente la nostra dashboard TextMap: aveva i pulsanti "Cos'è questo?" in diverse posizioni. Questo era un segno di debolezza: se la nostra grafica fosse stata davvero abbastanza intuitiva, non ne avremmo avuto bisogno. Anche se ho brontolato per la nuova interfaccia, senza dubbio mi sbagliavo. General Sentiment aveva un gruppo di analisti che usavano il nostro sistema tutto il giorno ed erano disponibili a parlare con i nostri sviluppatori. Presumibilmente l'interfaccia si è evoluta per servire meglio loro. Le migliori interfacce sono costruite da un dialogo tra sviluppatori e utenti. Quali sono le lezioni da trarre da questa esperienza? Trovo ancora molto interessante questa dashboard: la presentazione era abbastanza ricca da mettere in luce cose interessanti sul funzionamento del mondo. Ma non tutti erano d'accordo. Clienti diversi preferivano interfacce diverse, perché facevano cose diverse con esse. Una lezione da trarre è il potere di fornire visioni alternative degli stessi dati. La serie temporale di riferimento e la distribuzione del settore delle notizie erano esattamente gli stessi dati, ma hanno fornito intuizioni molto diverse quando sono stati presentati in modi diversi. Tutte le canzoni sono fatte con le stesse note, ma il modo in cui le si dispone diversi tipi di musica.

Capitolo 7
Modelli matematici

La *modellazione* è il processo di incapsulamento delle informazioni in uno strumento in grado di prevedere e fare previsioni. I modelli predittivi sono strutturati intorno a un'idea di ciò che fa accadere gli eventi futuri. L'estrapolazione dalle tendenze e dalle osservazioni recenti presuppone una visione del mondo secondo cui il futuro sarà come il passato. I modelli più sofisticati, come le leggi della fisica, forniscono nozioni di principio sulla causalità; i modelli fondamentali
spiegazioni sul perché cose.
Questo capitolo si concentrerà sulla progettazione e sulla validazione dei modelli. Formulare efficacemente i modelli richiede una comprensione dettagliata dello spazio

delle scelte possibili. Valutare con precisione le prestazioni di un modello può essere sorprendentemente difficile, ma è essenziale per sapere come interpretare le previsioni risultanti. Il miglior sistema di previsione non è necessariamente più accurato, ma il modello con il miglior senso dei suoi confini e limiti.

7.1 Valutare i modelli

Lo *sniff test* informale è forse il criterio più importante per valutare un modello.
Le valutazioni formali che verranno dettagliate di seguito riducono le prestazioni di un modello a poche statistiche sommarie, aggregate su molte istanze. Ma molti difetti di un modello possono essere nascosti quando si interagisce solo con questi punteggi aggregati. Non ha modo di sapere se ci sono bug nell'implementazione o nella normalizzazione dei dati, che hanno portato a prestazioni inferiori a quelle che avrebbe dovuto avere. Forse ha mescolato i dati di addestramento e di test, ottenendo punteggi molto migliori sul suo banco di prova di quanto meritasse. Per sapere davvero cosa sta succedendo, è necessario fare uno sniff test. Il mio sniff test personale consiste nell'esaminare attentamente alcuni esempi di casi in cui il modello ha ottenuto il risultato giusto e altri in cui ha sbagliato. L'obiettivo è assicurarmi di capire perché il modello ha ottenuto i risultati che ha ottenuto. Idealmente si tratterà di record di cui si conoscono i 'nomi', istanze in cui si ha una certa intuizione su quali dovrebbero essere le risposte giuste, come risultato dell'analisi esplorativa dei dati o della familiarità con il dominio.

Un'altra questione è il suo grado di sorpresa per l'accuratezza valutata del modello. Le prestazioni sono migliori o peggiori di quelle che si aspettava? Quanto pensa che sarebbe preciso nel compito dato, se dovesse usare il giudizio umano.

Una questione correlata è stabilire un senso di quanto sarebbe prezioso se il modello funzionasse solo un po' meglio. Un compito NLP che classifica correttamente le parole con un'accuratezza del 95% commette un errore circa una volta ogni due o tre frasi. È abbastanza buono? Quanto migliori sono le prestazioni attuali, tanto più difficile sarà apportare ulteriori miglioramenti.

Ma il modo migliore per valutare i modelli riguarda le previsioni *fuori campione*, ovvero i risultati su dati che non sono mai stati visti (o meglio ancora, non esistevano) quando si è costruito il modello. Una buona performance sui dati su cui ha addestrato i modelli è molto sospetta, perché i modelli possono essere facilmente sovra-adattati. Le previsioni fuori campione sono la chiave per essere onesti, a condizione che si disponga di dati sufficienti e di tempo per testarli.

7.1.1 Valutazione dei classificatori

Valutare un classificatore significa misurare la precisione con cui le nostre etichette previste corrispondono alle etichette gold standard nel set di valutazione. Nel caso comune di due etichette o classi distinte (classificazione binaria), in genere chiamiamo la più piccola e più interessante delle due classi come *positiva* e la più grande/altra classe come *negativa*. In un problema di classificazione dello spam, lo spam sarebbe tipicamente positivo e l'ham (non spam) sarebbe negativo. Questa etichettatura mira a garantire che
I quattro possibili risultati nella matrice di confusione riflettono quali istanze sono state classificate correttamente (TP e TN) e quali no (FN e FP). L'identificazione dei positivi è difficile almeno quanto quella dei negativi, anche se spesso le istanze di test sono selezionate in modo che le classi abbiano la stessa cardinalità. Ci sono quattro possibili risultati di ciò che il modello di classificazione potrebbe fare su una determinata istanza, che definisce la *matrice di confusione* o *tabella di contingenza*:

- *Veri positivi (TP)*: in questo caso il nostro classificatore etichetta un elemento positivo come positivo, con conseguente vittoria del classificatore.

- *Veri negativi (TN)*: qui il classificatore determina correttamente che un membro della classe negativa merita un'etichetta negativa. Un'altra vittoria.

- *Falsi positivi (FP)*: il classificatore chiama erroneamente un elemento negativo come positivo, dando luogo a un errore di classificazione di "tipo I".

- *Falsi negativi (FN)*: Il classificatore dichiara erroneamente un elemento positivo come negativo, dando luogo a un errore di classificazione di "tipo II".

Accuratezza, precisione, richiamo e F-Score

Esistono diverse statistiche di valutazione che possono essere calcolate dai conteggi positivi/negativi veri/falsi descritti sopra. Il motivo per cui abbiamo bisogno di così tante statistiche è che dobbiamo difendere il nostro classificatore contro due avversari di base, l'acuto e la scimmia.
L'*affilato* è l'avversario che conosce il sistema di valutazione che stiamo utilizzando e sceglie il modello di base che farà meglio in base ad esso. L'acuto cercherà di mettere in cattiva luce la statistica di valutazione, ottenendo un punteggio elevato con un classificatore inutile. Ciò potrebbe significare dichiarare tutti gli elementi positivi, o forse tutti negativi. Per interpretare le prestazioni del nostro modello, è importante stabilire di quanto batte sia l'acuto. La prima statistica misura l'*accuratezza* del classificatore, il rapporto tra il numero di predizioni corrette e le predizioni totali. Moltiplicando queste frazioni per 100, possiamo ottenere un punteggio di precisione percentuale. L'accuratezza è un numero ragionevole e relativamente facile da spiegare; quindi, vale la pena fornirlo in qualsiasi ambiente di valutazione. Quanto è accurata, quando metà delle istanze sono positive e metà negative? dovrebbe raggiungere un'accuratezza del 50% indovinando a caso. La stessa precisione del 50% sarebbe ottenuta dall'acuto, indovinando sempre il positivo o (equivalentemente) indovinando sempre il negativo. L'acuto otterrebbe una diversa metà delle istanze corrette in ogni caso. Tuttavia, l'accuratezza da sola ha dei limiti come metrica di valutazione, soprattutto quando la classe positiva è molto più piccola di quella negativa. Consideriamo lo sviluppo di un classificatore per diagnosticare se un paziente ha il cancro, dove la classe positiva ha la malattia (cioè risulta positiva) e la classe negativa è sana. La distribuzione precedente prevede che la stragrande maggioranza delle persone sia sana. L'accuratezza prevista equa sarebbe ancora di $0{,}5$: dovrebbe azzeccare in media la metà dei positivi e la metà dei negativi. Ma l'acuto dichiarerebbe che tutti sono sani, ottenendo un'accuratezza di $1\ p$. Supponiamo che solo il 5% dei partecipanti al test abbia davvero la malattia. L'acuta potrebbe vantarsi della sua precisione del 95%, condannando contemporaneamente tutti i membri della classe malata a una morte precoce. Pertanto, abbiamo bisogno di metriche di valutazione che siano più sensibili alla correttezza della classe positiva. *La precisione* misura la frequenza con cui questo classificatore è corretto quando osa dire positivo. Raggiungere un'elevata precisione è impossibile sia per un acuto, perché la frazione di positivi ($p = 0{,}05$) è così bassa. Se il classificatore emette troppe etichette positive, è condannato a una bassa precisione, perché molti proiettili mancano il bersaglio, dando luogo a molti falsi positivi. Ma se il classificatore è avaro etichette positive, è probabile che poche di esse si colleghino al raro positivo. istanze, quindi il classificatore ottiene un basso numero di veri positivi. Questi classificatori di base ottengono una precisione proporzionale alla probabilità di classe positiva $p = 0{,}05$, perché sono alla cieca. Nel caso della diagnosi del cancro, potremmo essere più disposti a tollerare i falsi posi- tivi (errori in cui spaventiamo una persona sana con una diagnosi sbagliata) che i falsi negativi (errori in cui uccidiamo un paziente malato sbagliando la diagnosi della sua malattia). *Il richiamo* misura la frequenza con cui si dimostra di avere ragione su tutte le istanze positive. Un richiamo elevato implica che il classificatore ha pochi falsi negativi. Il modo più semplice per ottenere questo risultato è dichiarare che *tutti* hanno il cancro, come avviene con una risposta sempre affermativa. Questo

classificatore ha un alto richiamo ma una bassa precisione: il 95% dei partecipanti al test riceverà uno spavento inutile. C'è un compromesso intrinseco tra precisione e richiamo quando si costruiscono i classificatori: più coraggiose sono le previsioni, meno è probabile che siano giuste. Ma le persone sono fortemente predisposte a desiderare un'unica misurazione che descriva le prestazioni del loro sistema. Il *punteggio F* (o talvolta il punteggio F1) è combinazione di questo tipo, che restituisce la media armonica di precisione e richiamo. Il punteggio F è una misura molto difficile da battere. La media armonica è sempre inferiore o uguale alla media aritmetica, e il numero più basso ha una sproporzione... effetto di grandi dimensioni. Il raggiungimento di un punteggio F elevato richiede sia un richiamo che una precisione elevati. Nessuno dei nostri classificatori di base riesce a ottenere un F-score decente, nonostante i valori elevati di accuratezza e richiamo, perché la loro precisione è troppo bassa. Il punteggio F e le relative metriche di valutazione sono state sviluppate per valutare i classificatori medi. Per capire come interpretarli, consideriamo una classe di classificatori magicamente *equilibrati*, che in qualche modo mostrano la stessa precisione sia sulle istanze positive che su quelle negative. Di solito non è così, ma i classificatori selezionati per ottenere punteggi F elevati devono bilanciare le statistiche di precisione e di richiamo, il che significa che devono mostrare prestazioni decenti sia sulle istanze positive che su quelle negative. Le lezioni da trarre sono:

- *La precisione è una statistica fuorviante quando le dimensioni delle classi sono sostanzialmente diverse*: Un classificatore di base che rispondeva "no" per ogni posizione in- terna ha ottenuto un'accuratezza del 95% sul problema del cancro, migliore anche di un classificatore bilanciato che ha azzeccato il 94% di ogni classe.

- *Il richiamo equivale all'accuratezza se e solo se i classificatori sono bilanciati* : Si verificano buone cose quando l'accuratezza per il riconoscimento di entrambe le classi è la stessa. Questo non accade automaticamente durante l'addestramento, quando le dimensioni delle classi sono diverse. Infatti, questo è uno dei motivi per cui è generalmente una buona pratica avere un numero uguale di esempi positivi e negativi nel set di formazione.

- *È molto difficile ottenere una precisione elevata in classi non bilanciate* : Persino un classificatore bilanciato che ottiene il 99% di precisione su esempi positivi e negativi non può ottenere una precisione superiore all'84% sul problema del cancro. Questo perché ci sono venti volte più istanze negative che positive. I falsi positivi derivanti dall'errata classificazione della classe più grande a un tasso dell'1% rimangono sostanziali sullo sfondo del 5% di veri positivi.

- *Il punteggio F fa il lavoro migliore di qualsiasi altra statistica, ma tutte e quattro lavorano insieme per descrivere le prestazioni di un classificatore*: La precisione del suo classificatore è superiore al suo richiamo? Allora sta etichettando un numero troppo basso di istanze come positive, e quindi forse può metterlo a punto meglio. Il richiamo è superiore alla precisione? Forse possiamo migliorare l'Fscore essendo meno aggressivi nel chiamare i positivi. La precisione è lontana dal richiamo? Allora il nostro classificatore non è molto equilibrato. Quindi verifichi quale lato sta andando peggio e come possiamo risolverlo.

Un trucco utile per aumentare la precisione di un modello a scapito del richiamo è quello di dargli il potere di dire "non so". I classificatori in genere fanno meglio nei casi facili che in quelli difficili, con la difficoltà definita da quanto l'esempio è lontano dall'assegnazione dell'etichetta alternativa. Definire una nozione *di fiducia* che la classificazione proposta sia corretta è la chiave per capire quando è il caso di lasciar perdere una domanda. Azzardi un'ipotesi solo quando la sua fiducia è superiore a una determinata soglia. I pazienti i cui punteggi del test sono vicini al limite preferirebbero in genere una diagnosi di "risultato borderline" piuttosto che "ha il cancro", soprattutto se il classificatore non è molto sicuro della sua decisione. La curva ROC ci aiuta a selezionare la soglia migliore da utilizzare in un classificatore, mostrando il compromesso tra veri positivi e falsi positivi in ogni possibile impostazione. Le nostre statistiche di precisione e di richiamo devono essere riconsiderate per

correttamente alla nuova classe indeterminata. Non è necessario modificare la formula della precisione: valutiamo solo le istanze che definiamo positive. Ma il denominatore per il richiamo deve tenere conto esplicitamente di tutti gli elementi che abbiamo rifiutato di etichettare. Supponendo di essere accurati nelle nostre misure di fiducia, la precisione aumenterà a spese del richiamo.

7.1.2 Curve caratteristiche del ricevitore-operatore (ROC)

Molti classificatori sono dotati di manopole naturali che possono essere modificate per alterare il trade-off tra precisione e richiamo. Ma dove tracciamo il confine tra positivo e negativo? Se il nostro punteggio "in classe" è accurato, allora dovrebbe essere generalmente più alto per gli articoli positivi rispetto a quelli negativi. Gli esempi positivi definiranno una distribuzione di punteggio diversa rispetto alle istanze negative. Sarebbe fantastico se queste distribuzioni fossero completamente disgiunte, perché allora ci sarebbe una soglia di punteggio t tale che tutte le istanze con punteggio t sono positive e tutte $< t$ sono negative. Questo definirebbe un classificatore perfetto. Ma è più probabile che le due distribuzioni si sovrappongano, almeno in una certa misura, trasformando il problema dell'identificazione della soglia migliore in un giudizio basato sulla nostra avversione relativa ai falsi positivi e ai falsi negativi. La curva ROC *(Receiver Operating Characteristic*) fornisce una rappresentazione visiva del nostro spazio completo di opzioni nel mettere insieme un classificatore. Il punto su questa curva rappresenta una particolare soglia di classificatore, definita dai suoi tassi di falsi positivi e falsi negativi. Questi tassi sono a loro volta definiti dal conteggio degli errori diviso per il numero totale di positivi nei dati di valutazione, e forse moltiplicati per cento per trasformarli in percentuali. Consideri cosa succede quando scorriamo la nostra soglia da sinistra a destra su queste distribuzioni. Ogni volta che passiamo sopra un altro esempio, aumentiamo il numero di veri positivi (se questo esempio era positivo) o di falsi positivi (se questo esempio era in realtà negativo). All'estrema sinistra, otteniamo tassi di veri/falsi positivi dello 0%, poiché il classificatore non ha etichettato nulla come positivo a quel cutoff. Spostandosi il più a destra possibile, tutti gli esempi saranno etichettati positivamente e quindi entrambi i tassi diventeranno del 100%. Ogni soglia intermedia definisce un possibile classificatore, e lo sweep definisce una curva a scala nello spazio dei tassi positivi veri/falsi che ci porta da (0%,0%) a (100%,100%). Supponiamo che la funzione di punteggio sia definita da una scimmia, cioè un valore di dominio arbitrario per ogni istanza. Quando scorriamo la nostra soglia verso destra, l'etichetta dell'esempio successivo dovrebbe essere positiva o negativa con la stessa probabilità. In questo modo abbiamo la stessa probabilità di aumentare il nostro tasso di veri positivi e falsi, e la curva ROC dovrebbe percorrere la diagonale principale. Fare meglio della scimmia implica una curva ROC che si trova sopra la diagonale. La migliore curva ROC possibile sale immediatamente da (0%,0%) a (0%,100%), cioè incontra tutte le istanze positive prima di quelle negative. Poi si sposta verso destra con ogni esempio negativo, fino a raggiungere l'angolo superiore destro. *L'area sotto la curva ROC* (AUC) è spesso utilizzata come statistica che misura la qualità della funzione di punteggio che definisce il classificatore. La migliore curva ROC possibile ha un'area pari a 100% 100% 1, mentre il triangolo della scimmia ha un'area pari a 1/2. Più l'area è vicina a 1, migliore la nostra funzione di classificazione.

7.1.3 Valutazione dei sistemi multiclasse

Molti problemi di classificazione sono non binari, ossia devono decidere tra più di due classi. Pertanto, il classificatore di articoli che regola il comportamento di questo sito deve assegnare ad ogni articolo un'etichetta tra otto classi diverse. Più sono le possibili etichette di classe, più è difficile azzeccare la classificazione. L'accuratezza prevista di una scimmia di classificazione con d etichette è $1/d$, quindi l'accuratezza diminuisce rapidamente con l'aumento della complessità della classe. Questo rende la valutazione corretta dei classificatori multiclasse una sfida, perché i numeri di successo bassi diventano scoraggianti. Una statistica migliore è la *percentuale di successo top-k*, che generalizza l'accuratezza per un valore specifico di k 1. Quanto spesso l'etichetta giusta era tra le prime k possibilità? Questa misura è buona, perché ci dà un credito parziale per essersi avvicinati alla risposta giusta. Il

grado di approssimazione è definito dal parametro k. Per k= 1, questo si riduce alla precisione. Per k= d, *qualsiasi* etichetta possibile è sufficiente e il tasso di successo è del 100% per definizione. I valori tipici sono 3, 5 o 10: abbastanza alti da permettere a un buon classificatore di raggiungere un'accuratezza superiore al 50% e di essere visibilmente migliore della scimmia. Ma non troppo migliore, perché una valutazione efficace dovrebbe lasciarci un margine sostanziale per fare meglio. In effetti, è una buona pratica calcolare il tasso top k per tutti i k da 1 a d, o almeno abbastanza alto da rendere il compito facile. Uno strumento di valutazione ancora più potente è la *matrice di confusione* C, ×una matrice d d dove $C[x, y]$ riporta il numero (o la frazione) di istanze della classe x che vengono etichettate come classe y. Come si legge una matrice di confusione? È tratta dall'ambiente di valutazione che abbiamo costruito per testare un classificatore di datazione dei documenti, che analizza i testi per prevedere il periodo di paternità. Tale datazione dei documenti sarà l'esempio continuo di valutazione nel resto di questo capitolo. La caratteristica più importante è la diagonale principale, $C[i, i]$, che conta quanti (o quale frazione di) articoli della classe i sono stati etichettati correttamente come classe i. Speriamo in una diagonale principale forte nella nostra matrice. Il nostro è un compito difficile. La matrice mostra una diagonale principale forte ma non perfetta. Ci sono diversi punti in cui i documenti sono più frequentemente classificati nel periodo vicino rispetto a quello corretto.Ma le caratteristiche più interessanti della matrice di confusione sono i grandi conteggi $C[i, j]$ che *non* si trovano lungo la diagonale principale. Questi rappresentano classi comunemente confuse. Nel nostro esempio, la matrice mostra un numero preoccupantemente alto di documenti (6%)
del 1900 classificati come 2000, mentre nessuno è classificato come 1800. Tali asimmetrie suggeriscono indicazioni per migliorare il classificatore. Ci sono due possibili spiegazioni per le confusioni di classe. La prima è un bug nel classificatore, il che significa che dobbiamo lavorare di più per fargli distinguere i da j. Ma la seconda implica l'umiltà, la consapevolezza che le classi i e j possono sovrapporsi a tal punto che non è ben definito quale dovrebbe essere la risposta giusta. La confusione frequente potrebbe suggerire di unire le due categorie, in quanto rappresentano una differenza senza distinzione. Le righe rade nella matrice di confusione indicano classi scarsamente rappresentate nei dati di formazione, mentre le colonne rade indicano etichette che il classificatore è riluttante ad assegnare. Entrambe le indicazioni sono un'argomentazione per cui forse dovremmo prendere in considerazione la possibilità di abbandonare questa etichetta e di unire le due categorie simili. Le righe e le colonne della matrice di confusione forniscono statistiche di performance analoghe a quelle della Sezione 7.4.1per le classi multiple, parametrizzate per classe. Precisione $_i$ è la frazione di tutti gli articoli dichiarati di classe i che erano effettivamente di classe i. Recall$_i$ è la frazione di tutti i membri della classe i che sono stati correttamente identificati come tali.

7.1.4 Valutazione dei modelli di previsione del valore

I problemi di previsione dei valori possono essere considerati come compiti di classificazione, ma su un numero infinito di classi. Tuttavia, esistono modi più diretti per valutare i sistemi di regressione, basati sulla distanza tra i valori previsti e quelli reali.

Statistiche sugli errori

Per i valori numerici, l'*errore* è una funzione della differenza tra una previsione y'= f (x) e il risultato effettivo y. Misurare le prestazioni di un sistema di previsione del valore comporta due decisioni: (1) fissare la funzione di errore individuale specifica e (2) selezionare la statistica che meglio rappresenta la distribuzione completa dell'errore. Le scelte principali per la funzione di errore individuale includono:

• *Errore assoluto*: Il valore $\Delta = y - y$ ha ´il pregio di essere semplice. e simmetrico, quindi il segno può distinguere il caso in cui $y' > y$ da $y > y'$. Il problema si presenta quando si aggregano questi valori in una statistica riassuntiva. Errori di compensazione come 1- e 1 significano che il sistema è perfetto? In genere, il valore assoluto dell'errore viene preso per cancellare il segno.
- *Errore relativo*: La grandezza assoluta dell'errore è priva di significato senza un senso delle unità coinvolte. Un errore assoluto di 1,2 nell'altezza prevista di una persona è buono se misurato in millimetri, ma terribile se misurato in miglia. Normalizzando l'errore in base alla grandezza dell'osservazione, si ottiene una quantità senza unità, che può essere interpretata sensibilmente come una frazione o (moltiplicata per il 100%) come una percentuale. L'errore assoluto valuta le istanze con valori più grandi di y come più importanti di quelle più piccole, un pregiudizio corretto quando si calcolano gli errori relativi.

• *Errore quadratico*: Il valore Δ è sempre positivo. questi valori possono essere sommati in modo significativo. I valori di errore più grandi contribuiscono in modo sproporzionato al totale quando si effettua la quadratura: Δ. Pertanto, i valori anomali possono facilmente arrivare a dominare la statistica dell'errore in un ensemble di grandi dimensioni. È un'ottima idea tracciare un istogramma della distribuzione dell'errore assoluto per qualsiasi predittore di valore, poiché si può imparare molto da esso. La distribuzione *dovrebbe* essere simmetrica e centrata intorno allo zero. *Dovrebbe essere* a forma di campana, il che significa che gli errori piccoli sono più comuni di grandi. E gli errori estremi *dovrebbero* essere rari. Se una delle condizioni è sbagliata, probabilmente esiste un modo semplice per migliorare la procedura di previsione. Ad esempio, se non è centrato intorno allo zero, l'aggiunta di un offset costante a tutte le previsioni migliorerà i risultati del consenso. Abbiamo bisogno di una statistica riassuntiva che riduca tali distribuzioni di errori a un unico numero, per poter confrontare le prestazioni di diverse modalità di predizione dei valori. Una statistica comunemente utilizzata è l'*errore quadratico medio* (MSE). Poiché pesa ogni termine in modo quadratico, i valori anomali hanno un effetto sproporzionato. Pertanto, l'errore quadratico mediano potrebbe essere una statistica più informativa per le istanze rumorose. L'errore quadratico medio (RMSD) è semplicemente la radice quadrata dell'errore quadratico medio: Il vantaggio dell'RMSD è che la sua grandezza è interpretabile sulla stessa scala dei valori originali, così come la deviazione standard è una quantità più interpretabile della varianza. Tuttavia, questo non elimina il problema che gli elementi anomali possono alterare in modo sostanziale il totale.

7.2 Ambienti di valutazione

Una parte sostanziale di qualsiasi progetto di scienza dei dati ruota intorno alla costruzione di un ambiente di valutazione ragionevole. In particolare, è necessario un *programma con un solo comando* per eseguire il modello sui dati di valutazione e produrre grafici/rapporti sulla sua efficacia. Perché un comando singolo? Se non è facile da eseguire, non lo proverà abbastanza spesso. Se i risultati non sono facili da leggere e da interpretare, non otterrà abbastanza informazioni da far valere lo sforzo. L'input di un ambiente di valutazione è un insieme di istanze con i risultati/etichette di output associati, oltre a un modello da testare. Il sistema esegue il modello su ogni istanza, confronta ogni risultato con questo gold standard e produce statistiche di riepilogo e grafici di distribuzione che mostrano le prestazioni ottenute su questo set di test. Un buon sistema di valutazione ha le seguenti proprietà:

- Produce distribuzioni di errori oltre ai risultati binari: quanto era vicina la sua previsione, non solo se era giusta o sbagliata.

- Produce automaticamente un rapporto con tracciati multipli su diverse dis- tribuzioni di ingresso, da leggere con attenzione a suo piacimento.

- Emette le statistiche di riepilogo rilevanti sulle prestazioni, in modo da poter valutare rapidamente la qualità. Sta facendo meglio o peggio dell'ultima volta?

Ricordiamo che il compito è quello di prevedere l'anno di paternità di un dato documento dall'uso delle parole. Cosa vale la pena notare?

- *Set di test suddivisi per tipo*: Osserviamo che l'ambiente di valutazione ha suddiviso gli input in nove sottoinsiemi separati, alcuni di notizie e altri di narrativa, e di lunghezza compresa tra 100 e 2000 parole. Così, a colpo d'occhio, possiamo vedere separatamente quanto siamo bravi in ciascuno di essi.

- *Progressioni logiche di difficoltà*: È ovviamente più difficile determinare l'età da documenti più brevi rispetto a quelli più lunghi. Separando i casi più difficili da quelli più piccoli, comprendiamo meglio la nostra fonte di errori. Vediamo un grande miglioramento in naive Bayes quando passiamo da 100 a 500 parole, ma questi guadagni si saturano prima di 2000 parole.

- *Statistiche appropriate al problema*: Non abbiamo stampato tutte le possibili metriche di errore, ma solo l'errore assoluto medio e mediano e l'accuratezza (quante volte abbiamo azzeccato l'anno?). Questi dati sono sufficienti per capire che le notizie sono più facili della finzione, che il nostro modello è molto migliore della scimmia e che le nostre possibilità di identificare correttamente l'anno reale (misurate dalla precisione) sono ancora troppo basse per .

Questa valutazione ci fornisce le informazioni di cui abbiamo bisogno per vedere come stiamo andando, senza sovraccaricarci di numeri che non mai veramente.

7.2.1 Igiene dei dati per la valutazione

Una valutazione è significativa solo quando non si inganna se stessi. Succedono cose terribili quando le persone valutano i loro modelli in modo indisciplinato, perdendo la distinzione tra dati di formazione, test e valutazione. Quando prende posizione su un set di dati con l'intenzione di costruire un modello predittivo, la sua prima operazione dovrebbe essere quella di suddividere l'input in tre parti:

- *Dati di allenamento*: Questo è ciò con cui è completamente libero di giocare. Li utilizzi per studiare il dominio e impostare i parametri del suo modello. In genere, circa il 60% dell'intero set di dati dovrebbe essere dedicato all'addestramento.

- *Dati di test*: Comprendendo circa il 20% del set di dati completo, questo è ciò che si utilizza per valutare la bontà del modello. In genere, si sperimentano più approcci di apprendimento automatico o impostazioni di parametri di base, quindi i test consentono di stabilire le prestazioni relative di tutti questi modelli diversi per lo stesso compito.

  Il test di un modello di solito rivela che non sta funzionando bene come vorremmo, innescando così un altro ciclo di progettazione e perfezionamento. Una scarsa performance sui dati di prova rispetto a quella ottenuta sui dati di addestramento suggerisce un modello che è stato adattato in modo eccessivo.

- *Dati di valutazione*: Il 20% finale dei dati dovrebbe essere messo da parte per un giorno di pioggia: per confermare le prestazioni del modello finale prima che venga messo in produzione. Questo funziona solo se non si aprono i dati di valutazione fino a quando non sono rea lmente necessari.

La ragione per applicare queste separazioni dovrebbe essere ovvia. Gli studenti farebbero molto meglio agli esami se avessero accesso in alla chiave delle risposte, perché saprebbero esattamente cosa studiare. Ma questo non rifletterebbe quanto hanno effettivamente imparato. Mantenere i dati dei test separati dalla formazione fa sì che i test misurino qualcosa di importante su ciò che il modello comprende. Inoltre, tenere

i dati di valutazione finale da utilizzare solo dopo che il modello si è stabilizzato, assicura che le specificità del set di test non siano trapelate nel modello attraverso ripetute iterazioni di test. Il set di valutazione serve come dati fuori campione per convalidare il modello finale. Nell'eseguire il partizionamento originale, bisogna fare attenzione a non creare artefatti non desiderabili o distruggere quelli desiderabili. La semplice suddivisione del file nell'ordine che è stato dato è pericolosa, perché qualsiasi differenza strutturale tra le popolazioni del corpus di formazione e di prova significa che il modello non funzionerà bene come dovrebbe. Ma supponiamo che lei stia costruendo un modello per prevedere i prezzi futuri delle azioni. Sarebbe pericoloso selezionare casualmente il 60% dei campioni di tutta la storia come dati di addestramento, invece di tutti i campioni del primo 60% del tempo. Perché? Supponiamo che il suo modello "impari" quali sono i giorni di rialzo e di ribasso del mercato dai dati di addestramento, e poi utilizzi questa intuizione per fare previsioni virtuali su altri titoli in questi stessi giorni. Questo modello otterrebbe risultati molto migliori nei test che nella pratica. Le tecniche di campionamento corrette sono piuttosto sottili. È essenziale mantenere il velo di ignoranza sui dati di valutazione il più a lungo possibile, perché li si rovina non appena li si utilizza. Le barzellette non sono mai divertenti la seconda volta che le si ascolta, dopo che si conosce già la battuta finale. Se si consuma l'integrità dei set di test e di valutazione, la soluzione migliore è quella di partire da dati nuovi, fuori dal campione, ma questo non è sempre disponibile. In caso contrario, è possibile ripartire in modo casuale l'intero set di dati in nuovi campioni di formazione, test e valutazione, e riqualificare tutti i suoi modelli da zero per riavviare il processo. Ma questo dovrebbe essere riconosciuto come un risultato infelice.

7.2.2 Amplificazione di piccoli set di valutazione

L'idea di suddividere rigidamente l'input in set di formazione, test e valutazione ha senso solo su set di dati sufficientemente grandi. Supponiamo di avere a 100.000 record. Non ci sarà una differenza qualitativa tra l'addestramento su 60.000 record invece che su 100.000, quindi è meglio facilitare una valutazione rigorosa. Problemi simili si presentano nelle sperimentazioni mediche, che sono molto costose da eseguire e che potenzialmente possono fornire dati su meno di un centinaio di pazienti. Qualsiasi applicazione in cui dobbiamo pagare per l'annotazione umana significa che ci ritroveremo con meno dati per la formazione di quanto vorremmo. Cosa può fare quando non può permettersi di rinunciare a una frazione dei suoi dati? per il test? *La convalida incrociata* suddivide i dati in *k* blocchi di uguali dimensioni, quindi addestra *k* modelli distinti. Il modello *i* viene addestrato sull'unione di tutti*/* i blocchi *x*= *i*, per un totale di (*k* 1)/kesimo dei dati, e testato sul blocco i esimo escluso. La prestazione media di questi *k* classificatori rappresenta l'accuratezza presunta del modello completo. Il caso estremo è quello della *convalida incrociata*, in cui *n* modelli distinti vengono addestrati su diversi insiemi di *n* 1 esempi, per determinare- se il classificatore è buono o meno. Questo massimizza la quantità di dati di addestramento, pur lasciando qualcosa da valutare. Un vantaggio reale della convalida incrociata è che produce una deviazione standard delle prestazioni, non solo una media. Ogni classificatore addestrato su un particolare sottoinsieme di dati differirà leggermente dai suoi colleghi. Inoltre, i dati di test per ogni classificatore saranno diversi, con conseguenti punteggi di performance diversi. Accoppiando la media con la deviazione standard e assumendo la normalità, si ottiene una distribuzione delle prestazioni e un'idea migliore di quanto fidarsi dei ri sultati. Questo fa sì che la convalida incrociata valga molto la pena di essere eseguita anche su *grandi* insiemi di dati, perché può permettersi di effettuare diverse partizioni e di riqualificare, aumentando così la fiducia nella bontà del suo modello. Dei *k* modelli risultanti dalla convalida incrociata, quale dovrebbe scegliere come prodotto finale? Forse potrebbe utilizzare quello che ha ottenuto le migliori prestazioni nella sua quota di test. Ma un'alternativa migliore è quella di riqualificare *tutti* i dati e confidare nel fatto che sarà almeno altrettanto buono dei modelli meno addestrati. Questo non è l'ideale, ma se non si possono ottenere dati sufficienti, bisogna fare del proprio meglio con quello che si ha. Ecco alcune altre idee che possono aiutare ad amplificare le piccole serie di dati per la formazione e la valutazione:

- *Creare esempi negativi da una distribuzione precedente:* Supponiamo di voler costruire un classificatore per identificare chi sarebbe qualificato per essere un candidato alla presidenza. Ci sono pochissimi esempi reali di candidati alla presidenza (istanze positive), ma presumibilmente il pool di eletti è così piccolo che una persona candidata sarà quasi certamente non qualificata. Quando gli esempi positivi sono rari, tutti gli altri sono molto probabilmente negativi e possono essere etichettati in modo da fornire dati di formazione, se necessario.

- *Perturbare gli esempi reali per creare esempi simili ma sintetici:* Un trucco utile per evitare l'overfitting crea nuove istanze di formazione aggiungendo del rumore casuale per distorcere gli esempi etichettati. Poi conserviamo l'etichetta del risultato originale con la nuova istanza.

  Ad esempio , supponiamo di voler addestrare un sistema di riconoscimento ottico dei caratteri (OCR) per riconoscere le lettere di un alfabeto nelle pagine scansionate. Inizialmente è stato affidato a un umano costoso il compito di etichettare alcune centinaia di immagini con i caratteri in esse contenuti. Possiamo amplificare questo compito a qualche milione di immagini aggiungendo del rumore a caso e ruotando/traducendo/dilatando la regione di interesse. Un classificatore addestrato su questi dati sintetici dovrebbe essere molto più robusto di uno limitato ai dati annotati originali.

- *Dia credito parziale quando può*: Quando ha un numero di esempi di formazione/test inferiore a quello desiderato, deve spremere quante più informazioni possibili da ciascuno di essi.

  Supponiamo che il nostro classificatore emetta un valore che misura la fiducia nella sua decisione, oltre all'etichetta proposta. Questo livello di fiducia ci fornisce un'ulteriore risoluzione con cui valutare il classificatore, al di là del fatto che abbia azzeccato l'etichetta. È un colpo maggiore contro il classificatore quando sbaglia una previsione fiduciosa, rispetto a un'istanza in cui pensava che la risposta fosse un lancio. Su un problema di dimensioni presidenziali, mi fiderei molto di più di un classificatore che ha ottenuto 30 risposte giuste e 15 sbagliate con valori di fiducia accurati, piuttosto che di uno con 32 risposte giuste e 13 sbagliate, ma con valori di fiducia che vanno da una parte all'altra.

7.3 Storia di guerra: 100% di precisione

La lezione fondamentale qui è che nessun sistema di riconoscimento dei modelli per qualsiasi problema ragionevole sbaglierà sempre al 100%. L'unico modo per non sbagliare mai è non fare mai una previsione. È necessaria un'attenta valutazione per misurare quanto il suo sistema sta funzionando e dove sta commettendo errori, al fine di migliorare il sistema.

7.4 Modelli di simulazione

Esiste un'importante classe di modelli di primo principio che non sono principalmente basati sui dati, ma che si rivelano molto preziosi per la comprensione di molto diversi. *Le simulazioni* sono modelli che cercano di replicare i sistemi e i processi del mondo reale, in modo da poter osservare e analizzare il loro comportamento. Le simulazioni sono importanti per dimostrare la validità della nostra comprensione di un sistema. Una semplice simulazione che cattura gran parte della complessità comportamentale di un sistema deve spiegare come funziona, secondo il rasoio di Occam. Le simulazioni *Monte Carlo* utilizzano numeri casuali per sintetizzare realtà alternative. La replica di un evento milioni di volte in condizioni

leggermente perturbate ci permette di generare una distribuzione di probabilità sull'insieme dei risultati. Questa è l'idea alla base dei test di permutazione per la significatività statistica. Abbiamo anche visto che i lanci casuali di monete possono sostituire quando un battitore viene colpito o eliminato, quindi possiamo simulare un numero arbitrario di carriere e osservare cosa succede nel corso di esse. La chiave per costruire una simulazione Monte Carlo efficace è la progettazione di un sistema di un appropriato modello a eventi discreti. Un nuovo numero casuale viene utilizzato dal modello per replicare ogni decisione o risultato dell'evento. In un modello di trasporto, potrebbe essere necessario decidere se andare a sinistra o a destra, quindi lanciare una moneta. Un modello sanitario o assicurativo potrebbe dover decidere se un determinato paziente avrà un infarto oggi, quindi lancia una moneta opportunamente ponderata. Il prezzo di un'azione in un modello finanziario può salire o scendere ad ogni tick, e anche in questo caso può lanciare una moneta. Un giocatore di pallacanestro potrà colpire o sbagliare un tiro, con una probabilità che dipende dalla sua abilità di tiro e dalla qualità del suo difensore.
L'accuratezza di tale simulazione si basa sulle probabilità che lei assegna a testa e croce. Questo regola la frequenza con cui si verifica ciascun risultato. Ovviamente, non è limitato all'utilizzo di una moneta giusta, cioè 50/50. Invece, le probabilità devono riflettere le ipotesi sulla probabilità dell'evento, dato lo stato del modello. Questi parametri sono spesso impostati utilizzando l'analisi statistica, osservando la distribuzione degli eventi come si sono verificati nei dati. Parte del valore delle simulazioni Monte Carlo è che ci permettono di giocare con realtà alternative, cambiando alcuni parametri e vedendo cosa succede. Un aspetto critico della simulazione efficace è la valutazione. Gli errori di programmazione e le inadeguatezze di modellazione sono abbastanza comuni che nessuna simulazione può essere accettata per fede. La chiave è trattenere una o più classi di osservazioni del sistema dall'incorporazione diretta nel modello. In questo modo si ottiene un comportamento fuori campione, quindi possiamo confrontare la distribuzione dei risultati della simulazione con queste osservazioni. Se non coincidono, la sua simulazione è solo un'accozzaglia. Non la lasci vivere.

Capitolo 8
Algebra lineare

La parte dei dati del suo progetto di scienza dei dati comporta la riduzione di tutte le informazioni rilevanti che può trovare in una o più matrici di dati, idealmente il più grande possibile. Le righe di ogni matrice rappresentano elementi o esempi, mentre le colonne rappresentano caratteristiche o attributi distinti.
L'algebra lineare è la matematica delle matrici: le proprietà delle disposizioni di numeri e le operazioni che agiscono su di esse. Questo la rende la lingua della scienza dei dati. Molti algoritmi di apprendimento automatico sono meglio compresi attraverso l'algebra lineare. Infatti, gli algoritmi per problemi come la regressione lineare possono essere ridotti a una singola formula, moltiplicando la giusta catena di prodotti matriciali per ottenere i risultati desiderati. Tali algoritmi possono essere contemporaneamente semplici e intimidatori, banali da implementare ma difficili da rendere efficienti e robusti.

8.1 Il potere dell'algebra lineare

Regola il funzionamento delle matrici e le matrici sono ovunque. Le rappresentazioni matriciali di oggetti importanti includono:

- *Dati*: La rappresentazione più utile degli insiemi di dati numerici è quella di matrici *n-m.* Le *n*× righe rappresentano oggetti, articoli o istanze, mentre le *m* colonne rappresentano ciascuna caratteristiche o dimensioni distinte.

- *Insiemi di punti geometrici*: Una× matrice *n m* può rappresentare una nuvola di punti nello spazio. Le *n* righe rappresentano ciascuna un punto geometrico, mentre le *m* colonne definiscono le dimensioni. Alcune operazioni matriciali hanno interpretazioni geometriche distinte, che ci permettono di generalizzare la geometria bidimensionale che possiamo effettivamente visualizzare in spazi di dimensioni superiori.

- *Sistemi di equazioni*: Un'equazione lineare è definita dalla somma di variabili ponderate da coefficienti costanti.

  Un sistema di *n* equazioni lineari può essere rappresentato come una× matrice *n m*, dove ogni riga rappresenta un'equazione, e ognuna delle *m* colonne è associata ai coefficienti di una particolare variabile (o la costante 'variabile' 1 nel caso di c_0). Questo viene tipicamente fatto utilizzando un array o un vettore separato *n*× 1 di valori di soluzione.

Grafici e reti: I grafi sono composti da vertici e bordi, dove i bordi sono definiti come coppie ordinate di vertici, come (*i, j*). Un grafico con *n* vertici e *m* bordi può essere rappresentato come una matrice.

Ci sono connessioni sorprendenti tra le proprietà combinatorie e l'algebra lineare, come la relazione tra i percorsi nei grafi e la moltiplicazione matriciale, e come i cluster di vertici si relazionano con gli autovalori/vettori di matrici appropriate.

- *Operazioni di riordino*: Le matrici possono *fare* delle cose. Le matrici progettate con cura possono eseguire operazioni geometriche sugli insiemi di punti, come traslazione, rotazione e scalatura. Moltiplicando una matrice di dati per una *matrice di permutazione* appropriata, si riordinano le righe e le colonne. I movimenti possono essere definiti da *vettori*, le matrici *n* 1 abbastanza potenti da codificare operazioni come la traslazione e la permutazione.

L'ubiquità delle matrici significa che è stata sviluppata una notevole infrastruttura di strumenti per manipolarle. In particolare, le librerie di algebra lineare ad alte prestazioni per il suo linguaggio di programmazione preferito significano *che non* dovrebbe *mai* implementare un algoritmo di base da solo. Le migliori implementazioni delle librerie ottimizzano le cose sporche, come la precisione numerica, i mancati salvataggi nella cache e l'uso di più core, fino al livello del linguaggio assembly. Il nostro compito è quello di formulare il problema utilizzando l'algebra lineare e lasciare l'algoritmo a queste librerie.

8.1.1 Interpretare le formule algebriche lineari

Le formule concise scritte come prodotti di matrici possono fornire il potere di fare cose sorprendenti, come la regressione lineare, la compressione di matrici e l'analisi geometrica.

La sostituzione algebrica, unita a un ricco insieme di identità, offre modi eleganti e meccanici per manipolare tali formule.
Per capire l'algebra lineare, il suo obiettivo dovrebbe essere quello di convalidare prima il caso interessante più semplice (tipicamente a due dimensioni), per costruire l'intuizione, e poi cercare di immaginare come potrebbe generalizzarsi a dimensioni superiori. Ci sono sempre casi speciali da tenere d'occhio, come la divisione per zero. Nell'algebra lineare, questi casi includono i disallineamenti dimensionali e matrici singolari (cioè non invertibili). La teoria dell'algebra lineare funziona tranne quando non funziona, ed è meglio pensare ai casi comuni piuttosto che a quelli patologici.

8.1.2 Geometria e vettori

C'è un'utile interpretazione dei "vettori", cioè delle matrici $1 \times d$, come *vettori* in senso geometrico, cioè raggi diretti dall'origine attraverso un determinato punto in d dimensioni.

Normalizzando ogni vettore v in modo che sia di lunghezza unitaria (dividendo ogni coordinato per la distanza da v all'origine), lo si colloca su una sfera *d-dimensionale*: un cerchio per i punti nel piano, una sfera reale per d= 3, e una qualche ipersfera non visualizzabile per d 4. Questa normalizzazione si rivela utile. Le distanza tra i punti diventano angoli tra i vettori, ai fini del confronto. Due punti vicini definiranno un piccolo angolo tra loro attraverso l'origine: piccole distanze implicano piccoli angoli. Ignorare le grandezze è una forma di scalatura, che rende tutti i p unti direttamente comparabili.

Il prodotto di punti è un'operazione utile per ridurre i vettori a quantità scalari. Si definisce il prodotto di punti di due vettori di lunghezza N, A e B.

Proviamo a interpretare questa formula. Il simbolo significa "la lunghezza di V". Per i vettori unitari, per definizione è uguale a 1. In generale, è la quantità per cui dobbiamo dividere V per renderlo un vettore unitario.
Ma qual è il legame tra il prodotto del punto e l'angolo? Consideriamo il caso più semplice di un angolo definito tra due raggi, A a zero gradi e B= (x, y). Quindi la semiretta unitaria è A= (1, 0). In questo caso, il prodotto del punto che è esattamente quello che dovrebbe essere il cos(ϑ) se B è un vettore unitario.

Possiamo credere che questo sia generalizzato per B generale e per dimensioni superiori.
Quindi un angolo più piccolo significa punti più vicini sulla sfera. Ma c'è un'altra connessione tra le cose che conosciamo. Ricordiamo i casi speciali della funzione coseno.

I valori della funzione coseno vanno da [1, 1], -esattamente lo stesso intervallo del coefficiente di correlazione. Inoltre, l'interpretazione è la stessa: due vettori identici sono perfettamente correlati, mentre i punti antipodali sono perfettamente correlati negativamente. I punti/vettori ortogonali hanno il meno possibile a che fare l'uno con l'altro.

La funzione coseno è esattamente la correlazione di due variabili a media zero. Per i vettori unitari, quindi l'angolo tra A e B è completamente definito dal prodotto di punti.

8.2 Visualizzazione delle operazioni di matrice

Ma per fornire una migliore intuizione, rappresenterò le matrici come immagini piuttosto che come numeri, in modo da poter *vedere* cosa succede quando operiamo su di esse. *Tenga presente che ridimensioneremo silenziosamente la matrice tra ogni operazione, quindi il colore assoluto non ha importanza.* I modelli interessanti si trovano nelle differenze tra chiaro e scuro, ossia i numeri più piccoli e più grandi nella matrice

corrente. Inoltre, si noti che l'elemento di origine della matrice M [1, 1] rappresenta l'angolo superiore sinistro dell'immagine.

8.2.1 Aggiunta di una matrice

L'addizione di matrici è un'operazione semplice. *La moltiplicazione scalare* offre un modo per modificare il peso di ogni elemento di una matrice simultaneamente, forse per normalizzarli. La combinazione dell'addizione matriciale con la moltiplicazione scalare ci dà la possibilità eseguire *combinazioni lineari* di matrici ci permette di sfumare in modo fluido tra A e B, come mostrato nella. Questo fornisce un modo per morfologizzare le immagini da A a B questo modo si ottiene un modo per modificare le immagini da A a B. La *trasposizione* di una matrice M scambia le righe e le colonne, trasformando un $a \times b$ matrice in una $b \times a$. La trasposizione di una matrice quadrata è una matrice quadrata, quindi M e M^T possono essere tranquillamente sommate o moltiplicate insieme. Più in generale, la trasposizione è un'operazione che viene utilizzata per orientare una matrice in modo che *possa* essere aggiunta o moltiplicata per il suo obiettivo. La trasposizione di una matrice in un certo senso la "ruota" di 180 gradi, quindi $(A^T)^T = A$. Nel caso di matrici quadrate, l'aggiunta di una matrice alla sua trasposizione è simmetrica.

8.2.2 Moltiplicazione di matrici

La moltiplicazione matriciale è una versione aggregata del *punto* vettoriale o del *prodotto interno*. Ricordiamo che per due vettori $di\ n$ elementi, X e Y, si definisce il prodotto di punti $X \cdot Y$. Il prodotto dei punti misura quanto sono "sincronizzati" i due vettori. Abbiamo già visto il prodotto di punti quando abbiamo calcolato la distanza del coseno e il coefficiente di correlazione. Si tratta di un'operazione che riduce una coppia di vettori a un singolo numero. Il prodotto matriciale XY^T di questi due vettori produce una matrice 1 $1 \times$ contenente il prodotto di punti $X \cdot Y$. Affinché questo funzioni, A e B devono condividere le stesse dimensioni interne. Ogni elemento della matrice di prodotto $C\ di\ n\ m$ è un prodotto di punti della iesima riga di A con la jesima colonna di B. Le proprietà più importanti della moltiplicazione matriciale sono:

- *Non è commutativa*: *La commutatività* è la notazione che l'ordine non conta, che $x \cdot y = y \cdot x$. Anche se diamo· per scontata- la commutatività quando moltiplichiamo gli interi, l'ordine conta nella moltiplicazione delle matrici. Per qualsiasi coppia di matrici non quadrate A e B, almeno una delle due matrici AB o BA è importante. ha dimensioni compatibili. Ma anche la moltiplicazione di matrici quadrate non è commutativa, come dimostrano i prodotti sottostanti e le matrici di covarianza.

- *La moltiplicazione di matrici è associativa*: *L'associatività* ci concede il diritto di mettere le parentesi come vogliamo, eseguendo le operazioni nell'ordine relativo che scegliamo. Nel calcolo del prodotto ABC, possiamo scegliere tra due operazioni: $(AB)C$ o $A(BC)$. Catene di matrici più lunghe consentono una libertà ancora maggiore, con il numero di parentesi possibili che cresce esponenzialmente con la lunghezza della catena.

Ci sono due ragioni principali per cui l'associatività è importante per noi. In algebrico, ci permette di identificare le coppie di matrici vicine in una catena e di sostituirle secondo un'identità, se ne abbiamo una. Ma l'altro problema è di tipo computazionale. Le dimensioni dei prodotti matriciali intermedi possono facilmente esplodere nel mezzo. L'algoritmo di moltiplicazione matriciale a loop annidato che le è stato insegnato alle scuole superiori è banalmente facile da programmare. Algoritmi molto più veloci e numericamente più stabili esistono nelle librerie di algebra lineare altamente ottimizzate associate al suo linguaggio di programmazione preferito. Formulare i suoi algoritmi come prodotti di matrici su grandi array, invece di usare una logica ad hoc, è

controintuitivo per la maggior parte degli informatici. Ma questa strategia può produrre grandi vantaggi in termini di prestazioni nella pratica.

8.2.3 Applicazioni della moltiplicazione di matrici

Matrici di covarianza

Moltiplicare una matrice A per la sua trasposizione A^T è un'operazione molto comune. Perché?

Per prima , *possiamo* : se A è una matrice $n \times d$, allora A^T è una matrice $d \times$

matrice n. Quindi è sempre compatibile moltiplicare AA^T . Sono ugualmente compatibili per moltiplicare nell'altro senso, ossia $A^T A$. Entrambi questi prodotti hanno interpretazioni importanti. Supponiamo che A sia una matrice$\times$ di caratteristiche n d, composta da n righe che rappresentano elementi o punti, e d colonne che rappresentano le caratteristiche osservate di questi elementi. Allora:

- $C = A - A^T$è una matrice $n \times n$ di prodotti di punti, che misura la "sincronizzazione". ness" tra i punti. In particolare, C_{ij} è una misura della somiglianza tra l'elemento i è all'articolo j.

- $D = A - {}^T A$ è una matrice $d \times$ d di prodotti di punti, che misura la "sincronia". tra le colonne o le caratteristiche. Ora D_{ij} rappresenta la similarità tra la caratteristica i e la caratteristica j.

Queste bestie sono abbastanza comuni da meritare un nome proprio, *ma- trici di covarianza*. Questo termine ricorre spesso nelle conversazioni tra gli scienziati dei dati, quindi si metta a proprio agio con esso. La formula di covarianza che abbiamo dato quando abbiamo calcolato il coefficiente di correlazione. quindi, a rigore, le nostre bestie sono matrici di covarianza solo se le righe o le colonne di A hanno media zero. Ma indipendentemente da ciò, le grandezze del prodotto matriciale catturano il grado in cui i valori di particolari coppie di righe o colonne si muovono insieme.

Moltiplicazione di matrici e percorsi

Le matrici quadrate possono essere moltiplicate per se stesse senza trasposizione. Infatti, $A^{(2)} A = \times$ A è chiamato il *quadrato* della matrice A. Più in generale, A^k è chiamato la *kesima* potenza della matrice.

I poteri della matrice A hanno un'interpretazione molto naturale, quando A rappresenta la *matrice di adiacenza* di un grafico o di una rete. In una matrice di adiacenza, $A[i, j] = 1$ quando (i, j) è un bordo della rete. In caso contrario, quando i e j non sono vicini diretti, $A[i, j] = 0$. Per tali matrici 0/1, il prodotto A^2 fornisce il numero di percorsi di lunghezza due in A.

Esiste esattamente un percorso di lunghezza due da i a j per ogni vertice intermedio k, tale che (i, k) e (k, j) siano entrambi bordi del grafico. La somma di questi percorsi è calcolata dal prodotto di punti di cui sopra.

Ma il calcolo dei poteri delle matrici ha senso anche per le matrici più generali. Simula gli effetti della diffusione, distribuendo il peso di ogni elemento tra elementi correlati. Questo accade nel famoso algoritmo PageRank di Google e in altri processi iterativi come la diffusione del contagio.

Moltiplicazione di matrici e permutazioni La moltiplicazione matriciale viene spesso utilizzata solo per riordinare l'ordine degli elementi di una determinata matrice. Ricordiamo che le routine di moltiplicazione matriciale ad alte prestazioni sono estremamente veloci, tanto che spesso possono eseguire tali operazioni più velocemente della logica di programmazione ad hoc. Inoltre, forniscono un modo per descrivere tali operazioni nella notazione delle formule algebriche, preservando così la compattezza e la leggibilità.

La matrice di riarrangiamento più famosa non fa nulla. L'*identità*

La matrice è una matrice $n \times n$ composta da tutti gli zeri, ad eccezione di quelli che si trovano lungo tutto il percorso. Si convinca che $AI = IA = A$, il che significa che la moltiplicazione per la matrice identità commuta.
Si noti che ogni riga e colonna di I contiene esattamente un elemento non nullo. Le matrici con questa proprietà sono chiamate *matrici di permutazione*, perché l'elemento non nullo in posizione (i, j) può essere interpretato nel senso che l'elemento i si trova nella posizione j di una permutazione. Osserviamo che la matrice identità corrisponde alla permutazione $(1, 2, \ldots, n)$. Il punto chiave qui è che possiamo moltiplicare A per la matrice di permutazione appropriata per riorganizzare le righe e le colonne, come desideriamo. Poiché la moltiplicazione delle matrici non è generalmente commutativa, otteniamo risultati diversi per $A \cdot r$ e $r A$. -Si convinca del perché.

Rotazione di punti nello spazio

Moltiplicare qualcosa per la matrice giusta può avere proprietà magiche. Abbiamo visto come un insieme di n punti nel piano (cioè due dimensioni) può essere rappresentato da una matrice $(n \times 2)$-dimensionale S Moltiplicando tali punti per la matrice giusta si possono ottenere trasformazioni geometriche naturali.
La *matrice di rotazione* R_θ esegue la trasformazione di rotazione dei punti intorno all'origine attraverso un angolo di ϑ. In due dimensioni. Esistono generalizzazioni naturali di R_θ per ruotare i punti in dimensioni arbitrarie. Inoltre, sequenze arbitrarie di trasformazioni successive possono essere realizzate moltiplicando catene di matrici di rotazione, dilatazione e riflessione, ottenendo così una descrizione completa di manipolazioni complesse.

8.2.4 Matrici di identità e inversione

Le operazioni di identità svolgono un ruolo importante nelle strutture algebriche. Per l'addizione numerica, lo zero è l'elemento di identità. Lo stesso ruolo è svolto da uno per la moltiplicazione, poiché $1 \times x = x \, 1 = x$.
Nella moltiplicazione matriciale, l'elemento identità è la matrice identità, con tutti gli uni lungo la diagonale principale. La moltiplicazione per la matrice identità commuta.
L'operazione *inversa* consiste nel ridurre un elemento x al suo elemento identico. Per l'addizione numerica, l'inverso di x è $(-x)$, perché $x + (-x) = 0$. L'operazione- inversa per la moltiplicazione si chiama *divisione*. Possiamo invertire un numero moltiplicandolo per il suo reciproco, dato che $x \, (1/x) = 1$.
In genere non si parla di dividere le matrici. Tuttavia, molto spesso si parla di . Diciamo che A^{-1} è l'*inversa moltiplicativa* della matrice A se $A \cdot A^{-1} = I$, dove I è la matrice identità. L'inversione è un caso speciale importante della divisione, poiché $A \cdot A = {}^{-1} I$ implica $A = I/A_{-1}$. Si tratta infatti di operazioni equivalenti, perché $A/B = A \cdot B$
Come possiamo calcolare l'inverso di una matrice? Esiste una forma chiusa per trovare l'inversa A^{-1} di una matrice-

Più in generale, esiste un approccio per invertire le matrici risolvendo un sistema lineare con l'eliminazione gaussiana.

Osserviamo che questa forma chiusa per l'inversione divide per zero ogni volta che i prodotti delle diagonali sono uguali, cioè $ad = bc$. Questo ci dice che tali matrici non sono invertibili o *singolari*, il che significa che non esiste un'inversione. Proprio come non possiamo dividere i numeri per zero, non possiamo

invertire le matrici singolari. Le matrici che possiamo invertire sono chiamate *non-singolari* e la vita è migliore quando le nostre matrici hanno questa proprietà. Il test per verificare se una matrice è invertibile è se il suo *determinante* non è zero. Per le matrici 2 2, il determinante è la differenza tra il prodotto delle sue diagonali, esattamente il denominatore nella formula di inversione. Inoltre, il determinante è definito solo per le matrici quadrate, quindi solo le matrici quadrate sono invertibili. Il costo del calcolo di questo determinante è $O(n^3)$, quindi è costoso su matrici di grandi dimensioni, in effetti tanto quanto il tentativo di invertire la matrice stessa utilizzando l'eliminazione gaussiana.

8.2.5 Inversione di matrice e sistemi lineari

Le equazioni lineari sono definite dalla somma di variabili ponderate da costanti.

Quindi i coefficienti che definiscono un sistema di n equazioni lineari possono essere rappresentati× come una matrice C *di* n m. Qui ogni riga rappresenta un'equazione, e ciascuna delle m colonne i coefficienti di una variabile distinta. Possiamo valutare ordinatamente tutte le n equazioni su un particolare vettore di ingresso m× 1 X moltiplicando C - X. Il risultato sarà un vettore n× 1, che riporterà il valore $f_i(X)$ per ciascuna delle n equazioni lineari, 1≤ i≤ n. Il caso speciale qui è il termine additivo c_0. Per una corretta interpretazione, la colonna associata in X deve contenere tutti gli uni. Se generalizziamo X come una matrice× m p contenente p punti distinti, il nostro prodotto C x -risulta in una matrice× n p, valutando ogni punto rispetto a ogni equazione in un'unica moltiplicazione matriciale. Ma l'operazione principale sui sistemi di n equazioni è quella di , cioè di identificare il vettore X necessario per ottenere un valore Y target per ogni equazione.

L'inversione matriciale può essere utilizzata per risolvere i sistemi lineari. Moltiplicando entrambi i lati.

L'eliminazione gaussiana è un altro approccio alla risoluzione dei sistemi lineari, che spero abbia già visto in precedenza. Ricordiamo che risolve le equazioni eseguendo operazioni di aggiunta/sottrazione di righe per semplificare la matrice di equazioni C fino a ridurla alla matrice identità. Questo rende banale la lettura dei valori delle matrici poiché ogni equazione è stata ridotta alla forma X_i= Y', dove Y' è il risultato dell'applicazione $_i$ di queste stesse operazioni di riga al vettore target originale Y .

Il calcolo dell'inverso della matrice può essere eseguito allo stesso modo. Eseguiamo operazioni di riga per semplificare la matrice dei coefficienti matrice di identità I, al fine di creare l'inverso. Lo considero come l'algoritmo di Dorian Gray: la matrice dei coefficienti C si abbellisce fino a diventare la matrice di identità, mentre l'obiettivo I invecchia nell'inverso.

Pertanto, possiamo utilizzare l'inversione di matrice per risolvere i sistemi lineari e i risolutori di sistemi lineari per invertire le matrici. Quindi i due problemi sono in un certo senso equivalenti. Il calcolo dell'inverso rende conveniente la valutazione di più vettori Y per un dato sistema C, riducendolo a una singola moltiplicazione di matrice. Ma questo può essere fatto in modo ancora più efficiente con la decomposizione LU. L'eliminazione gaussiana si dimostra più stabile numericamente rispetto all'inversione, e in genere è il metodo da scegliere quando si risolvono sistemi lineari.

8.2.6 Classifica della matrice

Un sistema di equazioni è *determinato* correttamente quando ci sono n equazioni linearmente indipendenti e n incognite.

Al contrario, i sistemi di equazioni sono *sottodeterminati* se ci sono righe (equazioni) che possono essere espresse come combinazioni lineari di altre righe.

è sottodeterminato, perché la seconda riga è il doppio della prima riga. Dovrebbe essere chiaro che non ci sono informazioni sufficienti per risolvere un sistema di equazioni lineari non determinato.

Il *rango* di una matrice misura il numero di righe linearmente indipendenti. Un $n\times$ matrice deve essere di rango n affinché tutte le operazioni siano definite correttamente su di essa. Il rango della matrice può essere calcolato eseguendo l'eliminazione gaussiana.

Se è sottodeterminato, alcune variabili scompariranno nel corso delle operazioni di riduzione delle righe. Esiste anche una connessione tra i sistemi sottodeterminati e le matrici singolari: ricordiamo che sono state identificate con un determinante pari a zero. Ecco perché la differenza nel prodotto incrociato qui (2 - 2 - 4 - 1) è uguale a zero. Le matrici di caratteristiche sono spesso di rango inferiore a quello desiderato. I file di esempi tendono a contenere voci duplicate, per cui due righe della matrice sarebbero identiche. È anche possibile che più colonne siano equivalenti: immaginiamo che ogni record contenga l'altezza misurata sia in piedi che in metri, per esempio. Queste cose accadono certamente e sono negative quando accadono. Alcuni algoritmi sulla nostra immagine del Memoriale di Lincoln hanno fallito numericamente. Si è scoperto che la nostra immagine di 512 512 aveva un rango di soli 508, quindi non tutte le righe erano linearmente indipendenti. Per renderla una matrice di rango completo, è possibile aggiungere una piccola quantità di rumore casuale a ciascun elemento, che aumenterà il rango senza gravi disturbi all'immagine. Questo espediente potrebbe far passare i suoi dati attraverso un algoritmo senza un messaggio di avvertimento, ma è indicativo dei problemi numerici che .

I sistemi lineari possono essere "quasi" di rango inferiore, il che comporta un pericolo maggiore di perdita di precisione a causa di problemi numerici. Questo è formalmente catturato da un invariante matriciale chiamato *numero di condizione*, che nel caso di un sistema lineare misura quanto è sensibile il valore di X a piccole variazioni di Y in $Y= AX$. Nella valutazione dei suoi risultati, tenga presente i capricci del calcolo numerico. Ad esempio, è una buona pratica calcolare AX per qualsiasi soluzione presunta X, e vedere quanto AX *sia realmente* paragonabile a Y . In teoria la differenza sarà pari a zero, ma in pratica potrebbe essere sorpreso di quanto sia approssimativo il calcolo.

8.3 Matrici di fattorizzazione

La fattorizzazione della matrice A in matrici B e C rappresenta un aspetto particolare della divisione. Abbiamo visto che qualsiasi matrice non singolare M ha un'inversa M^{-1}, quindi la matrice identità I può essere fattorizzata come $I= MM^{-1}$. Questo dimostra che alcune matrici (come I) possono essere fattorizzate, e inoltre che potrebbero avere molte fattorizzazioni distinte. In questo caso, ogni possibile M non-singolare definisce una fattorizzazione diversa. La fattorizzazione matriciale è un'astrazione importante nella scienza dei dati, che porta a rappresentazioni concise delle caratteristiche e a idee come il topic modeling. Svolge un ruolo importante nella risoluzione di sistemi lineari, attraverso fattorizzazioni speciali come la decomposizione LU.

8.3.1 Perché fattorizzare le matrici di caratteristiche?

Molti importanti algoritmi di apprendimento automatico possono essere visti in termini di fattorizzazione. Supponiamo di avere una matrice di caratteristiche A *di* $n\times m$.

Per convenzione, le righe rappresentano gli articoli/esempi e le colonne le caratteristiche degli esempi.

Ora supponiamo di poter *fattorizzare* la matrice A, ossia di esprimerla come il prodotto $A \approx B\text{-}C$, dove B è una matrice $\times n$ k e C una matrice $k \times$ m. Presumendo che $k<$ min(n, m), questa è una buona cosa per diverse ragioni:

- *Insieme, B e C forniscono una rappresentazione compressa della matrice A*: le matrici di dati sono generalmente grandi e sgraziate da lavorare. La fattorizzazione offre un modo per codificare tutte le informazioni della matrice grande in due matrici più piccole, che insieme saranno più piccole dell'originale.

- *B serve come matrice di caratteristiche più piccola sugli articoli, in sostituzione di A*: la matrice di fattori *B* ha *n* righe, proprio come la matrice *A* originale. Tuttavia, ha un numero sostanzialmente inferiore di colonne, poiché $k < m$. Ciò significa che la 'maggior parte' delle informazioni di *A* è ora codificata in *B*. Meno colonne significano una matrice più piccola e meno parametri da inserire in qualsiasi modello costruito utilizzando queste nuove caratteristiche. Queste caratteristiche più astratte possono essere interessanti anche per altre applicazioni, come descrizioni concise delle righe del set di dati.

• C^T *serve come una piccola matrice di caratteristiche sulle* caratteristiche^T*, sostituendo A* : La posa della matrice delle caratteristiche trasforma le colonne/caratteristiche in righe/voci. La matrice dei fattori C^T ha *m* righe e *k* colonne di proprietà che li rappresentano. In molti casi, le *m* "caratteristiche" originali meritano di essere modellate a sé stanti.

La matrice C^T può ora essere considerata come contenente un vettore di caratteristiche per ciascuna delle parole del vocabolario. Questo è interessante. Ci aspetteremmo che le parole che si applicano in contesti simili abbiano vettori di argomenti simili.

8.3.2 Decomposizione LU e determinanti

La decomposizione LU è una particolare fattorizzazione matriciale che fattorizza una matrice quadrata *A* in matrici triangolari inferiori e superiori *L* e *U*, in modo tale che $A = L\text{-}U$,

Una matrice è *triangolare* se contiene tutti i termini zero sopra o sotto diagonale principale. La matrice triangolare inferiore *L* ha tutti i termini non nulli sotto diagonale principale. L'altro fattore, *U* , è la matrice triangolare superiore. Poiché la diagonale principale di *L* è composta da tutti gli uni, possiamo racchiudere l'intera decomposizione nello stesso spazio della matrice originale $n \times n$.

Il valore principale della decomposizione LU è che si rivela utile per risolvere i sistemi lineari $AX = Y$, in particolare quando si risolvono problemi multipli con la stessa *A* ma *Y* diversi. La matrice *L* è il risultato della cancellazione di tutti i valori al di sopra della diagonale principale, tramite l'eliminazione gaussiana. Una volta in questa forma triangolare, le equazioni rimanenti possono essere semplificate direttamente. La matrice *U* riflette quali operazioni di riga si sono verificate nel corso della costruzione di *L*. Semplificare *U* e applicare *L* a *Y* richiede meno lavoro che risolvere *A* da zero.

L'altra importanza della decomposizione LU consiste nel fornire un algoritmo per calcolare il determinante di una matrice. Il *determinante* di *A* è il prodotto degli elementi diagonali principali di *U*. Come abbiamo visto, un determinante pari a zero significa che la matrice non è di rango pieno.

8.4 Autovalori e autovettori

La moltiplicazione di un vettore *U* per una matrice quadrata *A* può avere lo stesso effetto della multiplicazione per uno scalare *I*. Consideri questa coppia di esempi. Entrambe queste uguaglianze presentano prodotti con lo stesso vettore × 2 1 *U* a sinistra e a destra. Da un lato *U* è moltiplicato per una matrice *A* e dall'altro per uno scalare *λ*. In casi come questo, quando $AU = \lambda U$ diciamo che *λ* è un *autovalore* della matrice *A* e *U* è l'*autovettore* associato. Tali coppie autovettore-valore sono una cosa curiosa. Il fatto che lo scalare *λ* possa fare la stessa cosa con *U* e con l'intera matrice *A* ci dice che devono essere speciali. Insieme, l'autovettore *U* e l'autovalore *λ* devono codificare molte informazioni su *A*. Inoltre, in genere esistono più coppie autovettore-valore per qualsiasi matrice. Si noti che il secondo esempio sopra funziona sulla stessa matrice *A*, ma produce un *U* e un *λ* diversi.

8.4.1 Proprietà degli autovalori

La teoria degli autovalori ci porta più a fondo nel folto dell'algebra lineare di quanto io sia disposto a fare in questo testo. In generale, però, possiamo riassumere le proprietà che si riveleranno importanti per noi:

- Ogni autovalore ha un autovettore associato. Si presentano sempre in coppia.
- Ci sono, in generale, n coppie autovettore-autovalore per ogni rango completo $n \times n$ matrice
  - Ogni coppia di autovettori di una matrice simmetrica è reciprocamente *ortogonale*, allo stesso modo in cui gli *assi x* e y nel piano sono ortogonali. Due vettori sono ortogonali se il loro prodotto di punti è zero. Osserviamo che $(0, 1) \cdot (1, 0) = 0$, così come $(2, -1) \cdot (1, 2) = 0$ dall'esempio precedente.
  - Il risultato è che gli autovettori possono svolgere il ruolo di dimensioni o *basi* in uno spazio *ndimensionale*. Questo apre molte interpretazioni geometriche delle matrici. In particolare, qualsiasi matrice può essere codificata dove ogni autovalore rappresenta la grandezza dell'autovettore associato.

8.4.2 Calcolo degli autovalori

Gli n autovalori distinti di una matrice rank-n possono essere trovati attraverso la fattorizzazione della sua *equazione caratteristica*. Partiamo dall'uguaglianza di definizione $AU = \lambda U$. Si convinca che questa rimane invariata quando moltiplichiamo per la matrice identità I. Gli algoritmi più veloci per il calcolo degli autovalori/vettori si basano su un approccio di fattorizzazione della matrice chiamato *decomposizione QR*. Altri algoritmi cercano di evitare di risolvere il sistema lineare completo. Ad esempio, un approccio alternativo utilizza ripetutamente $U' = (AU)/\lambda$ per calcolare approssimazioni sempre migliori di U, fino a convergere. Quando le condizioni sono giuste, questo può essere molto più veloce della soluzione del sistema lineare completo.

8.5 Decomposizione degli autovalori

Qualsiasi matrice simmetrica M $n \times n$ può essere decomposta nella somma dei suoi n prodotti autogeni. Chiamiamo le n coppie di autovalori. Per convenzione ordiniamo per dimensione, quindi per tutti gli i.

Poiché ogni autovettore U_i è una matrice $n \times 1$, moltiplicandola per la sua trasposizione produce un prodotto matriciale $n \times n$, $U_i U_i$. Questo ha esattamente le stesse dimensioni come la matrice originale M. Possiamo calcolare la combinazione lineare di queste matrici ponderata per l'autovalore corrispondente.

Questo risultato vale solo per le matrici simmetriche, quindi non possiamo utilizzarlo per codificare la nostra immagine. Ma le matrici di covarianza sono sempre simmetriche e codificano le caratteristiche di base di ogni riga e colonna della matrice.

Quindi la matrice di covarianza può essere rappresentata dalla sua decomposizione degli autovalori. Questo richiede un po' più di spazio rispetto alla matrice iniziale: n autovettori di lunghezza n, più n autovalori rispetto agli $n(n+1)/2$ elementi nel triangolo superiore della matrice simmetrica più la diagonale principale.

Tuttavia, utilizzando solo i vettori associati agli autovalori più grandi, otteniamo una buona approssimazione della matrice. Le dimensioni più piccole contribuiscono molto poco ai valori della matrice e quindi possono essere escluse con un errore ridotto. Questo metodo di riduzione delle dimensioni è molto utile per produrre set di caratteristiche piccoli ed efficaci.

Le regioni di errore si riducono man mano che si ricostruiscono dettagli più fini, e l'entità degli errori si riduce. Si renda conto che anche cinquanta autovettori sono meno del 10% dei 512 necessari per ripristinare una matrice perfetta, ma questo è sufficiente per un'ottima approssimazione.

8.5.1 Decomposizione del valore singolare

La decomposizione degli autovalori è un'ottima cosa. Ma funziona solo su matrici simmetriche. La *decomposizione del valore singolare* è un approccio di fattorizzazione matriciale più generale, che riduce analogamente una matrice alla somma di altre matrici definite da vettori.

La *decomposizione del valore singolare* di una matrice reale M di $n \times m$ la fattorizza in tre matrici U, D e V, con dimensioni $n \times n$, $n \times m$ e $m \times m$ rispettivamente. La matrice centrale D ha la proprietà di essere una matrice *diagonale*, il che significa che tutti i valori non nulli si trovano sulla diagonale principale, come la matrice identità I.

Non si preoccupi di come trovare questa fattorizzazione. Concentriamoci invece sul suo significato. Il prodotto U D ha l'effetto di moltiplicare- U $[i, j]$ per $D[j, j]$, perché tutti i termini di D sono zero, tranne lungo la diagonale principale. Quindi D può essere interpretato come la misurazione dell'importanza relativa di ogni colonna di U, o attraverso D V^T, l'importanza di ogni riga di V^T. Questi valori di peso di D sono chiamati *valori singolari* di M. Sia X che Y sono vettori, di dimensione Nf $n \times$ 1 e $1 \times$ m, rispettivamente.
dove A_k è il vettore definito dalla kesima colonna di A, e B^T è il vettore k definito b y la
kesima riga di B. Mettendo insieme questi elementi, la matrice M può essere espressa come la somma dei prodotti esterni dei vettori risultanti dalla decomposizione dei valori singolari, ossia ($UD)_k$ e
$(V^T)_k$ per 1 k m. Inoltre, i valori singolari D definiscono il contributo che ogni prodotto esterno apporta a M , quindi è sufficiente prendere solo i vettori associati ai valori singolari più grandi per ottenere un'approssimazione a M . *L'analisi delle componenti principali* (PCA) è una tecnica strettamente correlata per ridurre la dimensionalità dei set di dati. Come SVD, definiremo dei vettori per rappresentare l'insieme di dati. Come la SVD, li ordineremo per importanza successiva, in modo da poter ricostruire una rappresentazione approssimativa utilizzando pochi componenti. La PCA e la SVD sono così strettamente correlate da essere indistinguibili per i nostri scopi. Fanno la stessa cosa nello stesso , ma da direzioni diverse. Le componenti principali definiscono gli assi di un ellissoide che meglio si adatta ai punti. L'origine di questa serie di assi è il centroide dei punti. La PCA inizia identificando la direzione su cui proiettare i punti per spiegare la massima quantità di varianza. Si tratta della linea passante per il centroide che, in un certo senso, si adatta meglio ai punti, rendendola analoga alla regressione lineare. Possiamo quindi proiettare ogni punto su questa linea, con questo punto di intersezione che definisce una posizione particolare sulla linea rispetto al centroide. Per ogni componente successivo, cerchiamo la linea l_k, che è ortogonale a tutte le linee precedenti e spiega la maggior parte della varianza rimanente. Il fatto che ogni dimensione sia ortogonale all'altra significa che agiscono come assi di coordinate, stabilendo la connessione con gli autovalori. Ogni dimensione successiva è progressivamente meno importante di quelle precedenti, perché abbiamo scelto prima le direzioni più promettenti. I componenti successivi contribuiscono solo a dettagli progressivamente più fini e quindi possiamo fermarci quando questi sono abbastanza piccoli. Supponiamo che le dimensioni x e y siano praticamente identiche. Ci aspetteremmo che la linea di regressione si proietti verso il basso fino a $y = x$ su queste due dimensioni, che quindi potrebbero essere ampiamente sostituite da un'unica dimensione. La PCA costruisce nuove dimensioni come combinazioni lineari di originali, facendo collassare quelle che sono altamente correlate in uno spazio di dimensio ni inferiori. *L'analisi statistica dei fattori* è una tecnica che identifica le dimensioni ortogonali più importanti (misurate dalla correlazione) che spiegano la maggior parte della varianza. Sono sufficienti relativamente pochi componenti per catturare la struttura di base dell'insieme di punti. Il residuo che rimane è probabilmente un rumore e spesso è meglio eliminarlo dai dati. Dopo la riduzione delle dimensioni tramite PCA (o SVD), dovremmo ottenere dati più puliti, non solo un numero inferiore di dimensioni.

Capitolo 9

Regressione lineare e logistica

La regressione lineare è il metodo di 'machine learning' più rappresentativo per costruire modelli di predizione e classificazione dei valori a partire dai dati di formazione. Offre uno studio sui contrasti:

- La regressione lineare ha un'ottima base teorica ma, nella pratica, questa formulazione algebrica viene generalmente scartata in favore di un'ottimizzazione più veloce ed euristica.

- I modelli di regressione lineare sono, per definizione, lineari. Ciò offre l'opportunità di testimoniare i limiti di tali modelli, nonché di sviluppare tecniche intelligenti per generalizzare ad altre forme.

- La regressione lineare incoraggia contemporaneamente la costruzione di modelli con centinaia di variabili e tecniche di regolarizzazione per garantire che la maggior parte di esse venga ignorata.

La regressione lineare è una tecnica di modellazione che dovrebbe servire come approccio di base per la costruzione di modelli basati sui dati. Questi modelli sono in genere facili da costruire, semplici da interpretare e spesso hanno un buon rendimento nella pratica. Con sufficiente abilità e fatica, le tecniche di apprendimento automatico più avanzate potrebbero fornire prestazioni migliori, ma il possibile guadagno non vale spesso lo sforzo. Costruisca prima i suoi modelli di regressione lineare, poi decida se vale la pena lavorare di più per ottenere risultati migliori.

9.1 Regressione lineare

Dato un insieme di *n* punti, la regressione lineare cerca di trovare la linea che meglio si approssima o si *adatta* ai punti. Ci sono molte ragioni per cui potremmo voler fare questo. Una classe di obiettivi riguarda la semplificazione e la compressione: possiamo sostituire una grande serie di punti di dati rumorosi nel *piano xy* con una linea ordinata che li descrive. Questa linea di regressione è utile per la visualizzazione. Questa linea di regressione è utile per la visualizzazione, mostrando la tendenza di fondo dei dati ed evidenziando la posizione e l'entità dei valori anomali. Tuttavia, saremo più interessati alla regressione come metodo per la previsione dei valori. Possiamo immaginare che ogni punto osservato $p = (x, y)$ sia il risultato di una funzione $y = f(x)$, dove x rappresenta le variabili caratteristiche e y la variabile target indipendente. Dato un insieme di n punti, cerchiamo la $f(x)$ che spiega meglio questi punti. Questa funzione $f(x)$ interpola o modella i punti, fornendo un modo per stimare il valore y' associato a qualsiasi possibile x', ossia che $y' = f(x')$.

9.1.1 Regressione lineare e dualità

Esiste un collegamento tra la regressione e la risoluzione di equazioni lineari, che è interessante esplorare. Quando risolviamo i sistemi lineari, cerchiamo il singolo punto che giace su *n* linee date. Nella regressione, invece, ci vengono dati *n* punti e cerchiamo la linea che giace su 'tutti' i punti. Le differenze sono due: (a) l'interscambio dei punti con le linee e (b) la ricerca del miglior adattamento sotto vincoli rispetto a un problema totalmente vincolato ("tutti" vs. tutti). La distinzione tra punti e linee si rivela banale, perché entrambi sono in realtà la stessa cosa. Nello spazio bidimensionale, sia i punti (s, t) che le linee $y = mx + b$ sono definiti da due parametri: s, t e m, b , rispettivamente. Inoltre, con un'appropriata trasformazione *di dualità*, queste rette sono equivalenti a

Il punto (4, 8) in rosso (a sinistra) corrisponde alla linea rossa $y = 4x - 8$ a destra. Entrambe le serie di tre punti allineati a sinistra corrispondono a tre linee che passano per lo stesso punto a destra. punti in un altro spazio.

In particolare. Ora, qualsiasi serie di punti che si trovano su una singola linea viene mappata su una serie di linee che intersecano un punto singolare - quindi trovare una linea che colpisce tutti i punti è algoritmicamente la stessa cosa che trovare un punto che colpisce tutte le linee.
Il punto di intersezione è $p = (4, 8)$ e corrisponde alla linea rossa $y = 4x$ 8 a destra. Questo- punto rosso p è definito dall'intersezione delle linee nere e blu. Nello spazio duale, queste linee si trasformano in punti neri e blu che giacciono sulla linea rossa. Tre punti allineati a sinistra (rosso con due neri o due blu) si trasformano in tre linee che passano per un punto comune a destra: una rossa e due dello stesso colore.
Questa trasformazione di dualità inverte i ruoli dei punti e delle linee in modo che tutto abbia senso.
La grande differenza nel definire la regressione lineare è che cerchiamo una linea che *si avvicini il più possibile* a colpire tutti i punti. Dobbiamo fare attenzione a misurare l'errore nel modo corretto per far sì che questo funzioni.

9.1.2 Errore nella regressione lineare

L'*errore residuo* di una linea adattata $f(x)$ è la differenza tra i valori previsti e quelli effettivi. per un particolare vettore di caratteristiche x_i e il corrispondente valore target y_i, definisce l'errore residuo. Questo è ciò che ci interessa, ma si noti che non è l'unico modo in cui l'errore potrebbe essere stato definito. La distanza più vicina alla linea è infatti definita dalla perpendicolare-bisettrice che passa per il punto di destinazione. Ma stiamo cercando di prevedere il valore di y_i da x_i, quindi il residuo è la giusta nozione di errore per i nostri scopi. *La regressione ai minimi quadrati* minimizza la somma dei quadrati dei residui di tutti i punti. Questa metrica è stata scelta perché (1) la quadratura del residuo ignora i segni degli errori, quindi i residui positivi e negativi non si compensano a vicenda, e (2) porta a una forma chiusa sorprendentemente piacevole per trovare i coefficienti della linea che si adatta meglio.

9.1.3 Trovare l'adattamento ottimale

La regressione lineare cerca la retta $y = f(x)$ che minimizza la somma degli errori al quadrato su tutti i punti di addestramento, cioè il vettore di coefficienti w che minimizza
Supponiamo di cercare di adattare un insieme di n punti, ognuno dei quali ha m dimensioni. Le prime m - 1 dimensioni di ogni punto sono il vettore di caratteristiche ($x_1, \ldots, x_{m-1}$), con l'ultimo valore $y = x_m$ che funge da variabile target o dipendente. Possiamo codificare questi n vettori di caratteristiche× come- una matrice n (m 1). Possiamo renderla× una matrice n m A aggiungendo una colonna di uno alla matrice. Questa colonna può essere considerata come una caratteristica 'costante', che moltiplicata per il coefficiente appropriato diventa l' *intercetta* y della linea adattata. Inoltre, gli n valori target possono essere ben rappresentati in un vettore n× 1 b. La linea di regressione ottimale $f(x)$ che cerchiamo è definita da un vettore m× 1 di coefficienti . La valutazione di questa funzione su questi punti è esattamente il prodotto A- w, che crea un vettore n× 1 di previsioni del valore target. Come possiamo trovare i coefficienti della linea più adatta? Innanzitutto, cerchiamo di capire questo concetto prima di cercare di comprenderlo. Le dimensioni del termine sulla destra sono che corrisponde esattamente alle dimensioni del vettore target w, quindi va bene. Inoltre, (A^T A) definisce la matrice di covarianza sulle colonne/caratteristiche della matrice di dati, e invertirla è simile a risolvere un sistema di equazioni. Il termine $A^T b$ calcola i prodotti di punti dei valori dei dati e dei valori target per ciascuna delle m caratteristiche, fornendo una misura della correlazione di ciascuna caratteristica con i risultati target. Non abbiamo ancora capito perché funziona, ma dovrebbe essere chiaro che questa equazione è composta da componenti significativi. Consideriamo il caso di una singola variabile x, in cui cerchiamo la linea di miglior adattamento della forma. Il collegamento con il coefficiente di correlazione (r_{xy}) qui è chiaro. Se x non fosse correlato con y (r_{xy}= 0), allora w_1 dovrebbe essere pari a zero. Anche se fossero perfettamente correlati (r_{xy}= 1), dobbiamo scalare x per portarlo nel giusto intervallo di dimensioni di y. Questo è il ruolo di σ_y/σ_x. Ora, da dove viene la formula di regressione lineare? Dovrebbe essere

chiaro che nella linea di miglior adattamento, non possiamo cambiare nessuno dei coefficienti w e sperare di ottenere un adattamento migliore. Ciò significa che il vettore di errore (b - Aw) deve essere ortogonale al vettore associato a ciascuna variabile x_i, altrimenti ci sarebbe un modo per modificare il coefficiente per adattarlo meglio. I vettori ortogonali hanno prodotti di punti pari a zero. Poiché l'iesima colonna di A^Tha un prodotto di punti pari

a zero con il vettore di errore.

9.2 Migliori modelli di regressione

Data una matrice A di n punti, ciascuno di m - 1 dimensioni, e una matrice target× b *di* n 1, possiamo invertire e moltiplicare le matrici appropriate per ottenere la matrice di coefficienti desiderata w. Questo definisce un modello di regressione. Fatto!

Tuttavia, ci sono diversi passi da compiere che possono portare a modelli di regressione migliori. Alcuni di questi comportano la manipolazione dei dati di ingresso per aumentare la probabilità di un modello accurato, ma altri richiedono questioni più concettuali su come *dovrebbe essere* il nostro modello.

9.2.1 Rimozione dei valori erratici

La regressione lineare cerca la retta y= f (x) che minimizza la somma degli errori al quadrato su tutti i punti di addestramento.

A causa del peso quadratico dei residui, i punti fuori norma possono influenzare notevolmente l'adattamento. Un punto a distanza 10 dalla sua previsione ha un impatto 100 volte maggiore sull'errore di formazione rispetto a un punto a solo 1 unità dalla linea adattata. Si potrebbe obiettare che questo è appropriato, ma dovrebbe essere chiaro che i punti outlier hanno un grande impatto sulla forma della linea di miglior adattamento. Questo crea un problema quando i punti anomali riflettono il rumore piuttosto che il segnale, perché la linea di regressione si sforza di adattarsi ai dati negativi invece di adattarsi a quelli buoni. L'adattamento a destra è molto migliore: con un r^2 di 0,917 senza l'outlier, rispetto a 0,548 con l'outlier. Pertanto, l'identificazione dei punti fuori norma e la loro rimozione in un modo basato su principi può produrre un adattamento più robusto. L'approccio più semplice consiste nell'adattare l'intera serie di punti e poi utilizzare la grandezza del residuo r_i= $(y_i - f(x_i))^2$ per decidere se il punto p_i è un outlier. Tuttavia, è importante convincersi che questi punti rappresentano *davvero* degli errori prima di . In caso contrario, si ritroverà con un adattamento straordinariamente lineare che funziona bene solo sugli esempi che lei non ha cancellato.

9.2.2 Adattamento di funzioni non lineari

Le relazioni lineari sono più facili da capire rispetto a quelle non lineari, e sono grossolanamente appropriate come ipotesi predefinita in assenza di dati migliori. Molti fenomeni *sono* di natura lineare, con la variabile dipendente che cresce in modo approssimativamente proporzionale alle variabili di ingresso:

- Il reddito cresce in modo approssimativamente lineare con la quantità di tempo lavorato.

- Il prezzo di una casa cresce in modo approssimativamente lineare con le dimensioni dell'area abitabile che contiene.

- Il peso delle persone aumenta in modo approssimativamente lineare con la quantità di cibo consumato.

La regressione lineare è ottima quando cerca di adattarsi a dati che in effetti hanno una relazione lineare sottostante. Ma, in generale, nessuna funzione interessante è *perfettamente* lineare. Infatti, c'è una vecchia

regola statistica che afferma che se si vuole che una funzione sia lineare, bisogna misurarla solo in due punti. Potremmo aumentare notevolmente il repertorio di forme che possiamo modellare se andiamo oltre le funzioni lineari. La regressione lineare si adatta alle linee, non alle curve di ordine superiore. Ma possiamo adattarci a funzioni quadratiche aggiungendo una variabile extra con il valore x^2 alla nostra matrice di dati, oltre a x. Il modello $y = w_{(0)} + w_1 x + w_2 x^2$ è quadratico, ma si noti che è una funzione lineare dei suoi valori di ingresso non lineari. Possiamo adattare funzioni arbitrariamente complesse aggiungendo le giuste variabili di ordine superiore alla nostra matrice di dati e formando combinazioni lineari di esse. Possiamo adattare polinomi ed esponenziali/logaritmi arbitrari includendo esplicitamente le variabili compone nti giuste nella nostra matrice di dati, come x, lg(x), x^3 e $1/x$. Si possono utilizzare anche ~~ca~~ratteristiche aggiuntive per catturare interazioni non lineari tra coppie di variabili di input. L'area di un rettangolo A è calcolata *in lunghezza× e larghezza*, il che significa che non si può ottenere un'approssimazione accurata di A come combinazione lineare di lunghezza e larghezza. Ma, una volta aggiunta una funzione di area alla nostra matrice di dati, questa interazione non lineare può essere catturata con un modello lineare. Tuttavia, l'inclusione esplicita di tutti i possibili termini non lineari diventa rapidamente intrattabile. L'aggiunta di tutte le potenze x^i per farà esplodere la matrice dei dati di un fattore k. Includere tutte le coppie di prodotti tra n variabili è ancora peggio, rendendo la matrice volte più grande. Bisogna essere cauti nel decidere quali termini non lineari considerare per un ruolo nel modello. Infatti, uno dei vantaggi dei metodi di apprendimento più potenti, come le macchine vettoriali di supporto, sarà che possono incorporare termini non lineari senza enumerazione esplicita.

9.2.3 Scala delle caratteristiche e degli obiettivi

In linea di principio, la regressione lineare può trovare il miglior modello lineare che si adatta a qualsiasi serie di dati. Ma dobbiamo fare tutto il possibile per aiutarla a trovare il modello giusto. In genere, questo comporta una pre-elaborazione dei dati per ottimizzare l'esprimibilità, l'interpretabilità e la stabilità numerica. Il problema è che le caratteristiche che variano su ampi intervalli numerici richiedono coefficienti su intervalli altrettanto ampi per riunirle. Supponiamo di voler costruire un modello per prevedere il prodotto nazionale lordo dei Paesi in dollari, in funzione delle dimensioni della popolazione x_1 e del tasso di alfabetizzazione x_2. Entrambi i fattori sembrano componenti ragionevoli di un tale modello. In effetti, entrambi i fattori potrebbero contribuire in egual misura alla quantità di attività economica. Ma operano su scale completamente diverse: le popolazioni nazionali variano da decine di migliaia a oltre un miliardo di persone, mentre la frazione di persone che sanno leggere è, per definizione, compresa tra zero e uno.

Questo è molto negativo, per diverse ragioni:

- *Coefficienti illeggibili*: Velocemente, qual è il coefficiente di x_2 nell'equazione precedente? Per noi è difficile gestire l'entità di questi numeri (si tratta di 10.000 miliardi), e ci è difficile dire quale variabile dia un contributo più importante al risultato, dati i loro intervalli. È x_1 o x_2?

- *Imprecisione numerica*: Gli algoritmi di ottimizzazione numerica hanno problemi quando i valori sono di molti ordini di grandezza. Non si tratta solo del fatto che i numeri in virgola mobile sono rappresentati da un numero finito di bit. Più importante è il fatto che molti algoritmi di apprendimento automatico vengono par- tecipati da costanti che devono valere contemporaneamente per tutte le variabili. Ad esempio, l'utilizzo di dimensioni fisse dei passi nella ricerca per discesa del gradiente (di cui si parlerà nella Sezione 9.4.2) potrebbe causare un superamento selvaggio in alcune direzioni e un superamento in altre.

- *Formulazioni inadeguate*: Il modello fornito sopra per prevedere il PIL è sciocco a vista. Supponiamo che io decida di formare il mio Paese, avrebbe esattamente una persona al interno, in grado di leggere. che può essere interpretato come ciascuna delle persone x_1 che creano ricchezza ad un tasso modulato dalla loro alfabetizzazione. In genere, questo richiede la semina della matrice di dati con i termini di prodotto appropriati. Ma c'è la possibilità con un'adeguata scalatura (logaritmica) dei target, questo modello possa uscire direttamente dalla regressione lineare.

Considereremo ora tre diverse forme di scalatura, che affrontano questi diversi tipi di problemi.

Scala delle caratteristiche: Punteggi Z

Abbiamo parlato in precedenza dei punteggi Z, che scalano i valori di ogni caratteristica individualmente, in modo che la media sia zero e gli intervalli siano comparabili. Facciamo in modo che μ sia il valore medio di una determinata caratteristica e σ la deviazione standard. L'utilizzo dei punteggi Z nella regressione risolve la questione dell'interpretabilità. Poiché ¯ tutte le caratteristiche avranno medie e varianze simili, la grandezza dei coefficienti Z determinerà l'importanza relativa di questi fattori per la previsione. Infatti, in condizioni adeguate, questi coefficienti rifletteranno il coefficiente di correlazione di ogni variabile con l'obiettivo. Inoltre, il fatto che queste variabili abbiano la stessa ampiezza semplifica il lavoro dell'algoritmo di ottimizzazione.

Scala delle caratteristiche sublineari

Consideriamo un modello lineare per prevedere il numero di anni di istruzione y che un bambino riceverà in funzione del reddito familiare. Un enorme divario tra i valori più grandi/piccoli e la mediana significa che nessun coefficiente può utilizzare la caratteristica senza un'esplosione sui valori grandi. Il livello di reddito è distribuito a legge di potenza e i punteggi Z di tali variabili a legge di potenza non possono aiutare, perché sono solo trasformazioni lineari. La chiave è posizionare nuovamente/aumentare tali caratteristiche x con funzioni sublineari come $\log(x)$ e $\sqrt{x}$. I punteggi Z di queste variabili trasformate si riveleranno molto più significativi per costruire modelli.

Scala target sublineare

Le variabili su piccola scala necessitano di obiettivi su piccola scala, per essere realizzati con coefficienti su piccola scala. Cercare di prevedere il PIL da variabili con punteggio Z richiederà coefficienti enormemente grandi. La soluzione qui è che cercare di prevedere il *logaritmo* ($\log_c(y)$) di un obiettivo a legge di potenza y è solitamente migliore della previsione di y stessa. Naturalmente, il valore $c^{f(x)}$ può essere utilizzato per stimare y, ma ora esiste il potenziale per fare previsioni significative sull'intera gamma di valori. Colpire una funzione di legge di potenza con un logaritmo produce generalmente una distribuzione più normale e con un comportamento migliore.

Questo non potrebbe mai essere realizzato con la regressione lineare senza variabili di interazione. In questo modo è possibile realizzare i logaritmi di prodotti di interazione arbitrari, a condizione che la matrice delle caratteristiche contenga anche i loghi delle variabili di ingresso originali.

9.2.4 Gestire le caratteristiche altamente correlate

È fantastico avere caratteristiche altamente correlate con l'obiettivo: queste ci permettono di costruire modelli altamente predittivi. Tuttavia, avere più caratteristiche altamente correlate *tra loro* può creare problemi. Supponiamo di avere due caratteristiche perfettamente correlate nella matrice di dati, ad esempio l'altezza

del soggetto in piedi (x_1) e la sua altezza in metri (x_2). Poiché 1 metro equivale a 3,28084 piedi, queste due variabili sono perfettamente correlate. Ma avere entrambe queste variabili non può aiutare il nostro modello, perché l'aggiunta di una caratteristica perfettamente correlata non fornisce informazioni aggiuntive per fare previsioni. Se queste caratteristiche duplicate avessero davvero un valor e per noi, implicherebbe che potremmo costruire modelli sempre più accurati semplicemente facendo copie aggiuntive di colonne da qualsiasi matrice di dati! Ma le caratteristiche correlate sono dannose per i modelli, non solo neutre. Supponiamo che la nostra variabile dipendente sia una funzione dell'altezza. Si noti che si possono costruire modelli altrettanto buoni dipendenti solo da x_1, o solo da x_2, o da qualsiasi combinazione lineare arbitraria di x_1 e x_2. Qual è il modello giusto da riportare come risposta? Questo crea confusione, ma possono accadere cose ancora peggiori. Le righe della matrice di co - varianza saranno reciprocamente dipendenti, quindi il calcolo di richiede ora l'inversione di una matrice singolare! I metodi numerici per calcolare la regressione possono fallire. La soluzione in questo caso è identificare le coppie di caratteristiche che hanno una correlazione eccessivamente forte, calcolando la matrice di covarianza appropriata. Se sono in agguato, si può eliminare una delle due variabili con poca perdita di potenza. La cosa migliore è eliminare completamente queste correlazioni, combinando le caratteristiche. Questo è uno dei problemi risolti dalla riduzione delle dimensioni, utilizzando tecniche come la decomposizione del valore singolare.

9.3 Regressione come adattamento dei parametri

La formula in forma chiusa per la regressione lineare è concisa ed elegante. Tuttavia, presenta alcuni problemi che la rendono subottimale per il calcolo nella pratica. L'inversione della matrice è lenta per i sistemi di grandi dimensioni e soggetta a instabilità numerica. Inoltre, la formulazione è fragile: la magia dell'algebra lineare è difficile da estendere a problemi di ottimizzazione più generali. Ma esiste un modo alternativo per formulare e risolvere i problemi di regressione lineare, che si rivela migliore nella pratica. Questo approccio porta ad algoritmi più veloci, a numeri più robusti e può essere facilmente adattato ad altri algoritmi di apprendimento. Modella la regressione lineare come un problema di *adattamento dei parametri* e impiega algoritmi di ricerca per trovare i migliori valori possibili per questi parametri. Per la regressione lineare, cerchiamo la linea che meglio si adatta ai punti, su tutti i set di coefficienti possibili. In particolare, cerchiamo la retta $y = f(x)$ che minimizza la somma degli errori al quadrato su tutti i punti di addestramento, cioè il vettore di coefficienti w

Per concretezza, iniziamo con il caso in cui stiamo cercando di modellare y come una funzione lineare di una singola variabile o caratteristica x, quindi $y = f(x)$. Per definire la nostra linea di regressione, cerchiamo la coppia di parametri (w_0, w_1) che minimizza l'errore o il costo o la *perdita*, ossia la somma dei quadrati di deviazione tra i valori dei punti e la linea. *Ogni* possibile coppia di valori per (w_0, w_1) definirà *una qualche* linea, ma in realtà vogliamo i valori che minimizzano l'errore o la *funzione di perdita.* Il risultato della discussione precedente è che la funzione di perdita $J(w_0, w_1)$ definisce una superficie nello spazio (w_0, w_1); il nostro interesse è il punto in questo spazio con il valore z più piccolo, dove $z = J(w_0, w_1)$. Cominciamo a semplificare ulteriormente, forzando la nostra linea di regressione a passare attraverso l'origine, impostando $w_0 = 0$. Questo ci lascia solo un parametro libero da trovare, ovvero la pendenza della linea w_1. L'aspetto interessante è che la funzione di errore ha la forma di una parabola. Raggiunge un unico valore minimo nella parte inferiore della curva. Il *valore* x di questo punto di minimo definisce la migliore pendenza w_1 per la linea di regressione, che si dà il caso sia $w_1 = 1$. Qualsiasi superficie *convessa* ha esattamente un minimo locale. Inoltre, per qualsiasi spazio di ricerca convesso, è abbastanza facile trovare questo minimo: basta continuare a camminare in una direzione verso il basso finché non lo si raggiunge. Da ogni punto della superficie, possiamo fare un piccolo passo verso un punto vicino della superficie. Alcune direzioni ci porteranno fino a un punto di riferimento.

valore più alto, ma altri ci porteranno in basso. Se riusciamo a identificare quale passo ci porterà più in basso, ci avvicineremo ai minimi. E c'è sempre questa direzione, tranne quando ci troviamo sul punto minimo stesso!
La funzione di perdita $J(w_0, w_1)$ assomiglia a una ciotola con un unico valore z- più piccolo, che definisce i valori ottimali per i due parametri della linea. Il bello è che questa funzione di perdita $J(w_0, w_1)$ è di nuovo convessa, e in effetti rimane convessa per qualsiasi problema di regressione lineare in qualsiasi numero di dimensioni. Come possiamo sapere se una determinata funzione è convessa? Ricorda quando hai studiato il calcolo in una variabile, x. Hai imparato a prendere la *derivata* $f'(x)$ di una funzione $f(x)$, che corrisponde al valore della pendenza superficie di $f(x)$ in ogni punto. Ogni volta che questa derivata era zero, significava che si era raggiunto un punto di interesse, sia esso un massimo o un minimo locale. Ricordiamo la *seconda derivata* $f''(x)$, che era la funzione derivata della derivata $f'(x)$. A seconda del segno di questa seconda derivata $f''(x)$, è possibile potrebbe identificare se ha colpito un massimo o un minimo. In conclusione, l'analisi di tali derivati può dirci quali funzioni sono e non sono convesse. Non approfondiremo in questa sede. Ma una volta stabilito che la nostra funzione di perdita è convessa, sappiamo che possiamo fidarci di una procedura come la *ricerca per discesa del gradiente*, che ci porta all'optima globale camminando verso il basso.

9.3.1 Ricerca per discesa del gradiente

Possiamo trovare i minimi di una funzione convessa semplicemente partendo da un punto arbitrario e camminando ripetutamente in una direzione verso il basso. Esiste un solo punto dove non c'è via d': il minimo globale stesso. Ed è questo punto definisce i parametri della linea di regressione più adatta. Ma come possiamo trovare una direzione che ci porti giù per la collina? Di nuovo, consideriamo prima il caso di una singola variabile, quindi cerchiamo la pendenza w_1 della linea più adatta dove $w_0 = 0$. Supponiamo che il nostro candidato attuale alla pendenza sia x_0. In questa impostazione restrittiva monodimensionale, possiamo solo spostarci a sinistra o a destra. Proviamo un piccolo passo in ogni direzione, allora dovremmo spostarci a destra. Se nessuno dei due casi è vero, significa che non abbiamo un posto dove andare per ridurre J, quindi dobbiamo aver trovato i minimi. La direzione verso il basso di $f(x_0)$ è definita dalla pendenza *della linea tangente* in questo punto. Una pendenza positiva significa che il minimo deve trovarsi a sinistra, mentre una pendenza negativa lo colloca a destra. L'entità di questa pendenza descrive la ripidità di questa discesa Questo è esattamente ciò che viene fatto nel calcolo della derivata, che in ogni punto specifica la tangente alla curva. Quando ci spostiamo oltre una dimensione, otteniamo la libertà di muoverci in una gamma più ampia di direzioni. I movimenti diagonali ci permettono di tagliare più dimensioni contemporaneamente. Ma in linea di principio, possiamo ottenere lo stesso effetto facendo più passi, lungo ogni dimensione distinta in una direzione orientata all'asse. Pensi alla griglia stradale di Manhattan, dove possiamo arrivare ovunque vogliamo muovendoci in una combinazione di passi nord-sud ed est-ovest. Per trovare queste direzioni è necessario calcolare la *derivata parziale* della funzione obiettivo lungo ogni dimensione, ossia: Ma zig-zagare lungo le dimensioni sembra lento e goffo. Come Superman, vogliamo saltare gli edifici in un solo colpo. L'ampiezza delle derivate parziali definisce la pendenza in ogni direzione, e il vettore risultante (ad esempio tre passi verso ovest per ogni passo verso nord) definisce il percorso più veloce per scendere da questo punto.

9.3.2 Qual è il giusto tasso di apprendimento?

La derivata della funzione di perdita ci indica la direzione giusta per camminare verso i minimi, che specificano i parametri per risolvere il nostro problema di regressione. Ma non ci dice quanto lontano camminare. Il valore di questa direzione diminuisce con la distanza. La ricerca per discesa graduale opera per turni: trova la direzione migliore, fa un passo e poi ripete fino a quando non raggiunge l'obiettivo. La dimensione del nostro passo si chiama *tasso di apprendimento* e definisce la velocità con cui troviamo i minimi. Facendo piccoli passi e consultando ripetutamente la mappa (cioè le derivate parziali) si arriva effettivamente a destinazione, ma solo molto lentamente. Tuttavia, le dimensioni non sono sempre migliori. Se il tasso di apprendimento è troppo alto, potremmo saltare oltre i minimi. Questo potrebbe significare un lento L'adozione di una dimensione di passo

troppo piccola richiede molte iterazioni per convergere, mentre una dimensione di passo troppo grande ci porta a superare i minimi. progresso verso il buco, in quanto rimbalziamo al di là di esso ad ogni passo, o addirittura un progresso negativo, in quanto ci ritroviamo con un valore di $J(w)$ più alto di quello che avevamo prima. In linea di principio, vogliamo un tasso di apprendimento elevato all'inizio della nostra ricerca, ma che diminuisca man mano che ci avviciniamo al nostro obiettivo. Dobbiamo monitorare il valore della nostra funzione di perdita nel corso dell'ottimizzazione. Se i progressi diventano troppo lenti, possiamo aumentare la dimensione del passo di un fattore moltiplicativo (ad esempio 3) o rinunciare: accettare i valori attuali dei parametri per la nostra linea di adattamento come sufficienti. Ma se il valore di $J(w)$ aumenta, significa che abbiamo superato il nostro obiettivo. Quindi la nostra dimensione di passo era troppo grande, quindi dovremmo diminuire il tasso di apprendimento di un fattore moltiplicativo: ad esempio di 1/3. I dettagli di questo sono disordinati, euristici e ad hoc. Ma fortunatamente le funzioni di libreria per la ricerca per discesa del gradiente hanno algoritmi integrati per regolare il tasso di apprendimento. Presumibilmente questi algoritmi sono stati altamente regolati e in genere dovrebbero fare ciò che dovrebbero fare. Ma la forma della superficie fa una grande differenza per quanto riguarda il successo della ricerca di discesa del gradiente nel trovare il minimo globale. Se la nostra superficie a forma di scodella fosse relativamente piatta, come un piatto, il punto veramente minimo potrebbe essere oscurato da una nuvola di rumore e di errori numerici. Anche se alla fine troviamo il minimo, potrebbe volerci molto tempo. Tuttavia, succede anche di peggio quando la nostra funzione di perdita non è convessa, il che significa che ci possono essere molti minimi locali. Ora, questo può essere il caso della regressione lineare, ma succede per molti altri problemi di apprendimento automatico interessanti che incontreremo. Ora, questo non può essere il caso della regressione lineare, ma accade per molti altri problemi interessanti di apprendimento automatico che incontreremo. L'ottimizzazione locale può facilmente bloccarsi in minimi locali per funzioni non convesse. Questo è il valore delle euristiche di ricerca come l'annealing simulato, che fornisce una via d'uscita da piccoli optima locali per continuare ad avanzare verso l'obiettivo globale.

9.3.3 Discesa del gradiente stocastica

La definizione algebrica della nostra funzione di perdita nasconde qualcosa di molto costoso. Si tratta di una somma. Per calcolare la migliore direzione e il tasso di variazione per ogni dimensione j, dobbiamo scorrere *tutti gli* n punti di formazione. La valutazione di ogni derivata parziale richiede un tempo lineare nel numero di esempi, per ogni fase! Per la regressione lineare sul nostro sontuoso set di dati taxi, questo significa 80 milioni di calcoli di differenza quadratica solo per identificare la direzione migliore in assoluto per avanzare di un passo verso l'obiettivo. Questa è una follia. Invece, possiamo provare un'approssimazione che utilizza solo un piccolo numero di esempi per stimare la derivata, sperando che la direzione risultante punti effettivamente verso il basso. In media dovrebbe, dato che ogni punto alla fine voterà sulla direzione. La *discesa stocastica del gradiente* è un approccio di ottimizzazione basato sul campionamento di un piccolo lotto di punti di formazione, idealmente a caso, e sul loro utilizzo per stimare la derivata nella nostra posizione attuale. Più piccola è la dimensione del lotto che utilizziamo, più veloce la valutazione, anche se dobbiamo essere più scettici sul fatto che la direzione stimata sia corretta. Ottimizzare il tasso di apprendimento e la dimensione del batch per il gradiente
La discesa porta a un'ottimizzazione molto veloce per le funzioni convesse, con i dettagli fortunatamente nascosti da una chiamata a una funzione di libreria. Può essere costoso fare scelte casuali in ogni fase della ricerca. La cosa migliore è randomizzare l'ordine degli esempi di addestramento una volta, per evitare artefatti sistematici nel modo in cui vengono presentati, e poi costruire i nostri batch semplicemente scorrendo l'elenco. In questo modo possiamo assicurarci che tutte le n istanze di addestramento alla fine contribuiscano alla ricerca, idealmente più volte, dato che ripetiamo tutti gli esempi nel corso dell'ottimizzazione.

9.4 Semplificare i modelli attraverso la regolarizzazione

La regressione lineare è felice di determinare il *miglior* adattamento lineare *possibile* a qualsiasi raccolta di n punti di dati, ciascuno specificato- da m 1 variabili indipendenti e da un determinato valore target. Ma il 'miglior' adattamento potrebbe non essere quello che vogliamo *veramente*. Il problema è questo. La maggior

parte delle- m 1 possibili caratteristiche può essere non correlata con l'obiettivo e quindi non ha un reale potere predittivo. In genere, queste si presenteranno come variabili con coefficienti piccoli. Tuttavia, l'algoritmo di regressione utilizzerà questi valori per spostare la linea in modo da ridurre l'errore quadratico minimo sugli esempi di formazione forniti. Usare il rumore (le variabili non correlate) per adattarsi al rumore (il residuo lasciato da un modello semplice sulle variabili realmente correlate) è un problema. Rappresentativa è la nostra esperienza con il modello di mancia dei taxi, come dettagliato nella storia di guerra. Il modello di regressione completo che utilizza dieci variabili ha avuto un errore quadratico medio di 1,5448. Il modello di regressione a singola variabile che opera solo sulla tariffa ha fatto leggermente peggio, con un errore di 1,5487. Ma questa differenza è solo rumore. Il modello a una sola variabile è ovviamente migliore, secondo il rasoio di Occam o di chiunque altro. Altri problemi sorgono quando si utilizza la regressione non vincolata. Abbiamo visto come le caratteristiche fortemente correlate introducano ambiguità nel modello. Se le caratteristiche A e B sono perfettamente correlate, l'utilizzo di entrambe produce la stessa precisione dell'utilizzo di una delle due, con il risultato di modelli più complicati e meno interpretabili. Fornire una ricca serie di caratteristiche alla regressione è positivo, ma ricordi che "la spiegazione più semplice è la migliore". La spiegazione più semplice si basa sul minor numero di variabili che fanno un buon lavoro di modellazione dei dati. Idealmente, la nostra regressione dovrebbe selezionare le variabili più importanti e adattarle, ma la funzione obiettivo che abbiamo discusso cerca solo di minimizzare l'errore della somma dei quadrati. Dobbiamo cambiare la nostra funzione obiettivo, attraverso la magia della regolarizzazione.

9.4.1 Regressione di cresta

La regolarizzazione è il trucco di aggiungere termini secondari alla funzione obiettivo per favorire i modelli che mantengono i coefficienti piccoli. Supponiamo di generalizzare la nostra funzione di perdita con una seconda serie di termini che siano una funzione dei coefficienti. In questa formulazione, paghiamo una penalità proporzionale alla somma dei quadrati dei coefficienti utilizzati nel modello. Elevando al quadrato i coefficienti, ignoriamo il segno e ci concentriamo sulla grandezza. La costante λ modula la forza relativa dei vincoli di regolarizzazione. Più λ è alto, più l'ottimizzazione lavorerà per ridurre la dimensione dei coefficienti, a scapito di un aumento dei residui. Alla fine, diventa più conveniente azzerare il coefficiente di una variabile non correlata, piuttosto che utilizzarla per sovra adattare il set di formazione. La penalizzazione della somma dei coefficienti al quadrato, come nella funzione di perdita di cui sopra, è chiamata *regressione di cresta* o *regolarizzazione di Tikhonov*. Supponendo che le variabili dipendenti siano state tutte normalizzate in modo appropriato alla media zero, la grandezza del loro coefficiente è una misura del loro valore per la funzione obiettivo. Come possiamo ottimizzare i parametri per la regressione ridge? Un'estensione naturale della formulazione dei minimi quadrati fa il lavoro. Lasci che Γ sia la nostra× matrice n n di "penalizzazione del peso del coefficiente". Per semplicità, lasciamo che $\Gamma = I$, la matrice di identità. La funzione di perdita somma dei quadrati che cerchiamo di minimizzare. La notazione v indica la *norma* di v, una funzione di distanza su un vettore o una matrice. La norma è esattamente la somma dei quadrati dei coefficienti quando $\Gamma = I$. Vista questo modo, la forma chiusa per ottimizzare per w è credibile. Quindi l'equazione della forma normale può essere generalizzata per gestire la regolarizzazione. Ma, in alternativa, possiamo calcolare le derivate parziali di questa funzione di perdita e utilizzare la ricerca per discesa del gradiente per svolgere il lavoro più velocemente su matrici di grandi dimensioni. In ogni caso, le funzioni di libreria per la regressione ridge e la cugina regressione LASSO saranno prontamente disponibili per essere utilizzate nel suo problema.

9.4.2 Regressione LASSO

La regressione Ridge ottimizza la selezione di coefficienti piccoli. A causa della funzione di costo della somma dei quadrati, punisce in particolare i coefficienti più grandi. Questo rende ideale evitare i modelli della forma, dove w_0

è un grande numero positivo e w_1 un grande numero negativo. Sebbene la regressione ridge sia efficace nel ridurre l'entità dei coefficienti, questo criterio non li spinge realmente a zero ed elimina totalmente la variabile dal modello. Una scelta alternativa è quella di cercare di minimizzare la somma dei valori assoluti dei coefficienti, che è altrettanto felice di ridurre i coefficienti più piccoli come quelli più grandi. La *regressione LASSO* (per "Least Absolute Shrinkage and Selection Operator") soddisfa questo criterio: minimizza la metrica L_1 sui coefficienti invece della metrica L della regressione ridge. Con LASSO, specifichiamo un vincolo esplicito t su quale può essere la somma dei coefficienti, e l'ottimizzazione minimizza la somma dei quadrati dell'errore sotto questo vincolo.
Specificando un valore più piccolo di t, il LASSO si restringe, vincolando ulteriormente le grandezze dei coefficienti w. Come esempio di come LASSO azzeri i coefficienti piccoli, osserviamo cosa ha fatto al modello di mancia dei taxi per un particolare valore di t: Come si può vedere, LASSO ha azzerato la maggior parte dei coefficienti, un modello più semplice e robusto, che si adatta ai dati quasi come regressione lineare non vincolata.
Ma perché LASSO guida attivamente i coefficienti a zero? Ha a che fare con la forma del cerchio della metrica L la forma del cerchio L (l'insieme di punti equidistanti dall'origine) non è rotonda, ma ha vertici e caratteristiche di dimensioni inferiori come bordi e facce. Vincolare i nostri coefficienti w a giacere sulla superficie di un cerchio L_1 di raggio-t significa che è probabile che colpisca una di queste caratteristiche di dimensioni inferiori, il che significa che le dimensioni inutilizzate ricevono coefficienti zero. Quale funziona meglio, LASSO o la regressione ridge? La risposta è che dipende da lei. Entrambi i metodi dovrebbero essere supportati dalla sua libreria di ottimizzazione preferita, quindi li provi e veda cosa succede.

9.4.3 Scambio tra adattamento e complessità

Come si imposta il valore giusto per il nostro parametro di regolarizzazione, che sia λ o t? L'utilizzo di un λ sufficientemente piccolo o di un t sufficientemente grande fornisce una piccola penalizzazione rispetto alla selezione dei coefficienti per minimizzare l'errore di formazione. Al contrario, l'utilizzo di un λ molto grande o di un t molto piccolo garantisce coefficienti piccoli, anche a costo di una sostanziale errore di modellazione. La regolazione di questi parametri ci permette di cercare il punto di equilibrio tra il sovra e il sotto-adattamento. Ottimizzando questi modelli su un'ampia gamma di valori per il parametro di regolarizzazione t appropriato, otteniamo un grafico dell'errore di valutazione in funzione di t. Un buon adattamento ai dati di formazione con pochi/piccoli parametri è più robusto di un adattamento leggermente migliore con molti parametri.
La gestione di questo compromesso è in gran parte una questione di gusto. Tuttavia, sono stati sviluppati diversi metodi per aiutare la selezione dei modelli. I più importanti sono i *criteri di informazione di Akaike* (AIC) e i *criteri di informazione baysiani* (BIC). Non approfondiremo oltre i loro nomi, quindi è giusto che lei queste metriche come voodoo a questo. Tuttavia, il suo sistema di ottimizzazione/valutazione potrebbe emetterli per i modelli adattati che produce, fornendo un modo per confrontare i modelli con un numero diverso di parametri. Anche se la regressione LASSO/ridge punisce i coefficienti basati sulla magni- tudine, non li azzera esplicitamente se vuole esattamente k parametri. Deve essere lei a rimuovere le variabili inutili dal suo modello. I metodi di selezione automatica delle caratteristiche potrebbero decidere di azzerare i coefficienti piccoli, ma costruire esplicitamente i modelli da tutti i possibili sottoinsiemi di caratteristiche è in genere computazionalmente impossibile. Le caratteristiche da rimuovere per prime dovrebbero essere quelle con (a) coefficienti piccoli, (b) bassa correlazione con la funzione obiettivo, (c) alta correlazione con un'altra caratteristica del modello e (d) nessuna relazione ovvia e giustificabile con l'obiettivo. Per esempio, un famoso studio una volta ha mostrato una forte correlazione tra la funzione obiettivo e l'obiettivo. Il prodotto nazionale lordo degli Stati Uniti e il volume annuale della produzione di burro in Bangladesh. Il saggio modellatore può rifiutare questa variabile come ridicola, in modi che i metodi automatizzati non possono fare.

9.5 Classificazione e regressione logistica

La classificazione è il problema di prevedere l'etichetta giusta per un dato record di input. Il compito si differenzia dalla regressione in quanto le etichette sono entità discrete, non valori di funzioni continue. Cercare di scegliere la risposta giusta tra due possibilità può sembrare più facile che prevedere quantità aperte, ma è anche molto più facile essere criticati per aver sbagliato. In questa sezione, verranno sviluppati gli approcci per costruire sistemi di classificazione utilizzando la regressione lineare, ma questo è solo l'inizio. La classificazione è un problema fondamentale nella scienza dei dati, e vedremo diversi altri approcci nei prossimi due capitoli.

9.5.1 Regressione per la classificazione

Possiamo applicare la regressione lineare ai problemi di classificazione convertendo i nomi delle classi degli esempi di formazione in numeri. Per ora, limitiamo la nostra attenzione ai problemi a due classi, o alla *classificazione binaria*. Generalizzeremo questo a problemi multiclasse. La numerazione di queste classi come 0/1 va bene per i classificatori binari. Per convenzione, la classe 'positiva' ottiene 0 e quella 'negativa' 1. La classe negativa/1 generalmente denota il caso più raro o più speciale. Non c'è alcun giudizio di valore inteso qui da positivo/negativo: infatti, quando le classi sono di dimensioni uguali, la scelta viene fatta arbitrariamente. Potremmo considerare di addestrare una linea di regressione $f(x)$ per il nostro vettore di caratteristiche x, dove i valori target sono queste etichette 0/1. C'è una certa logica qui. Le istanze simili agli esempi di formazione positivi dovrebbero ottenere punteggi inferiori rispetto a quelle più vicine alle istanze negative. Possiamo settare il valore restituito da $f(x)$ per interpretarlo come un'etichetta. Tuttavia, i separatori non lineari si adattano meglio a determinati set di formazione (destra). Ma ci sono problemi con questa formulazione. Supponiamo di aggiungere un certo numero di esempi "molto negativi" ai dati di formazione. La linea di regressione si inclinerà verso questi esempi, mettendo a rischio la corretta classificazione degli esempi più marginali. Questo è un peccato, perché comunque avremmo già classificato correttamente questi punti molto negativi. Vogliamo davvero che la linea tagli tra le classi e funga da confine, invece di attraversare queste classi come un marcatore.

9.5.2 Confini decisionali

Il modo giusto di pensare alla classificazione è quello di ritagliare lo spazio delle caratteristiche in regioni, in modo che tutti i punti all'interno di una determinata regione siano destinati ad essere assegnati alla stessa etichetta. Le regioni sono definite dai loro confini, quindi vogliamo che regressione trovi linee di separazione invece di un adattamento. Le nostre speranze per una classificazione accurata si basano sulla coerenza regionale tra i punti. Ciò significa che i punti vicini tendono ad avere etichette simili e che i confini tra le regioni tendono ad essere netti anziché sfumati. Idealmente, le nostre due classi saranno ben separate nello spazio delle caratteristiche, quindi una linea può facilmente dividerle. Ma più in generale, ci saranno degli outlier. Dobbiamo giudicare il nostro classificatore dalla 'purezza' della separazione risultante, penalizzando l'errata classificazione dei punti che si trovano sul lato sbagliato della linea. Qualsiasi insieme di punti può essere perfettamente suddiviso, se progettiamo un confine abbastanza complicato che si sposta all'interno e all'esterno per catturare tutte le istanze con una determinata etichetta. Separatori così complicati di solito riflettono un overfitting del set di formazione. I separatori lineari offrono la virtù della semplicità e della robustezza e, come vedremo, possono essere costruiti efficacemente utilizzando la *regressione logistica*. Più in generale, potremmo essere interessati a confini di decisione non lineari ma a bassa complessità, se separano meglio i confini delle classi. La separazione ideale

Tuttavia, la separazione perfetta ottenuta utilizzando confini complessi di solito riflette un adattamento eccessivo più che un'intuizione (a destra). Tuttavia, può essere trovata come una funzione lineare di caratteristiche

quadratiche. Possiamo utilizzare la regressione logistica per trovare i confini non lineari se la matrice di dati è alimentata con caratteristiche non lineari.

9.5.3 Regressione logistica

Ricordiamo la funzione logit $f(x)$. Questa funzione prende in ingresso un valore reale x e produce un valore che va da [0,1], cioè una probabilità. la funzione logit $f(x)$, che è una curva sigmoidale: piatta su entrambi i lati, ma con un'impennata al centro. La forma della funzione logit la rende particolarmente adatta all'interpretazione dei confini della classificazione. In particolare, lasciamo che x sia un punteggio che riflette la distanza che un particolare punto p si trova sopra/sotto o a sinistra/destra di una linea l che separa due classi. Vogliamo che $f(x)$ misuri la probabilità che p meriti un'etichetta negativa. La funzione logit mappa i punteggi in probabilità utilizzando un solo parametro. I casi importanti sono quelli del punto medio e dei punti finali. Logit dice che $f(0)= 1/2$, il che significa che l'etichetta di un punto sul confine è essenzialmente un lancio di moneta tra le due possibilità. Questo è come dovrebbe essere. Si possono prendere decisioni più disinvolte quanto maggiore è la distanza da questo confine. La nostra fiducia in funzione della distanza è modulata costante di scala c. Un valore di c vicino a zero determina una transizione molto graduale dal positivo al negativo. Al contrario, possiamo trasformare il logit in una scala assegnando un valore abbastanza grande a c, il che significa che piccole distanze dal confine si traducono in grandi aumenti della fiducia nella classificazione. Abbiamo bisogno di tre cose per utilizzare la funzione logit in modo efficace per la classificazione:
Si noti che i coefficienti di $h(x, w)$ sono abbastanza ricchi da codificare i parametri di soglia e di ripidità (c è essenzialmente la media di w_1 attraverso. L'unica domanda che rimane è come adattare il vettore di coefficienti w ai dati di formazione. Ricordiamo che ci viene data un'etichetta di classe zero/uno y_i per ogni vettore di input. Abbiamo bisogno di una funzione di penalità che attribuisca costi appropriati alla restituzione di $f(x_i)$ come probabilità che la classe y_i sia positiva, cioè $y_i= 1$. Consideriamo innanzitutto il caso in cui y_i *è* davvero 1. Idealmente $f(x_{(i)}) = 1$ in questo caso, quindi vogliamo penalizzarlo per essere più piccolo di 1. In effetti, vogliamo punirlo in modo aggressivo quando $f(y_i)$ 0, perché ciò significa che il classificatore sta affermando che l'elemento i ha poche possibilità di rientrare nella classe 1, quando in realtà è il caso.

La funzione logaritmica risulta essere una buona funzione di penalizzazione quando. Ci sono solo due valori possibili, ossia $y_i= 0$ o $y_i= 1$. Questo ha effetto desiderato, perché la penalità viene azzerata nel caso in cui non si applica. Allo stesso modo, la moltiplicazione ha l'effetto opposto: azzera la penalità quando $y_i= 1$, e la applica quando $y_i= 0$. La moltiplicazione dei costi per le variabili indicatrici appropriate ci permette di definire la funzione di perdita per la regressione logistica. L'aspetto meraviglioso di questa funzione di perdita è che è convessa, il che significa che possiamo trovare i parametri w che si adattano meglio agli esempi di formazione utilizzando la discesa del gradiente. Così possiamo utilizzare la regressione logistica per trovare il miglior separatore lineare tra due classi, fornendo un approccio naturale alla classificazione binaria.

9.6 Problemi nella classificazione logistica

Ci sono diverse sfumature nella costruzione di classificatori efficaci, questioni che riguardano sia la regressione logistica che gli altri metodi di apprendimento automatico che esploreremo nei prossimi due capitoli. Queste includono la gestione di classi sbilanciate, la classificazione multiclasse e la costruzione di vere distribuzioni di probabilità da classificatori indipendenti.

9.6.1 Corsi di formazione equilibrati

La morale è che in genere è meglio utilizzare un numero uguale di esempi positivi e negativi. Ma una classe può essere difficile da trovare esempi. Quindi, quali sono le nostre opzioni per produrre un classificatore migliore?

- *Forzare le classi equilibrate scartando i membri della classe più grande* : Questo è il modo più semplice per realizzare classi di allenamento equilibrate. È perfettamente giustificato se si hanno abbastanza

elementi di classe rara per costruire un classificatore rispettabile. Scartando le istanze in eccesso che non ci servono, creiamo un problema più difficile che non favorisce la classe maggioritaria.

- *Replicare gli elementi della classe più piccola, idealmente con una perturbazione* : Un modo semplice per ottenere più esempi di formazione è quello di clonare i terroristi, inserendo repliche perfette di loro nel set di formazione con nomi diversi. Questi esempi ripetuti assomigliano ai terroristi, dopotutto, e aggiungerne un numero sufficiente renderà le classi equilibrate.

  Questa formulazione è tuttavia fragile. Questi record di dati identici potrebbero creare instabilità numeriche e certamente hanno una tendenza all'adattamento eccessivo, poiché spostando un terrorista reale in più sul lato destro del limite si spostano anche tutti i suoi cloni. Sarebbe meglio aggiungere una certa quantità di rumore casuale a ciascun esempio clonato, in linea con la varianza della popolazione generale. Questo fa sì che il classificatore lavori di più per trovarli e quindi minimizza l'overfit ting.

- *Pesa maggiormente gli esempi di formazione rari rispetto alle istanze della classe maggiore*: La funzione di perdita per l'ottimizzazione dei parametri contiene un termine separato per l'errore di ogni istanza di formazione. L'aggiunta di un coefficiente per attribuire un peso maggiore alle istanze più importanti lascia un problema di ottimizzazione convesso, quindi può ancora essere ottimizzato con la discesa stocastica del gradiente.

Il problema di tutte e tre queste soluzioni è che influenziamo il classificatore modificando la distribuzione di probabilità sottostante. È importante che un classificatore sappia che i terroristi sono estremamente rari nella popolazione generale, magari specificando una distribuzione preventiva baysiana.

Naturalmente, la soluzione migliore sarebbe quella di raccogliere più esempi di formazione dalla classe più rara, ma questo non è sempre possibile. Queste tre tecniche sono il meglio che possiamo trovare come alternativa.

9.6.2 Classificazione multiclasse

Spesso i compiti di classificazione comportano la scelta di più di due etichette distinte.

Un approccio naturale, ma sbagliato, per rappresentare le classi k-distinte, aggiungerebbe numeri di classe oltre a 0/1. Poi potremmo eseguire una regressione lineare per prevedere il numero di classe.

Ma questa è generalmente una cattiva idea. Le scale *ordinali* sono definite da valori crescenti o decrescenti. A meno che l'ordinamento delle sue classi non rifletta una scala ordinale, la numerazione delle classi sarà un obiettivo privo di significato rispetto al quale regredire.

Alcuni insiemi di classi sono definiti correttamente da scale ordinali.

Le classi definite da tali *scale Likert* sono ordinali, e quindi tali numeri di classe sono una cosa perfettamente ragionevole per regredire. In particolare, assegnare erroneamente un elemento a una classe adiacente è molto meno problematico che assegnarlo all'estremità sbagliata della scala.

Ma in generale, le etichette di classe non sono ordinali. Un'idea migliore per la discriminazione multi-classe prevede la costruzione di molti classificatori uno-vs-tutti. Per ciascuna delle possibili classi C_i, dove addestriamo un classificatore logistico per distinguere gli elementi di C_i dall'unione degli elementi di tutte le altre classi combinate. Per identificare l'etichetta associata a un nuovo elemento x, lo testiamo rispetto a tutti i c di questi classificatori e restituiamo l'etichetta i che ha la più alta probabilità associata.

Questo approccio dovrebbe sembrare semplice e ragionevole, ma si noti che il problema della classificazione diventa tanto più difficile quante più classi si hanno. Consideriamo la scimmia. Lanciando una moneta, una scimmia dovrebbe essere in grado di etichettare correttamente il 50% degli esempi in qualsiasi problema di classificazione binaria. Ma ora supponiamo che ci siano cento classi. La scimmia indovinerà solo l'1% delle volte. Il compito è ora molto difficile e anche un classificatore eccellente avrà difficoltà a produrre buoni risultati.

9.6.3 Classificazione gerarchica

Quando il suo problema contiene un gran numero di classi, conviene in un albero o in una gerarchia, in modo da migliorare sia l'accuratezza che l'efficienza. Supponiamo che abbiamo costruito un albero binario, dove ogni singola categoria è rappresentata da un nodo foglia. Ogni nodo interno rappresenta un classificatore per distinguere tra i discendenti di sinistra e i discendenti di destra.

Per utilizzare questa gerarchia per classificare un nuovo elemento x, iniziamo dalla radice. L'esecuzione del classificatore della radice su x specificherà la sua appartenenza al sottoalbero sinistro o destro. Scendendo di un livello, confrontiamo x con il classificatore del nuovo nodo e continuiamo a ripetere fino a quando non raggiungiamo una foglia, che definisce l'etichetta assegnata a x. Il tempo necessario è proporzionale all'altezza dell'albero, idealmente logaritmico nel numero di classi c, invece di essere lineare in c se confrontiamo esplicitamente ogni classe.

Idealmente, questa gerarchia può essere costruita a partire dalla conoscenza del dominio, assicurando che le categorie che rappresentano classi simili siano raggruppate insieme. Questo ha due vantaggi. In primo luogo, rende più probabile che le classificazioni errate producano comunque etichette di classi simili. In secondo luogo, significa che i nodi intermedi possono definire concetti di ordine superiore, che possono essere riconosciuti con maggiore precisione.

C'è un altro pericolo indipendente con la classificazione, che diventa più acuto con l'aumento del numero di classi. I membri di alcune classi sono molto più numerosi di altri. La disparità relativa tra le dimensioni delle classi più grandi e quelle più piccole cresce in genere con il numero di classi.

I sistemi di classificazione che non hanno una percezione adeguata della distribuzione preventiva delle etichette sono destinati ad avere molti falsi positivi, assegnando troppo spesso etichette rare.

Questo è il cuore dell'*analisi baysiana*: aggiornare la nostra attuale (precedente) comprensione della distribuzione di probabilità di fronte a nuove prove. In questo caso, la prova è il risultato di un classificatore. Se incorporiamo una solida distribuzione preliminare nel nostro ragionamento, possiamo garantire che gli articoli richiedano prove particolarmente forti per essere assegnati a classi rare.

9.6.4 Funzioni di partizione e regressione multinomiale

Ricordiamo che il nostro metodo preferito di classificazione multiclasse prevedeva l'addestramento di classificatori indipendenti a classe singola contro tutti i classificatori logistici il numero di etichette distinte. Rimane un problema minore. Le probabilità che otteniamo dalla regressione logistica non sono realmente probabilità.

Trasformarle in probabilità reali richiede l'idea di una *funzione di partizione*.

Per qualsiasi elemento particolare x, la somma delle 'probabilità' su tutte le possibili etichette per x *dovrebbe dare* come risultato $T = 1$.

Ma dovrebbe non significa che lo sia. Tutti questi classificatori sono stati addestrati in modo indipendente, e quindi non c'è nulla che li costringa a sommare $T = 1$.

> Una soluzione consiste nel dividere tutte queste probabilità per la costante appropriata. Questo può sembrare una forzatura, perché lo è. Ma questo è essenzialmente ciò che fanno i fisici quando parlano di funzioni di fisica. Ma questo è essenzialmente ciò che fanno i fisici quando parlano di *funzioni di partizione*, che servono come denominatori per trasformare qualcosa di proporzionale alle probabilità in probabilità reali.

La regressione multinomiale è un metodo più principiale per addestrare classificatori indipendenti a classe singola rispetto a tutti i classificatori, in modo che le probabilità funzionino correttamente. Ciò comporta l'utilizzo della funzione di partizione corretta per i rapporti di probabilità logici, che vengono combinati con esponenziali dei valori risultanti. Non dirò di più, ma è ragionevole cercare una funzione di regressione

multinomiale nella sua libreria di apprendimento automatico preferita e vedere come si comporta di fronte a un problema di regressione multiclasse.
Una nozione correlata alla funzione di partizione emerge nell'analisi baysiana. Spesso ci troviamo di fronte alla sfida di identificare l'etichetta dell'articolo più probabile, ad esempio A, in funzione dell'evidenza E.
Il calcolo di questa probabilità reale richiede la conoscenza del denominatore $P(E)$, che può essere una cosa oscura da calcolare. Ma confrontare $P(A\ E|)$ con $P(B\ |E)$ per determinare se l'etichetta A è più probabile dell'etichetta B non richiede la conoscenza di $P(E)$, poiché è la stessa in entrambe le espressioni. Come un fisico, possiamo, borbottando sulla "funzione di partizione".

Capitolo 10
Metodi di distanza e di rete

Una matrice× di dati $n\ d$, composta da n esempi/frecce ciascuno definito da d caratteristiche/colonne, definisce naturalmente un insieme di n punti in uno spazio geometrico d-*dimensionale*.
Esiste una stretta connessione tra le collezioni di punti nello spazio e i vertici nelle reti. Spesso costruiamo reti da insiemi di punti geometrici, collegando coppie di punti vicini tramite bordi. Al contrario, possiamo costruire gli insiemi di punti dalle reti, incorporando i vertici nello spazio, in modo che le coppie di vertici collegati si trovino vicine tra loro nell'incorporazione.
Molti dei problemi importanti sui dati geometrici sono facilmente generalizzabili ai dati di rete, tra cui la classificazione e il clustering dei vicini più prossimi. Pertanto, in questo capitolo trattiamo entrambi gli argomenti insieme, per sfruttare meglio le sinergie tra loro.

10.1 Misurare le distanze

Il problema più elementare nella geometria dei punti p e q in d dimensioni è come misurare al meglio la distanza tra loro. Potrebbe non essere ovvio che ci sia un problema cui parlare, dal momento che la metrica *euclidea* tradizionale è *ovviamente* come si misurano le distanze. Ma ci sono altre nozioni ragionevoli di distanza da considerare. Infatti, che cos'è una metrica di distanza? In che modo si differenzia da una funzione di punteggio arbitraria?

10.1.1 Metriche della distanza

Le misure di distanza differiscono ovviamente dai punteggi di somiglianza, come il coefficiente di correlazione, per la loro direzione di crescita. Le misure di distanza si riducono man mano che gli elementi diventano più simili, mentre l'inverso è vero per le funzioni di somiglianza.
Ci sono alcune proprietà matematiche utili che assumiamo per qualsiasi misura di distanza ragionevole. Diciamo che una misura di distanza è una *metrica* se soddisfa le seguenti proprietà:

- *Positività*: $d(x, y) \geq 0$ per tutti gli x e gli y.
- *Identità*: $d(x, y) = 0$ se e solo se $x = y$.

- *Simmetria*: $d(x, y) = d(y, x)$ per tutti gli x e gli y.

- *Disuguaglianza del triangolo*: $d(x, y) \leq d(x, z) + d(z, y)$ per tutti gli x, y e z.

Queste proprietà sono importanti per il ragionamento sui dati. Infatti, molti algoritmi funzionano correttamente solo quando la funzione di distanza è una metrica.
La distanza euclidea è una metrica, ecco perché queste condizioni ci sembrano così naturali. Tuttavia, altre misure di somiglianza altrettanto naturali non sono metriche di distanza:

- *Coefficiente di correlazione*: Fallisce la positività perché va da 1 a 1. Fallisce - anche l'identità, poiché la correlazione di una sequenza con se stessa è pari a 1.

- *Similitudine del coseno/prodotto di punti*: Simile al coefficiente di correlazione, fallisce la positività e l'identità per lo stesso motivo.

- *Tempi di percorrenza in una rete diretta*: In un mondo con strade a senso unico, la distanza da x a y non è necessariamente la stessa di quella da y a x.

- *Biglietto aereo più economico*: Questo spesso viola la disuguaglianza del triangolo, perché il modo più economico per volare da x a y potrebbe comportare una deviazione attraverso z, a causa di bizzarre strategie tariffarie delle compagnie aeree.

Al contrario, non è immediatamente evidente che alcune note funzioni di distanza sono metriche, come la distanza di modifica utilizzata nella corrispondenza delle stringhe. Invece di fare ipotesi, dimostri o confuti ciascuna delle quattro proprietà di base, per essere sicuro di capire con cosa sta lavorando.

10.1.2 La metrica della distanza L_k

La distanza euclidea è solo un caso speciale di una famiglia più generale di funzioni di distanza, nota come metrica o norma di distanza. Il parametro k fornisce un modo per fare un compromesso tra le differenze dimensionali maggiori e quelle totali. Il valore di k può essere qualsiasi numero, con valori particolarmente popolari che includono:

- *Distanza da Manhattan ($k = 1$)*: Se ignoriamo le eccezioni come Broadway, tutte le strade di Manhattan corrono in direzione est-ovest e tutti i viali in direzione nordsud, definendo così una griglia regolare. La distanza tra due località è quindi la somma di questa differenza nord-sud e della differenza est-ovest, poiché gli edifici alti impediscono qualsiasi possibilità di scorciatoia.

  Allo stesso modo, la distanza L_1 o *Manhattan* è la somma totale delle deviazioni tra le dimensioni. Tutto è lineare, quindi una differenza di 1 in ciascuna delle due dimensioni è uguale a una differenza di 2 in una sola dimensione. Poiché non possiamo trarre vantaggio dalle scorciatoie diagonali, in genere esistono molti possibili percorsi più brevi tra due punti.

- *Distanza euclidea ($k = 2$)*: Questa è la metrica di distanza più popolare, che offre un peso maggiore alla deviazione dimensionale più grande, senza sovraccaricare le dimensioni minori.

- *Componente massima ($k = \infty$)*: Con l'aumento del valore di k, le differenze dimensionali più piccole diventano irrilevanti. Se $a > b$, allora $a^{k\ b(k)}$.

La metrica L restituisce la più grande differenza monodimensionale come distanza. Siamo a nostro agio con la distanza euclidea perché viviamo in un mondo euclideo.

Allo stesso modo, siamo a nostro agio con l'idea che i cerchi siano rotondi. Ricordiamo che un cerchio è definito come l'insieme di punti che si trovano a una distanza r da un punto di origine p. Cambiando la definizione di distanza, si cambia la forma di un cerchio.

La forma di un "cerchio" L_k regola quali punti sono vicini uguali intorno a un punto centrale p. Sotto la distanza di Manhattan (k= 1), il cerchio assomiglia a un diamante. Per k= 2, è l'oggetto rotondo che conosciamo bene. questo cerchio si estende fino a diventare una scatola orientata all'asse.

C'è una transizione graduale dal diamante alla scatola, al variare. Selezionare il valore di k equivale a scegliere quale cerchio si adatta meglio al nostro modello di dominio. Le distinzioni in questo caso diventano particolarmente importanti negli spazi dimensionali più elevati: ci interessano le deviazioni in tutte le dimensioni o principalmente quelle più grandi?

L'estrazione della radice k della somma dei termini di potenza k è necessaria affinché i valori di 'distanza' risultanti soddisfino la proprietà metrica. Tuttavia, in molte applicazioni utilizzeremo le distanze solo per il confronto: per verificare, invece di utilizzare i valori nelle formule o nell'isolamento. Poiché prendiamo il valore assoluto di ogni distanza dimensionale prima di elevarlo alla potenza k, la somma all'interno della funzione di distanza produce sempre un valore positivo. La funzione radice/potenza k è *monotona*. Pertanto, l'ordine di confronto delle distanze è invariato se non prendiamo la radice k della sommatoria. Evitare il calcolo della radice k risparmia tempo, che può rivelarsi non banale quando vengono eseguiti molti calcoli di distanza, come nella ricerca del vicino.

10.1.3 Lavorare in dimensioni superiori

Personalmente, non ho alcun senso geometrico per gli spazi a più alta dimensione, qualsiasi cosa in cui d> 3. Di solito, il meglio che possiamo fare è pensare alle geometrie superiori attraverso l'algebra lineare: le equazioni che regolano la nostra comprensione delle geometrie a due/tre dimensioni si generalizzano facilmente per d arbitrari, ed è proprio così che funzionano le cose.

Possiamo sviluppare un'intuizione sul lavoro con una serie di dati a più alte dimensioni attraverso i metodi di *proiezione*, che riducono la dimensionalità a livelli comprensibili. Spesso è utile visualizzare le proiezioni bidimensionali dei dati ignorando completamente le altre dimensioni d 2 e studiando invece i diagrammi a punti delle coppie dimensionali. Attraverso i metodi di riduzione delle dimensioni, come l'analisi delle componenti di principio possiamo combinare le caratteristiche altamente correlate per produrre una rappresentazione più pulita. Naturalmente, nel processo si perdono alcuni dettagli: se si tratta di rumore o di sfumature, dipende dalla sua interpretazione.

Dovrebbe essere chiaro che, aumentando il numero di dimensioni nella nostra serie di dati, stiamo implicitamente dicendo che ogni dimensione è una parte meno importante dell'insieme. Nel misurare la distanza tra due punti nello spazio delle caratteristiche, bisogna capire che una grande d significa che ci sono più modi per i punti di essere vicini (o lontani) l'uno dall'altro: possiamo immaginare che siano quasi identici lungo tutte le dimensioni tranne una.

Questo rende la scelta della metrica di distanza più importante negli spazi di dati ad alta dimensionalità. Naturalmente, possiamo sempre attenersi alla distanza L_2, che è una scelta sicura e standard. Ma se vogliamo premiare i punti di vicinanza su molte dimensioni, preferiamo una metrica più orientata verso L_1. Se invece le cose sono simili quando non ci sono singoli campi di grossolana dissomiglianza, forse dovremmo essere interessati a qualcosa di più vicino a L.

Un modo per pensare a questo è se siamo più preoccupati del rumore casuale aggiunto alle nostre caratteristiche, o di eventi eccezionali che portano a grandi artefatti. L_1 è indesiderabile nel primo caso, perché la metrica somma il rumore di tutte le dimensioni nella distanza. Ma gli artefatti rendono L_∞ sospetto, perché a L'errore sostanziale in una singola colonna arriverà a dominare l'intero calcolo della distanza.

10.1.4 Egualitarismo dimensionale

Le metriche di distanza L_k pesano tutte implicitamente ogni dimensione in modo uguale. Non deve essere necessariamente. A volte arriviamo a un problema con una comprensione specifica del dominio, secondo cui alcune caratteristiche sono più importanti per la somiglianza rispetto ad altre. Possiamo codificare questa informazione utilizzando un coefficiente c_i per specificare un peso diverso per ogni dimensione. Questa distanza ponderata per la dimensione soddisfa ancora le proprietà metriche. Se dispone di dati di verità sulla distanza desiderata tra determinate coppie di punti, può utilizzare la regressione lineare per adattare i coefficienti c_i alla migliore corrispondenza con il set di formazione. Ma, in generalità distanza ponderata per le dimensioni spesso non è una grande idea. A meno che non abbia una vera ragione per sapere che alcune dimensioni sono più importanti di altre, sta semplicemente codificando i suoi pregiudizi nella formula della distanza. Ma si insinuano pregiudizi molto più gravi se non si normalizzano le variabili prima di calcolare le distanze. L'approccio corretto consiste nel normalizzare i valori di ciascuna dimensione con i punteggi Z *prima di* calcolare la distanza. Sostituisca ogni valore x_i con il suo punteggio Z= $(x - \mu_i)/\sigma_i$, dove μ_i è il valore medio della dimensione i e σ_i la sua deviazione standard. Ora il valore atteso di x_i è zero per tutte le dimensioni e la diffusione è strettamente controllata se all'inizio erano distribuite normalmente. Si devono compiere sforzi più severi se una particolare dimensione è, ad esempio, distribuita a legge di potenza. Riveda la Sezione 4.3 sulla normalizzazione per le tecniche pertinenti, come ad esempio colpire prima con un logaritmo prima di calcolare lo Z-score.

10.1.5 Punti vs. Vettori

I vettori e i punti sono entrambi definiti da matrici di numeri, ma sono concepiti in modo diverso per rappresentare gli elementi nello spazio delle caratteristiche. I vettori disaccoppiano la direzione dalla magnitudine e quindi si può pensare che definiscano punti sulla superficie di una sfera unitaria.

Per capire perché questo è importante, consideriamo il problema di identificare i documenti più vicini dal conteggio delle parole-argomento. Supponiamo di aver suddiviso il vocabolario della lingua inglese in n sottoinsiemi diversi basati sugli argomenti, in modo che ogni parola del vocabolario si trovi esattamente in uno degli argomenti. Possiamo rappresentare ogni articolo A come un bagaglio di parole, come un punto p nello spazio *n-dimensionale*, dove p_i equivale al numero di parole che appaiono nell'articolo A e che provengono dall'argomento i.

Se vogliamo che un articolo lungo sul calcio sia vicino a un articolo breve sul calcio, la grandezza di questo vettore non ha importanza, ma solo la sua direzione. Senza la normalizzazione per la lunghezza, tutti i piccoli documenti di lunghezza pari a un tweet si raggrupperanno vicino all'origine, invece di raggrupparsi semanticamente nello spazio tematico, come desideriamo.

Le norme sono misure di grandezza vettoriale, essenzialmente funzioni di distanza che riguardano solo un punto, perché il secondo è considerato l'origine. I vettori sono essenzialmente punti normalizzati, dove dividiamo il valore di ogni dimensione di p per la sua L_2-norma $L_2(p)$, che è la distanza tra p e l'origine O. Dopo tale normalizzazione, la lunghezza di ogni vettore sarà 1, trasformandolo in un punto della sfera unitaria intorno all'origine. Abbiamo diverse possibili metriche di distanza da utilizzare per confrontare coppie di vettori. La prima classe è definita dalle metriche L_k, compresa la distanza euclidea. Questo funziona perché i punti sulla superficie di una sfera sono sempre punti nello spazio. Ma forse possiamo considerare in modo più significativo la distanza tra due vettori in termini di angolo definito tra loro. Abbiamo visto che la *somiglianza del coseno* tra due punti p e q è il loro prodotto di punti diviso per le loro norme. Per i vettori precedentemente normalizzati, queste norme sono uguali a 1, quindi l'unica cosa che conta è il prodotto del punto. La funzione coseno qui è una funzione di somiglianza, *non* una misura di distanza, perché valori più grandi significano una maggiore somiglianza. Definire una *distanza coseno* come $1 - \cos(p, q)$ produce una misura di distanza che soddisfa tre

delle proprietà metriche, tutte tranne la disuguaglianza del triangolo. Una vera metrica di distanza deriva da *distanza angolare*.

Qui arccos() è la funzione coseno inversa $\cos^{-1}()$, e π è l'intervallo angolare maggiore in radianti.

10.1.6 Distanze tra distribuzioni di probabilità

Ricordiamo il test di Kolmogorov-Smirnov, che ci ha permesso di determinare se due serie di campioni erano probabilmente tratte dalla stessa distribuzione di probabilità sottostante.

Questo suggerisce che spesso abbiamo bisogno di un modo per confrontare una coppia di distribuzioni e determinare una misura di somiglianza o di distanza tra loro. Un'applicazione tipica è la misurazione di quanto una distribuzione si avvicini ad un'altra, fornendo un modo per identificare il migliore di una serie di modelli possibili. Le misure di distanza che sono state descritte per i punti potrebbero, in linea di principio, essere applicate per misurare la somiglianza di due distribuzioni di probabilità P e Q su un determinato intervallo di variabili discrete R. Lo spettro dei valori p_i e q_i per $1\ i\ d$ può essere pensato come punti d-dimensionali che rappresentano P e Q, la cui distanza può essere calcolata utilizzando la metrica Eu- clidea.

Tuttavia, esistono misure più specializzate, che fanno un lavoro migliore per valutare la somiglianza delle distribuzioni di probabilità. Si basano sulla nozione teorica dell'informazione di *entropia*, che definisce una misura di incertezza per il valore di un campione estratto dalla distribuzione. Questo rende il concetto leggermente analogo alla varianza. L'entropia $H(P)$ di una distribuzione di probabilità P. Come la distanza, l'entropia è sempre una quantità non negativa. Le due somme di cui sopra differiscono solo per il modo in cui la raggiungono. Poiché p_i è una probabilità, generalmente è inferiore a 1, e quindi il $\log(p_i)$ è generalmente negativo. Quindi, sia che si prenda il reciproco delle probabilità prima di prendere il log, sia che si neghi ogni termine, è sufficiente per fare in modo che $H(P) \geq 0$ per tutte le P.

L'entropia è una misura dell'incertezza. Consideriamo la distribuzione in cui $p_{(1} = 1$ e $p_i = 0$, per $2 \leq i \leq d$. Questo è come lanciare un dado totalmente carico, quindi nonostante abbia d lati, non c'è incertezza sul risultato. Certo, $H(P) = 0$, perché o p_i o $\log_2(1)$ azzerano ogni termine della somma. Ora consideriamo la distribuzione in cui $q_i = 1/d$. Questo rappresenta un lancio di dadi equo, la distribuzione massimamente incerta in cui. Il rovescio della medaglia dell'incertezza è l'informazione. L'entropia $H(P)$ corrisponde a quante informazioni si apprendono dopo che un campione di P viene rivelato. Non si impara nulla quando qualcuno ci dice qualcosa che già conosciamo. Le misure di distanza standard sulle distribuzioni di probabilità si basano sull'entropia e sulla teoria dell'informazione. La divergenza di *Kullback-Leibler* (KL) misura l'incertezza guadagnata o l'informazione persa quando si sostituisce la distribuzione P con Q. Supponiamo che $P = Q$. Allora non si dovrebbe guadagnare o perdere nulla, e $KL(P, P) = 0$ perché $\lg(1) = 0$. Ma quanto peggiore è la sostituzione di Q per P, tanto più grande diventa $KL(P||Q)$, che arriva a quando $p_i > q_i = 0$. La divergenza KL assomiglia a una misura di distanza, ma non è una metrica, perché non è simmetrica ($KL(P\ Q) = /\ /KL(Q/\ P)$) e non$||$ soddisfa la disuguaglianza del triangolo. Tuttavia, costituisce la base della *divergenza Jensen-Shannon*. Inoltre $JS(P, Q)$ soddisfa magicamente la disuguaglianza del triangolo, trasformandosi in una vera metrica. Questa è la funzione giusta da utilizzare per misurare la distanza tra distribuzioni di probabilità.

10.2 Classificazione dei vicini più vicini

Le funzioni di distanza ci permettono di identificare i punti più vicini a un determinato obiettivo. Ciò offre un grande potere ed è il motore della *classificazione nearest neighbor*. Dato un insieme di esempi di formazione etichettati, cerchiamo l'esempio di formazione più simile a un punto non etichettato p, e poi prendiamo l'etichetta di classe per p dal suo vicino etichettato più vicino.

L'idea è semplice. Utilizziamo il vicino etichettato più vicino a un determinato punto di interrogazione q come suo rappresentante. Se si tratta di un problema di classificazione, assegneremo $a\ q$ la stessa etichetta dei suoi vicini più prossimi. Se si tratta di un problema di regressione, assegneremo $a\ q$ il valore medio/mediano dei suoi più prossimi. Queste previsioni sono facilmente difendibili, supponendo che (1) lo spazio delle caratteristiche catturi in modo coerente le proprietà degli elementi in questione e (2) la funzione di distanza riconosca in modo significativo righe/punti simili quando vengono incontrati.

Ci sono tre grandi vantaggi nei metodi di classificazione del vicino:

- *Semplicità*: I metodi di vicinanza non sono scienza missilistica; non c'è matematica più intimidatoria di una metrica di distanza. Questo è importante, perché significa che possiamo sapere esattamente cosa sta succedendo ed evitare di essere vittima di bug o idee sbagliate.

- *Interpretabilità*: Lo studio dei vicini più prossimi di un determinato punto di interrogazione q spiega esattamente perché il classificatore ha preso la decisione che ha preso. Se non è d'accordo con questo risultato, può un debug sistematico. I punti vicini sono stati etichettati in modo errato? La sua funzione di distanza non è riuscita a selezionare gli elementi che erano il gruppo di pari logico per q?
- *Non linearità*: I classificatori di prossimità hanno dei confini decisionali che sono frammentariamente lineari, ma possono incresparsi in modo arbitrario in seguito all'esperienza di addestramento. Dal calcolo sappiamo che le funzioni l ineari frammentarie si avvicinano a curve lisce quando i pezzi diventano abbastanza piccoli. Quindi i classificatori nearest neighbor ci permettono di realizzare confini di decisione molto complicati, anzi superfici così complesse da non avere una rappresentazione concisa.

Ci sono diversi aspetti nella costruzione di classificatori nearest neighbor efficaci, comprese le questioni tecniche relative alla robustezza e all'efficienza. Ma la cosa più importante è imparare ad apprezzare il potere dell'analogia. Discutiamo questi aspetti nelle sezioni seguenti.

10.2.1 Cercare buone analogie

Alcune discipline intellettuali si basano sul potere delle analogie. Gli avvocati non ragionano direttamente a partire dalle leggi, ma si basano sui precedenti: i risultati di
casi decisi in precedenza da giuristi rispettati. La decisione giusta per il caso attuale (vinco o perdo) è una funzione di quali casi precedenti possono essere dimostrati come fondamentalmente simili alla questione in questione. Per ottenere i massimi benefici dai metodi di prossimità, bisogna imparare a rispettare il ragionamento analogico. Oppure possiamo cercare "comparazioni", cercando proprietà comparabili in quartieri simili, e prevedere un prezzo simile a quello che vediamo. Il secondo approccio è il ragionamento analogico.

10.2.2 k-vicini *più vicini*

Per classificare un dato punto di interrogazione q, i metodi di prossimità restituiscono l'etichetta di q', il punto etichettato più vicino a q. Questa è un'ipotesi ragionevole, supponendo che la somiglianza nello spazio delle caratteristiche implichi la somiglianza nello spazio delle etichette. Tuttavia, questa classificazione si basa su un solo esempio di formazione, il che dovrebbe farci riflettere. Una classificazione o interpolazione più robusta deriva dalla votazione su più . Supponiamo di trovare i k punti più vicini alla nostra richiesta, dove k è tipicamente un valore che va da 3 a 50, a seconda della dimensione di n. La disposizione dei punti etichettati, abbinata alla scelta di k, scolpisce lo spazio delle caratteristiche in regioni, con tutti i punti in una determinata regione a cui viene assegnata la stessa etichetta. sull'altezza e sul peso. In generalele donne sono più basse e più leggere degli uomini, ma ci sono molte eccezioni, soprattutto in prossimità del confine decisionale. l'aumento di k tende a produrre regioni più ampie con confini più morbidi, che rappresentano decisioni più robuste. Tuttavia, quanto più grande è k, tanto più generiche sono le nostre decisioni. La scelta di $k= n$ è semplicemente un altro nome per il classificatore di maggioranza, in cui assegniamo a ogni punto l'etichetta

più comune, indipendentemente dalle sue caratteristiche individuali. Il modo giusto per impostare *k* è assegnare una frazione di esempi di formazione etichettati come set di valutazione, e poi sperimentare diversi valori del parametro *k* per vedere dove si ottiene la migliore performance. Questi valori di valutazione possono poi essere reinseriti nell'insieme di addestramento/target, una volta *k*. Per un problema di classificazione binaria, vogliamo che *k* sia un numero dispari, in modo che la decisione non sia mai un pareggio. In generale, la differenza tra il numero di voti positivi e negativi può essere interpretata come una misura della nostra fiducia nella decisione. Ci sono potenziali asimmetrie per quanto riguarda i geometrici. Ogni punto ha un più vicino, ma per i punti outlier questi vicini più vicini potrebbero non essere particolarmente vicini. Questi punti anomali, infatti, possono avere un ruolo esagerato nella classificazione, definendo il vicino più prossimo in un volume enorme dello spazio delle caratteristiche.

Tuttavia, se gli esempi di formazione sono stati scelti correttamente, questo dovrebbe essere un territorio in gran parte disabitato, una regione dello spazio delle caratteristiche in cui i punti si presentano raramente. L'idea della classificazione del vicino più prossimo può essere generalizzata all'interpolazione di funzioni, facendo una media dei valori *dei k* punti più vicini. Questo è presumibilmente fatto da siti web immobiliari come www.zillow.com, per prevedere i prezzi delle abitazioni da più vicini. Tali schemi di mediazione possono essere generalizzati da schemi di mediazione non uniformi. pesi, valutando i punti in modo diverso in base al rango o alla grandezza della distanza. Idee simili funzionano per tutti i metodi di classificazione.

10.2.3 Trovare i vicini più vicini

Forse il limite maggiore dei metodi di classificazione nearest neighbor è il loro costo di esecuzione. Il confronto di un punto di interrogazione *q* in *d* dimensioni con *n* punti di addestramento di questo tipo avviene ovviamente eseguendo *n* confronti di distanza espliciti, con un costo di $O(nd)$. Con migliaia o addirittura milioni di punti di addestramento disponibili, questa ricerca può introdurre un ritardo notevole in qualsiasi sistema di classificazione.

Un approccio per accelerare la ricerca prevede l'utilizzo di strutture dati geometriche. Le scelte più popolari includono:

- *Diagrammi di Voronoi*: Per un insieme di punti target, vorremmo suddividere lo spazio intorno ad essi in celle, in modo che ogni cella contenga esattamente un punto target. Inoltre, vogliamo che il punto target di ogni cella sia il vicino target più vicino per tutte le posizioni nella cella. Una tale partizione si chiama *diagramma di Voronoi*.

  I confini dei diagrammi di Voronoi sono definiti dalle bisettrici perpendicolari tra coppie di punti (a, b). Ogni bisettrice taglia lo spazio a metà: una metà contenente a e l'altra contenente b, in modo tale che tutti i punti sulla metà di a siano più vicini ad a che a b, e viceversa.

  I diagrammi di Voronoi sono uno strumento meraviglioso per pensare ai dati e hanno molte proprietà interessanti. Esistono algoritmi efficienti per costruirli e ricercarli, soprattutto in due dimensioni. Tuttavia, queste procedure diventano rapidamente più complesse con l'aumentare della dimensionalità, rendendole generalmente poco pratiche oltre le due o tre dimensioni.

- *Indici di griglia*: Possiamo ritagliare lo spazio in scatole *d-dimensionali*, dividendo l'intervallo di ogni dimensione in *r* intervalli o secchi. Per esempio,

Una struttura di dati con indice a griglia fornisce un accesso rapido ai più prossimi quando i punti sono distribuiti in modo uniforme, ma può essere inefficiente quando i punti in alcune regioni sono densamente raggruppati.

Consideriamo uno spazio bidimensionale in cui ogni asse rappresenta una probabilità, che va quindi da 0 a 1. Questo intervallo può essere suddiviso in *r* intervalli di uguali dimensioni, in modo che l'intervallo *iesimo* sia compreso tra [(*i* - 1)*/r, i/r*].

Questi intervalli definiscono una griglia regolare sullo spazio, quindi possiamo associare ciascuno dei punti di formazione alla cella della griglia a cui appartiene. La ricerca diventa ora il problema di identificare la giusta cella della griglia per il punto *q* attraverso la ricerca di array o la ricerca binaria, e poi confrontare *q* con tutti i punti di questa cella per identificare il vicino più prossimo.

Tali indici di griglia possono essere efficaci, ma ci sono dei problemi potenziali. In primo luogo, i punti di addestramento potrebbero non essere distribuiti in modo uniforme e molte celle potrebbero essere vuote. Stabilire una griglia non uniforme potrebbe portare a una disposizione più equilibrata, ma rende più difficile trovare rapidamente la cella che contiene *q*. Ma non c'è nemmeno la garanzia che il vicino più prossimo di *q* si trovi effettivamente nella stessa cella di *q*, soprattutto se *q* si trova molto vicino al confine della cella. Ciò significa che dobbiamo cercare anche nelle celle vicine, per assicurarci di trovare il vicino più vicino in assoluto.

- *Alberi Kd*: Esiste un'ampia classe di strutture di dati ad albero che par- tecipano allo spazio utilizzando una gerarchia di divisioni che facilita la ricerca. Partendo da una dimensione arbitraria come radice, ogni nodo dell'*albero* kd definisce una linea/piano mediano che divide equamente i punti in base a quella dimensione. La costruzione ricorre su ogni lato utilizzando una dimensione diversa, e così via fino a quando la regione definita da un nodo contiene un solo punto di formazione.
Questa gerarchia di costruzione è ideale per supportare la ricerca. Partendo dalla radice, verifichiamo se il punto di interrogazione *q* si trova a sinistra o a destra della linea/piano mediano. Questo identifica il lato su cui si trova *q* e quindi il lato dell'albero su cui ricorre. Il tempo di ricerca è log *n*, poiché dividiamo l'insieme di punti a metà ad ogni passo lungo l'albero. una varietà di strutture ad albero di ricerca a partizione spaziale, con una o più probabilmente implementate nella libreria di funzioni del suo linguaggio di programmazione preferito. Alcune offrono tempi di ricerca più rapidi su problemi come il nearest neighbor, con forse un compromesso tra precisione e velocità. Sebbene queste tecniche possano effettivamente accelerare la ricerca dei vicini in un numero modesto di dimensioni, diventano meno efficaci con l'aumento della dimensionalità. Il motivo è che il numero di modi in cui due punti possono essere vicini aumenta rapidamente con la dimensionalità, rendendo più difficile tagliare le regioni che non hanno alcuna possibilità di contenere il vicino più prossimo a

q. La ricerca deterministica del vicino più prossimo si riduce alla fine alla ricerca lineare, per dati di dimensionalità sufficientemente elevata.

10.2.4 Hashing sensibile alla località

Per ottenere tempi di esecuzione più rapidi, dobbiamo abbandonare l'idea di trovare il vicino esatto e accontentarci di una buona ipotesi. Vogliamo raggruppare i punti vicini in bucket in base alla somiglianza e trovare rapidamente il bucket *B* più appropriato per il nostro punto di interrogazione *q*. Calcolando solo la distanza tra *q* e i punti nel bucket, risparmiamo tempo di ricerca quando *B* n. Questa era l'idea di base dell'indice a griglia, descritto // nella sezione precedente, ma le strutture di ricerca diventano ingombranti e sbilanciate nella pratica. Un approccio migliore si basa sull'hashing.

L'hashing sensibile alla località (LSH) è definito da una funzione di hash $h(p)$ che prende in ingresso un punto o un vettore e produce in uscita un numero o un codice tale che è probabile che $h(a) = h(b)$ se a e *b* sono vicini tra loro, e $h(a) = h(b)$ se sono lontani.

Queste funzioni di hash sensibili alla localizzazione svolgono facilmente lo stesso ruolo dell'indice di

griglia, senza il problema. Possiamo semplicemente mantenere una tabella di punti raggruppati in base a questo valore hash unidimensionale, e poi cercare le potenziali corrispondenze per il punto di query q cercando $h(q)$.

Come possiamo costruire funzioni hash sensibili alla località? L'idea è più facile da capire all'inizio quando ci si limita ai vettori invece che ai punti. Ricordiamo che gli insiemi di vettori d-dimensionali possono essere pensati come punti sulla superficie di una sfera, cioè di un cerchio quando $d = 2$.

Consideriamo una linea arbitraria l_1 passante per l'origine di questo cerchio, che taglia il cerchio a metà. Infatti, possiamo selezionare casualmente l_1 semplicemente scegliendo un angolo casuale $0 \leq \vartheta_1 < 2\pi$. Questo angolo definisce la pendenza di una linea che passa per l'origine O, e insieme ϑ_1 e O specificano completamente l_1. Se scelto a caso, l_1 dovrebbe dividere grossolanamente i vettori, mettendo circa la metà di essi a sinistra e il resto a destra.

I codici hash sensibili alla località per ogni punto possono essere composti come una sequenza di test di lateralità (sinistra o destra) per qualsiasi sequenza specifica di linee.

Ora aggiungiamo un secondo divisore casuale l_2, che dovrebbe condividere le stesse proprietà. In questo modo, tutti i vettori vengono suddivisi in quattro regioni, *{LL, LR, RL, RR}*, definite dal loro stato rispetto a questi divisori l_1 e l_2.

Il vicino più prossimo di qualsiasi vettore v dovrebbe trovarsi nella stessa regione di v, a meno che non siamo stati sfortunati e li abbia separati. Ma la probabilità $p(v_1, v_2)$ che sia v_1 che v_2 si trovino sullo stesso lato di l dipende dall'angolo tra v_1 e v_2.

Così possiamo calcolare la probabilità esatta che i vicini siano conservati per n punti e m piani casuali. Lo schema di L e R su questi m piani definisce un codice hash sensibile alla località a m bit $h(v)$ per qualsiasi vettore v. Quando ci spostiamo oltre i due piani del nostro esempio verso codici più lunghi, il numero previsto di punti in ogni bucket scende a $n/2^m$, anche se con un rischio maggiore che uno degli m piani separi un vettore dal suo vero vicino.

Si noti che questo approccio può essere facilmente generalizzato oltre le due dimensioni. L'iperpiano è definito dal suo vettore normale r, che è perpendicolare in direzione del piano. Il segno di $s = v\,r$ determina su quale lato si trova un vettore di interrogazione v. Ricordiamo che il prodotto di punti di due vettori ortogonali è 0, quindi $s = 0$ se v si trova esattamente sul piano di separazione. Inoltre, s è positivo se v si trova al di sopra di questo piano e negativo se v si trova al di sotto. Quindi l'iperpiano iesimo contribuisce esattamente con un bit al codice hash, dove $h_i(q) = 0$ iff $v - r_i \leq 0$.

Tali funzioni possono essere generalizzate al di là dei vettori, a insiemi di punti arbitrari. Inoltre, la loro precisione può essere migliorata costruendo più serie di parole codice per ogni elemento, coinvolgendo diverse serie di iperpiani casuali. Fintanto che q condivide almeno una parola di codice con il suo vero , alla fine incontra un secchio che contiene entrambi questi punti.

Si noti che LSH ha esattamente l'obiettivo opposto rispetto alle funzioni hash tradizionali utilizzate per le applicazioni crittografiche o per gestire le tabelle hash. Le funzioni hash tradizionali cercano di garantire che coppie di elementi simili producano valori hash molto diversi, in modo da poter riconoscere i cambiamenti e utilizzare l'intera gamma della tabella. Al contrario, LSH vuole che gli elementi simili ricevano esattamente lo stesso codice hash, in modo da poter riconoscere la somiglianza per collisione. Con LSH, i vicini più prossimi appartengono allo stesso bucket.

L'hashing sensibile alla località ha altre applicazioni nella scienza dei dati, oltre alla ricerca del vicino. Forse la più importante è la costruzione di rappresentazioni di caratteristiche compresse da oggetti complicati, come i flussi video o musicali. I codici LSH costruiti da intervalli di questi flussi definiscono valori numerici potenzialmente adatti come caratteristiche per il pattern matching o la costruzione di modelli.

10.3 Grafici, reti e distanze

Un *grafo* $G = (V, E)$ è definito su un insieme di *vertici* V e contiene un insieme di *bordi* E di coppie ordinate o non ordinate di vertici di V. Nella modellazione di una rete stradale, i vertici possono rappresentare le città o gli incroci, alcune coppie dei quali sono direttamente collegate da strade/bordi. Nell'analisi delle interazioni umane, i vertici rappresentano tipicamente le persone, con bordi che collegano coppie di anime correlate. Molti altri set di dati moderni sono naturalmente modellati in termini di grafi o reti:

- *Il Web mondiale (WWW)*: Qui c'è un vertice del grafico per ogni pagina web, con un bordo diretto (x, y) se la pagina web x contiene un collegamento ipertestuale alla pagina web y.

- *Reti di prodotti/clienti*: Queste si presentano in qualsiasi azienda che ha molti clienti e tipi di prodotti: che si tratti di Amazon, Netflix o anche del negozio di alimentari all'angolo. Ci sono due tipi di vertici: un insieme per i clienti e un altro per i prodotti. Il bordo (x, y) denota un prodotto y acquistato dal cliente x.

- *Reti genetiche*: Qui i vertici rappresentano i diversi geni/proteine di un particolare organismo. Pensate a questo come a un elenco di parti della bestia. Il bordo (x, y) indica che ci sono interazioni tra le parti x e y. Forse il gene x regola il gene y, oppure le proteine x e y si legano per formare un complesso più grande. Tali reti di interazione codificano informazioni considerevoli sul funzionamento del sistema sottostante.

I grafici e gli insiemi di punti sono oggetti strettamente correlati. Entrambi sono composti da entità discrete (punti o vertici) che rappresentano elementi di un insieme. Entrambi codificano importanti nozioni di distanza e di relazioni, sia vicine che lontane o connesse e indipendenti. Gli insiemi di punti possono essere rappresentati in modo significativo da grafici e i grafici da insiemi di punti.
La sogliatura mediante un taglio di distanza elimina tutti i bordi lunghi, lasciando un grafo rado che cattura la struttura dei punti (destra).

10.3.1 Grafi ponderati e reti indotte

I bordi nei grafi catturano le *relazioni binarie*, dove ogni bordo (x, y) rappresenta che esiste una relazione tra x e y. L'esistenza di questa relazione a volte è tutto ciò che c'è da sapere, come nel caso della connessione tra le pagine web o il fatto che qualcuno abbia acquistato un determinato prodotto.
Ma spesso esiste una misura intrinseca della forza o della vicinanza della relazione. Certamente lo vediamo nelle reti stradali: ogni segmento stradale ha una lunghezza o un tempo di percorrenza, che è essenziale conoscere per trovare il percorso migliore da percorrere tra due punti. Diciamo che un grafico è *ponderato* se ogni bordo ha un valore numerico associato.
E questo peso è spesso (ma non sempre) interpretato naturalmente come una distanza. In effetti, si può interpretare un insieme di dati di n punti nello spazio come un grafico ponderato completo su n vertici, dove il peso del bordo (x, y) è la distanza geometrica tra i punti x e y nello spazio. Per molte applicazioni, questo grafo codifica tutte le informazioni rilevanti sui punti.
I grafi sono rappresentati in modo più naturale da× *matrici di adiacenza* n n. Definiamo un simbolo di non bordo x. La matrice M rappresenta il grafo $G = (V, E)$ quando $M[i, j] = x$ se e solo se i vertici i, j V sono collegati da un bordo (i, j) E. Per le reti non ponderate, in genere il simbolo di bordo è 1 mentre $x = 0$. Per i grafi ponderati per la distanza, il peso del bordo (i, j) è il costo del viaggio tra di loro, quindi l'impostazione di denota l'assenza di qualsiasi collegamento diretto tra i e j.
Questa rappresentazione matriciale delle reti ha un potere considerevole, perché possiamo introdurre tutti i nostri strumenti dell'algebra lineare per lavorare con esse. Sfortunatamente, ha un

costo, perché può essere irrimediabilmente costoso memorizzare *n* matrici quando le reti superano qualche centinaio di vertici. Esistono modi più efficienti per memorizzare grandi *grafi sparsi*, con molti vertici ma relativamente poche coppie collegate da bordi. Non discuterò qui i dettagli degli algoritmi dei grafi, ma la rimando con fiducia al mio libro *The Algorithm Design Manual* [Ski08] per saperne di più.

Le immagini di grafici/reti sono spesso realizzate assegnando a ciascun vertice un punto nel piano e tracciando linee tra questi punti-verticali per rappresentare i bordi. Tali *diagrammi nodo-inchiostro* sono estremamente utili per visualizzare la struttura di le reti con cui si lavora. Possono essere costruite algoritmicamente utilizzando la *disposizione diretta dalla forza*, dove i bordi agiscono come molle per avvicinare le coppie di vertici adiacenti, mentre i vertici non adiacenti si respingono.

Tali disegni stabiliscono la connessione tra le strutture del grafo e le posizioni dei punti. Un *embedding* è una rappresentazione puntiforme dei vertici di un grafo che cattura un aspetto della sua struttura. L'esecuzione di una compressione delle caratteristiche, come la decomposizione degli autovalori o dei valori singolari, sulla matrice di adiacenza di un grafo, produce una rappresentazione di dimensioni inferiori che serve come rappresentazione puntuale di ogni vertice. Altri ap procci alle incorporazioni di grafi includono DeepWalk.

I grafi geometrici definiti dalle distanze tra i punti sono rappresentativi di una classe di grafi che chiamerò *reti indotte*, dove i bordi sono definiti in modo meccanico da una fonte di dati esterna. Si tratta di una fonte comune di reti nella scienza dei dati, quindi è importante tenere d'occhio i modi in cui il suo set di dati potrebbe essere trasformato in un grafo.

Le funzioni di distanza o di somiglianza sono comunemente utilizzate per costruire reti su insiemi di elementi. In genere siamo interessati ai bordi che collegano ogni vertice ai suoi *k* più vicini/più simili. Otteniamo un grafo rado mantenendo *k* modesto, ad esempio *k* 10, il che significa che può essere facilmente elaborato anche per valori elevati di *n*.

Ma ci sono altri tipi di reti indotte. La tipica consiste nel collegare i vertici *x* e *y* ogni volta che hanno un attributo significativo in comune. Per esempio, possiamo costruire una rete sociale indotta sulle persone a partire dalle loro rime, collegando due persone che hanno lavorato nella stessa azienda o frequentato la stessa scuola in un periodo simile. Tali reti tendono ad avere una struttura a blocchi, dove ci sono ampi sottoinsiemi di vertici che formano cricche completamente connesse. Dopotutto, se *x* si è laureato nella stessa università di *y* e *y* si è laureato nella stessa università di *z*, questo implica che (*x, z*) deve essere un bordo del grafico.

10.3.2 Parlare di grafici

Esiste un vocabolario sui grafici che è importante conoscere per lavorare con loro. Parlare di ciò che si dice è una parte importante del cammino. Diverse proprietà fondamentali dei grafi hanno un impatto su ciò che rappresentano e su come possiamo . Pertanto, il primo passo in qualsiasi problema di grafi è determinare i sapori dei grafi con cui si ha a che fare:

- *Non diretto vs. Diretto*: Un grafo *G*= (*V, E*) è *indiretto* se il bordo (*x, y*) *E* implica che∈ (*y, x*) sia anche in *E*. In caso contrario, diciamo che il grafo è *diretto*. Le reti stradali *tra le* città sono tipicamente non dirette, poiché qualsiasi strada di grandi dimensioni ha corsie che vanno in entrambe le direzioni. Le reti stradali *all'interno delle* città sono quasi

I grafi delle pagine web sono tipicamente diretti, perché il collegamento dalla pagina *x* alla pagina *y* non deve essere necessariamente reciproco.

- *Ponderato vs. Non ponderato*: Come discusso nella Sezione 10.3.1, ad ogni bordo (o vertice) di un grafico *ponderato* G viene assegnato un valore numerico, o peso. I bordi di un grafo di rete stradale possono essere ponderati con la loro lunghezza, il tempo di guida o il limite di velocità, a seconda dell'applicazione. Nei grafi *non ponderati*, non esiste una distinzione di costo tra i vari bordi e vertici.

  I grafi a distanza sono intrinsecamente ponderati, mentre le reti sociali/web sono generalmente non ponderate. La differenza determina se i vettori di caratteristiche associati ai vertici sono 0/1 o valori numerici di importanza, che potrebbero dover essere normalizzati.

- *Semplice vs. non semplice*: Alcuni tipi di bordi complicano il compito di lavorare con i grafici. Un *auto-loop* è un bordo (x, x), che coinvolge un solo vertice. Un bordo (x, y) è un *bordo multiplo* se si verifica più di una volta nel grafico.

  Entrambe queste strutture richiedono un'attenzione particolare nella pre-elaborazione per la generazione delle caratteristiche. Pertanto, qualsiasi grafo che le eviti è definito *semplice*. Spesso cerchiamo di eliminare sia gli auto-loop che i multiedges all'inizio dell'analisi.

  - *Sparse vs. Dense*: I grafici sono *sparsi* quando solo una piccola frazione di il totale delle possibili coppie di $_2$ per un grafo semplice e non diretto su n vertici (vertici) hanno effettivamente dei bordi definiti tra loro. I grafici in cui una grande frazione delle coppie di vertici definisce bordi sono chiamati *densi*. Non esiste un confine ufficiale tra ciò che viene definito rado e ciò che viene definito denso, ma in genere i grafi densi hanno un numero quadratico di bordi, mentre i grafi radi sono di dimensioni lineari.

  I grafi sparsi sono solitamente sparsi per motivi specifici dell'applicazione. Le reti stradali devono essere grafi sparsi a causa delle intersezioni stradali. L'intersezione più orribile che abbia mai sentito era il punto finale di solo nove strade diverse. grafi knearest neighbor hanno gradi dei vertici di esattamente k. I grafi sparsi rendono possibili rappresentazioni molto più efficienti dal punto di vista spaziale rispetto alle matrici di adiacenza, consentendo la rappresentazione di reti molto più grandi.

- *Incorporato vs. Topologico* - Un grafo è *incorporato* se ai vertici e ai bordi vengono assegnate posizioni geometriche. Pertanto, qualsiasi disegno di un grafo è un *incorporamento*, che può avere o meno un significato algoritmico.

  A volte, la struttura di un grafo è completamente definita dalla geomatica del suo incorporamento, come abbiamo visto nella definizione del grafo a distanza, dove i pesi sono definiti dalla distanza euclidea tra ogni coppia di punti. Le rappresentazioni a bassa dimensione delle matrici di adiacenza mediante SVD si qualificano anche come incorporazioni, rappresentazioni di punti che catturano gran parte delle informazioni di connettività del grafo.

- *Etichettato vs. Non etichettato* - In un grafo *etichettato*, ad ogni vertice viene assegnato un nome o un identificatore unico per distinguerlo da tutti gli altri vertici. Nei grafi *non etichettati* non viene fatta questa distinzione.

  I grafici che nascono nelle applicazioni di scienza dei dati sono spesso etichettati in modo naturale e significativo, come i nomi delle città in una rete di trasporti. Questi sono utili come identificatori per gli esempi rappresentativi e anche per fornire collegamenti a fonti di dati esterne, ove opportuno.

10.3.3 Teoria dei grafici

La teoria dei grafi è un'area importante della matematica che si occupa delle proprietà divertenti delle reti e di come calcolarle. La maggior parte degli studenti di informatica viene esposta alla teoria dei grafi attraverso i corsi di strutture discrete o di algoritmi.

Gli algoritmi classici per trovare i percorsi più brevi, le componenti connesse, gli alberi di spanning, i tagli, le corrispondenze e l'ordinamento topologico possono essere applicati a qualsiasi grafo ragionevole. Tuttavia, non ho visto questi strumenti applicati in modo così generale nella scienza dei dati, come credo dovrebbero essere. Uno dei motivi è che i grafi nella scienza dei dati tendono ad essere molto grandi, limitando la complessità di ciò che si può fare con essi. Ma molto è dovuto semplicemente alla miopia: le persone non vedono che una matrice di distanza o di somiglianza è in realtà solo un grafico che può trarre vantaggio da altri strumenti.

- *Percorsi più brevi*: Per una 'matrice' di distanza m, il valore di $m[i, j]$ deve riflettere il percorso di lunghezza minima tra i vertici i e j. Si noti che le stime indipendenti della distanza a coppie sono spesso incoerenti e non soddisfano necessariamente la disuguaglianza del triangolo. Ma quando $m'[i, j]$ riflette la *distanza del percorso più breve* da i a j in qualsiasi matrice m, *deve* soddisfare le proprietà metriche. Questo potrebbe rappresentare una matrice migliore per l'analisi rispetto all'originale .

- *Componenti collegate*: Ogni pezzo disgiunto di un grafo è chiamato *componente connesso*. Identificare se il suo grafo è costituito da un singolo componente o da più pezzi è importante. In primo luogo, qualsiasi algoritmo che eseguirà otterrà prestazioni migliori se tratterà i componenti in modo indipendente. I componenti separati possono essere indipendenti per motivi validi, ad esempio non c'è un passaggio stradale tra gli Stati Uniti e l'Europa a causa dell'oceano. Ma i componenti separati possono indicare problemi, come artefatti di elaborazione o connettività insufficiente per lavorare.

Alberi a scansione minima: Un albero di spanning è l'insieme minimo di bordi che collegano tutti i vertici di un grafo, essenzialmente una prova che il grafo è connesso. L'albero di spanning di peso minimo serve come rappresentazione più scarna possibile della struttura del grafo, rendendolo utile per la visualizzazione. Infatti, nella Sezione 10.5 .mostreremo che gli alberi a scansione minima hanno un ruolo importante negli algoritmi di clustering

- *Tagli dei bordi*: Un cluster in un grafico è definito da un sottoinsieme di vertici c, con la proprietà che (a) c'è una notevole somiglianza tra le coppie di vertici all'interno di c, e (b) c'è una debole connettività tra i vertici in c e fuori da c. I bordi (x, y) dove x c e y definiscono un taglio che separa il cluster dal resto del grafico, rendendo la ricerca di tali tagli un aspetto importante dell'analisi dei cluster.
- *Abbinamenti*: Abbinare ogni vertice con un partner simile e fedele può utile in molti modi. Interessanti tipi di confronti diventano possibili dopo un tale *abbinamento*. Ad esempio, l'esame di tutte le coppie vicine che differiscono in un attributo (ad esempio il sesso) potrebbe far luce su come questa variabile impatta su una particolare variabile di risultato (ad esempio il reddito o la durata della vita). Gli abbinamenti offrono anche la possibilità di ridurre le dimensioni effettive di una rete. Sostituendo ogni coppia abbinata con un vertice che rappresenta il suo centroide, possiamo costruire un grafico con la metà dei vertici, ma comunque rappresentativo dell'insieme.
- *Ordinamento topologico*: I problemi di classificazione (si ricordi il Capitolo 4) impongono un ordine di selezione su una collezione di elementi in base ad alcuni criteri di merito. L*'ordinamento topo- logico* classifica i vertici di un grafo aciclico diretto (DAG) in modo che il bordo (i, j) implichi che i si colloca al di sopra di j nell'ordine di classifica. Dato un insieme di vincoli osservati della forma " i dovrebbe essere superiore a j, l'ordinamento topologico definisce un ordine degli articoli coerente con queste osservazioni.

10.4 PageRank

Spesso è utile classificare l'importanza relativa dei vertici in un grafo. Forse la nozione più semplice si basa sul *grado* del vertice, il numero di bordi che collegano il vertice v al resto del grafico. Più un vertice è connesso, più è importante.

Il grado del vertice v è una buona caratteristica per rappresentare l'elemento associato a v. Ma ancora meglio è *PageRank* . Il PageRank ignora il contenuto testuale delle pagine web, per concentrarsi solo sulla struttura dei collegamenti ipertestuali tra le pagine. Le pagine più importanti (vertici) dovrebbero avere un indegree più alto rispetto alle pagine meno importanti, sicuramente. Ma anche l'importanza delle pagine che la linkano è importante.

Il PageRank è meglio compreso nel contesto delle passeggiate casuali lungo una rete. Supponiamo di partire da un vertice arbitrario e di selezionare a caso un collegamento in uscita in modo uniforme dall'insieme delle possibilità. Ora ripetiamo il processo da qui, saltando a un vicino casuale della nostra posizione attuale ad ogni passo. Il PageRank del vertice v è una misura della probabilità che, partendo da un vertice random, si arrivi a v dopo una lunga serie di passi casuali.

Si tratta di una formula ricorsiva, con j come numero di iterazione. Inizializziamo $PR_0(v_i) = 1/n$ per ogni vertice v_i nella rete. I valori di PageRank cambieranno ad ogni iterazione, ma convergeranno sorprendentemente velocemente verso valori stabili. Per i grafi non diretti, questa probabilità è essenzialmente uguale all'in-degree di ogni vertice, ma con i grafi diretti accadono cose molto più interessanti.

In sostanza, il PageRank si basa sull'idea che se tutte le strade portano a Roma, Roma deve essere un luogo piuttosto importante. Sono i percorsi *che portano alla* sua pagina che contano. Questo è ciò che rende il PageRank difficile da giocare: altre persone devono collegarsi alla sua pagina web, e qualsiasi grido su di lei è irrilevante.

Ci sono diverse modifiche che si possono apportare a questa formula di base del PageRank per rendere i risultati più interessanti. Possiamo consentire alla camminata di saltare a un vertice arbitrario (invece che a un vicino collegato) per consentire una diffusione più rapida nella rete. Lasciare che p sia la probabilità di seguire un link nel passo successivo, noto anche come *fattore di smorzamento*. dove n è il numero di vertici del grafico. Altri miglioramenti comportano modifiche alla rete stessa.

Aggiungendo bordi da ogni vertice a un singolo super-vertice, assicuriamo che le passeggiate casuali non possano rimanere intrappolate in un piccolo angolo della rete. Gli auto-loop e i bordi paralleli (multiedges) possono essere eliminati per evitare pregiudizi dovuti alla ripetizione.

Esiste anche un'interpretazione algebrica lineare del PageRank. Lasciando che M sia una matrice di probabilità di transizione vertice-vertice, M_{ij} è la probabilità che il nostro prossimo passo da i sia verso j. Chiaramente se c'è un bordo diretto da i a j, e zero in caso contrario.

Dopo che questa stima converge, il vettore PageRank. Questa è l'equazione di definizione degli autovalori, quindi il vettore n 1 dei valori di PageRank risulta essere l'autovettore principale della matrice di probabilità di transizione definita dai link. Pertanto, i metodi iterativi per il calcolo degli autovalori e la moltiplicazione veloce delle matrici portano a calcoli efficienti del PageRank.

Quanto funziona bene il PageRank nell'individuare i vertici più centrali? Per avere un'idea, abbiamo eseguito il PageRank sulla rete di collegamenti dell'edizione inglese di Wikipedia, concentrandoci sulle pagine associate alle persone.

Queste figure ad alto PageRank sono tutte facilmente riconoscibili come persone molto significative. Il meno conosciuto tra loro è probabilmente *Carl Linnaeus*, il "padre della tassonomia" della biologia, il cui sistema linneo è utilizzato per classificare tutta la vita sulla terra. Era un grande scienziato, ma perché è *così* altamente considerato da PageRank? Le pagine di Wikipedia di tutte le specie vegetali e animali che ha classificato per la prima volta rimandano a lui, quindi migliaia di forme di vita contribuiscono con percorsi di rilievo alla sua

pagina. L'aggiunta e l'eliminazione di gruppi di bordi da una determinata rete dà origine a reti diverse, alcune delle quali rivelano meglio il significato sottostante utilizzando il PageRank. Supponiamo di calcolare il PageRank (indicato con PR2) utilizzando solo i bordi di Wikipedia che collegano le persone. Questo calcolo ignorerebbe qualsiasi contributo di luoghi, organizzazioni e organismi inferiori.

10.5 Raggruppamento

Il clustering è il problema di raggruppare i punti in base alla somiglianza. Spesso gli elementi provengono da un piccolo numero di "fonti" o "spiegazioni" logiche, e il clustering è un buon modo per rivelare queste origini. I modelli su un diagramma a punti bidimensionale sono generalmente abbastanza facili da vedere, ma spesso abbiamo a che fare con dati di dimensioni superiori che gli esseri umani non possono visualizzare in modo efficace. Abbiamo bisogno di algoritmi che trovino questi modelli per noi. Il clustering è forse la prima cosa da fare con qualsiasi serie di dati interessanti. Le applicazioni includono:

- *Sviluppo di ipotesi*: Apprendere che sembrano esserci (ad esempio) quattro popolazioni distinte rappresentate nel suo set di dati dovrebbe far nascere la domanda sul perché ci sono. Se questi cluster sono abbastanza compatti e ben separati, ci deve essere una ragione ed è suo compito trovarla. Una volta assegnato a ciascun elemento un'etichetta di cluster, può studiare più rappresentanti dello stesso cluster per capire che cosa hanno in comune, oppure esaminare coppie di elementi di cluster diversi e identificare perché sono diversi.

- *Modellazione su sottoinsiemi più piccoli di dati*: Gli insiemi di dati spesso contengono un numero molto elevato di righe (n) rispetto al numero di colonne di caratteristiche (m): si pensi ai dati dei taxi di 80 milioni di viaggi con dieci campi registrati per viaggio. Il clustering fornisce un modo logico per suddividere un grande insieme di record in (diciamo) un centinaio di sottoinsiemi distinti, ciascuno ordinato per somiglianza. Ognuno di questi cluster contiene ancora più di un numero sufficiente di record su cui adattare un modello di previsione, e il modello risultante può essere più preciso su questa classe ristretta di articoli rispetto a un modello generale addestrato su tutti gli articoli. Per fare una previsione, ora è necessario identificare il cluster appropriato.
  L'elemento della query q appartiene a questo cluster, attraverso una ricerca di prossimità, e poi utilizzando il modello appropriato per quel cluster per effettuare la chiamata su q.

Riduzione dei dati: Trattare milioni o miliardi di record può essere eccessivo, sia per l'elaborazione che per la visualizzazione. Si consideri il costo computazionale per identificare il vicino più prossimo a un determinato punto della query, o per cercare di capire un grafico a punti con un milione di punti. Una tecnica consiste nel raggruppare i punti in base alla somiglianza, e poi nominare il *centroide* di ogni cluster per rappresentare l'intero cluster. Questi modelli di prossimità possono essere molto robusti, perché si riporta l'etichetta di consenso del cluster e si ottiene una misura naturale di fiducia: l'accuratezza di questo consenso sull'intero cluster.

- *Rilevamento degli outlier*: Alcuni elementi risultanti da qualsiasi procedura di raccolta dati saranno diversi da tutti gli altri. Forse riflettono errori di inserimento dei dati o misurazioni sbagliate. Forse segnalano bugie o altri comportamenti scorretti. O forse derivano da una miscela inaspettata di popolazioni, con alcune mele strane che possono rovinare l'intero cesto.

Il rilevamento degli outlier è il problema di liberare un insieme di dati da elementi discordanti, in modo che il resto rifletta meglio la popolazione desiderata. Il clustering è primo passo utile per trovare gli outlier. Gli elementi del cluster più lontani dal centro del cluster assegnato non si adattano bene a quel centro, ma non si adattano meglio nemmeno altrove. Questo li rende candidati ad essere degli outlier. Poiché gli invasori provenienti da un'altra popolazione tenderebbero a raggrupparsi insieme, potremmo sospettare dei piccoli cluster i cui centri si trovano insolitamente lontani da tutti gli altri centri di cluster.

Il clustering è un problema intrinsecamente poco definito, poiché i cluster adeguati dipendono dal contesto e dall'occhio di chi guarda.

Il numero di raggruppamenti che vede dipende in qualche modo da quanti raggruppamenti vuole vedere. Le persone possono essere raggruppate in due gruppi, quelli *che si dividono* e quelli che *si dividono*, a seconda della loro inclinazione a fare distinzioni sottili. Gli scissionisti guardano i cani e vedono barboncini, terrier e cocker spaniel. Gli scissionisti guardano i cani e vedono i mammiferi. Gli scissionisti traggono conclusioni più entusiasmanti, mentre i lumpers sono meno propensi a sovrastimare i dati. Quale sia la mentalità più appropriata dipende dal suo compito.

Sono stati sviluppati molti algoritmi di clustering diversi e nelle sezioni seguenti esamineremo i metodi più importanti (*k-means*, clustering agglomerativo e clustering spettrale). Ma è facile farsi prendere troppo dalle differenze tra i metodi. Se i dati presentano cluster abbastanza forti, qualsiasi metodo troverà qualcosa di simile. Ma quando un algoritmo restituisce cluster con una coerenza molto scarsa, di solito la colpa è più del set di dati che dellalgoritmo stesso.

10.5.1 Clustering k-means

Siamo stati un po' permissivi nel definire esattamente *ciò che* un algoritmo di clustering dovrebbe restituire come risposta. Una possibilità è quella di etichettare ogni punto con il nome del cluster in cui si trova. Se ci sono k cluster, queste etichette possono essere gli interi da 1 a k, dove etichettare il punto p con i significa che si trova nel cluster iesimo. Una rappresentazione di output equivalente potrebbe essere costituita da k elenchi separati di punti, dove l'elenco i rappresenta tutti i punti nel cluster *iesimo*.

Ma una nozione più astratta riporta il *punto centrale* di ogni cluster. In genere pensiamo ai cluster naturali come regioni compatte, di tipo gaussiano, dove esiste un centro ideale che definisce la posizione in cui i punti 'dovrebbero' trovarsi. Dato l'insieme di questi centri, il raggruppamento dei punti diventa facile: basta assegnare ogni punto p al punto centrale C_i più vicino. Il cluster *iesimo* consiste in tutti i punti il cui centro più vicino è C_i.

Il *clustering k-means* è un approccio al clustering veloce, semplice da capire e generalmente efficace. Inizia facendo un'ipotesi su dove potrebbero essere i centri del cluster, valuta la qualità di questi centri e poi li affina per ottenere stime migliori del centro.

L'algoritmo inizia assumendo che ci saranno esattamente k cluster nei dati, e poi procede a scegliere i centri iniziali per ogni cluster. Forse questo significa selezionare casualmente k punti dall'insieme di n punti S e chiamarli centri, oppure selezionare k punti casuali dal rettangolo di selezione di S. Ora verifica ciascuno degli n punti rispetto a tutti i k centri, e assegna ogni punto in S al suo centro attuale più vicino. Ora possiamo calcolare una stima migliore del centro di ogni cluster, come centroide dei punti assegnati ad esso. Ripetere fino a quando le assegnazioni dei cluster sono sufficientemente stabili, presumibilmente quando non sono cambiate dalla generazione precedente. Le ipotesi iniziali per i centri dei cluster sono davvero pessime el'assegnazione iniziale dei punti ai centri divide i cluster reali invece di rispettarli. Ma la situazione migliora rapidamente, con i centroidi che si spostano in posizioni che separano i punti nel modo desiderato. Si noti che la procedura k-means non termina necessariamente con il miglior insieme possibile di k centri, ma solo con una soluzione localmente ottimale che fornisce un punto di arresto logico. È una buona idea ripetere l'intera procedura più volte con diverse inizializzazioni casuali e accettare il miglior raggruppamento trovato su tutti. L'*errore quadratico medio* è la somma dei quadrati della distanza tra ogni punto P_i e il suo centro C_j, divisa per il numero di punti n. Il migliore dei due raggruppamenti può essere identificato come quello che ha un errore quadratico medio inferiore, o un'altra statistica di errore ragionevole. **Centri o centroidi?** Ci sono almeno due criteri possibili per calcolare una nuova stima del punto centrale come funzione dell'insieme S' di punti assegnati ad esso. Il *centroide* C di un insieme di punti viene calcolato prendendo il valore medio di ogni dimensione. Sono necessarie completamente sette iterazioni, a causa del posizionamento non fortunato dei tre centri iniziali del cluster vicino al centro logico.

Il centroide funge da *centro di massa* di S', il luogo in cui i vettori definiti attraverso questo punto si sommano a zero. Questo criterio di equilibrio definisce un centro naturale e unico per qualsiasi S'. La velocità di calcolo è un altro aspetto positivo dell'utilizzo del centroide. Per n punti d-dimensionali in S', questo richiede un tempo $O(nd)$, ossia lineare rispetto alla dimensione dei punti in ingresso. Per i punti dati numerici, l'utilizzo del centroide su una metrica L appropriata (come la distanza euclidea) dovrebbe funzionare bene. Tuttavia, i centroidi non sono ben definiti quando si raggruppano record di dati con attributi non numerici, come i dati categorici. Abbiamo discusso come costruire funzioni di distanza significative su record categoriali. Il problema qui non è tanto la misurazione della somiglianza, quanto la costruzione di un centro rappresentativo. Esiste una soluzione naturale, talvolta chiamata algoritmo *k-mediods*. Al posto del centroide, definiamo il punto più centrale C in S' come rappresentante del cluster. Si tratta del punto che minimizza la somma delle distanze a tutti gli altri punti del cluster.

Un vantaggio dell'utilizzo di un punto centrale per definire il cluster è che dà al cluster un nome e un'identità potenziali, supponendo che i punti di ingresso corrispondano a elementi con nomi identificabili.
Usare l'esempio di input più centrale come centro significa che possiamo eseguire kmeans, a condizione di avere una funzione di distanza significativa. Inoltre, non perdiamo molta precisione scegliendo il punto più centrale invece del centroide. Infatti, la somma delle distanze attraverso il punto centrale è al massimo il doppio di quella del centroide, su esempi numerici in cui il centroide può essere calcolato. La grande vittoria del centroide è che può essere calcolato più velocemente del vertice più centrale, di un fattore n.
L'utilizzo dei vertici centrali per rappresentare i cluster permette di estendere k -means naturalmente ai grafi e alle reti. Per i grafi ponderati, è naturale impiegare un algoritmo di percorso più breve per costruire una matrice D tale che $D[i, j]$ sia la lunghezza del percorso più breve nel grafo dal vertice i al vertice j. Una volta costruita D, *k-means* può procedere leggendo le distanze da questa matrice, invece di chiamare una funzione di distanza. Per i grafi non ponderati, si può utilizzare in modo efficiente un algoritmo a tempo lineare come la ricerca di ampiezza iniziale per calcolare le distanze del grafo su richiesta. **Quanti cluster?**

Nell'interpretazione del clustering k-means è insita l'idea di un *modello misto*. Invece di tutti i nostri dati osservati provenienti da un'unica fonte, presumiamo che i nostri dati provengano *da k* popolazioni o fonti diverse. Ogni fonte genera punti che assomigliano al suo centro, ma con un certo grado di variazione o errore. La questione del numero di cluster di un insieme di dati è fondamentale: a quante popolazioni diverse si è attinto quando si è selezionato il campione?
Il primo passo dell'algoritmo k-means consiste nell'inizializzare k, il numero di cluster nel set di dati dato. A volte abbiamo un preconcetto sul numero di cluster che vogliamo vedere: forse due o tre per l'equilibrio o la visualizzazione, o forse 100 o 1000 come proxy per "molti" quando si suddivide un file di input di grandi dimensioni in set più piccoli per la modellazione separata. Ma in generale questo è un problema, perché il numero "giusto" di cluster è solitamente sconosciuto. In effetti, la ragione principale per clusterizzare in primo luogo è la nostra comprensione limitata della struttura dell'insieme di dati.
Il modo più semplice per trovare il k giusto è provarli tutti e poi scegliere il migliore. Partendo da $k2$ fino al massimo del tempo a disposizione, esegua k-means e valuti il raggruppamento risultante in base all'errore quadratico medio (MSE) dei punti dai loro centri. Tracciando questo dato si ottiene una curva di errore. Viene fornita anche la curva di errore per i centri casuali. Entrambe le curve di errore mostrano che l'MSE dei punti dai loro centri diminuisce man mano che ammettiamo sempre più centri di cluster. Ma l'interpretazione sbagliata sarebbe Il clustering "giusto" viene trovato per k= 3, ma l'algoritmo non è in grado di distinguere correttamente tra cluster circolari annidati e cluster lunghi e sottili per k grandi. A titolo di confronto, viene mostrata la curva di errore per i centri di cluster casuali.
suggerisce che abbiamo bisogno di k il più grande possibile, perché l'MSE *dovrebbe* diminuire quando si ammettono più centri. In effetti, l'inserimento di un nuovo centro in una posizione casuale r in una soluzione k-means precedente può solo diminuire l'errore quadratico medio, perché si trova più vicino ad alcuni punti di

ingresso rispetto al loro centro precedente. Questo ritaglia un nuovo cluster intorno a *r*, ma presumibilmente un clustering ancora migliore sarebbe stato trovato eseguendo *k-means* da zero su (*k*+ 1) centri.

Vogliamo che *k* si trovi esattamente al gomito. Questo punto potrebbe essere più facile da identificare rispetto a un grafico simile dell'errore MSE per i centri casuali, poiché il tasso relativo di riduzione dell'errore per i centri casuali dovrebbe essere analogo a quello che vediamo dopo il gomito. La lenta deriva verso il basso ci dice che i cluster extra non stanno facendo nulla di speciale per noi.

Ogni nuovo centro di cluster aggiunge *d* parametri al modello, dove *d* è la dimensionalità dell'insieme di punti. Il rasoio di Occam ci dice che il modello più semplice è il migliore, che è la base filosofica per utilizzare la piega del gomito per selezionare *k*. Esistono criteri di merito formali che incorporano sia il numero di parametri che l'errore di previsione per valutare i modelli, come il *criterio di informazione di Akaike* (AIC). Tuttavia, nella pratica, dovrebbe sentirsi sicuro nel fare una scelta ragionevole per *k* in base alla forma della curva di errore.

Massimizzazione delle aspettative

L'algoritmo k-means è l'esempio più importante di una classe di algoritmi di apprendimento basati sulla *massimizzazione delle aspettative* (EM). I dettagli richiedono statistiche più formali di quelle che sono disposta ad approfondire in questa sede, ma il principio può essere osservato nelle due fasi logiche dell'algoritmo k means: (a) assegnare i punti al centro del cluster stimato che è più vicino a loro e (b) utilizzare queste assegnazioni di punti per migliorare la stima del centro del cluster. L'operazione di assegnazione è il *passo* dell'aspettativa o *E* dell'algoritmo, mentre il calcolo del centroide è il *passo* della massimizzazione dei parametri o *M*. I nomi "aspettativa" e "massimizzazione" non hanno una particolare risonanza per me in termini di algoritmo *k-means*. Tuttavia, la forma generale di un algoritmo di adattamento dei parametri iterativo, che migliora i parametri in serie in base agli errori dei modelli precedenti, sembra una cosa sensata da fare. Ad esempio, potremmo avere dati di classificazione parzialmente etichettati, dove ci sono pochissimi esempi di formazione assegnati con sicurezza alla classe corretta. Possiamo costruire dei classificatori basati su questi esempi di formazione e utilizzarli per assegnare i punti non etichettati alle classi candidate. In questo modo si definiscono presumibilmente set di formazione più ampi, quindi dovremmo essere in grado di adattare un modello migliore per ogni classe. Ora, riassegnando i punti e iterando di nuovo, si dovrebbe convergere su un modello migliore.

10.5.2 Clustering agglomerativo

Molte fonti di dati sono generate da un processo definito da una gerarchia o tassonomia sottostante. Spesso si tratta del risultato di un processo evolutivo: All'inizio c'era una cosa, che si è ripetutamente biforcata per creare un ricco universo di elementi. Idealmente, nel corso della clusterizzazione degli articoli, ricostruiremo queste storie evolutive. Questo obiettivo è esplicito nel *clustering agglomerativo*, un insieme di metodi bottom-up che uniscono ripetutamente i due cluster più vicini in un super-cluster di grandi dimensioni, definendo un albero radicato le cui foglie sono i singoli elementi e la cui radice definisce l'universo.

il clustering agglomerativo applicato ai dati di espressione genica. Qui ogni colonna rappresenta un particolare gene e ogni riga i risultati di un esperimento che misura l'attività di ciascun gene in una particolare condizione. Dall'ispezione risulta chiaro che ci sono blocchi di colonne che si comportano tutti in modo simile, attivandosi e disattivandosi in condizioni simili. La scoperta di questi blocchi si riflette nell'albero sopra la matrice: le regioni di grande somiglianza sono associate a piccole ramificazioni. Ogni nodo dell'albero rappresenta la fusione di due cluster. L'altezza del nodo è proporzionale alla distanza tra i due cluster che vengono uniti. Quanto più alto è il bordo, tanto più è discutibile l'idea che questi cluster *debbano* essere fusi. Le colonne della matrice sono state permutate per riflettere questo albero. ci permette di visualizzare centinaia di geni quantificati in quattordici dimensioni (con ogni riga che definisce
una dimensione distinta).

I raggruppamenti biologici sono spesso associati a tali *dendogrammi* o *alberi filogenici*, perché sono il risultato di un processo evolutivo. Infatti, i raggruppamenti di comportamenti di espressione genica simili che vediamo qui sono il risultato dell'evoluzione di una nuova funzione dell'or- ganismo, che cambia la risposta di alcuni geni a una particolare condizione.

Utilizzo di alberi agglomerativi

Il clustering agglomerativo restituisce un albero in cima ai raggruppamenti di elementi. Dopo aver tagliato i bordi più lunghi di questo albero, ciò che rimane sono i gruppi disgiunti di articoli prodotti dagli algoritmi di clustering come *k-means*. Ma questo albero è una cosa meravigliosa, con poteri che vanno ben oltre la suddivisione degli articoli:

- *Organizzazione dei cluster e dei sottocluster*: Ogni nodo interno dell'albero definisce un cluster particolare, composto da tutti gli elementi dei nodi-foglia che si trovano al di sotto di esso. Ma l'albero descrive la gerarchia tra questi cluster, dai cluster più raffinati/specifici vicino alle fogli e ai cluster più generali vicino alla radice. Idealmente, i nodi di un albero definiscono concetti nominabili: raggruppamenti naturali che un esperto del dominio potrebbe spiegare se gli venisse chiesto. Questi vari livelli di granularità sono importanti, perché definiscono concetti strutturali che potremmo non aver notato prima di fare il clustering.

- *Visualizzazione del processo di clustering*: Un disegno di questo albero di agglomerazione ci dice molto sul processo di clustering, soprattutto se il disegno riflette il costo di ogni fase di fusione. Idealmente, ci saranno bordi molto lunghi vicino alla radice dell'albero, a dimostrazione che i c luster di livello più alto sono ben separati e appartengono a raggruppamenti distinti. Possiamo capire se i raggruppamenti sono equilibrati o se i raggruppamenti di alto livello hanno dimensioni sostanzialmente diverse. Lunghe catene di fusione di piccoli cluster in un grande cluster sono generalmente un segnale negativo, anche se la scelta dei criteri di fusione (di cui parleremo più avanti) può influenzare la forma dell'albero. Gli outlier sono ben visibili su un albero filogenico, come elementi singoli o piccoli cluster che si collegano alla radice attraverso lunghi bordi.

- *Misura naturale della distanza tra i cluster*: Una proprietà interessante di qualsiasi albero T è che esiste esattamente un percorso in T tra due qualsiasi nodi x e y. Ogni vertice interno di un albero di clustering agglomerativo ha un peso associato, il costo della fusione dei due sottoalberi sottostanti. Possiamo calcolare una 'distanza del cluster' tra due foglie qualsiasi, attraverso la somma dei costi di fusione sul percorso tra di esse. Se l'albero è buono, questo può essere più significativo della distanza euclidea tra i record associati a x e a y.

- *Classificazione efficiente di nuovi articoli*: Un'applicazione importante per la clusterizzazione è la classificazione. Supponiamo di aver clusterizzato in modo agglomerativo i prodotti... in un negozio, per costruire una tassonomia di cluster. Ora un nuovo pezzo. In quale categoria dovrebbe essere classificato?

Per *k-means*, ciascuno dei c cluster è classificato dal suo centroide, quindi la classificazione di un nuovo elemento q si riduce al calcolo della distanza tra q e tutti i centriidi c per identificare il cluster più vicino. Un albero gerarchico offre metodo potenzialmente più veloce. Supponiamo di aver precompilato i centroidi di tutte le foglie dei sottoalberi destro e sinistro sotto ogni nodo. L'identificazione della posizione giusta nella gerarchia per un nuovo elemento q inizia confrontando q con i centroidi dei sottoalberi sinistro e destro della radice. Il più vicino dei due centroidi a q definisce il lato appropriato dell'albero, quindi riprendiamo la ricerca da lì, un livello più in basso. Questa ricerca richiede un tempo proporzionale all'altezza dell'albero, invece che al numero di foglie. In genere si

tratta di un miglioramento da n a log n, che è molto meglio. Capire che gli alberi a fusione binaria possono essere disegnati in molti modi diversi che riflettono esattamente la stessa struttura, perché non esiste una nozione intrinseca di quale sia il figlio sinistro e quale il figlio destro. Ciò significa che sono possibili permutazioni distinte delle n foglie, invertendo la direzione di qualsiasi sottoinsieme degli n 1 nodi interni dell'albero. Si renda conto di questo quando cerca di leggere una tassonomia di questo tipo: due elementi che sembrano lontani nell'ordine sinistra-destra potrebbero essere vicini se questo capovolgimento fosse stato fatto in modo diverso. E il nodo più a destra del sottoalbero sinistro potrebbe essere vicino al nodo più a sinistra del sottoalbero destro, anche se in realtà sono molto distanti nella tassonomia.

Costruzione di alberi di cluster agglomerativi

L'algoritmo di clustering agglomerativo di base è abbastanza semplice da essere descritto in due frasi. Inizialmente, ogni elemento viene assegnato al proprio cluster. Unisce i due cluster più vicini in uno mettendo una radice su di essi, e ripete fino a quando rimane solo cluster. Non resta che specificare come calcolare la distanza tra i cluster. Quando i cluster contengono singoli elementi, la risposta è semplice: utilizzare la metrica di distanza preferita, come L_2. Ma ci sono diverse risposte ragionevoli per la distanza tra due cluster non banali, che portano ad alberi diversi sullo stesso input e possono avere un profondo impatto sulla forma dei cluster risultanti.

- *Vicino più vicino (collegamento singolo)*: In questo caso, la distanza tra i cluster C_1 e C_2 è definito dalla coppia di punti più vicina che li attraversa:

  L'utilizzo di questa metrica si chiama *clustering a legame singolo*, perché la decisione di unire si basa esclusivamente sul singolo legame più vicino tra i cluster.

  Il *minimum spanning tree* di un grafo G è un albero disegnato dai bordi di G che collegano tutti i vertici al costo totale più basso. Il clustering agglomerativo con il criterio del legame singolo è essenzialmente lo stesso dell'algoritmo di Kruskal, che crea il *minimum spanning tree* (MST) di un grafico aggiungendo ripetutamente il bordo di peso più basso rimasto che non crea un ciclo nell'albero emergente.
  Il collegamento tra l'MSTe l'albero dei cluster è piuttosto— sottile: l'ordine dei bordi -di inserimento nell'MST, dal più piccolo al più grande, descrive l'ordine di fusione nell'albero dei cluster.

  L'ideale platonico dei cluster è rappresentato da regioni circolari compatte, che in genere si irradiano dai centroidi, come nel clustering *k-means*. Al contrario, il clustering singlelink tende a creare cluster relativamente lunghi e sottili, perché la decisione di fusione si basa solo sulla vicinanza dei punti di confine. Il clustering single link è veloce, ma tende ad essere soggetto ad errori, in quanto i punti più lontani possono facilmente risucchiare due cluster ben definiti.

- *Collegamento medio*: In questo caso, calcoliamo la distanza tra tutte le coppie di punti di cluster spanning e ne facciamo una media per ottenere un criterio di fusione più robusto rispetto al legame singolo

Questo tenderà ad evitare i cluster magri del collegamento singolo, ma con un costo computazionale maggiore. L'implementazione semplice del clustering a link medio è $O(n^3)$, perché ognuna delle n fusioni richiederà potenzialmente di toccare $O(n^2)$ bordi per ricompilare il cluster rimanente più vicino. Questo è n volte più lento del clustering a link singolo, che può essere implementato in tempo $O(n^2)$.

- *Centroide più vicino*: qui manteniamo il centroide di ogni cluster e uniamo la coppia di cluster con i centroidi più vicini. Questo ha due vantaggi principali. In primo luogo, tende a produrre cluster simili al legame medio, perché i punti anomali in un cluster vengono sopraffatti all'delle dimensioni del cluster

(numero di punti). In secondo luogo, è molto più veloce confrontare i centroidi dei due cluster che testare tutte le coppie di punti C_1 C_2 nell'implementazione più semplice. Naturalmente, i centroidi possono essere calcolati solo per i record con tutti i valori numerici, ma l'algoritmo può essere adattato per utilizzare il punto più centrale di ogni cluster (medioide) come rappresentante nel caso generale.

- *Collegamento più lontano*: Qui il costo dell'unione di due cluster è la coppia di punti più lontani tra loro Sembra una follia, ma questo è il criterio che funziona meglio per mantenere i cluster rotondi, penalizzando le fusioni con elementi distanti e anomali.

Quale di questi è il migliore? Come sempre in questo settore, dipende. Per gli insiemi di dati molto grandi, ci preoccupiamo soprattutto di utilizzare gli algoritmi più veloci, che in genere sono il single linkage o il nearest centerid con strutture di dati appropriate. Per gli insiemi di dati di piccole e medie dimensioni, ci preoccupiamo soprattutto della qualità, rendendo interessanti i metodi più robusti.

10.5.3 Confronto dei raggruppamenti

È una pratica comune provare diversi algoritmi di clustering sullo stesso set di dati e utilizzare quello che sembra migliore per i nostri scopi. I raggruppamenti prodotti da due algoritmi diversi dovrebbero essere abbastanza simili se entrambi gli algoritmi stanno facendo cose ragionevoli, ma spesso è interessante misurare esattamente quanto sono simili. Ciò significa che dobbiamo definire una misura di somiglianza o di distanza sui raggruppamenti.

Ogni cluster è definito da un sottoinsieme di elementi, siano essi punti o record. *La somiglianza Jaccard* degli insiemi s_1 e s_2 è definita come il rapporto tra la loro intersezione e unione.

Poiché l'intersezione di due insiemi è sempre non più grande dell'unione dei loro elementi. La somiglianza di Jaccard è una misura generalmente utile conoscere, ad esempio per confrontare la somiglianza dei k più prossimi di un punto in base a due diverse metriche di distanza, o la frequenza con cui i primi elementi di un criterio corrispondono ai primi elementi di un'altra metrica. Questa misura di somiglianza può essere trasformata in una vera e propria metrica di distanza $d(s_1, s_2)$ chiamata *distanza di Jaccard*. Questa funzione di distanza assume solo valori compresi tra 0 e 1, ma soddisfa tutte le proprietà di una metrica, compresa la disuguaglianza del triangolo. Ogni clustering è descritto da una partizione dell'insieme universale e può avere molte parti. L' *indice Rand* è una misura naturale della somiglianza tra due raggruppamenti c_1 e c_2. Se i raggruppamenti sono compatibili, qualsiasi coppia di elementi nello stesso sottoinsieme di c_1 dovrebbe essere nello stesso sottoinsieme di c_2, e qualsiasi coppia in cluster diversi di c_1 dovrebbe essere separata in c_2. L'indice Rand conta il numero di coppie di articoli così coerenti e lo divide per il numero totale di elementi. coppie per creare un rapporto da 0 a 1, dove 1 denota raggruppamenti identici. 2

10.5.4 Grafici di somiglianza e clustering basato sul taglio

Ricordiamo la nostra discussione iniziale sul clustering, in cui le ho chiesto quanti cluster ha visto nell'insieme di punti ripetuto. Per ottenere la risposta ragionevole di nove cluster, il suo algoritmo di clustering interno doveva gestire trucchi come la classificazione di un anello intorno a un blob centrale come due cluster distinti, ed evitare di unire due linee che si muovono sospettosamente vicine l'una all'altra. *k-means* non ha alcuna possibilità, perché cerca sempre cluster circolari ed è felice di dividere cluster lunghi e filiformi. Tra le procedure di clustering agglomerativo, solo single-link con la soglia giusta potrebbe avere la possibilità di fare la cosa giusta, ma è facile che venga ingannato nel fondere due cluster da un'unica coppia di punti vicini.

I cluster non sono sempre rotondi. Per riconoscere quelli che non lo sono, è necessaria una *densità* di punti abbastanza elevata e sufficientemente *contigua* da non essere.

La clusterizzazione spettrale trova correttamente i cluster collegati che *k-means* non riesce a trovare.

Cerchiamo i cluster che sono *collegati* in un grafico di somiglianza appropriato.

Una *matrice di somiglianza* $n \times$ n segna quanto sono simili le coppie di elementi p_i e $p_{(j)}$. La somiglianza è essenzialmente l'inverso della distanza: quando p_i è vicino a p_j, allora l'elemento associato a p_i deve essere simile a quello di p_j. È naturale misurare la somiglianza su una scala da 0 a 1, dove 0 rappresenta una differenza totale e 1 significa identico. Questo può essere realizzato rendendo $S[i, j]$ una funzione esponenziale inversa della distanza, regolata da un parametro.

Un *grafico di somiglianza* ha un bordo ponderato (i, j) tra ogni coppia di vertici i e j che riflette la somiglianza di p_i e p_j. Si tratta esattamente della matrice di somiglianza descritta in precedenza. Tuttavia, possiamo rendere questo grafico più rado impostando tutti i termini piccoli a zero. Questo riduce notevolmente il numero di bordi del grafico. Possiamo anche trasformarlo in un grafico non ponderato, impostando il peso a 1 per tutti i termini.

Tagli nei grafici

I cluster reali nei grafici di somiglianza hanno l'aspetto di regioni dense che sono solo debolmente collegate al resto del grafico. Un cluster C ha un peso che è una funzione dei bordi all'interno del cluster. I bordi che collegano C al resto del grafico definiscono un *taglio*, ossia l'insieme di bordi che hanno un vertice in C e l'altro nel resto del grafico (V C). Il peso di questo taglio $W(C)$.

Idealmente, i cluster avranno un peso elevato $W(C)$ ma un taglio piccolo $W(C)$. conduttanza del cluster C è il rapporto tra il peso interno e il peso interno. *La conduttanza* del cluster C è il rapporto tra il peso del taglio e il peso interno ($W(C)/W(C)$), con cluster migliori che hanno una conduttanza inferiore.

Trovare cluster a bassa conduttanza è una sfida. L'aiuto viene, sorprendentemente, dall'algebra lineare. La matrice di somiglianza S è una matrice simmetrica, il che significa che ha una decomposizione agli autovalori, come discusso nella Sezione 8.5. Abbiamo visto che l'autovalore principale determina un'approssimazione a blocco di S, con il contributo di altri autovalori. Abbiamo visto che l'autovalore principale determina un'approssimazione a blocchi di S, con il contributo di altri autovalori che migliorano gradualmente l'approssimazione. L'eliminazione degli autovalori più piccoli elimina i dettagli o il rumore, a seconda dell'interpretazione.

Si noti che la matrice di somiglianza ideale *è* una matrice a blocchi, perché all'interno di ogni cluster ci aspettiamo una connessione densa di coppie altamente simili, con pochi collegamenti incrociati con i vertici di altri cluster. Questo suggerisce di utilizzare gli autovettori di S per definire caratteristiche robuste cui raggruppare i vertici. Eseguendo il clustering k-means su questo spazio di caratteristiche trasformato, si recupereranno buoni cluster.

Questo approccio si chiama *clustering spettrale*. Costruiamo una matrice di somiglianza opportunamente normalizzata, chiamata *Laplaciano*.

Gli autovalori importanti di L definiscono una matrice di caratteristiche di n k. Curiosamente, gli autovalori più preziosi per il clustering risultano avere i *più piccoli* autovalori non nulli, grazie a particolari proprietà della matrice laplaciana. L'esecuzione di un clustering a k significati in questo spazio di caratteristiche genera cluster altamente collegati.

Capitolo 11

Apprendimento automatico

Detto questo, sono stati sviluppati una serie di algoritmi di apprendimento automatico interessanti e importanti. In questo passeremo in rassegna questi metodi, con l'obiettivo di comprendere i punti di forza e di debolezza di ciascuno, lungo diverse dimensioni rilevanti delle prestazioni:

- *Potenza ed esprimibilità*: I metodi di apprendimento automatico si differenziano per la ricchezza e la complessità dei modelli che supportano. La regressione lineare si adatta a funzioni lineari, mentre i metodi nearest neighbor definiscono confini di separazione lineare con pezzi sufficienti per approssimare curve arbitrarie. Una maggiore potenza espressiva offre la possibilità di modelli più accurati, ma anche i pericoli di overfitting.

- *Interpretabilità*: Metodi potenti come l'apprendimento profondo spesso producono modelli che sono completamente impenetrabili. Potrebbero fornire una classificazione molto accurata nella pratica, ma senza una spiegazione leggibile per l'uomo del motivo per cui prendono le decisioni che prendono. Al contrario, i coefficienti più grandi in un modello di regressione lineare identificano le caratteristiche più potenti, e le identità dei vicini più prossimi ci permettono di determinare in modo indipendente la nostra fiducia in queste analogie.

  Personalmente ritengo che l'interpretabilità sia una proprietà importante di un modello e sono generalmente più felice di accettare un modello meno performante che comprendo rispetto a uno leggermente più preciso che non comprendo. Questa potrebbe non essere un'opinione universalmente condivisa, ma lei ha la sensazione di capire veramente il suo modello e il suo particolare dominio di applicazione.

Facilità d'uso: Alcuni metodi di apprendimento automatico presentano relativamente pochi parametri o decisioni, il che significa che funzionano subito. Sia la regressione lineare che la classificazione nearest neighbor sono piuttosto semplici a questo proposito. Al contrario, metodi come le macchine a vettori di supporto (SVM) offrono una possibilità molto più ampia di ottimizzare le prestazioni dell'algoritmo con le impostazioni appropriate. La mia sensazione è che gli strumenti disponibili per l'apprendimento automatico continueranno a migliorare: più facili da usare e più potenti. Ma per il momento, alcuni metodi consentono all'utente di avere abbastanza corda per impiccarsi se non sa cosa sta facendo.

- *Velocità di addestramento*: I metodi si differenziano notevolmente per la velocità con cui si adattano ai parametri necessari del modello, il che determina la quantità di dati di addestramento che si può permettere di utilizzare nella pratica. I metodi tradizionali di regressione lineare possono essere costosi da adattare a modelli di grandi dimensioni. Al contrario, la ricerca per non richiede quasi alcun tempo di addestramento, se non quello necessario per costruire la struttura di dati di ricerca appropriata.

- *Velocità di previsione*: i metodi si differenziano per la velocità con cui effettuano le decisioni di classificazione su una nuova query q. La regressione lineare/logistica è veloce, in quanto calcola semplicemente una somma ponderata dei campi nei record di input. Al contrario, la ricerca del vicino richiede di testare esplicitamente q rispetto a una quantità sostanziale di test di formazione. In generale, esiste un compromesso con la velocità di formazione: può pagarmi ora o pagarmi dopo.

Si spera che il panorama degli algoritmi di apprendimento automatico venga analizzato in modo utile. Certamente nessun metodo di apprendimento automatico domina tutti gli altri. Questa osservazione è formalizzata nel *teorema no free lunch*, opportunamente chiamato, che dimostra che non esiste un singolo algoritmo di apprendimento automatico migliore di tutti gli altri su tutti i problemi. Un ultimo commento. Gli scienziati dei dati tendono ad avere un approccio di apprendimento automatico preferito, che sostengono in modo simile alla loro lingua preferita o alla loro squadra sportiva. Gran parte di questo è dovuto all'esperienza, ossia alla familiarità con una particolare

implementazione che funziona meglio nelle loro mani. Ma in parte si tratta di pensiero magico, del fatto che hanno notato una libreria leggermente superiore ad altre su alcuni esempi e hanno generalizzato in modo inappropriato. Non cada in questa trappola. Scelga i metodi che meglio si adattano alle esigenze della sua applicazione in base ai criteri di cui sopra, e acquisisca sufficiente esperienza con le varie manopole e leve per ottimizzare le prestazioni.

11.1 Baia ingenua

Ricordiamo che due eventi A e B sono *indipendenti*. Se A è l'evento "la mia squadra sportiva preferita vince oggi" e B è "il mercato azionario sale oggi", allora presumibilmente A e B sono indipendenti. Ma questo non è vero in generale. Consideriamo il caso in cui A è l'evento "prendo una A in Data Science questo semestre" e B è "prendo una A in un altro corso questo semestre". Ci sono delle dipendenze tra questi eventi: il rinnovato entusiasmo per lo studio o per il bere influenzerà il rendimento del corso in correlato.

Se tutto fosse indipendente, il mondo della probabilità sarebbe un luogo molto più semplice. L'algoritmo di classificazione di Bayes ingenuo incrocia le dita e assume l'indipendenza, per evitare di dover calcolare queste complicate probabilità condizionali.

11.1.1 Formulazione

Supponiamo di voler classificare il vettore in una delle m classi. Cerchiamo di calcolare la probabilità di ogni possibile classe data X, in modo da poter assegnare $a\ X$ l'etichetta della classe con la più alta probabilità. Per il teorema di Bayes.

Analizziamo questa equazione. Il termine $p(C_i)$ è la probabilità *precedente*, la probabilità dell'etichetta di classe senza alcuna prova specifica. Il denominatore $P\ (X)$ indica la probabilità di vedere il dato vettore di ingresso X su tutti i possibili vettori di ingresso. Stabilire il valore esatto di $P\ (X)$ sembra un po' rischioso, ma fortunatamente di solito non è necessario. Osserviamo che questo denominatore è lo stesso per tutte le classi. Cerchiamo solo di stabilire un'etichetta di classe per X, quindi il valore di $p(X)$ non ha alcun effetto sulla nostra decisione. Selezionare la classe con la probabilità più alta significa. Il termine rimanente è la probabilità di vedere il vettore di input X, dato che sappiamo che la classe dell'articolo è C_i. Anche questo sembra un po' rischioso.

Ora, chiunque creda davvero in un mondo di probabilità indipendenti è piuttosto ingenuo, da cui il nome *ingenuo di Bayes*. Ma questo presupposto rende davvero i calcoli molto più semplici.

Infine, dovremmo colpire il prodotto con un log per trasformarlo in una somma, per una migliore stabilità numerica. I log delle probabilità saranno numeri negativi, ma gli eventi meno probabili sono più negativi di quelli comuni.

Come si calcola la probabilità dell'osservazione x_j data l'etichetta di classe i? Questo è facile dai dati di formazione, soprattutto se x_j è una variabile categorica. Possiamo semplicemente selezionare tutte le istanze di classe i nel set di addestramento e calcolare la frazione di essi che hanno la proprietà x_j. Questa frazione definisce una stima ragionevole di p. È necessaria un po' più di immaginazione quando x_j è una variabile numerica, ma in linea di principio viene calcolata in base alla frequenza con cui questo valore viene osservato nell'insieme di formazione.

11.1.2 Trattare con i conteggi zero (attualizzazione)

C'è un problema sottile ma importante di preparazione delle caratteristiche, particolarmente associato all'algoritmo di Bayes ingenuo. I conteggi osservati non catturano accuratamente la frequenza degli eventi rari, per i quali c'è in genere una lunga coda. Ci possono sempre essere eventi che non sono ancora stati osservati in un insieme di dati finito. *L'attualizzazione* è una tecnica statistica per aggiustare i conteggi per gli eventi non ancora visti, lasciando esplicitamente la massa di probabilità disponibile per essi. La tecnica più semplice e popolare è l'attualizzazione *con l'aggiunta di uno*, in cui si aggiunge uno alla frequenza di tutti i risultati, compresi quelli non visti. Ad esempio, supponiamo di estrarre delle palline da un'urna. Dopo aver visto cinque

rossi e tre verdi, qual è la probabilità di vedere un nuovo colore alla prossima estrazione? Se utilizziamo l'attualizzazione con l'aggiunta di uno. Per un numero ridotto di campioni o un numero elevato di classi conosciute, il disconto provoca uno smorzamento non banale delle probabilità. La nostra stima della probabilità di vedere una palla rossa cambia da 5/8= 0,625 a 6/11= 0,545 quando impieghiamo l'attualizzazione per addizione. Ma questa è una stima più sicura e onesta, e le differenze scompariranno nel nulla dopo aver visto un numero sufficiente di campioni. Deve essere consapevole che sono stati sviluppati altri metodi di attualizzazione e l'aggiunta di uno potrebbe non essere il miglior stimatore possibile in tutte le situazioni. Detto questo, *non* scontare i conteggi significa creare problemi, e nessuno verrà licenziato per aver utilizzato il metodo dell'aggiunta. L'attualizzazione diventa particolarmente importante nell'elaborazione del linguaggio naturale, dove la rappresentazione tradizionale del *bag of words* modella un documento come un vettore di conteggio della frequenza delle parole sull'intero vocabolario della lingua, ad esempio 100.000 parole. Poiché la frequenza d'uso delle parole è regolata da una legge di potenza (legge di Zipf), le parole nella coda di sono piuttosto rare. Ha mai visto la parola inglese *defenestrare*?[16] Ancora peggio, i documenti di lunghezza inferiore a quella di un libro sono troppo brevi per contenere 100.000 parole, quindi siamo destinati a vedere degli zeri ovunque guardiamo. L'attualizzazione dell'aggiunta trasforma questi vettori di conteggio in vettori di probabilità sensati, con probabilità non nulle di vedere parole rare e finora non incontrate.

11.2 Classificatori ad albero decisionale

Un *albero decisionale* è una struttura a ramificazione binaria utilizzata per classificare un vettore di input arbitrario X. Ogni nodo dell'albero contiene un semplice confronto di caratteristiche rispetto a qualche campo X.

confronto è vero o falso e determina se dobbiamo procedere verso il figlio sinistro o destro del nodo in questione. Queste strutture sono talvolta chiamate *alberi di classificazione e regressione* (CART) perché possono essere applicate a una classe più ampia di problemi. L'albero decisionale suddivide gli esempi di formazione in gruppi con una composizione di classi relativamente univoca, per cui la decisione diventa facile. Ogni riga/istanza percorre un percorso unico dalla radice alla foglia per arrivare alla classificazione. Il test della radice riflette la tradizione navale di dare la precedenza alle donne e ai bambini: il 73% delle donne è sopravvissuto, quindi questa caratteristica da sola è sufficiente per fare una previsione per le donne. Il secondo livello dell'albero riflette prima i bambini: qualsiasi maschio di 10 anni o più è considerato sfortunato. Anche i più giovani devono superare un ultimo ostacolo: in genere riuscivano a raggiungere una scialuppa di salvataggio solo se avevano fratelli e sorelle che facevano pressione per loro. Qual è l'accuratezza di questo modello sui dati di formazione? Dipende da quale fazione degli esempi finisce su ogni foglia e da quanto sono puri questi campioni di foglie. Un'accuratezza del 78,86% non è male per una procedura decisionale così semplice. Avremmo potuto portarla al 100% completando l'albero in modo che ognuno dei 1317 passeggeri avesse una foglia a sé, etichettando quel nodo con il suo destino finale. Forse i maschi di 23 anni in seconda classe avevano maggiori probabilità di sopravvivere rispetto ai maschi di 22 o 24 anni, un'osservazione che l'albero avrebbe potuto sfruttare per una maggiore accuratezza della formazione. Ma un albero così complicato sarebbe eccessivamente adattato, trovando una struttura che non è significativa. L'albero è interpretabile, robusto e ragionevolmente accurato. Al di là di questo, ognuno pensa a sé stesso. I vantaggi degli alberi decisionali includono:

- *Non linearità*: Ogni foglia rappresenta una porzione dello spazio decisionale, ma raggiungibile attraverso un percorso potenzialmente complicato. Questa catena di logica permette agli alberi decisionali di rappresentare confini decisionali molto complicati.
- *Supporto per le variabili categoriche*: Gli alberi decisionali utilizzano in modo naturale le variabili categoriali, come "se il colore dei capelli = è rosso", oltre ai dati numerici. Le variabili categoriali si adattano meno comodamente alla maggior parte degli altri metodi di apprendimento automatico.

- *Interpretabilità*: Gli alberi decisionali sono spiegabili; si possono leggere e capire il loro ragionamento. Pertanto, gli algoritmi degli alberi decisionali possono dirle qualcosa sulla sua serie di dati che potrebbe non aver visto prima. Inoltre, l'interpretabilità le consente di valutare se si fida delle decisioni che prenderà: sta prendendo decisioni per le giuste ragioni?

- *Robustezza*: Il numero di alberi decisionali possibili cresce esponenzialmente nel numero di caratteristiche e di test possibili, il che significa che possiamo costruirne quanti ne vogliamo. Costruire molti alberi decisionali casuali (CART) e considerare il risultato di ciascuno come un voto per l'etichetta data aumenta la robustezza e ci permette di valutare la fiducia della nostra classificazione.

Applicazione alla regressione: I sottoinsiemi di elementi che seguono un percorso simile lungo un albero decisionale sono probabilmente simili per proprietà diverse dalla semplice etichetta. Per ciascuno di questi sottoinsiemi, possiamo utilizzare la regressione lineare per costruire un modello di previsione speciale per i valori numerici di questi elementi della foglia. Questo avrà presumibilmente prestazioni migliori rispetto a un modello più generale addestrato su tutte le istanze.

Lo svantaggio maggiore degli alberi decisionali è una certa mancanza di eleganza. I metodi di apprendimento come la regressione logistica e le macchine vettoriali di supporto utilizzano la *matematica*. Teoria della probabilità avanzata, algebra lineare, geometria dimensionale superiore. Insomma, la *matematica*.

Al contrario, gli alberi decisionali sono un gioco da hacker. Ci sono molte manopole interessanti da girare nella procedura di formazione, e relativamente poca teoria per aiutarla a girarle nel modo giusto.

Ma il fatto è che i modelli di alberi decisionali funzionano molto bene nella pratica. *Gli alberi decisionali potenziati dal gradiente* (GBDT) sono attualmente il metodo di apprendimento automatico più utilizzato per vincere le competizioni Kaggle. Lavoreremo su questo argomento in più fasi. Prima gli alberi decisionali, poi il boosting nella sezione successiva.

11.2.1 Costruire alberi decisionali

Gli alberi decisionali sono costruiti in top-down. Si parte da una data collezione di istanze di formazione, ciascuna con n caratteristiche ed etichettata con una delle m classi $C_1, \ldots C_m$. Ogni nodo dell'albero decisionale contiene un predicato binario, una condizione logica derivata da una determinata caratteristica.

Le caratteristiche con un insieme discreto di valori v_i possono essere facilmente trasformate in predicati binari attraverso il test di uguaglianza: "La caratteristica x_i è= v_{ij}?". Quindi ci sono v predicati distinti associati a x_i. Le caratteristiche numeriche possono essere trasformate in predicati binari con l'aggiunta di una soglia t.

L'insieme di soglie potenzialmente interessanti t sono definite dagli scarti tra i valori osservati che x_i assume nel set di formazione. Se l'insieme completo delle osservazioni di x_i sono i valori significativi possibili per t.

Entrambi gli insiemi di soglie producono le stesse partizioni delle osservazioni, ma l'utilizzo dei punti medi di ogni divario sembra essere più sicuro quando si generalizza a valori futuri non visti nell'addestramento.

Abbiamo bisogno di un modo per valutare ogni predicato per quanto bene contribuirà a suddividere l'insieme S di esempi di formazione raggiungibili da questo nodo. Un predicato ideale p sarebbe una *partizione pura* di S, in modo che le etichette di classe siano disgiunte. In questo sogno, tutti i membri di S di ogni classe C_i appariranno esclusivamente su un lato dell'albero, ma tale purezza non è solitamente possibile. Vogliamo anche dei predicati che producano *delle suddivisioni bilanciate* di S, il che significa che il sottoalbero di sinistra contiene circa lo stesso numero di elementi di S del sottoalbero di destra. Le suddivisioni bilanciate consentono di progredire più rapidamente nella classificazione e sono anche potenzialmente più robuste. Impostando la soglia t al valore minimo di $x_{(i),}$ si preleva un elemento solitario da S e si ottiene una suddivisione perfettamente pura ma massimamente squilibrata.

Quindi i nostri criteri di selezione dovrebbero premiare sia l'equilibrio che la purezza, per massimizzare ciò che impariamo dal test. Un modo per misurare la purezza di un sottoinsieme di elementi S è l'inverso del disordine, o *entropia*. Lasciare che f_i indichi la frazione di S che appartiene alla classe C_i. Si può quindi calcolare l'entropia teorica dell'informazione di S, $H(S)$.

Il segno negativo qui esiste per rendere positiva l'intera quantità, poiché il logaritmo di una frazione propria è sempre negativo.

Analizziamo questa formula. Il contributo più puro possibile si verifica quando tutti gli elementi appartengono a una sola classe, ovvero $f_j = 1$ per qualche classe j. Il contributo della classe j a $H(S)$ è $1 \log_2(1) = 0$, identico a quello di tutte le altre classi. La versione più disordinata si ha quando tutte le m classi sono rappresentate in modo uguale, cioè $f_i = 1/m$. Allora $H(S) = \log_2(m)$ in base alla definizione precedente. Più piccola è l'entropia, migliore è il nodo per la classificazione.

Il valore di una potenziale suddivisione applicata a un nodo dell'albero è quanto riduce l'entropia del sistema. Supponiamo che un predicato booleano p divida S in due sottoinsiemi disgiunti. Allora il *guadagno di informazioni* di p. Cerchiamo il predicato p' che massimizza questo guadagno di informazioni, come miglior splitter per S. Questo criterio preferisce implicitamente splitter bilanciati, poiché vengono valutati entrambi i lati dell'albero. Sono state definite misure alternative di purezza che vengono utilizzate nella pratica.

L'*impurità Gini* si basa su un'altra quantità che è zero in A sinistra, presentiamo quattro cluster naturali nello spazio. Questo dimostra la completa incapacità della regressione logistica di trovare un separatore significativo, anche se un piccolo albero decisionale svolge facilmente il lavoro (a destra). in entrambi i casi di scissione pura.

I criteri di selezione dei predicati per ottimizzare l'impurità di Gini possono essere definiti in modo simile.

Abbiamo bisogno di una condizione di arresto per completare l'euristica. Quando un nodo è sufficientemente puro per essere considerato una foglia? Impostando una soglia ϵ sul guadagno di informazioni, possiamo smette di dividere quando la ricompensa di un altro test è inferiore a ϵ.

Una strategia alternativa consiste nel costruire l'albero completo fino a quando tutte le foglie sono completamente pure, e poi potarlo eliminando i nodi che contribuiscono al minor guadagno di informazioni. È abbastanza comune che un universo di grandi dimensioni possa non avere buoni splitter vicino alla radice, ma ne emergono di migliori man mano che l'insieme di elementi vivi si riduce. Questo approccio ha il vantaggio di non rinunciare troppo presto al processo.

11.2.2 Realizzare l'esclusiva o

Alcune forme di confine decisionale possono essere difficili o addirittura impossibili da adattare a un particolare approccio di apprendimento automatico. La cosa più nota è che i classificatori lineari non possono essere utilizzati per adattarsi ad alcune semplici funzioni non lineari, come l'OR esclusivo (XOR). La funzione logica $A \oplus B$ è definita come per i punti (x, y) in due dimensioni, possiamo definire dei predicati tali che A significa "$x \geq 0$?" e B significa "$y \geq 0$?". Quindi ci sono due distinti regioni dove $A \oplus$ B vero, quadranti opposti in questo *piano* xy. La necessità di ritagliare due regioni con una linea spiega perché lo XOR è impossibile per i classificatori lineari. Gli alberi decisionali sono abbastanza potenti da riconoscere lo XOR. Dopo che la radice verifica se A è vero o falso, i test di secondo livello per B sono già condizionati da A, quindi ognuna delle quattro foglie può essere associata a un quadrante distinto, consentendo una classificazione corretta. Sebbene gli alberi decisionali siano in grado di riconoscere lo XOR, ciò non significa che sia facile trovare l'albero che lo fa. Ciò che rende XOR difficile da gestire è che non è possibile vedere i progressi verso una migliore classificazione, anche se si sceglie il nodo radice corretto. Nell'esempio precedente, la scelta di un nodo radice di "è $x > 0$?" non provoca alcun arricchimento apparente della purezza della classe su entrambi i lati. Il valore di questo test diventa evidente solo se guardiamo avanti di un altro livello, poiché il guadagno di informazioni è pari a zero.

L'euristica di costruzione di alberi decisionali avidi fallisce in problemi come lo XOR. Questo suggerisce il valore di procedure di costruzione di alberi più sofisticate e computazionalmente costose nei casi difficili, che guardano avanti come i programmi di scacchi del computer, valutando il valore della mossa p non ora, ma come appare diverse mosse dopo.

11.2.3 Ensemble di alberi decisionali

Esiste un numero enorme di possibili alberi decisionali che possono essere costruiti su qualsiasi serie di formazione S. Inoltre, ognuno di essi classificherà perfettamente *tutti gli* esempi di formazione, se continuiamo a raffinare fino a quando tutte le foglie sono pure. Questo suggerisce di costruire centinaia o addirittura migliaia di alberi diversi e di valutare un elemento della query q rispetto a ciascuno di essi per restituire una possibile etichetta. Lasciando che ogni albero esprima il proprio voto indipendente, otteniamo la certezza che l'etichetta più comunemente vista sarà quella giusta.

Per evitare il pensiero di gruppo, è necessario che gli alberi siano diversi. L'utilizzo ripetuto di una procedura di costruzione deterministica che trova l'albero migliore è inutile, perché saranno tutti identici. Sarebbe meglio selezionare in modo casuale una nuova dimensione di divisione in ogni nodo dell'albero, e poi trovare la migliore soglia possibile per questa variabile per definire il predicato.

Ma anche con una selezione casuale delle dimensioni, gli alberi risultanti sono spesso altamente correlati. Un approccio migliore è il *bagging*, che consiste nel costruire i migliori alberi possibili su sottoinsiemi casuali relativamente piccoli di elementi. Se fatto correttamente, gli alberi risultanti dovrebbero essere relativamente indipendenti l'uno dall'altro, fornendo una diversità di classificatori con cui lavorare, facilitando la saggezza delle folle.

L'utilizzo di ensemble di alberi decisionali presenta un altro vantaggio oltre alla robustezza. Il grado di consenso tra gli alberi offre una misura di fiducia per qualsiasi decisione di classificazione. C'è una grande differenza nella presenza dell'etichetta di maggioranza in 501 alberi su 1000 rispetto a 947 alberi.

Questa frazione può essere interpretata come una probabilità, ma ancora meglio potrebbe essere inserire questo numero nella regressione logistica per una misura di fiducia meglio motivata. Supponendo di avere un problema di classificazione binaria, lasciamo che f_i denoti la

frazione di alberi che scelgono la classe C_1 sul vettore di input X_i. Eseguire l'intero set di formazione attraverso l'ensemble di alberi decisionali. Ora definiamo un problema di regressione logistica dove f_i è la variabile di ingresso e la classe di X_i la variabile di uscita. La funzione logit risultante determinerà un livello di fiducia appropriato, per qualsiasi frazione di accordo osservata.

11.3 Boosting e Apprendimento Ensemble

L'idea di aggregare un gran numero di "predittori" rumorosi in un classificatore più forte si applica sia agli algoritmi che alle folle. Spesso accade che molte caratteristiche diverse siano tutte debolmente correlate con la variabile dipendente. Qual è quindi il modo migliore per combinarle in un classificatore più forte?

11.3.1 Votare con i classificatori

L'apprendimento in ensemble è la strategia di combinare molti classificatori diversi in un'unica unità predittiva. L'approccio naive Bayes della Sezione 11.1 ha un po' questo sapore, perché utilizza ogni caratteristica come un classificatore separato relativamente debole, poi li moltiplica insieme. La regressione lineare/logistica ha un'interpretazione simile, in quanto assegna un peso a ciascuna caratteristica per massimizzare il potere predittivo dell'insieme.

Ma più in generale, l'apprendimento in ensemble ruota attorno all'idea del *voto*. Abbiamo visto che gli alberi decisionali possono essere più potenti in aggregato, costruendo centinaia o migliaia di essi su sottoinsiemi

casuali di esempi. La saggezza delle folle deriva dal trionfo della diversità di pensiero rispetto all'individuo con maggiore competenza.

La democrazia si basa sul principio "un uomo, un voto". Il suo giudizio educato e ragionato sul miglior corso d'azione conta tanto quanto il voto di quell'idiota che parla a voce alta in fondo al corridoio. La democrazia ha senso in termini di dinamiche della società: le decisioni condivise in genere riguardano l'idiota tanto quanto lei, quindi l'uguaglianza impone che tutte le persone abbiano la stessa voce in .

Ma lo stesso argomento non si applica ai classificatori. Il più naturale di utilizzare più classificatori dà a ciascuno un voto e prende l'etichetta della maggioranza. Ma perché ogni classificatore dovrebbe avere lo stesso voto?

L'esempio consiste in cinque votanti, ognuno dei quali classifica cinque articoli. Tutti i votanti sono abbastanza bravi, ognuno ottiene il 60% di errori, ad eccezione di v_1, che ha ottenuto l'80%. Tuttavia, l'opzione di maggioranza non si dimostra migliore del peggior classificatore individuale, con il 60%. Ma si ottiene un classificatore perfetto se eliminiamo i votanti v_4 e v_5 e pesiamo i rimanenti in modo uguale. Ciò che rende preziosi v_2 e v_3 non è la loro precisione complessiva, ma le loro prestazioni sui problemi più difficili (D e soprattutto E).

Sembrano esserci tre modi principali per assegnare i pesi ai classificatori/votanti. Il più semplice potrebbe essere quello di attribuire un peso maggiore ai voti dei classificatori che hanno: La ponderazione uniforme dei voti non produce sempre il miglior classificatore possibile, anche quando i votanti sono ugualmente precisi, perché alcune istanze del problema sono più difficili di altre (qui, D ed E). Il simbolo "*" indica che il votante in questione ha classificato correttamente l'elemento in questione.si è dimostrato accurato in passato, forse assegnando a v_i il peso moltiplicativo dove t_i è il numero di volte che v_i è stato classificato correttamente e T= t_i. Un secondo approccio potrebbe essere quello di utilizzare la regressione lineare/logistica per trovare i migliori pesi possibili. In un problema di classificazione binaria, le due classi sarebbero indicate come 0 e 1, rispettivamente. I risultati 0-1 di ciascun classificatore possono essere utilizzati come caratteristica per prevedere il valore effettivo della classe. Questa formulazione troverebbe pesi non uniformi che favoriscono i classificatori correlati con le risposte corrette, ma non cerca esplicitamente di massimizzare il numero di classificazioni corrette.

11.3.2 Algoritmi di boosting

La terza idea è il *boosting*. Il punto chiave è pesare gli esempi in base a quanto è difficile azzeccarli, e premiare i classificatori in base al peso degli esempi azzeccati, non solo al numero.

Per impostare i pesi del classificatore, regoleremo i pesi degli esempi di formazione. Gli esempi di addestramento facili saranno classificati correttamente dalla maggior parte dei classificatori: premiamo maggiormente i classificatori per aver azzeccato i casi difficili.

Un algoritmo di boosting rappresentativo è *AdaBoost*. Non sottolineeremo i dettagli in questa sede, in particolare le specifiche dei pesi aggiunti in ogni round. Presumiamo che il nostro classificatore sarà costruito come l'unione di classificatori non lineari della forma, cioè utilizzando caratteristiche soglie come classificatori. L'algoritmo procede in T round, per t= {0, . . . , T}. Inizialmente tutti gli esempi di addestramento (punti) devono avere lo stesso peso, quindi $w_{i,0}$= $1/n$ per tutti i punti x_1, . . . x_n. Consideriamo tutti i possibili classificatori di caratteristiche/soglia e identifichiamo il $f_t(x)$ che minimizza$_{\epsilon(t)}$, la somma dei pesi dei punti mal classificati. Il peso α_t del nuovo classificatore dipende da quanto è accurato sul punto corrente impostato. I pesi dei punti sono normalizzati in modo che w_i= 1, quindi ci deve essere sempre una $_{i=1}$ classificatore con errore.

Nel round successivo, i pesi dei punti mal classificati vengono incrementati per renderli più importanti.

Facciamo che $h_t(x_i)$ sia la classe (-1 o 1) prevista per x_i, e y_i la classe corretta o quel punto. Il segno di $h_t(x_i)$ y riflette se le classi sono d'accordo (positive) o in disaccordo (negative).

alberi decisionali, ciascuno con un solo nodo. I pesi assegnati da AdaBoost non sono uniformi, ma non sono così follemente distorti in questo caso particolare da comportarsi diversamente da un classificatore di maggioranza. Osservare il confine decisionale non lineare, risultante dalla natura discreta dei test/alberi decisionali con soglia. Il boosting è particolarmente prezioso quando viene applicato agli alberi decisionali come classificatori elemen- tari. Il popolare approccio *gradient boosted decision trees* (GBDT) inizia tipicamente con un universo di piccoli alberi, con forse quattro-dieci nodi ciascuno. Questi alberi codificano ciascuno una logica abbastanza semplice da non adattarsi eccessivamente ai dati. I pesi relativi assegnati a ciascuno di questi alberi derivano da una procedura di addestramento, che cerca di adattare gli errori dei cicli precedenti (residui) e aumenta i pesi degli alberi che hanno classificato correttamente i dati. esempi più difficili. Il boosting si sforza di classificare correttamente ogni istanza di formazione, il che significa che si sforza particolarmente di classificare le istanze più difficili. Un adagio dice che "i casi difficili fanno cattiva legge", suggerendo che i casi difficili da decidere sono dei cattivi precedenti per le analisi successive. Questo è un argomento importante contro il boosting, perché il metodo sembra incline all'overfitting, anche se in genere si comporta bene nella pratica. Il pericolo di overfitting è particolarmente grave quando i dati di formazione non sono un gold standard perfetto. Le annotazioni di classe umane sono spesso soggettive e incoerenti, il che porta il boosting ad amplificare il rumore a scapito del segnale. I migliori algoritmi di boosting affronteranno l'overfitting attraverso la regolarizzazione. L'obiettivo sarà quello di ridurre al minimo il numero di coefficienti non nulli e di evitare coefficienti di grandi dimensioni, che danno troppa fiducia a un solo classificatore dell'ensemble.

11.4 Macchine vettoriali di supporto

Le macchine vettoriali di supporto (SVM) sono un modo importante per costruire classificatori non lineari. Possono essere considerate come un parente della regressione logistica, che cercava di la linea/piano l che separa meglio i punti con due classi di etichette. La regressione logistica ha assegnato a un punto di query q la sua etichetta di classe, a seconda che q si trovasse al di sopra o al di sotto di questa linea l. Inoltre, ha utilizzato la funzione logit per trasformare la distanza da q a l probabilità che q appartenga alla classe identificata. La considerazione di ottimizzazione nella regressione logistica comportava laminimizzazione della somma delle probabilità di misclassificazione su tutti i punti. Al contrario, le macchine a vettori di supporto lavorano cercando i separatori *lineari* a margine massimo tra le due classi. Questa linea cerca di massimizzare la distanza d dal punto di formazione più vicino, il margine massimo di separazione tra rosso e blu. Si tratta di un obiettivo naturale nella costruzione di un confine decisionale tra due classi, in quanto più grande è il margine, più lontani sono i nostri punti di addestramento dall'essere classificati in modo errato. Il classificatore con il margine massimo dovrebbe essere il separatore più robusto tra le due classi. Ci sono diverse proprietà che aiutano a definire il margine massimo di separazione tra gli insiemi di punti rossi e blu:

- La linea ottimale deve trovarsi nel punto medio del canale, a una distanza d sia dal punto rosso più vicino che dal punto blu più vicino. Se così non fosse, potremmo spostare la linea fino a che non biforcasse questo canale, allargando così il margine nel processo.

- Il canale di separazione effettivo è definito dal suo contatto con un piccolo numero di punti rossi e blu, dove "un piccolo numero" significa al massimo due volte il numero di dimensioni dei punti, per insiemi di punti ben educati, evitando d+ 1 punti che giacciono su qualsiasi faccia d-dimensionale. Questo è diverso da quello che avviene con la regressione logistica, dove *tutti* i punti contribuiscono all'adattamento del la migliore posizione della linea. Questi punti di contatto sono i *vettori di sostegno* che definisce il canale.
- I punti all'interno della carena convessa dei punti rossi o blu non hanno assolutamente alcun effetto sul separatore del margine massimo, poiché abbiamo bisogno che tutti i punti dello stesso colore si trovino

sullo stesso lato del confine. Possiamo eliminare questi punti interni o spostarli, ma il separatore del margine massimo non cambierà fino a quando uno dei punti non lascerà la carena ed entrerà nella striscia di separazione.

- Non è sempre possibile separare perfettamente il rosso dal blu utilizzando una linea retta. Immagini un punto blu che si trova da qualche parte all'interno dello scafo convesso dei punti rossi. Non c'è modo di separare questo punto blu dal rosso utilizzando solo una linea.

La regressione logistica e le macchine vettoriali di supporto producono entrambe linee di separazione tra gli insiemi di punti. Queste sono ottimizzate per criteri diversi, e quindi possono essere diverse. La regressione logistica cerca il separatore che massimizza la fiducia totale nella nostra classificazione sommata su tutti i punti, mentre il separatore ad ampio margine di SVM fa il meglio che può con i punti più vicini tra gli insiemi.
Entrambi i metodi producono generalmente classificatori simili.

11.4.1 SVM lineari

Queste proprietà definiscono l'ottimizzazione delle *macchine vettoriali di supporto lineari*. La linea/piano di separazione, come qualsiasi altra linea/piano. per un vettore di coefficienti w punteggiato con un vettore di variabili di ingresso x. Il canale che separa le due classi sarà definito da due linee parallele a questo ed equidistanti su entrambi i lati. L'effettiva separazione geometrica tra le linee dipende da w, ossia $2/w$. Per intuizione, pensi//// alla pendenza in due dimensioni: queste linee saranno distanti 2 per le linee orizzontali, ma trascurabilmente distanti se sono quasi verticali. Questo canale di separazione deve essere privo di punti e separare i punti rossi da quelli blu. Quindi dobbiamo aggiungere dei vincoli. Per ogni punto rosso (classe 1) x_i, insistiamo sul fatto che $w - x$ mentre ogni punto blu (classe -1) x_i.
Se lasciamo che $-y_i$ [1, 1] denoti la classe di x_i, allora questi possono essere combinati per ottenere il problema di ottimizzazione. Questo può essere risolto utilizzando tecniche simili alla programmazione lineare. Si noti che il canale deve essere definito dai punti di contatto con i suoi confini. Questi vettori 'supportano' il canale, da cui il nome provocatorio di *macchine vettoriali supporto*. L'algoritmo di ottimizzazione di risolutori efficienti come LibLinear e LibSVM cerca tra i piccoli sottoinsiemi rilevanti di vettori di supporto che potenzialmente definiscono canali di separazione per trovare più ampio. Si noti che esistono criteri di ottimizzazione più generali per le SVM, che cercano la linea che definisce un canale ampio e penalizza (ma non vieta) i punti che vengono classificati in modo errato. Questo tipo di funzione a duplice obiettivo (rendere il canale ampio e sbagliare la classificazione di pochi punti) può essere considerata come una forma di regolarizzazione, con una costante di compromesso tra i due obiettivi. La ricerca per discesa del gradiente può essere utilizzata per risolvere questi problemi generali.

11.4.2 SVM non lineari

Le SVM definiscono un iperpiano che separa i punti delle due classi. I piani sono linee in dimensioni superiori, facilmente definibili utilizzando l'algebra lineare. Quindi, come può questo metodo lineare produrre un confine decisionale non lineare?

Affinché un dato insieme di punti abbia un margine massimo di separazione, i due colori devono prima essere linearmente separabili. Ma come abbiamo visto, questo non è sempre il caso. Consideriamo il caso patologico, dove il cluster di rosso è circondato da un gruppo ad anello di punti neri. Come potrebbe nascere una cosa del genere? Supponiamo di suddividere le destinazioni di viaggio in *viaggi di un giorno* o *viaggi lunghi*, a seconda che siano abbastanza vicine alla nostra posizione. La longitudine e la latitudine di ogni possibile destinazione produrranno dati con la stessa struttura.

L'idea chiave è che possiamo proiettare i nostri punti d-dimensionali in uno spazio dimensionale superiore, dove ci saranno più possibilità di . Per n punti rossi/blu lungo una linea in una dimensione, ci sono solo n 1 modi potenzialmente interessanti per, in particolare con un taglio tra lith e il (i+ 1)st punti. Ma questo aumenta

a circa n modi quando passiamo a due dimensioni, perché c'è più libertà di partizione quando aumentiamo la dimensionalità.

Se eleviamo abbastanza la dimensionalità di qualsiasi insieme di punti a due classi, ci sarà sempre una linea di separazione tra i punti rossi e neri. Infatti, se mettiamo gli n punti in n dimensioni attraverso una trasformazione ragionevole, saranno sempre linearmente separabili in un modo molto semplice. Per avere un'idea, pensiamo al caso speciale di due punti (uno rosso e uno blu) in due dimensioni: ovviamente ci deve essere una linea che li separa. Proiettando questo piano di separazione verso il basso spazio originale, si ottiene una forma di confine decisionale curvo, e quindi la non linearità delle SVM dipende esattamente dal modo in cui l'input è stato proiettato in uno spazio di dimensioni superiori.

Un modo simpatico per trasformare n punti in d dimensioni in n punti in n dimensioni potrebbe essere quella di rappresentare ogni punto con le sue distanze da tutti gli n punti di ingresso. In particolare, per ogni punto p_i possiamo creare un vettore v_i tale che la distanza da p_i a p_j. Il vettore di tali distanze dovrebbe servire come un potente insieme di caratteristiche per classificare qualsiasi nuovo punto q, dal momento che le distanze dai membri della classe effettiva dovrebbero essere piccole rispetto a quelle dell'altra classe. Questo spazio di caratteristiche è davvero potente e si può facilmente immaginare di scrivere una funzione per trasformare la matrice di caratteristiche originale in una nuova matrice di caratteristiche n n per la classificazione. Il problema in questo caso è lo spazio, perché il numero di punti di input n è solitamente molto più grande della dimensione d cui si trovano. Una trasformazione di questo tipo sarebbe fattibile solo per insiemi di punti piuttosto piccoli, ad esempio n 1000. Inoltre, lavorare con punti di dimensioni così elevate dovrebbe essere molto≤ costoso, dal momento che ogni singola valutazione della distanza richiede un tempo lineare per il numero di punti n, anziché per la dimensione dei dati d.

11.4.3 I gherigli

La magia delle SVM consiste nel fatto che questa matrice distanza-caratteristica non deve essere calcolata esplicitamente. l'ottimizzazione inerente alla ricerca del separatore di margine massimo esegue solo i prodotti di punti con altri punti e vettori. Quindi potremmo immaginare di eseguire l'espansione della distanza al volo, quando il punto associato viene utilizzato in un confronto. Quindi non ci sarebbe bisogno di precompilare la matrice di distanza: possiamo espandere i punti da d a n dimensioni come necessario, eseguire il calcolo della distanza e poi buttare via le espansioni.

Questo funzionerebbe per eliminare il collo di bottiglia dello spazio, ma pagheremmo comunque un prezzo pesante in termini di tempo di calcolo. La cosa davvero sorprendente è che esistono funzioni, chiamate *kernel*, che restituiscono quello che è essenzialmente il calcolo della distanza sul vettore più grande senza mai costruire il vettore più grande. L'utilizzo di SVM con i kernel ci dà la possibilità di trovare il miglior separatore su una varietà di funzioni non lineari senza costi aggiuntivi.

Le macchine vettoriali di supporto richiedono esperienza per essere utilizzate in modo efficace. Sono disponibili molte funzioni kernel diverse, oltre al kernel di distanza che ho presentato qui. Ognuna di esse presenta dei vantaggi su determinati set di dati, per cui è necessario armeggiare con le opzioni di strumenti come LibSVM per ottenere le migliori prestazioni. Funzionano meglio su set di dati di medie dimensioni, con migliaia ma non milioni di punti.

11.5 Gradi di supervisione

Esiste una distinzione naturale tra gli approcci di apprendimento automatico, basata sul grado e sulla natura della *supervisione* impiegata nella raccolta dei dati di formazione e di valutazione. Come ogni tassonomia, c'è un po' di sfumatura ai margini, il che rende un esercizio insoddisfacente cercare di etichettare esattamente ciò che un dato sistema sta e non sta facendo.

Tuttavia, come ogni *buona* tassonomia, fornisce una cornice per guidare il suo pensiero e suggerisce approcci che potrebbero portare a risultati migliori. I metodi discussi finora in questo capitolo presuppongono che ci vengano forniti dati di addestramento con etichette di classe o variabili target, lasciando il nostro compito di addestrare sistemi di classificazione o regressione. Ma arrivare al punto di avere dati etichettati è di solito la parte difficile. Gli algoritmi di apprendimento automatico in genere funzionano meglio quanto più dati si possono fornire loro, ma l'annotazione è spesso difficile e costosa. La modulazione del grado di supervisione offre un modo per aumentare il volume di dati. *L'apprendimento supervisionato* è il paradigma di base per i problemi di classificazione e di regressione. Ci vengono forniti vettori di caratteristiche x_i, ciascuno con un'etichetta di classe associata o un valore target y_i. Le annotazioni y_i rappresentano la supervisione, in genere derivata da un processo manuale che limita la quantità potenziale di dati di addestramento. In alcuni problemi, le annotazioni dei dati di addestramento provengono da osservazioni nell'interazione con il mondo, o almeno da una sua simulazione. Il programma AlphaGo di Google è stato il primo programma informatico a battere il campione mondiale di Go. Una funzione di valutazione della posizione è una funzione di punteggio che prende una posizione sulla tavola e calcola un numero che stima la forza. La funzione di valutazione della posizione di AlphaGo è stata addestrata su tutte le partite pubblicate da maestri umani, ma erano necessari molti più dati. La soluzione era, essenzialmente, costruire un valutatore di posizione allenandosi contro se stesso. La valutazione della posizione è sostanzialmente migliorata dalla ricerca guardando diverse mosse in avanti prima di richiamare la funzione di valutazione su ogni foglia. Cercare di prevedere il punteggio successivo alla ricerca senza la ricerca produce una funzione di valutazione più forte. E la generazione di questi dati di formazione è solo un risultato del calcolo: il programma gioca contro se stesso. Questa idea di apprendimento dall'ambiente si chiama *apprendimento per rinforzo*. Non può essere applicata ovunque, ma vale sempre la pena di cercare approcci intelligenti per generare dati di formazione annotati meccanicamente.

11.5.1 Apprendimento non supervisionato

I metodi non supervisionati cercano di trovare una struttura nei dati, fornendo etichette (cluster) o valori (classifiche) senza alcuno standard affidabile. Sono utilizzati al meglio per l'esplorazione, per dare un senso a un insieme di dati altrimenti non toccati da mani umane.

La madre di tutti i metodi di apprendimento non supervisionato è il *clustering*. Si noti che il clustering può essere utilizzato per fornire dati di formazione per la classificazione anche in assenza di etichette. Se presumiamo che i cl usters trovati rappresentino un fenomeno reale, possiamo utilizzare l'ID del cluster come etichetta per tutti gli elementi del cluster dato. Questi possono ora servire come dati di addestramento per costruire un classificatore per prevedere l'ID cluster. La previsione degli ID cluster può essere utile anche se questi concetti non hanno un nome associato, fornendo un'etichetta ragionevole per qualsiasi record di input q.

Argomento Modellazione

Un'altra classe importante di metodi non supervisionati è la *modellazione per argomenti*, tipicamente associata a documenti estratti da un determinato vocabolario. I documenti sono scritti su argomenti, di solito un mix di argomenti. Questo libro è suddiviso in capitoli, ognuno dei quali riguarda un argomento diverso, ma tocca anche temi che vanno dal baseball ai matrimoni. Ma cos'è un argomento? In genere, ogni argomento è associato a un particolare insieme di parole del vocabolario. Gli articoli sul baseball parlano di *battute*, *lanciatori*, *strikeout*, *basi* e *slugging*. *Sposato*, *fidanzato*, *sposo*, *sposa*, *amore* e *festeggiare* sono parole associate al tema del matrimonio. Alcune parole possono rappresentare più argomenti. Ad esempio, *l'amore* è associato anche al tennis e i *colpi* ai gangster.

Una volta che si dispone di un insieme di argomenti ($t_1, \ldots t_k$) e le parole che li definiscono, il problema di identificare gli argomenti specifici associati a un dato documento d sembra abbastanza semplice. Contiamo il numero di occorrenze di parole di d in comune con t_i e segnaliamo il successo quando questo è sufficientemente alto. Se si riceve un insieme di documenti etichettati manualmente con argomenti, sembra ragionevole contare

la frequenza di ogni parola su ogni classe di argomenti, per costruire l'elenco delle parole più fortemente associate a ciascun argomento. Ma tutto questo è molto supervisionato. *La modellazione topica* è un approccio non supervisionato, che infonde gli argomenti e gli elenchi di parole da zero, solo in presenza di documenti non etichettati. Possiamo rappresentare questi testi con una matrice di frequenza, dove w è la dimensione del vocabolario e d il numero di documenti e $F[i, j]$ riflette quante volte il lavoro i appare nel documento j. Una tale fattorizzazione rappresenterebbe una forma di apprendimento completamente non supervisionata, ad eccezione della specificazione del numero desiderato di argomenti t. Sembra un processo complicato costruire una tale fattorizzazione approssimativa, ma ci sono diversi approcci per cercare di farlo. Forse il metodo più popolare per la modellazione dei temi è un approccio chiamato *allocazione latente di Dirichlet* (LDA), che produce un insieme simile di matrici W e D, anche se non strettamente prodotto dalla fattorizzazione. L'algoritmo LDA ha definito questi argomenti in modo non supervisionato, assegnando a ciascuna parola dei pesi per il suo contributo a ciascun argomento. I risultati sono generalmente efficaci: il concetto di ogni argomento emerge dalle sue parole più importanti (a destra). E la distribuzione delle parole all'interno di ogni libro può essere facilmente suddivisa tra i tre argomenti latenti (a sinistra). Si noti che questa mentalità di fattorizzazione può essere applicata al di là dei documenti, a qualsiasi matrice di caratteristiche F. Gli approcci di decomposizione della matrice di cui abbiamo parlato in precedenza, come la decomposizione del valore singolare e l'analisi delle componenti principali, sono ugualmente non supervisionati, inducendo la struttura inerente ai set di dati senza vantaggio.

11.5.2 Apprendimento semi-supervisionato

Il divario tra l'apprendimento supervisionato e quello non supervisionato viene colmato dai metodi di *apprendimento semi-supervisionato*, che amplificano piccole quantità di dati di formazione etichettati in un numero maggiore. Trasformare piccoli numeri di esempi in numeri più grandi è spesso chiamato *bootstrapping*, dal concetto di "tirarsi su dalle proprie gambe". Gli approcci semi-supervisionati personificano l'astuzia che deve essere impiegata per costruire serie di addestramenti sostanziali.

Supponiamo che ci venga fornito un piccolo numero di esempi etichettati come coppie, affiancati da un gran numero di input x_i di etichetta sconosciuta. Invece di costruire direttamente il nostro modello dal set di formazione, possiamo utilizzarlo per classificare la massa di istanze non etichettate. Forse utilizziamo un approccio di prossimità per classificare queste istanze sconosciute, o uno qualsiasi degli altri approcci che abbiamo discusso qui. Ma una volta classificate, assumiamo che le etichette siano corrette e ci riqualifichiamo sull'insieme più grande.

Questi approcci traggono grande vantaggio dall'avere un set di valutazione affidabile. Dobbiamo stabilire che il modello addestrato sugli esempi bootstrap ha prestazioni migliori rispetto a quello addestrato su ciò che abbiamo iniziato. L'aggiunta di miliardi di esempi di addestramento è inutile se le etichette non sono affidabili.

Ci sono altri modi per generare dati di addestramento senza annotazioni. Spesso sembra più facile trovare esempi positivi che negativi. Si consideri la problema di addestramento di un correttore grammaticale, nel senso che distingue i pezzi di scrittura corretti da quelli malformati. È facile ottenere grandi quantità di esempi di inglese corretto: tutto ciò che viene pubblicato nei libri e nei giornali è generalmente considerato buono. Ma sembra più difficile ottenere un grande corpora di scrittura scorretta. Tuttavia, possiamo osservare che l'aggiunta, l'eliminazione o la sostituzione casuale di parole arbitrarie a qualsiasi testo lo peggiora quasi sempre.[18] Etichettando tutti i testi pubblicati come corretti e tutte le perturbazioni casuali come errate, possiamo creare un set di addestramento tanto grande quanto desideriamo, senza dover assumere qualcuno per annotarlo.

Come possiamo valutare un classificatore di questo tipo? Di solito è possibile ottenere un numero sufficiente di dati annotati autentici per la valutazione, perché i dati necessari per la valutazione sono in genere molto più piccoli di quelli per l'addestramento. Possiamo anche utilizzare il nostro classificatore per suggerire cosa annotare. Gli esempi più preziosi da sottoporre all'annotatore sono quelli su cui il nostro classificatore commette errori: le frasi pubblicate contrassegnate come errate o le mutazioni casuali che superano il test sono degne di essere sottoposte a un giudice umano.

11.5.3 Ingegneria delle caratteristiche

L*'ingegneria delle caratteristiche* è l'arte raffinata di applicare la conoscenza del dominio per facilitare il lavoro degli algoritmi di apprendimento automatico. Nel contesto della nostra tassonomia, l'ingegneria delle caratteristiche può essere considerata una parte importante dell'apprendimento supervisionato, dove la supervisione si applica ai vettori di caratteristiche x_i invece che alle annotazioni target associate y_i.

È importante garantire che le caratteristiche siano presentate ai modelli in modo che il modello possa correttamente. Incorporare la conoscenza specifica dell'applicazione nei dati, invece di impararla, sembra un imbroglio, per i dilettanti. Ma i professionisti capiscono che ci sono cose che non possono essere apprese facilmente e che quindi è meglio inserire esplicitamente nel set di caratteristiche.

Consideri un modello per prezzare l'arte alle aste. Le case d'asta guadagnano i loro soldi addebitando una commissione all'offerente vincente, oltre a quanto pagano al proprietario. Le case d'asta applicano tariffe diverse, ma possono arrivare a un conto sostanzioso. Poiché il costo totale per il vincitore è diviso tra prezzo d'acquisto e commissione, le commissioni più alte possono ridurre il prezzo d'acquisto, tagliando quanto l'offerente può permettersi di pagare al proprietario.

Quindi, come si può rappresentare il prezzo della commissione in un modello di prezzo dell'arte? Mi vengono in mente almeno tre approcci diversi, alcuni dei quali possono avere esiti disastrosi:

- *Specificare la percentuale di commissione come caratteristica* : Rappresentare il taglio della casa (ad esempio il 10%) come colonna nel set di caratteristiche potrebbe non essere utilizzabile in un modello lineare. Il colpo subito dall'offerente è il prodotto dell'aliquota fiscale e del prezzo finale. Ha un effetto moltiplicativo, non additivo, e quindi non può essere sfruttato in modo significativo se la fascia di prezzo dell'arte va da 100 a 1.000.000 di dollari.
- *Includere la commissione effettivamente pagata come caratteristica* : Imbroglione. . . Se si include la commissione *pagata* alla fine come caratteristica, si inquinano le caratteristiche con dati non noti al momento dell'asta. Infatti, se tutti i dipinti dovessero essere soggetti ad un'imposta del 10% e l'imposta pagata fosse una caratteristica, un modello perfettamente accurato (e completamente inutile) prevederebbe il prezzo pari a dieci volte l'imposta pagata!

- *Impostate la variabile target di regressione come l'importo totale pagato*: Poiché i tassi di commissione della casa e le spese aggiuntive sono noti all'acquirente prima che faccia l'offerta, la variabile target giusta dovrebbe essere l' totale pagato. Qualsiasi previsione data del prezzo di acquisto totale può essere scomposta in seguito in prezzo di acquisto, commissioni e tasse, secondo le regole della casa.

L'ingegneria delle caratteristiche può essere considerata come una versione dipendente dal dominio della pulizia dei dati, quindi le tecniche discusse nella Sezione 3.3 si applicano tutte qui. Le più importanti saranno esaminate qui nel contesto, ora che abbiamo finalmente raggiunto il punto di costruire effettivamente modelli guidati dai dati:

- *Punteggi Z e normalizzazione*: I valori distribuiti in modo normale su intervalli numerici comparabili costituiscono le caratteristiche migliori, in generale. Per rendere gli intervalli comparabili, trasformi i valori in punteggi Z, sottraendo la media e dividendo per la deviazione standard, $Z = (x - \mu)/\sigma$. Per rendere più normale una variabile a legge di potenza, sostituisca x nel set di caratteristiche con $\log x$.
- *Imputare i valori mancanti*: Si assicuri che non ci siano valori mancanti nei suoi dati e, in tal caso, li sostituisca con un'ipotesi o una stima significativa. Registrare che il peso di una persona è uguale a 1 è un modo semplice per rovinare qualsiasi modello. Il metodo

-

di imputazione più semplice sostituisce ogni valore mancante con la media della colonna in questione e in genere è sufficiente, ma i metodi più efficaci addestrano un modello per prevedere il valore mancante in base alle altre variabili presenti nel record. Riveda la Sezione 3.3.3 per i dettagli.

- *Riduzione delle dimensioni*: Ricordiamo che *la regolarizzazione* è un modo per costringere i modelli a scartare le caratteristiche irrilevanti per evitare l'overfitting. È ancora più efficace eliminare le caratteristiche irrilevanti prima di adattare i modelli, spostandole dal set di dati. Quando una caratteristica x è probabilmente irrilevante per il suo modello? La scarsa correlazione con la variabile target y e l'assenza di una ragione qualitativa che possa spiegare perché x *potrebbe avere* un impatto su y sono entrambi indicatori eccellenti.

  Le tecniche di riduzione della dimensione, come la decomposizione del valore singolare, sono un modo eccellente per ridurre vettori di caratteristiche di grandi dimensioni in rappresentazioni più potenti e concise. I vantaggi includono tempi di formazione più rapidi, meno overfitting e riduzione del rumore dalle osservazioni.
- *Inclusione esplicita di combinazioni non lineari*: Alcuni prodotti o rapporti di variabili caratteristiche hanno interpretazioni naturali nel contesto. L'area o il volume sono prodotti di lunghezza, larghezza e altezza, ma non possono far parte di alcun modello lineare, a meno che non vengano esplicitamente inseriti come colonne nel ma- trix delle caratteristiche. I totali aggregati, come i punti segnati in carriera nello sport o i dollari totali guadagnati con lo stipendio, sono solitamente incomparabili tra elementi di età o durata diversa. Ma convertendo i totali in tassi (come i punti per partita giocata o i dollari per ora) si ottengono di solito caratteristiche più significative.

  La definizione di questi prodotti e rapporti richiede informazioni specifiche del dominio e un'attenta riflessione durante il processo di ingegnerizzazione delle caratteristiche. È molto più probabile che lei conosca le combinazioni giuste, piuttosto che il suo classificatore non lineare le trovi da solo.

La differenza tra un buon modello e un cattivo modello di solito si riduce alla qualità dell'ingegneria delle caratteristiche. Gli algoritmi avanzati di apprendimento automatico sono affascinanti, ma è la preparazione dei dati che produce i risultati.

11.6 Apprendimento profondo

Gli algoritmi di apprendimento automatico che abbiamo studiato qui non scalano bene su serie di dati *enormi*, per diversi motivi. I modelli come la regressione lineare in genere hanno relativamente pochi parametri, ad esempio un coefficiente per colonna, e quindi non possono beneficiare di un numero enorme di esempi di addestramento. Se i dati hanno un buon adattamento lineare, sarà possibile trovarlo con un piccolo insieme di dati. E se non è così, beh, non voleva comunque trovarlo.

L'apprendimento profondo è uno sviluppo recente incredibilmente entusiasmante nell' automatico. Si basa sulle *reti neurali*, un approccio popolare degli anni '80 che poi è passato sostanzialmente di moda. Ma negli ultimi cinque anni è successo qualcosa e improvvisamente le reti multistrato (profonde) hanno iniziato a superare selvaggiamente gli approcci tradizionali su problemi classici di computer vision e di elaborazione del linguaggio naturale.

Il motivo esatto per cui questo è accaduto rimane un po' un mistero. Non sembra che ci sia stata una svolta algoritmica fondamentale, quanto piuttosto che il volume dei dati e la velocità di calcolo abbiano superato una soglia in cui la capacità di sfruttare enormi quantità di dati di formazione ha prevalso su metodi più efficaci nel gestire una risorsa scarsa. Ma l'infrastruttura si sta sviluppando rapidamente per sfruttare questo vantaggio: i

nuovi framework software open source come *Tensor Flow* di Google rendono facile specificare le architetture di rete ai processori speciali progettati per accelerare la formazione di ordini di grandezza.

Ciò che distingue l'apprendimento profondo da altri approcci è che in genere evita l'ingegneria delle caratteristiche. Ogni strato di una rete neurale generalmente accetta come input l'output dello strato precedente, producendo caratteristiche di livello progressivamente più elevato man mano che si sale verso la parte superiore della rete. Ciò serve a definire una gerarchia di comprensione dall'input grezzo al risultato finale, e in effetti il penultimo livello di una rete progettata per un compito spesso fornisce utili caratteristiche di alto livello per compiti correlati.

Perché le reti neurali hanno così tanto successo? Nessuno lo sa veramente. Ci sono indicazioni che per molti compiti il peso di queste reti non è davvero necessario; che ciò che stanno facendo sarà fatto alla fine con metodi meno opachi. Le reti neurali sembrano funzionare con l'overfitting, trovando un modo per utilizzare milioni di esempi per adattarsi a milioni di parametri. Tuttavia, in genere riescono a evitare il comportamento peggiore dell'overfitting, forse utilizzando modi meno precisi per codificare la conoscenza. Un sistema che memorizza esplicitamente lunghe stringhe di testo da suddividere su richiesta sembrerà fragile e sovraadattato, mentre un sistema che rappresenta tali frasi in modo meno rigido potrebbe essere più flessibile e generalizzabile.

11.6.1 Reti e profondità

Ogni nodo x rappresenta un'unità computazionale, che calcola il valore di una determinata funzione simi- lizzata $f(x)$ su tutti gli ingressi. Per ora, possiamo considerarlo come un semplice sommatore che somma tutti gli ingressi e poi emette la somma. Ogni bordo diretto (x, y) collega l'uscita del nodo x all'ingresso di un nodo y più in alto nella rete. Inoltre, ogni bordo ha un coefficiente moltiplicatore associato $w_{x,y}$. Il valore effettivamente passato a y è $w_{x,y} f(x)$, il che significa che il nodo y calcola una somma ponderata dei suoi ingressi.

un insieme di variabili di ingresso, i cui valori cambiano ogni volta che chiediamo alla rete di fare una previsione. Pensi a questo come interfaccia della rete. I collegamenti da qui al livello successivo propagano questo valore di ingresso a tutti i nodi che lo calcoleranno. Sul lato destro si trovano una o più variabili di uscita, che presentano i risultati finali di questo calcolo. Tra questi strati di ingresso e di uscita si trovano *strati nascosti* di nodi. Dati i pesi di tutti i coefficienti, la struttura della rete e i valori delle variabili di ingresso, il calcolo è semplice: calcola i valori del livello più basso della rete, li propaga in avanti e ripete dal livello successivo fino a raggiungere la cima.

Apprendere la rete significa impostare i pesi dei parametri dei coefficienti $w_{x,y}$. Più bordi ci sono, più parametri dobbiamo imparare. In , l'apprendimento significa analizzare un corpus di formazione di coppie (x_i, y_i) e regolare i pesi dei parametri dei bordi in modo che i nodi di uscita generino qualcosa di simile a y_i quando viene alimentato l'input x_i. **Profondità della rete**

La profondità della rete dovrebbe, in un certo senso, corrispondere alla gerarchia concettuale associata agli oggetti da modellare. L'immagine che dovremmo avere è quella dell'input che viene successivamente trasformato, filtrato, ridotto e trasformato in una forma sempre migliore man mano che si sale nella rete. In generale il numero di nodi dovrebbe diminuire progressivamente man mano che si sale verso i livelli superiori. Possiamo pensare che ogni livello fornisca un livello di astrazione. Consideriamo un problema di classificazione di immagini, magari per decidere se l'immagine contiene o meno la foto di un gatto. Pensando in termini di livelli successivi di astrazione, si può dire che le immagini sono composte da pixel, zone vicine, bordi, texture, regioni, oggetti semplici, oggetti composti e scene. Si tratta di un'argomentazione secondo cui almeno otto livelli di astrazione potrebbero potenzialmente essere riconoscibili e utilizzabili dalle reti sulle immagini. Gerarchie simili esistono nella comprensione dei documenti (caratteri, parole, frasi, frasi, paragrafi, sezioni, documenti) e qualsiasi altro artefatto di complessità simile. In effetti, le reti di apprendimento profondo addestrate per compiti specifici possono produrre caratteristiche di valore per uso generale, esponendo le uscite dei

livelli inferiori della rete come caratteristiche potenti per i classificatori convenzionali. Ad , *Imagenet* è una rete popolare per il riconoscimento di oggetti dalle immagini. Uno strato di alto livello di 1000 nodi misura la fiducia che l'immagine contenga oggetti di ciascuno dei 1000 tipi diversi. I modelli di quali oggetti si illuminano in che misura sono generalmente utili per altri compiti, come la misurazione della somiglianza delle immagini.

Non imponiamo una visione reale di ciò che ciascuno di questi livelli dovrebbe rappresentare, ma solo di collegarli in modo che esista il potenziale per riconoscere tale complessità. I patch di vicinato sono funzioni di piccoli gruppi di pixel collegati, mentre le regioni saranno costituite da piccoli numeri di patch collegati. Un certo senso di ciò che stiamo cercando di riconoscere va nella progettazione di questa topologia, ma la rete fa ciò che ritiene di dover fare durante l'addestramento per minimizzare l'errore di addestramento, o la *perdita*. Lo svantaggio delle reti più profonde è che diventano più difficili da addestrare.

più diventano grandi e profondi. Ogni nuovo strato aggiunge una nuova serie di parametri di peso dei bordi, aumentando il rischio di overfitting. Attribuire correttamente l'effetto degli errori di previsione ai pesi dei bordi diventa sempre più difficile, man mano che cresce il numero di strati intermedi tra il bordo e il risultato osservato. Tuttavia, sono state addestrate con successo reti con oltre dieci strati e milioni di parametri e, in generalele prestazioni di riconoscimento aumentano con la complessità della rete.

Le reti diventano anche più costose dal punto di vista computazionale per fare previsioni con l'aumentare della profondità, poiché il calcolo richiede un tempo lineare per il numero di bordi della rete. Questo non è terribile, soprattutto perché tutti i nodi di un dato livello possono essere valutati in parallelo su più core per ridurre il tempo di predizione. Il tempo di addestramento è il punto in cui si verificano i veri colli di bottiglia computazionali.

Non linearità L'immagine del riconoscimento di livelli crescenti di astrazione lungo gli strati nascosti di una rete certamente convincente. Tuttavia, è lecito chiedersi se sia reale. Gli strati aggiuntivi in una rete ci danno *davvero* una potenza computazionale supplementare per fare cose che non possiamo fare con meno?

Mostra reti di addizione costruite con due e tre strati di nodi, rispettivamente, ma entrambe eseguono esattamente la stessa funzione su tutti gli ingressi. Ciò suggerisce che lo strato aggiuntivo non era necessario, tranne forse per ridurre il vincolo ingegneristico del grado del nodo, il numero di bordi che entrano come input.

Ciò che dimostra in realtà è che abbiamo bisogno di funzioni *di accesso ai* nodi $\varphi(v)$ più complicate e non lineari per sfruttare la profondità. Le funzioni non lineari non possono essere composte nello stesso modo in cui l'addizione può essere composta per ottenere l'addizione. Questa funzione di attivazione non lineare $\varphi(v_i)$ opera tipicamente su una somma ponderata degli ingressi x.

Qui b è una costante per il nodo dato, forse da apprendere durante la formazione. Si chiama *bias* del nodo perché definisce l'attivazione in assenza altri input. Il fatto che il calcolo dei valori di uscita del livello l comporti l'applicazione della funzione di attivazione φ alle somme ponderate dei valori del livello- l 1 ha un'implicazione importante sulle prestazioni. In particolare, la valutazione della rete neurale comporta fondamentalmente solo una moltiplicazione matriciale per livello, dove le somme ponderate si ottengono moltiplicando una matrice di pesi W per un vettore di uscita V_{l-1}. Ogni elemento del vettore V_l 1 risultante viene poi colpito con la funzione φ per preparare i valori di uscita per quel livello. Le librerie veloci per la moltiplicazione di matrici possono eseguire il cuore di questa valutazione in modo molto efficiente.

Una serie di interessanti funzioni di attivazione non lineari sono state impiegate nelle reti di costruzione.

- *Logit*: Abbiamo già incontrato la *funzione logistica* o logit, nella nostra discussione sulla regressione logistica per la classificazione.

  Questa unità ha la proprietà che l'uscita è vincolata all'intervallo [0,1], dove $f(0) = 1/2$. Inoltre, la funzione è differenziabile, quindi si può utilizzare la retropropagazione per addestrare la rete risultante. Inoltre, la funzione è differenziabile, quindi si può utilizzare la propagazione all'indietro per addestrare la rete risultante.

- *Unità lineari rettificate (ReLU)*: Un *raddrizzatore* o un diodo in un circuito elettrico lascia scorrere la corrente in una sola direzione. La sua funzione di risposta $f(x)$ è lineare quando x è positivo, ma nulla quando x è negativo.

Questo punto di snodo a $x = 0$ è sufficiente per eliminare la linearità dalla funzione e fornisce un modo naturale per spegnere l'unità guidandola in negativo. La funzione ReLU rimane differenziabile, ma ha una risposta molto diversa rispetto al logit, aumentando monotonicamente e non avendo limiti su un lato.

In generale, l'aggiunta di uno strato nascosto aggiunge una potenza considerevole alla rete, mentre gli strati aggiuntivi soffrono di rendimenti decrescenti. La teoria mostra che le reti senza strati nascosti hanno la capacità di riconoscere classi linearmente separabili, ma ci siamo rivolti alle reti neurali per costruire classificatori più potenti.

11.6.2 Retropropagazione

La retropropagazione è la procedura di addestramento principale per le reti neurali, che raggiunge risultati molto impressionanti adattando un gran numero di parametri in modo crescente su grandi serie di addestramento. Ricorda molto la discesa stocastica del gradiente.

Il nostro problema di base è questo. Ci viene data una rete neurale con valori preliminari per ogni parametro w^l, cioè il moltiplicatore che l'uscita del nodo. Ci viene dato anche un set di formazione prima di essere aggiunto al nodo v costituito da n coppie di valori vettoriali di ingresso. Nel nostro modello di rete, il vettore x_i rappresenta i valori da assegnare allo strato di ingresso v^1 e y_i la risposta desiderata dallo strato di uscita v_l. Valutando la rete attuale su x_i, si otterrà un vettore di uscita v_l. L'errore E_l della rete allo strato l può essere misurato.

Vorremmo migliorare i valori dei coefficienti di peso w^l in modo che siano prevedere meglio y_i e minimizzare E_l. L'equazione di cui sopra definisce la perdita E_l come una funzione dei coefficienti di peso, poiché i valori di ingresso del precedente strato è fisso. Come nella discesa stocastica del gradiente, il valore corrente del $w_i^l{}_j$definisce un punto p su questa superficie di errore, e la derivata di E_l in questo punto definisce la direzione di discesa più ripida che riduce gli errori. Percorrendo una distanza d in questa direzione definita dalla dimensione del passo o dal *tasso di apprendimento* corrente, si ottengono valori aggiornati dei coefficienti, il cui v_l fa un lavoro migliore nel predire y_a da x_a.

Ma questo cambia solo i coefficienti nel livello di uscita. Per passare allo strato precedente, si noti che la valutazione precedente della rete ha fornito un'uscita per ciascuno di questi nodi in funzione dell'ingresso. Per ripetere la stessa procedura di addestramento, abbiamo bisogno di un valore target per ogni nodo del livello l - 1, per svolgere il ruolo di y_a dal nostro esempio di addestramento. Dato y_a e i nuovi pesi per calcolare v^l, possiamo calcolare i valori per le uscite di questi strati che prevedano perfettamente y_i. Con questi obiettivi, possiamo modificare i pesi dei coefficienti a questo livello e continuare la propagazione all'indietro fino a raggiungere la parte inferiore della rete, al livello di ingresso.

11.6.3 Incorporamenti di parole e grafici

C'è una particolare applicazione non supervisionata della tecnologia di apprendimento profondo che ho trovato facilmente applicabile a diversi problemi di interesse. Questo ha l'ulteriore vantaggio di essere accessibile a un pubblico più ampio che non ha familiarità con le reti neurali. *Le incorporazioni di parole* sono rappresentazioni distribuite di ciò che le parole effettivamente *significano* o *fanno*.

Ogni parola è indicata da un singolo punto nello spazio, ad esempio, di 100 dimensioni, in modo che le parole che svolgono ruoli simili tendano ad essere rappresentate da punti vicini. Il valore principale delle incorporazioni di parole è come caratteristiche generali da applicare in applicazioni specifiche di apprendimento automatico. Riconsideriamo il problema di distinguere lo spam dai messaggi e - mail

significativi. Nella tradizionale rappresentazione a sacchetto di parole, ogni messaggio potrebbe essere rappresentato come un vettore rado b, dove $b[i]$ potrebbe riportare il numero di volte in cui la parola del vocabolario w_i compare nel messaggio. Una dimensione ragionevole del vocabolario v per l'inglese è di 100.000 parole, trasformando b in una rappresentazione orribile di 100.000 dimensioni che non cattura la tra termini correlati. Le rappresentazioni vettoriali delle parole si rivelano molto meno fragili, grazie alla minore dimensionalità.

Abbiamo visto come algoritmi come la decomposizione del valore singolare (SVD) o l'analisi delle componenti principali possono essere utilizzati per comprimere una matrice in modo tale che M' caratteristiche n m in una matrice n k M mantenga la maggior× parte delle informazioni di M . Allo stesso modo, possiamo pensare alle incorporazioni di parole come una compressione di una matrice di incidenza parola -testo v t M , dove t è il numero di ×documenti nel corpus, e M $[i, j]$ misura la pertinenza della parola i al documento j. Comprimendo questa matrice a v× k si otterrebbe una forma di incorporazione di parole.

Detto questo, le reti neurali sono l'approccio più popolare per costruire le incorporazioni di parole. Immaginiamo una rete in cui lo strato di ingresso accetta gli embeddings attuali di (diciamo) cinque parole, $w_1, \ldots w_5$, che corrispondono a una particolare frase di cinque parole del nostro corpus di formazione dei documenti. Il compito della rete potrebbe essere quello di predettare l'incorporamento della parola centrale w_3 dagli incorporamenti delle quattro parole che la affiancano. Attraverso la retropropagazione, possiamo regolare i pesi dei nodi della rete in modo da migliorare la precisione su questo particolare esempio. La chiave qui è che continuiamo la retropropagazione oltre il livello più basso, in modo da modificare i parametri di ingresso effettivi! Questi parametri rappresentano le incorporazioni per le parole nella frase data, quindi questo passaggio migliora l'incorporazione per il compito di predizione. Ripetendo questa operazione su un gran numero di esempi di formazione, si ottiene un'incorporazione significativa per l'intero vocabolario.

Uno dei motivi principali della popolarità delle incorporazioni di parole è *word2vec*, un'implementazione terrifica di questo algoritmo, che può addestrare rapidamente le incorporazioni per centinaia di migliaia di parole del vocabolario su gigabyte di testo in modo totalmente non supervisionato. Il parametro più importante da impostare è il numero desiderato di dimensioni d. Se d è troppo piccolo, l'incorporamento non ha la libertà di catturare completamente il significato del simbolo dato. Se d è troppo grande, la rappresentazione diventa ingombrante e sovra adattata. In generale, il punto di forza è compreso tra 50 e 300 dimensioni.

Incorporamenti di grafici

Supponiamo di avere una matrice di somiglianza a coppie S definita su un universo di n articoli. Possiamo costruire la matrice di adiacenza del grafo di somiglianza G dichiarando un bordo (x, y) ogni volta che la somiglianza di x e y in S è sufficientemente alta. Questa matrice G di grandi dimensioni potrebbe essere compressa utilizzando la decomposizione del valore singolare (SVD) o l'analisi delle componenti principali (PCA), ma ciò si rivela costoso su reti di grandi dimensioni. Programmi come *word2vec* fanno un lavoro eccellente nel costruire rappresentazioni da sequenze di simboli in un corpus di formazione. La chiave per applicarli in nuovi. Il dominio è la mappatura del vostro particolare set di dati in stringhe su un interessante vocabolo. *DeepWalk* è un approccio alla costruzione di *embeddings di grafi*, rappresentazioni di punti per ogni vertice, in modo che i vertici "simili" siano posizionati vicini nello spazio. Il nostro vocabolario può essere scelto come l'insieme di ID di vertice distinti, da 1 a n. Ma qual è il testo che può rappresentare il grafico come una sequenza di simboli? Possiamo costruire delle passeggiate casuali sulla rete, in cui partiamo da un vertice arbitrario e saltiamo ripetutamente a un vicino casuale. Queste passeggiate possono essere considerate come 'frasi' sul nostro vocabolario di parole-verticali. Le incorporazioni risultanti, dopo aver eseguito *word2vec* su queste passeggiate casuali, si rivelano caratteristiche molto efficaci nelle applicazioni. DeepWalk è un'eccellente illustrazione di come le incorporazioni di parole possano essere utilizzate per catturare il

significato da qualsiasi corpus di sequenze su larga scala, indipendentemente dal fatto che siano tratte da un linguaggio naturale. La stessa idea gioca un ruolo importante nella seguente storia di guerra.

Capitolo 12

Big Data: Raggiungere la scala

Una volta sono stato intervistato in un programma televisivo e mi è stata chiesta la differenza tra *dati* e *big data*. Dopo averci pensato un po', ho dato una risposta che mantengo ancora oggi: "La dimensione". *Bupkis* è una meravigliosa parola yiddish che significa "troppo piccolo per avere importanza". Usata in una frase come "Lo pagarono *bupkis*", è una lamentela su una misera somma di denaro. Forse l'analogia più vicina nel vernacolo inglese è la parola "noccioline". In generale, i volumi di dati che abbiamo trattato finora in questo libro sono tutti pari a *zero*. Gli insiemi di formazione annotati dall'uomo si aggirano tra le centinaia e le migliaia di esempi, ma tutto ciò che si deve pagare per creare ha difficoltà a raggiungere i milioni. Il registro di tutte le corse dei taxi di New York per diversi anni, discusso nella Sezione 1.6, è arrivato a 80 milioni di record. Non male, ma pur sempre *bupkis*: può memorizzarlo facilmente sul suo computer portatile e fare una scansione del file per tabulare le statistiche in pochi minuti. La parola d'ordine *big data* sta forse raggiungendo la sua data di scadenza, ma presuppone l'analisi di serie di dati veramente enormi. Il significato di *"big"* aumenta con il tempo, ma al momento traccerei la linea di partenza a circa 1 terabyte. Non è così impressionante come può sembrare. Dopotutto, al momento in cui scriviamo, un disco di dimensioni terabyte le costerà solo 100 dollari, cioè un bel po' di soldi. Ma l'acquisizione di un set di dati significativo per riempirlo richiederà una certa iniziativa, forse un accesso privilegiato all'interno di una grande azienda Internet o grandi volumi di video. Ci sono molte organizzazioni che lottano regolarmente con petabyte e persino exabyte di dati. I big data richiedono un'infrastruttura di dimensioni maggiori rispetto ai progetti che abbiamo considerato finora. Lo spostamento di enormi volumi di dati tra le macchine richiede reti veloci e pazienza. Dobbiamo abbandonare l'elaborazione sequenziale, anche al di là dei core multipli, per arrivare a un gran numero di macchine flottanti. Questi calcoli scalano fino al punto in cui dobbiamo considerare la robustezza, a causa della quasi certezza che qualche componente hardware si guasterà prima di ottenere la nostra risposta. Lavorare con i dati diventa generalmente più difficile con le dimensioni. In questa sezione, cercherò di sensibilizzarla sui problemi generali associati alle serie di dati enormi. È importante capire perché le dimensioni sono importanti, in modo da poter contribuire a progetti che operano su quella scala.

12.1 Che cosa sono i Big Data?

Le dimensioni sono importanti: possiamo fare cose incredibili con questa roba. Ma contano anche altre cose. Questa sezione esaminerà alcune delle complessità tecniche e concettuali della gestione dei big data.

12.1.1 Big Data come Bad Data

Le serie di dati enormi sono in genere il risultato di un'opportunità, anziché di un progetto. Nella scienza tradizionale guidata dalle ipotesi, progettiamo un esperimento per raccogliere esattamente i dati di cui abbiamo bisogno per rispondere alla nostra domanda specifica. Ma i big data sono più tipicamente il prodotto di un processo di registrazione di eventi discreti, o forse di contributi distribuiti da milioni di persone sui social media. I dati

Lo scienziato in genere ha un controllo minimo o nullo del processo di raccolta, solo un vago incarico di trasformare tutti quei bit in denaro. Consideri il compito di misurare l'opinione popolare dai post di una piattaforma di social media o di un sito di recensioni online. I big data possono essere una risorsa meravigliosa. Ma sono particolarmente soggetti a pregiudizi e limitazioni che rendono difficile trarre conclusioni accurate, tra cui:

- *Partecipazione non rappresentativa*: Ci sono pregiudizi di campionamento inerenti a qualsiasi fonte di dati ambientali. I dati di un particolare sito di social media non riflettono le persone che non lo utilizzano - e bisogna fare attenzione a non generalizzare troppo.

  Gli utenti di Amazon acquistano molti più libri rispetto agli acquirenti di Walmart. Anche le loro affiliazioni politiche e il loro status economico differiscono. Si ottengono visioni del mondo ugualmente parziali ma molto diverse se si analizzano i dati di Instagram (troppo giovani), del *New York Times* (troppo liberali), di *Fox News* (troppo conservatori) o del *Wall Street Journal* (troppo ricchi).

- *Spam e contenuti generati dalle macchine*: Le fonti di big data sono peggio che poco rappresentative. Spesso sono state progettate per essere deliberatamente fuorvianti.

  Qualsiasi piattaforma online abbastanza grande da generare enormi quantità di dati è abbastanza grande da avere incentivi economici per pervertirli. Eserciti di recensori pagati lavorano ogni giorno scrivendo recensioni di prodotti falsi e fuorvianti. I bot sfornano in massa tweet scritti meccanicamente e testi più lunghi, e ne sono addirittura i principali consumatori: una frazione considerevole delle visite riportate su qualsiasi sito web proviene da crawler meccanici, anziché da persone. Il 90% di tutte le e-mail inviate attraverso le reti è spam: l'efficacia dei filtri antispam in diverse fasi della pipeline è l'unico motivo per cui non se ne vedono di più.
  Il filtraggio dello spam è una parte essenziale del processo di pulizia dei dati in qualsiasi analisi dei social media. Se non rimuove lo spam, questo mentirà invece di ingannarla.

- *Troppa ridondanza*: Molte attività umane seguono una distribuzione a legge di potenza, il che significa che una percentuale molto piccola di elementi rappresenta una grande percentuale dell'attività totale. Le notizie e i social media si concentrano molto sugli ultimi passi falsi delle Kardashian e di altre celebrità simili, coprendoli con articoli a migliaia. Molti di questi saranno dei duplicati quasi esatti di altri articoli. Quanto di più le dice l' di questi articoli rispetto a uno solo di essi?

  Questa legge di copertura ineguale implica che molti dei dati che vediamo attraverso le fonti ambientali sono già stati visti in precedenza. La rimozione di questa duplicazione è una fase di pulizia essenziale per molte applicazioni. Qualsiasi sito di condivisione di foto conterrà migliaia di immagini dell'Empire State Building, ma nessuna dell'edificio in cui lavoro. L'addestramento di un classificatore con queste immagini producono caratteristiche favolose per i punti di riferimento, che possono o meno essere utili per compiti più generali.

- *Suscettibilità alle distorsioni temporali*: I prodotti cambiano in risposta alla concorrenza e ai cambiamenti della domanda dei consumatori. Spesso questi miglioramenti cambiano il modo in cui le persone utilizzano questi prodotti. Una serie temporale risultante dalla raccolta di dati ambientali potrebbe codificare diverse transizioni prodotto/interfaccia, che rendono difficile distinguere l'artefatto dal segnale.

12.1.2 I tre V

I tipi di consulenza manageriale si sono aggrappati alla nozione *delle tre V dei big data* come mezzo per: le proprietà di *volume*, *varietà* e *velocità*. Esse forniscono una base per parlare di ciò che rende i big data diversi. Le V sono:

- *Volume*: È ovvio che i big data sono più grandi dei piccoli dati. La distinzione è di classe. Lasciamo il mondo in cui possiamo rappresentare i nostri dati in un foglio di calcolo o elaborarli su una singola macchina. Questo richiede lo sviluppo di un'infrastruttura computazionale più sofisticata e la limitazione dell'analisi ad algoritmi a tempo lineare per garantire l'efficienza.

- *Varietà*: La raccolta di dati ambientali in genere va oltre la matrice per accumulare dati eterogenei, che spesso richiedono tecniche di integrazione ad hoc.

  Consideri i social media. I post possono includere testo, link, foto e video. A seconda del nostro compito, tutti questi elementi possono essere rilevanti, ma l'elaborazione del testo richiede tecniche molto diverse rispetto ai dati di rete e ai media multipli. Anche le immagini e i video sono bestie molto diverse, che non possono essere elaborate con la stessa pipeline. Integrare in modo significativo questi materiali in un unico set di dati per l'analisi richiede una riflessione e uno sforzo notevoli.
- *Velocità*: La raccolta di dati da fonti ambientali implica che il sistema sia *vivo*, ossia che sia sempre acceso e che raccolga sempre dati. Al contrario, i set di dati che abbiamo studiato finora sono stati generalmente *morti*, ossia raccolti una volta e inseriti in un file per un'analisi successiva.

  Dati in tempo reale significa che devono essere costruite infrastrutture per la raccolta, l'indicizzazione, l'accesso e la visualizzazione dei risultati, in genere attraverso un sistema di dashboard. Dati in tempo reale significa che i consumatori vogliono accedere in tempo reale ai risultati più recenti, attraverso grafici, diagrammi e API.

  A seconda del settore, l'accesso in tempo reale può comportare l'aggiornamento dello stato del database entro pochi secondi o addirittura millisecondi dagli eventi reali. In particolare, i sistemi finanziari associati al trading ad alta frequenza richiedono un accesso immediato alle informazioni più recenti. Lei è in una gara contro l'altro, e può trarre profitto solo se vince.

  La velocità dei dati è forse il punto in cui la scienza dei dati si differenzia in modo più sostanziale dalla statistica classica. È ciò che alimenta la domanda di architetture di sistema avanzate, che richiedono ingegneri che costruiscano per la scala utilizzando le tecnologie più recenti.

Il set di gestione a volte definisce una quarta V: la *veridicità*, una misura di quanto ci fidiamo dei dati sottostanti. Qui ci troviamo di fronte al problema di eliminare lo spam e altri artefatti derivanti dal processo di raccolta, oltre il livello di pulizia normale.

12.2 Algoritmica per i Big Data

I big data richiedono algoritmi efficienti per lavorare su di essi. In questa sezione, approfondiremo brevemente le questioni algoritmiche di base associate ai big data: complessità asintotica, hashing e modelli di streaming per ottimizzare le prestazioni di I/O nei file di grandi dimensioni.

Non ho il tempo o lo spazio per fornire un'introduzione completa alla progettazione e all'analisi degli algoritmi combinatori. Tuttavia, posso raccomandare con fiducia *The Algorithm Design Manual* [Ski08] come un libro eccellente su questi argomenti, se le capita di cercarne uno.

12.2.1 Analisi di Big Oh

L'analisi tradizionale degli algoritmi si basa su un computer astratto chiamato *Random Access Machine* o *RAM*. Su tale modello:

- Ogni semplice operazione richiede esattamente un passo.

- Ogni operazione di memoria richiede esattamente un passo.

Quindi, contando le operazioni eseguite nel corso dell'algoritmo, si ottiene il suo tempo di esecuzione.

In generale, il numero di operazioni eseguite da qualsiasi algoritmo è una funzione della dimensione dell'ingresso n: una matrice con n righe, un testo con n parole, un insieme di punti con n punti. L'analisi dell'algoritmo è il processo di stima o di delimitazione del numero di passi che l'algoritmo compie in funzione di n.

Per gli algoritmi definiti da **cicli for**, tale analisi è abbastanza semplice. La profondità dell'annidamento di questi loop definisce la complessità dell'algoritmo. Un singolo ciclo da 1 a n definisce un algoritmo *in tempo lineare* o $O(n)$, mentre due cicli annidati definiscono un algoritmo *in tempo quadratico* o $O(n^2)$. Due cicli for sequenziali che non si annidano sono ancora lineari, perché vengono utilizzati $n+ n= 2n$ passaggi invece di tali operazioni.

Esempi di algoritmi di base con struttura a loop includono:

- *Trovare il vicino più prossimo del punto* p: Dobbiamo confrontare p con tutti gli n punti di una determinata matrice a. Il calcolo della distanza tra p e il punto $a[i]$ richiede la sottrazione e la quadratura di d termini, dove d è dimensionalità di p. Il ciclo di tutti gli n punti e la traccia del punto più vicino richiede $O(d\ n)$ tempo. Poiché d è in genere abbastanza piccolo da essere considerato una costante, questo è considerato un algoritmo a tempo lineare. -

La coppia di punti più vicina in un insieme: Dobbiamo confrontare ogni punto $a[i]$ con ogni altro punto $a[j]$, dove 1 i = j n.≤ In / base al≤ ragionamento precedente, questo richiede un tempo $O(d\ n^2)$ e sarebbe considerato un algoritmo a tempo quadratico.

- *Moltiplicazione di matrici*: La moltiplicazione di× una matrice $x\ y$ per una× matrice $y\ z$ dà come risultato× una matrice x z, dove ciascuno -dei termini $x\ z$ è il prodotto di punti di due vettori *di lunghezza* y.
  Questo algoritmo richiede $-x - y\ z$ passi. Se $n = \max(x, y, z)$, allora questo richiede al massimo $O(n^3)$ passi, e sarebbe considerato un algoritmo a tempo cubo.

Per gli algoritmi definiti dai cicli condizionali **while** o dalla ricorsione, l'analisi richiede spesso una maggiore sofisticazione. Gli esempi, con spiegazioni molto concise, includono:

- *Aggiunta di due numeri*: Le operazioni molto semplici potrebbero non avere condizionali, come l'addizione di due numeri. In questo caso non c'è un valore reale di n, ma solo due, quindi l'operazione richiede un tempo costante o $O(1)$.

- *Ricerca binaria*: Cerchiamo di individuare una determinata chiave di ricerca k in un'ar- raia ordinata A, contenente n elementi. Pensi alla ricerca di un nome nell'elenco telefonico. Confrontiamo k con l'elemento centrale $A[n/2]$ e decidiamo se ciò che stiamo cercando si trova nella metà superiore o in quella inferiore. Il numero di dimezzamenti fino ad arrivare a 1 è $\log_2(n)$, come abbiamo discusso nella Sezione 2.4. Pertanto, la ricerca binaria viene eseguita in tempo $O(\log n)$.

- *Mergesort*: Due elenchi ordinati con un totale di n elementi possono essere fusi in un unico elenco ordinato in tempo lineare: estragga il più piccolo dei due elementi di testa come primo in ordine ordinato, e ripeta. Mergesort divide gli n elementi in due metà, ordina ciascuno e poi li unisce. Il numero di dimezzamenti fino ad arrivare a 1 è di nuovo $\log_2(n)$ (vedere la Sezione 2.4), e l'unione di tutti gli elementi a tutti i livelli produce un algoritmo di ordinamento $O(n \log n)$.

Si è trattato di una rassegna algoritmica molto veloce, forse troppo rapida per essere considerata, ma è riuscita a fornire i rappresentanti di sei diverse classi di complessità degli algoritmi. Queste funzioni di complessità definiscono uno spettro dal più veloce al più lento, definito dal seguente ordine:

12.2.2 Hashing

L'hashing è una tecnica che spesso può trasformare gli algoritmi quadratici in algoritmi a tempo lineare, rendendoli praticabili per gestire la scala di dati con cui speriamo di lavorare. Abbiamo parlato per la prima volta delle funzioni di hash nel contesto dell'hashing sensibile alla località (LSH) nella Sezione 10.2.4. Una *funzione hash* h prende un oggetto x e lo mappa in un numero intero specifico $h(x)$. L'idea chiave è che ogni volta che x= y, allora $h(x) = h(y)$. Quindi possiamo utilizzare $h(x)$ come un intero per indicizzare un array e raccogliere tutti gli oggetti simili nello stesso posto. Oggetti diversi vengono *solitamente* mappati in posti diversi, supponendo una funzione hash ben progettata, ma non ci sono garanzie.Gli oggetti che cerchiamo di sottoporre a hashish sono spesso sequenze di elementi più semplici. Ad esempio, i file o le stringhe di testo sono solo sequenze di caratteri elementari. Questi componenti elementari di solito hanno una mappatura naturale con i numeri: i codici dei caratteri come Unicode, per definizione, mappano i simboli con i numeri, ad esempio. Il primo passo per l'hash x è rappresentarlo come una sequenza di tali numeri, senza perdita di informazioni. Supponiamo che ciascuno dei $n = |S|$ numeri di caratteri di x siano numeri interi compresi tra 0 e α - 1. Trasformare il vettore di numeri in un singolo numero rappresentativo è il compito della funzione hash $h(x)$. Un buon modo per farlo è pensare al vettore come ad un numero di *base-*α.

La funzione mod (x mod m) restituisce il resto di x diviso per m, e quindi restituisce un numero compreso tra 0 e m 1. Questo- numero di n cifre, con base a 1, è destinato ad essere enorme. Questo numero *a n cifre*, di base α, è destinato ad essere enorme, quindi il resto ci offre un modo per ottenere un codice rappresentativo di dimensioni modeste. Il principio è lo stesso di una roulette per il gioco d'azzardo: il lungo percorso della pallina intorno alla ruota termina in una delle m= 38 fessure, determinate dal resto della lunghezza del percorso diviso per la circonferenza della ruota. Queste funzioni hash sono incredibilmente utili. Le applicazioni principali includono:

- *Manutenzione del dizionario*: Una tabella hash è una struttura di dati basata su array che utilizza $h(x)$ per definire la posizione dell'oggetto x, abbinata a un metodo di risoluzione delle collisioni appropriato. Se implementate correttamente, tali tabelle hash producono tempi di ricerca costanti (o $O(1)$) nella pratica.

Questo è molto meglio della ricerca binaria, e quindi le tabelle hash sono ampiamente utilizzate nella pratica. Infatti, Python utilizza l'hashing per collegare i nomi delle variabili ai valori che memorizzano. L'hashing è anche il metodo fondamentale idea che sta alla base dei sistemi di calcolo distribuito come MapReduce, di cui si parlerà nella Sezione 12.6.

- *Conteggio delle frequenze*: Un compito comune nell'analisi dei registri è quello di tabulare le frequenze di determinati eventi, come il conteggio delle parole o delle pagine visitate. L'approccio più semplice e veloce consiste nell'impostare una tabella hash con i tipi di evento come chiave e nell'incrementare il contatore associato per ogni nuovo evento. Se implementato correttamente, questo algoritmo è lineare rispetto al numero totale di eventi da analizzare.

- *Rimozione dei duplicati*: Un'importante operazione di pulizia dei dati consiste nell'identificare i record doppi in un flusso di dati e nel rimuoverli. Forse si tratta di tutti gli indirizzi e -mail dei nostri clienti e vogliamo assicurarci di inviare spam a ciascuno di loro solo una volta. In alternativa, potremmo cercare di costruire il vocabolario completo di una determinata lingua da grandi volumi di testo.

  L'algoritmo di base è semplice. Per ogni elemento nel flusso, controlla se è già presente nella tabella hash. Se non lo è, lo inserisce, altrimenti ignora. Se implementato correttamente, questo algoritmo richiede un tempo lineare rispetto al numero totale di record da analizzare.

- *Canonizzazione*: Spesso lo stesso oggetto può essere indicato con più nomi diversi. Le parole del vocabolario sono generalmente insensibili alle maiuscole, il che significa che "Il" è equivalente a "la". Per determinare il vocabolario di una lingua è necessario unificare le forme alternative, mappandole su un'unica chiave.

  Questo processo di costruzione di una *rappresentazione canonica* può essere interpretato come hashing. In genere, ciò richiede una funzione di semplificazione specifica del dominio, che esegue operazioni come la riduzione a lettere minuscole, la rimozione degli spazi bianchi, l'eliminazione delle parole di stop e l'espansione delle abbreviazioni. Queste chiavi canoniche possono poi essere sottoposte a hashing, utilizzando funzioni hash convenzionali.

- *Hashing crittografico*: costruendo rappresentazioni concise e *invertibili*, l'hashing può essere utilizzato per monitorare e limitare il comportamento umano. Come può dimostrare che un file di input è rimasto invariato dall'ultima volta che lo ha analizzato? Costruisca un codice hash o un *checksum* per il file quando ci ha lavorato, e salvi questo codice per confrontarlo con l'hash del file in qualsiasi momento futuro. I due codici saranno uguali se il file è invariato e quasi sicuramente differiranno se si sono verificate delle modifiche.

  Supponiamo che voglia impegnarsi a fare un'offerta su un articolo specifico, ma non rivelare il prezzo effettivo che pagherà fino a quando tutte le offerte non saranno state fatte. Esegua l'hash della sua offerta utilizzando una determinata funzione di hash crittografico e invii il codice hash risultante. Dopo la scadenza, invii nuovamente la sua offerta, questa volta senza crittografia. Qualsiasi mente sospettosa può eseguire l'hash della sua offerta ora aperta e confermare che il valore corrisponde al codice hash precedentemente inviato. La chiave è che sia difficile produrre collisioni con la funzione di hash data, ossia che non sia possibile costruire facilmente un altro messaggio che abbia lo stesso codice di hash. Altrimenti, potrebbe inviare il secondo messaggio invece del primo, modificando la sua offerta dopo la scadenza.

12.2.3 Sfruttare la gerarchia di archiviazione

Gli algoritmi dei big data sono spesso *legati allo storage* o alla *larghezza di banda*, piuttosto che al *calcolo*. Ciò significa che il costo dell'attesa per l'arrivo dei dati dove sono necessari supera quello della manipolazione algoritmica per ottenere i risultati desiderati. Ci vuole ancora mezz'ora solo per leggere 1 terabyte di dati da un disco moderno. Il raggiungimento di buone prestazioni può basarsi più su una gestione intelligente dei dati che su algoritmi sofisticati. Per essere disponibili per l'analisi, i dati devono essere archiviati da qualche parte in un sistema informatico. Ci sono diversi tipi di dispositivi possibili in cui inserirli, che differiscono notevolmente per velocità, capacità e latenza. Le differenze di prestazioni tra i diversi livelli della *gerarchia di archiviazione* sono così enormi che non possiamo ignorarle nella nostra astrazione della macchina RAM. Infatti, il rapporto tra la velocità di accesso dal disco e la memoria cache è approssimativamente uguale (10^6) alla velocità di una tartaruga rispetto alla velocità di uscita della terra! I livelli principali della gerarchia di archiviazione sono:

- *Memoria cache*: Le moderne architetture dei computer presentano un complesso sistema di registri e cache per memorizzare copie di lavoro dei dati attivamente. Una parte di questa memoria viene utilizzata per il prefetching: la cattura di blocchi di dati più grandi intorno a posizioni di memoria che sono state accedute di recente, in previsione di una loro successiva necessità. Le dimensioni della cache sono tipicamente misurate in megabyte, con tempi di accesso da cinque a cento volte più veloci rispetto alla memoria principale. Queste prestazioni rendono molto vantaggiose le computazioni che sfruttano *la località*, per utilizzare in modo intensivo particolari elementi di dati in raffiche concentrate, piuttosto che in modo intermittente nel corso di una lunga computazione.

- *Memoria principale*: È quella che contiene lo stato generale della computazione e dove vengono ospitate e mantenute le strutture di dati di grandi dimensioni. La memoria principale è generalmente misurata in gigabyte e funziona da centinaia a migliaia di volte più velocemente dell'archiviazione su disco. Per quanto possibile, abbiamo bisogno di strutture di dati che si adattino alla memoria principale e che evitino il comportamento di paginazione della memoria virtuale.

- *Memoria principale su un'altra macchina*: I tempi di latenza su una rete locale sono dell'ordine dei millisecondi, il che la rende generalmente più veloce dei dispositivi di archiviazione secondaria come i dischi. Ciò significa che le strutture di dati distribuite, come le tabelle hash *, possono* essere mantenute in modo significativo su reti di macchine, ma con tempi di accesso che possono essere centinaia di volte più lenti della memoria principale.

- *Archiviazione su disco*: I dispositivi di archiviazione secondaria possono essere misurati in terabyte, fornendo la capacità che consente ai big data di diventare grandi. I dispositivi fisici come i dischi rotanti richiedono un tempo considerevole per spostare la testina di lettura nella posizione in cui si trovano i dati. Una volta lì, la lettura di un grande blocco di dati è relativamente veloce. Questo motiva il pre-fetching, ovvero la copia di grandi blocchi di file nella memoria con il presupposto che saranno necessari in seguito.

I problemi di latenza in genere agiscono come uno sconto sul volume: paghiamo molto per il primo elemento a cui accediamo, ma poi ne otteniamo altri a prezzi molto bassi. Dobbiamo organizzare le nostre computazioni per trarre vantaggio da questo, utilizzando tecniche come:

- *Elaborare file e strutture di dati in flussi*: È importante accedere ai file e alle strutture di dati in modo sequenziale, quando possibile, per sfruttare il pre-fetching. Ciò significa che gli array sono migliori delle strutture collegate, perché gli elementi logicamente vicini si trovano vicini sul dispositivo di archiviazione. Significa effettuare interi passaggi sui file di dati che leggono ogni elemento una volta, per poi eseguire tutte le computazioni necessarie prima di passare oltre. Gran parte del v antaggio dell'ordinamento dei dati è che possiamo saltare alla posizione appropriata in questione. Si renda conto che questo tipo di accesso casuale è costoso: pensi a spazzare invece di cercare.

- *Pensate a grandi file invece che a directory*: Si può organizzare un corpus di documenti in modo che ognuno di essi si trovi in un proprio file. Questo è logico per gli esseri umani, ma lento per le macchine, quando ci sono milioni di piccoli file. Molto meglio è organizzarli in un unico grande file, in modo da esaminare in modo efficiente tutti gli esempi, invece di richiedere un accesso separato al disco per ciascuno di essi.
- *Impacchettare i dati in modo conciso*: Il costo della decompressione dei dati conservati nella memoria principale è generalmente molto inferiore ai costi di trasferimento aggiuntivi per i file più grandi. Questa è un'argomentazione secondo cui conviene rappresentare i file di dati di grandi dimensioni in modo conciso, quando è possibile. Questo potrebbe significare schemi di compressione dei file espliciti, con dimensioni dei file sufficientemente piccole da poter essere espanse in memoria.

Significa però progettare formati di file e strutture di dati da codificare in modo conciso. Consideriamo la rappresentazione delle sequenze di DNA, che sono lunghe stringhe su un alfabeto di quattro lettere. Ogni lettera/base può essere rappresentata con 2 bit, il che significa che quattro basi possono essere rappresentate in un singolo byte da 8 bit e trentadue basi in una parola da 64 bit. Queste riduzioni delle dimensioni dei dati possono ridurre notevolmente i tempi di trasferimento e valgono lo sforzo computazionale di impacchettare e disimballare. Abbiamo già sottolineato l'importanza della leggibilità nei formati di file nella Sezione 3.1.2 e manteniamo questa opinione anche in questa sede. Riduzioni minori dimensioni probabilmente non valgono

la perdita di leggibilità o di facilità di parsing. Ma dimezzare le dimensioni di un file equivale a raddoppiare la velocità di trasferimento, il che può essere importante in un ambiente di big data.

12.2.4 Algoritmi di streaming e a passaggio singolo

I dati non vengono necessariamente conservati per sempre. O addirittura per sempre. Nelle applicazioni con un volume molto elevato di aggiornamenti e attività, può essere utile calcolare le statistiche al volo, man mano che i dati emergono, in modo da poter poi buttare via l'originale. In un algoritmo *di streaming* o a *passaggio singolo*, abbiamo solo una possibilità di visualizzare ogni elemento dell'input. Possiamo ipotizzare un po' di memoria, ma non abbastanza per memorizzare i dati di dei singoli record. Dobbiamo decidere cosa fare con ogni elemento quando lo vediamo, e poi non c'è più. Ad esempio, supponiamo di voler calcolare la media di un flusso di numeri al suo passaggio. Non è un problema difficile: possiamo mantenere due variabili: s che rappresenta la somma corrente fino ad oggi, e n il numero di elementi che abbiamo visto . Per ogni nuova osservazione a_i, la aggiungiamo a s e incrementiamo n. Ogni volta che qualcuno ha bisogno di conoscere l'attuale media del flusso A, riportiamo la media del flusso A. Il problema è che la media della sequenza non può essere conosciuta fino alla fine del flusso, a quel punto A abbiamo perso gli elementi originali da sottrarre alla media. Ma non tutto è perduto.Quindi, tenendo traccia di una somma corrente dei quadrati degli elementi, oltre a n e s, abbiamo tutto il materiale necessario per calcolare la varianza su richiesta. Molte quantità non possono essere calcolate esattamente con il modello di streaming. Un esempio potrebbe essere la ricerca dell'elemento *mediano* di una lunga sequenza. Supponiamo di non avere abbastanza memoria per memorizzare la metà degli elementi del flusso completo. Il primo elemento che abbiamo scelto di eliminare, qualunque esso sia, potrebbe essere reso mediano da un flusso accuratamente progettato di elementi non ancora visti. Abbiamo bisogno di avere a disposizione tutti i dati contemporaneamente per risolvere alcuni problemi. Ma anche se non possiamo calcolare qualcosa con esattezza, spesso possiamo una stima che è sufficiente per il lavoro governativo. Problemi importanti di questo tipo includono l'identificazione degli elementi più frequenti in un flusso, il numero di elementi distinti, o anche la stima della frequenza degli elementi quando non abbiamo abbastanza memoria per tenere un contatore esatto. *L'abbozzo* comporta l'utilizzo della memoria che abbiamo per tenere traccia di una rappresentazione parziale della sequenza. Forse si tratta di un istogramma di frequenza di elementi suddivisi per valore, o di una piccola tabella hash dei valori che abbiamo visto finora. La qualità della nostra stima aumenta con la quantità di memoria che abbiamo per memorizzare il nostro schizzo. *Il campionamento casuale* è uno strumento immensamente utile per la costruzione di schizzi ed è l'oggetto della Sezione 12.4.

12.3 Filtraggio e campionamento

Un vantaggio importante dei big data è che, con un volume sufficiente, ci si può permettere di buttare via la maggior parte dei dati. E questo può essere molto vantaggioso, per fare
la sua analisi è più pulita e più facile.

Distinguo due modi distinti di eliminare i dati: il filtraggio e il campionamento. *Filtrare* significa selezionare un sottoinsieme rilevante di dati in base a un criterio specifico. Ad esempio, supponiamo di voler costruire un modello linguistico per un'applicazione negli Stati Uniti e di volerlo addestrare sui dati di Twitter. L'inglese rappresenta solo un terzo di tutti i tweet su Twitter, per cui filtrando tutte le altre lingue rimane un numero sufficiente per un'analisi significativa. Possiamo pensare al filtraggio come a una forma speciale di pulizia, in

cui rimuoviamo i dati non perché sono errati, ma perché distraggono dall'argomento in questione. Filtrare i dati irrilevanti o difficili da interpretare richiede una conoscenza specifica dell'applicazione. L'inglese è effettivamente la lingua principale in uso negli Stati Uniti, il che rende la decisione di filtrare i dati in questo modo perfettamente ragionevole.

Ma il filtraggio introduce dei pregiudizi. Oltre il 10% della popolazione statunitense parla spagnolo. Non dovrebbero essere rappresentati nel modello linguistico, *amigo*? È importante selezionare i giusti criteri di filtraggio per ottenere il risultato che cerchiamo. Forse sarebbe meglio filtrare i tweet in base alla località di origine, invece che alla lingua.

Al contrario, *campionare* significa selezionare un sottoinsieme di dimensioni adeguate in arbi- co, senza criteri specifici del dominio. Ci sono diverse ragioni per cui potremmo voler sotto-campionare dei dati buoni e rilevanti:

- *Dimensionamento dei dati di addestramento*: I modelli semplici e robusti hanno in genere pochi parametri, rendendo superfluo 'utilizzo di grandi dati per adattarli. Il sottocampionamento dei dati in modo imparziale porta a un adattamento efficiente del modello, ma è comunque rappresentativo dell'intero set di dati.
- *Suddivisione dei dati*: L'igiene della costruzione del modello richiede una separazione netta dei dati di formazione, di test e di valutazione, in genere in un mix di 60%, 20% e 20%. La costruzione di queste partizioni in modo imparziale è necessaria per la veridicità di questo processo.
- *Analisi e visualizzazione esplorativa dei dati*: Le serie di dati in formato foglio elettronico sono veloci e facili da esplorare. Un campione imparziale è rappresentativo dell'insieme, pur rimanendo comprensibile.

Campionare *n* record in modo efficiente e imparziale è un compito più delicato di quanto possa sembrare all'inizio. Esistono due approcci generali, deterministico e randomizzato, che vengono descritti in dettaglio nelle sezioni seguenti.

12.4.1 Algoritmi di campionamento deterministico

Il nostro algoritmo di campionamento dell'uomo di paglia sarà il *campionamento per troncamento*, che prende semplicemente i primi *n* record del file come campione desiderato. Questo è semplice e ha la proprietà di essere facilmente *riproducibile*, il che significa che qualcun altro con il file di dati completo potrebbe facilmente ricostruire il campione.

Tuttavia, l'ordine dei record in un file spesso codifica informazioni semantiche, il che significa che i campioni troncati spesso contengono effetti sottili dovuti a fattori quali:

- *Pregiudizi temporali*: I file di registro sono in genere costruiti aggiungendo nuovi record alla fine del file. Pertanto, i primi *n* record saranno i più vecchi disponibili e non rifletteranno i recenti cambiamenti di regime.
- *Pregiudizi lessicografici*: Molti file sono ordinati in base alla chiave primaria, il che significa che i primi *n* record sono orientati verso una particolare popolazione. Immaginiamo un registro del personale ordinato per nome. I primi *n* record potrebbero essere costituiti solo dal nome, il che significa che probabilmente sovra-campioneremo i nomi arabi della popolazione generale e sotto- campioneremo quelli cinesi.
- *Pregiudizi numerici*: Spesso i file vengono ordinati in base ai numeri di identità, che possono sembrare definiti in modo arbitrario. Ma i numeri identificativi possono codificare un significato. Consideri l'ordinamento dei registri del personale in base ai numeri di previdenza sociale degli Stati Uniti. Infatti,

le prime cinque cifre dei numeri di previdenza sociale sono generalmente una funzione dell'anno e del luogo di nascita. Quindi il troncamento porta a un campione geograficamente e anagraficamente distinto.

Spesso i file di dati sono costruiti concatenando insieme file più piccoli, alcuni dei quali possono essere molto più ricchi di esempi positivi rispetto ad altri. In casi particolarmente patologici, il numero di record potrebbe codificare completamente la variabile di classe, il che significa che un classificatore accurato ma totalmente inutile potrebbe derivare dall'utilizzo dell'ID di classe come caratteristica.

Quindi il troncamento è generalmente una cattiva idea. Un approccio migliore è il *campionamento uni- forme*. Supponiamo di voler campionare *n/m* record su *n* da un determinato file. Un approccio semplice consiste nel partire dal record ith, dove *i* è un valore compreso tra 1 e *m*, e poi campionare ogni mth record a partire da *i*. Un altro modo per dirlo è quello di produrre il jesimo record se *j*(mod *m*)= i. Questo campionamento uniforme offre un modo per bilanciare molte preoccupazioni:

- Otteniamo esattamente il numero di record desiderato per il nostro campione.
  - È veloce e riproducibile da chiunque, dato il file e i valori di *i* e *m*.
  - È facile costruire più campioni disgiunti. Se ripetiamo il processo con un offset *i* diverso, otteniamo un campione indipendente.

Twitter utilizza questo metodo per governare i servizi API che forniscono accesso ai tweet. Il livello di accesso gratuito (il tubo spruzzatore) distribuisce l'1% del flusso, distribuendo ogni 100 tweet. I livelli di accesso professionali distribuiscono ogni 10 tweet o anche di più, a seconda di ciò che si è disposti a pagare.

Questo è generalmente migliore della troncatura, ma esistono ancora potenziali pregiudizi temporali periodici. Se si campiona ogni *mth* record nel registro, forse ogni elemento che si vede sarà associato a un evento di martedì, o alle 23 di ogni sera. Sui file ordinati per numeri, si rischia di ritrovarsi con elementi con le stessecifre di ordine inferiore. Numeri di telefono che terminano con "000" o che si ripetono. Cifre come "8888" sono spesso riservate all'uso lavorativo invece che residenziale, e quindi influenzano il campione. Si possono minimizzare le possibilità di questo fenomeno obbligando *m* a essere un numero primo abbastanza grande, ma l'unico modo certo per evitare le distorsioni del campionamento è utilizzare la randomizzazione.

12.4.2 Campionamento randomizzato e di flusso

Il campionamento casuale dei record con una probabilità *p* comporta una selezione *p n* elementi attesi,- senza alcuna distorsione esplicita. I tipici generatori di numeri casuali restituiscono un valore compreso tra 0 e 1, estratto da una distribuzione uniforme. Possiamo utilizzare la probabilità di campionamento *p* come soglia. Quando scansioniamo ogni nuovo record, generiamo un nuovo numero casuale *r*. Quando *r p*, accettiamo questo record nel nostro campione, ma quando *r*> p lo ignoriamo.≤

Il campionamento casuale è una metodologia generalmente solida, ma presenta alcune stranezze tecniche. Le discrepanze statistiche assicurano che alcune regioni o aree demografiche saranno sovracampionate rispetto alla popolazione, ma in modo imparziale e in misura prevedibile. I campioni casuali multipli non saranno disgiunti e il campionamento casuale non è riproducibile senza il seme e il generatore casuale.

Poiché il numero finale di record campionati dipende dalla casualità, potremmo ritrovarci con un numero leggermente eccessivo o insufficiente di elementi. Se abbiamo bisogno di *esattamente k* elementi, possiamo costruire una permutazione casuale degli elementi e troncarla dopo i primi *k*. Gli algoritmi per la costruzione di permutazioni casuali sono stati discussi nella Sezione 5.5.1. Questi sono semplici, ma richiedono grandi quantità di movimento irregolare dei dati, il che li rende potenzialmente dannosi per i file di grandi dimensioni. Sono semplici, ma richiedono grandi quantità di movimento irregolare dei dati, il che li rende potenzialmente negativi per i file di grandi dimensioni. Un approccio più semplice consiste nell'aggiungere un nuovo campo

di numeri casuali a ciascun record e nell'ordinarlo come chiave. Prendere i primi *k* record da questo file ordinato equivale a campionare casualmente esattamente *k* record.

Ottenere un campione casuale di dimensioni fisse da un flusso è un problema più difficile, perché non possiamo memorizzare tutti gli elementi fino alla fine. In effetti, non nemmeno quanto sarà grande *n* alla fine. Per risolvere questo problema, manterremo un campione uniformemente selezionato in un array di dimensioni *k*, aggiornato quando ogni nuovo elemento arriva dal flusso. La probabilità che il nono elemento del flusso appartenga al campione è *k/n*, e quindi lo inseriremo nel nostro array se il numero casuale *r* k/n. In questo modo, un residente attuale viene espulso dalla tabella e la selezione dell'elemento dell'array attuale che ne è vittima può essere effettuata con un'altra chiamata al generatore di numeri casuali.

12.5 Parallelismo

Due teste sono meglio di una, e cento teste meglio di due. La tecnologia di elaborazione è maturata in modo tale da rendere sempre più fattibile il comando di più elementi di elaborazione su richiesta per la sua applicazione. I microprocessori hanno abitualmente 4 core e oltre, il che rende utile pensare al parallelismo anche su macchine singole. L'avvento dei data center e del cloud

L'informatica ha reso facile il noleggio di un gran numero di macchine su richiesta, consentendo anche ai piccoli operatori di trarre vantaggio dalle grandi infrastrutture distribuite.

Esistono due approcci distinti per l'elaborazione simultanea con macchine multiple, ossia l'elaborazione parallela e distribuita. La distinzione consiste nell'accoppiamento stretto delle macchine e nel fatto che i compiti siano legati alla CPU o alla memoria/IO. All'incirca:

- L'*elaborazione parallela* avviene su una macchina, coinvolgendo più core e/o processori che comunicano attraverso i thread e le fonti del sistema operativo. Questo tipo di calcolo strettamente accoppiato è spesso legato alla CPU, limitato più dal numero di cicli che dal movimento dei dati attraverso la macchina. L'enfasi è sulla risoluzione di un particolare problema di calcolo più velocemente di quanto si potrebbe fare in modo sequenziale.
    - L'*elaborazione distribuita* avviene su molte macchine, utilizzando la comunicazione di rete. Il potenziale di scala in questo caso è enorme, ma è più adatto a lavori accoppiati in modo lasco che non comunicano molto. Spesso l'obiettivo dell'elaborazione distribuita prevede la condivisione di risorse come la memoria e l'archiviazione secondaria su più macchine, più che lo sfruttamento di più CPU. Quando la velocità di lettura dei dati da un disco è il collo di bottiglia, è meglio avere molte macchine che leggono il maggior numero possibile di dischi diversi, simultaneamente. In questa sezione, introduciamo i principi di base del calcolo parallelo e due modi relativamente semplici per sfruttarlo: il parallelismo dei dati e la ricerca in rete. MapReduce è il paradigma principale per il calcolo distribuito sui Big Data e sarà l'argomento della Sezione 12.6.

12.5.1 Uno, due, molti

Le culture primitive non erano molto esperte di numeri e, presumibilmente, contavano solo con le parole *uno*, *due* e *molti*. Questo è in realtà un ottimo modo di pensare all'informatica parallela e distribuita, perché la complessità aumenta molto rapidamente con il numero di macchine:

- *Uno*: cerchi di tenere occupati tutti i core del suo computer, ma sta lavorando su un solo computer. Non si tratta di informatica distribuita.
- *Due*: forse cercherà di dividere manualmente il lavoro tra alcune macchine della sua rete locale. Questo è a malapena un calcolo distribuito e generalmente viene gestito con tecniche ad hoc.
- *Molti*: Per sfruttare decine o addirittura centinaia di macchine, magari nel cloud, non abbiamo altra scelta che impiegare un sistema come MapRe- duce, in grado di gestire in modo efficiente queste risorse.

La complessità aumenta di pari passo con il numero di agenti che vengono coordinati verso un compito. Consideri cosa cambia quando le riunioni sociali aumentano di dimensioni. C'è una tendenza continua ad

accontentarsi di una coordinazione più debole con l'aumentare delle dimensioni, e una maggiore Ma forse questo fa luce su alcune delle sfide della parallelizzazione e dell'informatica distribuita:

- *Coordinamento*: Come assegniamo le unità di lavoro ai processori, in particolare quando abbiamo più unità di lavoro che lavoratori? Come aggregare o combinare gli sforzi di ciascun lavoratore in un unico risultato?
- *Comunicazione*: In che misura i lavoratori possono condividere i risultati parziali? Come possiamo sapere quando tutti i lavoratori hanno terminato i loro compiti?
- *Tolleranza ai guasti*: Come riassegnare i compiti se i lavoratori si licenziano o muoiono? Dobbiamo proteggerci da attacchi maligni e sistematici, o solo da guasti casuali?

12.5.2 Parallelismo dei dati

Il parallelismo dei dati implica il partizionamento e la replica dei dati tra più processori e dischi, l'esecuzione dello stesso algoritmo su ogni pezzo, e poi la collazione dei risultati per produrre i risultati finali. Si ipotizza una macchina *master* che distribuisce i compiti a un gruppo di slave e raccoglie i risultati.

Un compito rappresentativo è l'aggregazione di statistiche da un'ampia raccolta di file, ad esempio il conteggio della frequenza di comparsa delle parole in un corpus di testo massivo. I conteggi per ogni file possono essere calcolati in modo indipendente come risultati parziali verso l'insieme, e il compito di unire questi file di conteggio risultanti può essere facilmente calcolato da un'unica macchina alla fine. Il vantaggio principale è la semplicità, perché tutti i processi di conteggio eseguono lo ste sso programma. La comunicazione tra i processori è semplice: spostare i file sulla macchina appropriata, avviare il lavoro e quindi riportare i risultati alla macchina master.

L'approccio più semplice al calcolo multicore prevede il parallelismo dei dati. I dati formano naturalmente partizioni stabilite dal tempo, da algoritmi di clustering o da categorie naturali. Per la maggior parte dei problemi di aggregazione, i record possono essere suddivisi in modo arbitrario, a condizione che tutti i sottoproblemi vengano uniti alla fine.

Per i problemi più complicati, è necessario un lavoro supplementare per combinare insieme i risultati di queste esecuzioni in un secondo momento. Ricordiamo l'algoritmo di clustering kmeans (Sezione 10.5.1), che ha due fasi:

1. Per ogni punto, identifica il centro del cluster attuale più vicino.

2. Calcola il nuovo centroide dei punti ora associati.

Supponendo che i punti siano stati distribuiti su più macchine, il primo passo richiede che il master comunichi tutti i centri attuali a ogni macchina, mentre il secondo passo richiede che ogni slave riferisca al master i nuovi centri dei punti nella sua partizione. Il master calcola quindi in modo appropriato le medie di questi centri per concludere l'iterazione.

12.5.3 Ricerca nella griglia

Un secondo approccio per sfruttare il parallelismo prevede più esecuzioni indipendenti sugli stessi dati. Abbiamo visto che molti metodi di apprendimento automatico coinvolgono parametri che hanno un impatto sulla qualità del risultato finale, come la selezione del giusto numero di cluster *k* per il clustering *k-means*. Scegliere il migliore significa provarli tutti, e ciascuna di queste esecuzioni può essere condotta simultaneamente su macchine diverse.

La *ricerca della griglia* è la ricerca dei meta-parametri giusti nell'addestramento. È difficile prevedere con esattezza come la variazione del tasso di apprendimento o della dimensione del lotto nella discesa stocastica

del gradiente influisca sulla qualità del modello finale. Si possono eseguire più adattamenti indipendenti in parallelo e alla fine si sceglie il migliore in base alla nostra valutazione.
Ricercare efficacemente lo spazio su k parametri diversi è difficile a causa delle interazioni:
identificare il miglior valore singolo di ogni parametro separatamente
non produce necessariamente il miglior set di parametri quando viene combinato. In genere, l'utente stabilisce valori minimi e massimi ragionevoli per ogni parametro p_i, nonché il numero di valori t_i per questo parametro da testare. Ogni intervallo viene suddiviso in valori equidistanti regolati da questo t_i. Proviamo quindi tutti i set di parametri che possono essere formati scegliendo un valore per intervallo, stabilendo la griglia nella ricerca a griglia.

Quanto dobbiamo credere che il modello migliore in una ricerca a griglia sia *davvero* migliore degli altri? Spesso c'è una semplice varianza che spiega le piccole differenze di prestazioni su un determinato set di test, trasformando la ricerca a griglia in una selezione del numero che fa sembrare migliori le nostre prestazioni. Se ha a disposizione le risorse computazionali per condurre una ricerca a griglia per il suo modello, si senta libero di procedere, ma riconosca i limiti di ciò che può fare il trial-and-error.

12.5.4 Servizi di cloud computing

Piattaforme come Amazon AWS, Google Cloud e Microsoft Azure consentono di noleggiare facilmente un numero elevato (o ridotto) di macchine per lavori a breve (o lungo) . Le offrono la possibilità di accedere esattamente alle risorse informatiche giuste quando ne ha bisogno, a condizione che sia in grado di pagarle, ovviamente.
I modelli di costo di questi fornitori di servizi sono però un po' complicati. In genere si tratta di tariffe orarie per ogni macchina virtuale, in funzione del tipo di processore, del numero di core e della memoria principale coinvolti. Le macchine ragionevoli saranno noleggiate a un prezzo compreso tra 10 e 50 centesimi all'ora. Pagherà la quantità di spazio di archiviazione a lungo termine in funzione dei gigabyte/mese, con diversi livelli di costo a seconda dei modelli di accesso. Inoltre, pagherà dei costi di larghezza di banda che coprono il volume di trasferimento dei dati tra le macchine e sul web.
I prezzi spot e le istanze riservate possono portare a costi orari inferiori per modelli di utilizzo speciali, ma con ulteriori avvertenze. Con il *prezzo spot*, le macchine vanno al miglior offerente, quindi il suo lavoro rischia di essere interrotto se qualcun altro ne ha più bisogno di . Con le *istanze riservate*, paga un certo importo in anticipo per ottenere un prezzo orario inferiore. Questo ha senso se ha bisogno di un computer 24 ore su 24, 7 giorni su 7, per un anno, ma non se ha bisogno di cento computer ciascuno per un giorno particolare.

Fortunatamente, la sperimentazione può essere gratuita. Tutti i principali fornitori di cloud offrono un po' di tempo gratuito ai nuovi utenti, in modo che lei possa giocare con la configurazione e decidere a loro spese se è adatta a lei.

12.6 MapReduce

Il paradigma MapReduce di Google per il calcolo distribuito si è diffuso ampiamente attraverso le implementazioni open-source come Hadoop e Spark. Offre un modello di programmazione semplice con diversi vantaggi, tra cui la possibilità di scalare facilmente a
centinaia o addirittura migliaia di macchine, e la tolleranza ai guasti attraverso la ridondanza.
Il livello di astrazione dei modelli di programmazione aumenta costantemente nel tempo, come testimoniano strumenti e sistemi più potenti che nascondono i dettagli di implementazione all'utente. Se si occupa di scienza dei dati su scala multipla, il calcolo MapReduce è probabilmente in corso sotto il cofano, anche se non lo sta programmando esplicitamente.

Una classe importante di compiti di scienza dei dati su larga scala ha la seguente struttura di base:

- Iterare su un gran numero di elementi, siano essi record di dati, stringhe di testo o directory di file.

- Estrae qualcosa di interessante da ogni elemento, che sia il valore di un campo particolare, il conteggio della frequenza di ogni parola o la presenza/assenza di particolari modelli in ogni file.
- Aggrega questi risultati intermedi su tutti gli articoli e genera un risultato combinato adeguato.

I rappresentanti di questa classe di problemi includono il conteggio della frequenza delle parole, il clustering *k-means* e i calcoli di PageRank. Tutti sono risolvibili attraverso semplici algoritmi iterativi, i cui tempi di esecuzione scalano linearmente in base alle dimensioni dell'input. Ma questo può essere inadeguato per gli input di dimensioni enormi, dove i file non si adattano naturalmente alla memoria di una singola macchina. Pensiamo ai problemi su scala web, come il conteggio della frequenza delle parole su miliardi di tweet, il clustering k-means su centinaia di milioni di profili Facebook e il PageRank su tutti i siti web di Internet. La soluzione tipica in questo caso è il *divide et impera*. Si suddividono i file di input tra m macchine diverse, si eseguono i calcoli in parallelo su ciascuna di esse e poi si combinano i risultati sulla macchina appropriata. Una soluzione di questo tipo In linea di principio, funziona per il conteggio delle parole, perché anche gli enormi corpora di testo alla fine si ridurranno a file relativamente piccoli di parole di vocabolario distinte con conteggi di frequenza as - sociati, che possono poi essere facilmente sommati per produrre i conteggi totali. Ma consideriamo un calcolo di PageRank, dove per ogni nodo v sommare il PageRank di tutti i nodi x in cui x punta a v. Non c'è modo di tagliare il grafo in pezzi separati in modo che tutti questi vertici x si trovino sulla stessa macchina di v. Mettere le cose nel posto giusto per è il cuore di MapReduce.

12.6.1 Programmazione Map-Reduce

La chiave per distribuire tali calcoli è l'impostazione di una tabella hash distribuita di bucket, dove tutti gli elementi con la stessa chiave vengono mappati nello stesso bucket:

- *Conteggio delle parole*: Per contare la frequenza totale di una particolare parola w in un insieme di file, dobbiamo raccogliere i conteggi di frequenza per tutti i file in un unico bucket associato a w. Da lì possono essere per produrre il totale finale.

- *clustering k-means*: La fase critica del clustering k-means è l'aggiornamento del nuovo centroide c' dei punti più vicini al centroide corrente c. Dopo aver fatto l'hashing di tutti i punti p più vicini a c in un unico bucket associato a c, possiamo calcolare c' in un unico sweep attraverso questo bucket.

- *PageRank*: Il nuovo PageRank del vertice v è la somma dei vecchi PageRank per tutti i vertici vicini x, dove (x, v) è un bordo diretto nel grafico. L'hashing del PageRank di x al bucket per tutti i vertici adiacenti v raccoglie tutte le informazioni rilevanti nel posto giusto, in modo da poter aggiornare il PageRank in un unico passaggio.

Questi algoritmi possono essere specificati attraverso due scritte dal programmatore, *map* e *reduce*:

- *Mappa*: Esegue una scansione di ogni file di input, eseguendo l'hashing o *emettendo* coppie chiave-valore come appropriato.
- *Ridurre*: Effettua un'analisi dell'insieme di valori v associati ad una chiave specifica k, aggregando ed elaborando di conseguenza.

L'efficienza di un programma MapReduce dipende da molti fattori, ma un obiettivo importante è mantenere il numero di emissioni al minimo. L'emissione di un conteggio per ogni parola attiva un messaggio tra le macchine, e questa comunicazione e le scritture associate al bucket si rivelano costose in grandi quantità. Più cose vengono mappate, più devono essere ridotte. L'ideale è *combinare* i conteggi da particolari flussi di input localmente prima, e poi emettere solo il totale per ogni parola distinta per file. Questo potrebbe essere fatto

aggiungendo strutture logiche/dati supplementari alla funzione di mappatura. Un'idea alternativa è quella di eseguire i mini-riduttori in memoria dopo la fase di mappatura, ma prima della comunicazione tra processori, come ottimizzazione per ridurre il traffico di rete. Notiamo che l'ottimizzazione per il calcolo in memoria è uno dei principali vantaggi prestazionali di Spark rispetto a Hadoop per la programmazione in stile MapReduce. La combinazione è stata eseguita localmente, quindi i conteggi per ogni parola utilizzata più di una volta in un file di input (qui *doc* e *be*) sono stati tabulati prima di emetterli ai riduttori. Un problema illustrato è quello dello *skew* di mappatura, lo squilibrio naturale nella quantità di lavoro assegnato a ciascuna attività di riduzione. In questo esempio giocattolo, al top reducer sono stati assegnati file di mappa con il 33% di parole in più e conteggi più grandi del 60% rispetto al suo partner. Per un compito con un tempo di esecuzione seriale di T, la parallelizzazione perfetta con n processori produrrà un tempo di esecuzione di T/n. Ma il tempo di esecuzione di un lavoro MapReduce è determinato dal pezzo più grande e più lento. La distorsione del mapper ci condanna a un pezzo più grande che spesso è sostanzialmente superiore alla dimensione media. Una fonte di skew del mapper è la fortuna del sorteggio, ovvero che è raro lanciare n monete e finire con un numero di teste esattamente uguale a quello delle code. Ma un problema più serio è che la frequenza delle chiavi è spesso distribuita a legge di potenza, quindi la chiave più frequente arriverà a dominare i conteggi. Consideriamo il problema del conteggio delle parole e supponiamo che la frequenza delle parole osservi la legge di Zipf della Sezione 5.1.5. Quindi la frequenza della parola più popolare (*il*) dovrebbe essere maggiore della somma delle mille parole classificate da 1000 a 2000. Qualunque sia la classifica in cui finirà *la parola*, probabilmente sarà la più difficile da digerire.[20]

12.6.2 MapReduce sotto il cofano

Tutto questo va bene. Ma come fa un'implementazione MapReduce come Hadoop a garantire che tutti gli elementi mappati vadano nel posto giusto? E come fa ad assegnare il lavoro ai processori e a sincronizzare le operazioni MapReduce, il tutto con una tolleranza agli errori?

Ci sono due componenti principali: la tabella hash distribuita (o file system) e il sistema di run-time che gestisce il coordinamento e la gestione delle risorse. Entrambi sono descritti in dettaglio qui di seguito.

Sistemi di file distribuiti

Grandi collezioni di computer possono contribuire con il loro spazio di memoria (RAM) e l'archiviazione su disco locale per attaccare un lavoro, non solo con le loro CPU. Un file system distribuito come Hadoop Distributed File System (HDFS) può essere implementato come una tabella hash distribuita. Dopo che un insieme di macchine registra la propria memoria disponibile con il sistema runtime di coordinamento, a ciascuna di esse può essere assegnato un certo intervallo di tabella hash cui sarà responsabile. Ogni processo che esegue la mappatura può quindi assicurarsi che gli elementi emessi siano inoltrati al bucket appropriato sulla macchina appropriata.

Poiché un gran numero di elementi potrebbe essere mappato in un singolo bucket, possiamo scegliere di rappresentare i bucket come file su disco, con i nuovi elementi aggiunti alla fine. L'accesso al disco è lento, ma il throughput del disco è ragionevole, quindi le scansioni lineari dei file sono generalmente gestibili.

Un problema con una tabella hash distribuita di questo tipo è la tolleranza ai guasti: un singolo arresto della macchina potrebbe perdere abbastanza valori da invalidare l'intero calcolo. La soluzione consiste nel replicare tutto su per garantire l'affidabilità su hardware di base. In particolare, il sistema di runtime replicherà ogni elemento su tre macchine diverse, per ridurre al minimo le possibilità di perdere i dati in caso di guasto hardware. Quando il sistema di runtime rileva che una macchina o un disco sono fuori uso, si mette al lavoro per replicare i dati persi da queste copie per ripristinare la salute del file system. **Sistema di runtime MapReduce**

L'altro componente principale degli ambienti MapReduce per Hadoop o Spark è il loro *sistema di runtime*, il livello di software che regola compiti :

- *Pianificazione del processore*: Quali core vengono assegnati all'esecuzione di quali compiti di mappatura e riduzione e su quali file di input? Il programmatore può aiutare suggerendo quanti mappatori e riduttori devono essere attivi in qualsiasi momento, ma l'assegnazione dei lavori ai core spetta al sistema di runtime.

- *Distribuzione dei dati*: Questo potrebbe comportare lo spostamento dei dati verso un processo disponibile che possa gestirli, ma ricordiamo che le operazioni tipiche di map e reduce richiedono semplici scansioni lineari attraverso file potenzialmente grandi. Pertanto, spostare un file potrebbe essere più costoso che eseguire il calcolo desiderato localmente.

  Quindi è meglio *spostare i processi verso i dati*. Il sistema di runtime deve avere la configurazione di quali risorse sono disponibili su quale macchina e il layout generale della rete. Può prendere una decisione appropriata su quali processi devono essere eseguiti dove.

Sincronizzazione: I riduttori non possono essere eseguiti fino a quando non viene mappato qualcosa su di loro, e non possono essere completati fino a quando non viene eseguita la mappatura. Spark consente flussi di lavoro più complicati, oltre ai cicli sincronizzati di map e reduce. È il sistema di runtime che gestisce questa sincronizzazione.

- *Tolleranza agli errori e ai guasti*: L'affidabilità di MapReduce richiede il recupero con grazia dai guasti hardware e di comunicazione. Quando il sistema di runtime rileva un guasto di un worker, tenta di riavviare la comu- nicazione. Quando questo non riesce, trasferisce i compiti non completati ad altri worker. Il fatto che tutto questo avvenga senza soluzione di continuità, senza il coinvolgimento del programmatore, ci consente di scalare le computazioni su grandi reti di macchine, sulla scala in cui gli intoppi diventano probabili invece che eventi rari.

Strati su strati

Sistemi come HDFS e Hadoop sono solo strati di software su cui altri sistemi possono basarsi. Anche se Spark può essere considerato un concorrente di Hadoop, in realtà può sfruttare il file system distribuito di Hadoop e spesso è più efficiente quando fa. Al giorno d'oggi, i miei studenti sembrano dedicare meno tempo alla scrittura di lavori MapReduce di basso livello, perché utilizzano invece livelli di software che lavorano a livelli di astrazione più elevati.

L'ecosistema completo dei Big Data è composto da molte specie diverse. Una classe importante è costituita dai database *NoSQL*, che consentono la distribuzione di dati strutturati su una rete distribuita di macchine, permettendo di combinare la RAM e il disco di più macchine. Inoltre, questi sistemi sono in genere progettati in modo da poter aggiungere altre macchine e risorse quando ne ha bisogno. Il costo di questa flessibilità è che di solito supportano linguaggi di interrogazione più semplici rispetto all 'SQL completo, ma comunque abbastanza ricchi per molte applicazioni.

L'ecosistema software dei Big Data si evolve molto più rapidamente rispetto alle questioni fondazionali trattate in questo libro. Le ricerche su Google e una scansione del catalogo di libri di O'Reilly dovrebbero rivelare le ultime tecnologie quando sarà pronto a mettersi al lavoro.

12.7 Implicazioni sociali ed etiche

La nostra capacità di metterci in guai seri aumenta con le dimensioni. Un'automobile può causare un incidente più grave di una bicicletta, e un aereo una carneficina più grave di un'automobile.

I big data possono fare grandi cose per il mondo, ma hanno anche il potere di danneggiare gli individui e la società in generale. Comportamenti che sono innocui su piccola scala, come lo scraping, diventano furto di proprietà intellettuale su larga scala. Descrivere l'accuratezza del suo modello in una luce eccessivamente favorevole è comune per le presentazioni PowerPoint, ma ha implicazioni reali quando il suo modello governa l'autorizzazione al credito o l'accesso alle cure mediche. Perdere l'accesso al proprio account di posta

elettronica è una mossa da idioti, ma non proteggere adeguatamente i dati personali di 100 milioni di clienti diventa potenzialmente criminale.

Concludo questo libro con una breve rassegna di preoccupazioni etiche comuni nel mondo dei big data, per aiutarla a sensibilizzarsi sul tipo di cose di cui il pubblico si preoccupa o dovrebbe preoccuparsi:

- *Integrità nella comunicazione e nella modellazione*: Lo scienziato dei dati funge da tramite tra la sua analisi e il suo datore di lavoro o il pubblico in generale. C'è una grande tentazione di far sembrare i nostri risultati più forti di quanto non in realtà, utilizzando una serie di tecniche collaudate nel tempo:
  - Possiamo riportare un livello di correlazione o di precisione, senza con una linea di base o riportare un *valore p*.
  - Possiamo scegliere tra più esperimenti e presentare solo i risultati migliori che otteniamo, invece di presentare un quadro più accurato.
  - Possiamo usare le tecniche di visualizzazione per oscurare le informazioni, invece di rivelarle.

All'interno di ogni modello ci sono presupposti e punti deboli. Un buon modellista sa quali i limiti del suo modello: cosa si fida che sia in grado di fare e dove inizia a sentirsi meno sicuro. Un modellista onesto comunica il quadro completo del suo lavoro: ciò che sa e ciò di cui non è sicuro.

I conflitti di interesse sono una vera preoccupazione nella scienza dei dati. Spesso sa qual è la "risposta giusta" prima dello studio, in particolare il risultato che il capo vuole sentire. Forse i suoi risultati saranno utilizzati per influenzare l'opinione pubblica, o per apparire in una testimonianza davanti ad autorità legali o governative. La segnalazione e la diffusione accurata dei risultati sono comportamenti essenziali per gli scienziati dei dati etici.

- *Trasparenza e proprietà*: In genere, le aziende e le organizzazioni di ricerca pubblicano le politiche di utilizzo e conservazione dei dati per dimostrare che ci si può fidare dei dati dei loro clienti. Questa trasparenza è importante, ma ha dimostrato di essere soggetta a cambiamenti non appena il valore commerciale dei dati diventa evidente. Spesso è più facile ottenere il perdono che il permesso.

  In che misura gli utenti sono proprietari dei dati che hanno generato? La proprietà significa che dovrebbero avere il diritto di vedere quali informazioni sono state raccolte su di loro, e la possibilità di impedire l'uso futuro di questo materiale. Queste questioni possono essere difficili, sia dal punto di vista tecnico che etico. Un criminale dovrebbe poter chiedere che tutti i riferimenti al suo crimine vengano eliminati da un motore di ricerca come Google? Mia figlia dovrebbe poter richiedere la rimozione delle immagini che la ritraggono pubblicate da altri senza il suo permesso?

  Gli errori nei dati possono propagarsi e danneggiare gli individui, senza consentire alle persone di accedere e capire quali informazioni sono state raccolte su di loro. Informazioni finanziarie errate o incomplete possono rovinare il rating di una persona, ma le agenzie di credito sono obbligate per legge a rendere disponibili i dati di ogni persona e a fornire un meccanismo per correggere gli errori. Tuttavia, la provenienza dei dati viene generalmente persa nel corso dell'unione dei file, per cui questi aggiornamenti non necessariamente arrivano a tutti i prodotti derivati che sono stati costruiti a partire da dati difettosi. Senza di essa, come possono i suoi clienti scoprire e correggere le informazioni errate che avete su di loro?

- *Decisioni non correggibili e cicli di feedback*: Impiegare i modelli come criteri di selezione rigidi può essere pericoloso, in particolare nei domini in cui il modello è solo un proxy di ciò che si vuole realmente misurare. La correlazione non è causale. Ma consideriamo un modello che suggerisce che è rischioso assumere un particolare candidato al lavoro perché le persone come lui che vivono in quartieri di classe inferiore hanno maggiori probabilità di essere arrestate. Se tutti i datori di lavoro utilizzano tali modelli,

queste persone semplicemente non verranno assunte e saranno spinte ancora di più nella povertà, senza alcuna colpa.

Questi problemi sono particolarmente insidiosi perché in genere non sono correggibili. La vittima del modello non ha in genere alcun mezzo di ricorso. E il proprietario del modello non ha modo di sapere cosa sta sbagliando, cioè quanti buoni candidati sono stati scartati senza ulteriore considerazione.

- *Pregiudizi e filtri basati su modelli*: I big data consentono di personalizzare i prodotti per adattarli al meglio a ogni singolo utente. Google, Facebook e altri analizzano i suoi dati in modo da mostrarle i risultati che i loro algoritmi pensano che lei voglia vedere.

  Ma questi algoritmi possono contenere pregiudizi involontari raccolti da algoritmi di apprendimento ma - chino su set di formazione dubbi. Forse il motore di ricerca mostrerà buone opportunità di lavoro agli uomini molto più spesso che alle donne, o discriminerà su altri criteri.

Mostrarle esattamente ciò che dice di voler vedere può impedirle di vedere le informazioni che ha veramente bisogno di vedere. Tali filtri possono avere una certa responsabilità nella polarizzazione politica della nostra società: vede punti di vista opposti o solo una camera d'eco per i suoi pensieri?

- *Mantenere la sicurezza delle grandi serie di dati*: I big data rappresentano un bersaglio più grande per gli hacker rispetto a un foglio di calcolo sul disco rigido. Abbiamo dichiarato che i file con 100 milioni di record *non* sono *nulla*, ma potrebbero rappresentare dati personali sul 30% della popolazione degli Stati Uniti. Le violazioni di dati di questa portata si verificano con una frequenza sconcertante.

Far cambiare la password a 100 milioni di persone costa 190 anni di lavoro sprecato, anche se ogni correzione richiede solo un minuto. Ma la maggior parte delle informazioni non può essere cambiata così rapidamente: gli indirizzi, i numeri di identificazione e le informazioni sul conto corrente rimangono per anni, se non per tutta la vita, rendendo il danno derivante dal rilascio in blocco dei dati impossibile da mitigare completamente.

Gli scienziati dei dati hanno l'obbligo di aderire pienamente alle pratiche di sicurezza delle loro organizzazioni e di identificare i potenziali punti deboli. Hanno anche la responsabilità di ridurre al minimo i pericoli di violazione della sicurezza attraverso la crittografia e l'anonimizzazione. Ma forse la cosa più importante è evitare di richiedere campi e record di cui non si ha bisogno e (questa è in assoluto la cosa più difficile da fare) cancellare i dati una volta che il progetto ne ha esaurito la necessità.

- *Mantenere la privacy nei dati aggregati*: Non è sufficiente eliminare nomi, indirizzi e numeri di identità per mantenere la privacy in un insieme di dati. Anche i dati anonimizzati possono essere efficacemente deanonimizzati in modi intelligenti, utilizzando fonti di dati ortogonali. Consideria mo il set di dati sui taxi che abbiamo presentato nella Sezione 1.6. Non ha mai contenuto alcuna informazione sull'identificazione dei passeggeri. Tuttavia, fornisce le coordinate GPS del prelievo a una risoluzione che potrebbe individuare una particolare casa come origine e un particolare locale di spogliarelli come destinazione. Ora abbiamo un'idea abbastanza buona di chi ha fatto quel viaggio e un'idea altrettanto buona di chi potrebbe essere interessato a queste informazioni, se l'uomo fosse sposato.

Un esperimento correlato ha identificato particolari corse di taxi prese da celebrità, in modo da capire la loro destinazione e l'entità della mancia [Gay14]. Utilizzando Google per trovare le fotografie dei paparazzi delle celebrità che salgono sui taxi ed estraendo l'ora e il luogo in cui sono state scattate, è stato facile identificare il record corrispondente a quell'esatto prelievo come contenente l'obiettivo desiderato.

I problemi etici nella scienza dei dati sono abbastanza seri che organizzazioni professionali si sono pronunciate sulle migliori pratiche, tra cui il *Codice di condotta professionale per la scienza dei dati* della Data Science Association e le *Linee guida etiche per le pratiche statistiche* dell'Associazione statistica americana. La invito a leggere questi documenti per aiutarla a sviluppare il suo senso delle questioni etiche e

degli standard di comportamento professionale. Ricordiamo che le persone si rivolgono agli scienziati dei dati per avere saggezza e consulenza, più che per il codice. Faccia il per dimostrare di essere degno di questa fiducia.

Capitolo 13

È giusto dire che esistono diversi tipi di lavoro legati alla scienza dei dati, che si distinguono per l'importanza relativa della conoscenza delle applicazioni e della forza tecnica. Vedo i seguenti percorsi di carriera di base legati alla scienza dei dati:

- *Ingegneria del software per la scienza dei dati*: Una parte sostanziale delle posizioni di sviluppo software di alto livello si trova nelle aziende di big data come Google,
  Facebook e Amazon, o aziende incentrate sui dati nel finanziario, come banche e hedge fund. Questi lavori ruotano attorno alla costruzione di infrastrutture software su larga scala per la gestione dei dati, e in genere richiedono una laurea in informatica per acquisire le competenze tecniche e l'esperienza necessarie.

- *Statistici/ scienziati dei dati*: C'è sempre stato un mercato del lavoro diversificato per gli statistici qualificati, soprattutto nei settori della sanità, dell'industria, dell'impresa, dell'istruzione e del governo/non profit. Questo mondo continuerà a crescere e a prosperare, anche se sospetto che richiederà maggiori competenze computazionali rispetto al passato. Questi analisti statistici orientati al calcolo avranno una formazione o un'esperienza nella scienza dei dati, partendo da una solida base di statistica.

- *Analisti aziendali quantitativi*: Un'ampia schiera di professionisti aziendali lavora nel marketing, nelle vendite, nella pubblicità e nella gestione, svolgendo funzioni essenziali in qualsiasi azienda di prodotti o di consulenza. Queste carriere richiedono un grado maggiore di conoscenza del dominio aziendale rispetto alle due categorie precedenti, ma richiedono sempre più competenze quantitative. Potrebbero assumerla per lavorare nel marketing, ma richiedere un background o un'esperienza nella scienza dei dati/analitica. Oppure la assumono per lavorare nelle risorse umane, ma si aspettano che sia in grado di sviluppare metriche per le prestazioni e la soddisfazione sul lavoro.