Appunti di Modelli Stocastici – Silvio Moioli

Definizioni

- Varianza: $V[X]=E[(X-E[X])^2]=E[X^2]-E[X]^2$;
- Deviazione standard: $STD[X]=\sqrt{V[X]}$;
- Covarianza: $COV[X,Y]=E[(X-E[X])(Y-E[Y])]=E[XY]-E[X]E[Y]$;
- Correlazione: $\rho[X,Y]=\frac{COV[X,Y]}{STD[X]STD[Y]}$;
- Autocovarianza di un processo stazionario a ritardo h: $\gamma(h)=COV[Y_t, Y_t+h]$ (nota che $\gamma(0)=V[Y]$);
- Autocorrelazione di un processo stazionario a ritardo h: $\rho(h)=\rho[Y_t, Y_{t+h}]=\frac{\gamma(h)}{\gamma(0)}$;
- Correlogramma: diagramma ritardo-autocorrelazione;
- Funzione indicatrice: $I(condizione)=\begin{cases}1\ se\ condizione\ è\ vera\\0\ se\ condizione\ è\ falsa\end{cases}$;

Stima di parametri (o identificazione)

Problema generale

Date n osservazioni casuali $y_1, y_2, \dots y_n$ provenienti da un modello $M(\theta^0)$ stimare il parametro ignoto θ^0 con uno **stimatore** $\hat{\theta}$ calcolato dalle osservazioni. Calcolare inoltre, se possibile, l'incertezza (varianza) dello stimatore, e degli intervalli di confidenza intorno ad esso.

- Uno stimatore è **corretto** quando vale $E[\hat{\theta}]=\theta^0$, altrimenti è **distorto** (e $E[\hat{\theta}]-\theta^0$ è la **distorsione**);
- uno stimatore è **consistente** quando vale $\lim_{n\to\infty}\hat{\theta}=\theta^0$ in senso stocastico, ossia $P(\hat{\theta}\neq\theta^0)\to 0$ per $n\to\infty$;
- uno stimatore è **ottimo** quando è quello a minima varianza (non esiste uno stimatore con varianza minore di quello ottimo);
- uno stimatore si dice **asintoticamente normale** quando vale $\sqrt{n}(\hat{\theta}-\theta^0)\sim N(0,V)$, dove V è un valore di varianza opportuno.

Nota: y_i , θ^0 e $\hat{\theta}$ possono essere vettoriali.

Metodologie

Minimi quadrati lineari

Idea

Dato un modello $M(\theta^0)$, calcolare $\hat{\theta}$ come il valore che minimizza la somma dei quadrati dei **residui** (differenze tra i valori prodotti dal modello e le osservazioni "vere"). Quindi la cifra di merito da minimizzare, convenzionalmente indicata con $Q(\theta)$, è quindi $\sum_{i=0}^{n} r_i^2$ dove r_i

sono i residui. Riassumendo:

$$\hat{\theta} = \underset{\theta}{argmin}(Q(\theta))$$

Applicabilità

- Modelli AR o ARX;
- Modelli lineari del tipo $y = \beta x + e$ in cui la variabile e è incorrelata dalle x. In altre parole, indicando con δ la correlazione tra x ed e $E[xe]$, modelli che rispettano la condizione $\delta = 0$;
- In particolare nel caso in cui $\delta = 0$ ed e è un rumore gaussiano bianco, $e\ iid\ N(\mu, \sigma^2)$, il metodo prende il nome di **regressione lineare**;

Validità

Nelle ipotesi di applicabilità, in particolare quando $\delta = 0$, lo stimatore non è distorto ed è ottimo.

Nel caso in cui, inoltre, $e\ iid\ N(\mu, \sigma^2)$ lo stimatore è anche asintoticamente normale.

Al contrario, se $\delta \neq 0$ (ad esempio se l'errore è autocorrelato, modelli con parte MA), possono verificarsi diversi casi:

- lo stimatore può perdere l'ottimalità ma rimanere corretto;
- lo stimatore può diventare distorto. In particolare si dimostra che la distorsione è pari a $\Sigma^{-1}\delta$, dove Σ è la matrice di varianze e covarianze dei dati.

Implementazione

Nelle ipotesi di applicabilità si dimostra che esiste unico il minimo della cifra di merito e vale la formula:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

o equivalentemente:

$$\hat{\beta} = (\sum_i x_i x_i^T)^{-1} \sum_i x_t y_t \quad .$$

Metodo dei Momenti

Idea

Sia $M(\theta^0)$ un modello parametrico rispetto a un parametro che può essere calcolato a partire dai primi n momenti della popolazione, ossia $\theta^0 = h(\mu_1, \mu_2, \ldots \mu_n)$ dove $\mu_1 \mu_2 \ldots \mu_n$ sono i primi momenti della popolazione ed h è una funzione nota. θ^0 si può stimare calcolando h con i momenti campionari ($m_1, m_2, \ldots m_n$) anziché con quelli della popolazione ($\mu_1, \mu_2, \ldots \mu_n$):

$$\hat{\theta} = h(m_1, m_2, \ldots m_n)$$

Applicabilità

- Modelli i cui parametri siano esprimibili come $h(\mu_1, \mu_2, \ldots \mu_n)$ con h nota e $\mu_j < \infty \ \forall\ j$;

Validità

Nelle ipotesi di applicabilità, lo stimatore è consistente poiché le quantità m_j sono stime consistenti delle corrispondenti μ_j (non è invece garantita a priori la correttezza).

Inoltre, se h è continua, derivabile e sviluppabile in serie in un intorno di μ, lo stimatore è anche asintoticamente normale (ciò dipende dal fatto che i momenti campionari m_j sono asintoticamente normali per il teorema del limite centrale, dunque se h è "abbastanza regolare" lo stimatore mantiene la proprietà). Si dimostra che:

$\hat{\theta} \sim N(\theta^0, JV[X]J^T)$ dove $J = \frac{\partial h(\mu)}{\partial \mu^T}|_{\mu}$.

Lo stimatore ottenuto, in generale, non è ottimo.

Implementazione

I momenti campionari si calcolano tramite la formula:

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

Massima Verosimiglianza

Idea

Sia $x_1, x_2, \ldots x_n$ un campione di dati indipendenti e identicamente distribuiti provenienti da uno stesso modello $M(\theta^0)$ e sia $f(x|\theta^0)$ la funzione di densità di probabilità relativa. La probabilità che si presenti esattamente il campione in esame in funzione del parametro è detta verosimiglianza ed è data da:

$L(x_1, x_2, \ldots x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ (per l'indipendenza delle variabili).

Fissato un campione, tale probabilità varia con il parametro del modello θ ed è massima in $\theta = \theta^0$. Si prende quindi come stimatore di θ^0 quel valore $\hat{\theta}$ che massimizza la verosimiglianza . Quindi:

$$\hat{\theta} = \underset{\theta}{argmax}(L(\theta))$$

Per semplicità di calcolo si usa normalmente la funzione di log-verosimiglianza, che ha gli stessi punti stazionari:

$l(\theta) = \log(L(\theta))$ quindi la cifra di merito è:

$Q(\theta) = l(\theta)$

e lo stimatore è:

$$\hat{\theta} = \underset{\theta}{argmax}(Q(\theta))$$

Applicabilità

- Modelli i cui campioni sono indipendenti, identicamente distribuiti e con distribuzione nota. Il metodo si può estendere, sotto opportune ipotesi, anche al caso di modelli a memoria corta come gli ARMAX con variabili non indipendenti o identicamente distribuite;

- In generale devono valere alcune ipotesi per la dimostrazione dei risultati di validità del metodo che in alcuni casi, a seconda del modello, possono anche essere rilassate:
  - il dominio di X deve essere θ -indipendente, ossia devo poter scambiare i simboli di derivata e integrale nei calcoli della log-verosimiglianza;
  - devono esistere su tutto il dominio le prime tre derivate della log-verosimiglianza;
  - le prime due derivate della log-verosimiglianza devono anche essere maggiorabili da una funzione g(x) integrabile, con integrale finito;
  - la terza derivata della log-verosimiglianza deve essere maggiorabile da una funzione H(x) che abbia valore atteso finito;
- l'**informazione di Fisher** $i(\theta)$ calcolata nell'intorno di θ^0 deve assumere valore finito e non nullo, ossia $0<i(\theta^0)<\infty$ (nel caso di θ vettoriale, $i(\theta^0)$ deve essere invertibile). L'informazione di Fisher è definita come la varianza dello **score,** $\frac{\partial}{\partial\theta}l(\theta)$, e rappresenta una misura dell'informazione su θ contenuta nei dati osservati. Si dimostra che $i(\theta)=V[\frac{\partial}{\partial\theta}l(\theta)]=-E[\frac{\partial^2}{\partial\theta^2}l(\theta)]$.

Validità

Si può dimostrare che l'algoritmo a massima verosimiglianza, così come gli algoritmi derivati da esso (minimi quadrati non lineari, EM) ha le seguenti proprietà:

- forniscono uno stimatore consistente: $\hat{\theta}\to\theta^0$ per $n\to\infty$;
- forniscono uno stimatore asintoticamente normale, il che può essere usato per costruire intervalli di confidenza: $\sqrt{(n)}(\hat{\theta}-\theta^0)\to N(0,\frac{1}{i(\theta^0)})$;
- forniscono uno stimatore ottimo. In particolare si dimostra che la varianza di un qualsiasi stimatore $\hat{\theta}$ corretto non può essere inferiore all'inverso dell'informazione di Fisher calcolata in θ^0 : $V[\hat{\theta}]\geq i(\theta^0)^{-1}$ (disuguaglianza di Rao-Cramer). Per quanto detto al punto precedente l'algoritmo a massima verosimiglianza raggiunge esattamente questo limite.

Implementazione

Sotto le ipotesi di applicabilità, $\hat{\theta}$ si calcola risolvendo l'equazione:

$\frac{\partial l(\theta)}{\partial\theta}|_{\hat{\theta}}=0$, ossia $l'(\hat{\theta})=0$.

Generalmente l'equazione non è risolvibile analiticamente, quindi occorre un algoritmo numerico. Una possibilità, illustrata nel caso di θ scalare, è di approssimare l'equazione con il suo sviluppo di Taylor al primo ordine nell'intorno di $\hat{\theta}$:

$l'(\theta)\approx l'(\hat{\theta})+(\hat{\theta}-\theta)l''(\hat{\theta})=0$ da cui si ottiene:

$\theta=\hat{\theta}-\frac{l'(\hat{\theta})}{l''(\hat{\theta})}$

che si può interpretare come passo elementare dell'iterazione per avvicinarsi a $\hat{\theta}$:

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{l'(\hat{\theta}_{t-1})}{l''(\hat{\theta}_{t-1})}$$

Fissati dei valori iniziali di $\hat{\theta}$ opportuni (ad esempio ricavati con altri metodi), ad ogni ciclo si calcola un nuovo $\hat{\theta}$ sulla base di quello precedente. L'iterazione si arresta quando il valore dello stimatore si è sufficientemente stabilizzato (la differenza tra i valori di iterazioni diverse è sufficientemente piccola). Questo algoritmo è detto di **Newton-Raphson**.

E' possibile, a causa delle approssimazioni effettuate, che il valore calcolato a una certa iterazione sia più lontano dal valore vero rispetto al precedente (fenomeno di **instabilità**). In questo caso, oppure nel caso in cui la derivata seconda non fosse calcolabile, è possibile usare una variante dell'algoritmo che definisce il passo di aggiornamento come segue (**metodo del gradiente**):

$$\hat{\theta}_t = \hat{\theta}_{t-1} - k\, l'(\hat{\theta}_{t-1})$$

con k costante opportuna. Un'altra variante, detta **scoring**, usa il valore atteso della derivata seconda:

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{l'(\hat{\theta}_{t-1})}{E[l''(\hat{\theta}_{t-1})]}$$

A seconda dei casi, le varianti descritte possono convergere più velocemente.

Minimi quadrati non lineari

Idea

Applicazione del metodo di massima verosimiglianza al caso in cui i campioni provengono da una sequenza di distribuzioni normali. Tali distribuzioni hanno tutte la stessa varianza, che si suppone nota, mentre la media varia in modo dipendente (anche non linearmente) dal parametro θ^0 :

$$x_i \, iid \, N(g_i(\theta^0), \sigma^2)$$

si ha

$$\hat{\theta} = \underset{\theta}{argmax}\,(l(\theta)) = \underset{\theta}{argmin}\,(-l(\theta))$$

$$l(\theta) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma\pi}} e^{\frac{(x_i - g_i(\theta))^2}{\sigma^2}}\right) = \log\left(\left(\frac{1}{\sqrt{2\sigma\pi}}\right)^n \prod_{i=1}^{n} e^{\frac{(x_i - g_i(\theta))^2}{\sigma^2}}\right)$$

$$= -\frac{n}{2}\log(2\sigma\pi) + \sum_{i=1}^{n} \log\left(e^{\frac{(x_i - g_i(\theta))^2}{\sigma^2}}\right) = -\frac{n}{2}\log(2\sigma\pi) + \sum_{i=1}^{n} \frac{(x_i - g_i(\theta))^2}{\sigma^2}$$

A parte il valore costante $-\frac{n}{2}\log(2\sigma\pi)$ e la costante moltiplicativa $\frac{1}{\sigma^2}$, l'oggetto da minimizzare è la forma quadratica $\sum_{i=1}^{n}(x_i - g_i(\theta))^2$, che posso prendere quindi come cifra di merito:

$$Q(\theta) = \sum_{i=1}^{n}(x_i - g_i(\theta))^2$$

Quindi:

$$\hat{\theta} = \underset{\theta}{argmin}\,(Q(\theta))$$

In altre parole massimizzare la verosimiglianza, in questo caso, equivale a minimizzare la somma dei quadrati degli scarti dalla media (da cui il nome del metodo). Questo include, come caso particolare, anche i minimi quadrati lineari quando g_i è lineare $\forall i$.

Applicabilità, validità e implementazione

Valgono le stesse considerazioni del metodo a massima verosimiglianza.

Algoritmo EM

Idea

Essenzialmente la stessa dell'algoritmo a massima verosimiglianza, ma si applica quando le osservazioni sono incomplete.

Più esattamente, l'algoritmo EM è un metodo per determinare stimatori a massima verosimiglianza nel caso in cui non tutto il campione sia osservabile: si suppone che occorra stimare il parametro θ^0 del modello che genera il campione $z_1, z_2, \dots z_n$ solo parzialmente noto. In particolare si suppone di conoscere alcune osservazioni dette **parziali** e denotate con $z_i^{(p)}$ e di ignorarne altre, dette **mancanti** e denotate con $z_i^{(m)}$. Detta f la densità di probabilità delle z_i si può calcolare la verosimiglianza del campione completo che dipenderà da $z_i^{(p)}$, $z_i^{(m)}$ e θ , ossia:

$$L(\theta) = f(z_i^{(p)}, z_i^{(m)} | \theta)$$

Come per la massima verosimiglianza, per semplicità di calcolo, si può definire la log-verosimiglianza:

$$l(\theta) = \log(L(\theta))$$

Idealmente, quindi, $l(\theta)$ sarebbe la cifra di merito da massimizzare. Il problema è che una parte del campione, $z_i^{(m)}$, non è osservabile quindi $l(\theta)$ non è calcolabile. Ci si "accontenta", quindi, di usare come cifra di merito il valore atteso di $l(\theta)$ anziché il suo valore vero. Tale valore atteso deve essere calcolato al variare dei valori che i dati mancanti $z_i^{(m)}$ possono prendere:

$$E_{z_i^{(m)}}[l(\theta)] = \int_{z_i^{(m)}} f(z_i^{(m)}) l(\theta) d\, z_i^{(m)}$$

Questo valore atteso, però, ha due difetti: primo, è difficile da calcolare analiticamente e secondo, non tiene conto del fatto che le osservazioni $z_i^{(p)}$ si sono effettivamente presentate. La cifra di merito può essere quindi ulteriormente "raffinata" condizionandola ai valori osservati di $z_i^{(p)}$:

$$Q(\theta) = E_{z_i^{(m)}}[l(\theta) | z_i^{(p)}]$$

La cifra di merito dell'algoritmo EM è quindi *il valore atteso della log-verosimiglianza calcolato su tutti i valori assumibili dai dati mancanti, dato che si sono osservati i dati* $z_i^{(p)}$. Si dimostra che, nonostante l'apparente complessità, è relativamente facile trovare espressioni in forma analitica e facilmente ottimizzabile di questa formula.

Ciò detto, lo stimatore si calcola con la formula usuale:

$$\hat{\theta} = \underset{\theta}{argmax}(Q(\theta))$$

Applicabilità

- Modelli in cui è applicabile la massima verosimiglianza;
- Modelli con dati parzialmente osservabili;
- Modelli in cui sia facile calcolare $Q(\theta)$ e massimizzarla;
- Modelli in cui sia stimabile un valore iniziale "buono" di θ (questo occorre perché l'implementazione funzioni, vedi oltre);

Validità

Nelle ipotesi di applicabilità, valgono le stesse considerazioni del metodo a massima verosimiglianza.

Implementazione

L'implementazione è iterativa. Infatti, per calcolare l'integrale che compare nella cifra di merito:

$$Q(\theta)=E_{z_i^{(m)}}[l(\theta)|z_i^{(p)}]=E_{z_i^{(m)}}[\log(f(z_i^{(p)},z_i^{(m)}|\theta))|z_i^{(p)}]=\int_{z_i^{(m)}} f(z_i^{(m)}|z_i^{(p)})\log(f(z_i^{(p)},z_i^{(m)}|\theta))\,d\,z_i^{(m)}$$

occorre calcolare i valori di $f(z_i^{(m)}|z_i^{(p)})$ i quali implicitamente dipendono dal valore di θ^0. Più esattamente si può quindi indicare $f(z_i^{(m)}|z_i^{(p)})$ con $f(z_i^{(m)}|z_i^{(p)},\theta^0)$ e la cifra di merito con $Q(\theta,\theta^0)$, per sottolineare la dipendenza dal parametro "vero" che ha generato i dati.

Ovviamente θ^0 non è noto (è proprio quello che si vorrebbe stimare) quindi l'unica possibilità è calcolare Q con una stima $\hat{\theta}_{t-1}$ che si suppone in qualche modo nota, ossia stimare $Q(\theta,\theta^0)$ con $Q(\theta,\hat{\theta}_{t-1})$. Ragionando ricorsivamente si può impostare un algoritmo che ad ogni iterazione t produce un nuovo e più accurato $\hat{\theta}_t$ utilizzando la stima dell'iterazione precedente $\hat{\theta}_{t-1}$:

$$\hat{\theta}_t=\underset{\theta}{argmax}\,Q(\theta,\hat{\theta}_{t-1})$$

Si può dimostrare che ogni iterazione dell'algoritmo EM produce un $\hat{\theta}_t$ effettivamente più vicino all'ottimo rispetto al precedente (proprietà di **monotonicità**), e in particolare l'algoritmo non presenta mai fenomeni di oscillazioni o instabilità (a differenza dell'algoritmo a massima verosimiglianza). Sorgono però alcuni problemi:

- l' inizializzazione: occorre un valore iniziale $\hat{\theta}_0$ da cui partire. E' anche necessario che sia un "buon valore", in caso contrario l'algoritmo può convergere a un minimo locale o in un punto di sella della funzione Q;
- la lentezza: in alcuni casi l'algoritmo produce stimatori poco distanti l'uno dall'altro, richiedendo quindi un gran numero di iterazioni (esistono varianti più veloci).

Algoritmo di Baum-Welch

Idea

Applicazione dell'algoritmo EM per la stima dei parametri di un modello markoviano latente (vedi oltre per la notazione e i risultati utilizzati in questo paragrafo). Le variabili osservabili $z_i^{(p)}$ corrispondono alle y_i del modello markoviano e le variabili non osservabili $z_i^{(m)}$ corrispondono allo stato nascosto x_i. Si suppone di avere a disposizione un totale di n osservazioni.

I parametri da stimare sono:

1. le **probabilità iniziali** di trovarsi in ogni stato $P(X_0 = x_i)$;
2. le **probabilità di transizione** da uno stato i a uno stato j in due istanti di tempo successivi $P(X_{t+1} = x_j | X_t = x_i)$;
3. il **modello dei sensori**, ossia la probabilità di osservare y_i dato che lo stato in quell'istante è x_j $P(Y_t = y_i | X_t = x_j)$.

Si può dimostrare che le tre quantità hanno stimatori in forma chiusa che massimizzano la cifra di merito $Q(\theta) = E_{x_{0:n}}[l(\theta)|y_{1:n}]$ e che possono essere calcolati a partire dalle probabilità forward e backward.

Applicabilità

Modelli markoviani in cui si abbia una ragionevole stima iniziale del parametro $\hat{\theta}$, in particolare per cui si ha un'idea di quali e quanti potrebbero essere gli stati. Teoricamente questa conoscenza non è strettamente necessaria, ma in pratica è richiesta per la convergenza numerica dell'algoritmo.

Validità

Valgono considerazioni analoghe a quelle per l'algoritmo EM.

Implementazione

Gli stimatori che massimizzano la cifra di merito $Q(\theta) = E_{x_{0:n}}[l(\theta)|y_{1:n}]$ sono:

- $\hat{P}(X_{t+1} = x_j | X_t = x_i) = \frac{\hat{n}_{x_i, x_j}}{\hat{n}_{x_i}}$ dove $\hat{n}_{x_i, x_j}$ è la stima del numero di passaggi dallo stato x_i allo stato x_j in due istanti successivi, e $\hat{n}_{x_i}$ è la stima del del numero di istanti in cui lo stato è x_i ;
- $\hat{P}(Y_t = y_i | X_t = x_j) = \frac{\hat{m}_{y_i, x_j}}{\hat{n}_{x_i}}$ dove $\hat{m}_{y_i, x_j}$ è la stima del numero di volte in cui si è osservato y_i quando lo stato era x_j .

Non siamo interessati, al momento, alla stima di $P(x_0)$ quindi restano da calcolare gli stimatori $\hat{n}_{x_i, x_j}$, $\hat{n}_{x_i}$ ed $\hat{m}_{y_i, x_j}$.

$\hat{n}_{x_i, x_j}$ può essere stimato con il suo valore atteso calcolato come segue:

$$\hat{n}_{x_i, x_j} = \sum_{t=0}^{n-1} P(X_{t+1} = x_j, X_t = x_i | Y_{1:n} = y_{1:n})$$

La probabilità contenuta nella sommatoria di passare da x_i a x_j in istanti successivi date le osservazioni può essere espressa come segue:

$$P(X_{t+1} = x_j, X_t = x_i | Y_{1:n} = y_{1:n}) =$$
$$= P(X_t = x_i | Y_{1:t} = y_{1:t}) \; P(X_{t+1} = x_j | X_t = x_i) \; P(Y_{t+1} = y_{t+1} | X_{t+1} = x_j) \; P(Y_{t+1:n} = y_{t+1:n} | X_{t+1} x_j)$$

Ossia, è la probabilità che avvengano i quattro eventi aventi le seguenti probabilità:

1. $P(X_t = x_i | Y_{1:t} = y_{1:t})$: essere al tempo t nello stato x_i , date tutte le osservazioni fino a t. La probabilità è calcolabile dividendo la probabilità forward per la verosimiglianza delle

osservazioni fino a t: $P(X_t=x_i|Y_{1:t}=y_{1:t})=\frac{P(X_t=x_i, Y_{1:t}=y_{1:t})}{P(Y_{1:t}=y_{1:t})}$;

2. $P(X_{t+1}=x_j|X_t=x_i)$: passare dallo stato x_i allo stato x_j . La probabilità è proprio la probabilità di transizione che si cerca di stimare. Ciò non deve sorprendere, analogamente al caso dell'algoritmo EM originale questa probabilità si può stimare con il valore ottenuto al passo precedente;
3. $P(Y_{t+1}=y_{t+1}|X_{t+1}=x_j)$: osservare y_{t+1} dato che lo stato è x_j , valgono le stesse considerazioni del punto precedente;
4. $P(Y_{t+1:n}=y_{t+1:n}|X_{t+1}=x_j)$: osservare tutte le altre y_i fino a n dato che lo stato è x_j , la probabilità coincide con la probabilità backward.

Dal momento che i quattro eventi sono tutti indipendenti, la probabilità $P(X_{t+1}=x_j, X_t=x_i|y_{1:n})$ si può calcolare per semplice prodotto, e da quella calcolare $\hat{n}_{x_i,x_j}$ per somma su t.

Noto $\hat{n}_{x_i,x_j}$ si può ottenere $\hat{n}_{x_i}$ sommando $\hat{n}_{x_i,x_j}$ su tutte le x_j .

L'ultimo stimatore $\hat{m}_{y_i,x_j}$ può essere calcolato come:

$$\hat{m}_{y_i,x_j}=\sum_{t|Y_t=y_i} P(X_t=x_j|Y_{1:n}=y_{1:n})$$

ma

$$P(X_t=x_j|Y_{1:n}=y_{1:n})=P(X_t=x_i|Y_{1:t}=y_{1:t})\,P(Y_{t+1:n}=y_{t+1:n}|X_t=x_i)$$

che è il prodotto della probabilità forward per la backward diviso per la verosimiglianza delle osservazioni fino a t, analogamente ai punti 1 e 4 del caso precedente.

Stima bayesiana

Idea

Si suppone che, a differenza dei casi precedenti, θ non sia un valore fissato ma una variabile casuale con distribuzione generica e nota $G(\theta)$, con densità di probabilità $f_G(\theta)$. Si vuole quindi ottenere una cifra di merito per stimare θ che tenga conto dei seguenti due fatti:

1. il parametro θ è casuale e
2. la forma di G è nota.

Si può manipolare la scrittura della verosimiglianza dei dati $x_{1,}x_{2,}\dots x_n$, ossia

$L(\theta)=f(x_{1,}x_{2,}\dots x_n|\theta)$, utilizzando la formula di Bayes ed il teorema delle probabilità totali:

$$P(B_i|A)=\frac{P(A|B_i)\,P(B_i)}{P(A)}=\frac{P(A|B_i)\,P(B_i)}{\sum_j P(A|B_j)\,P(B_j)}$$

scrivendo:

$$f(\theta|x_{1,}x_{2,}\dots x_n)=\frac{f(x_{1,}x_{2,}\dots x_n|\theta)\,f_G(\theta)}{\int_{-\infty}^{+\infty} f(x_{1,}x_{2,}\dots x_n|\theta')\,f_G(\theta')\,d\theta'}$$

Quest'ultima formula rappresenta la probabilità che il parametro, che è una variabile casuale, abbia assunto valore θ dato che si è presentato il campione $x_{1,}x_{2,}\dots x_n$ ed è indicata per definizione

con $h(\theta)$ o **probabilità a posteriori**. La scrittura tiene conto dei due fatti di cui sopra, in particolare suppone che $f_G(\theta)$ sia calcolabile e viene assunta come cifra di merito per la stima bayesiana, quindi:

$$Q(\theta) = h(\theta)$$

$$\hat{\theta} = \underset{\theta}{argmax}\, Q(\theta)$$

Validità

Si può dimostrare che, per $n \to \infty$, la stima bayesiana converge alla stima di massima verosimiglianza. Questo dipende dal fatto che la stima bayesiana tiene conto sia dei dati che della conoscenza di G ma, per campioni molto grandi, quest'ultima diventa sempre meno importante perché i dati contengono già tutta l'informazione necessaria.

Metodo Monte Carlo

Idea

Siano $y_i = f(x_i)$ variabili generate dall'applicazione di una funzione f nota e facilmente computabile ad una sequenza di variabili x_i che hanno modello $M(\theta^0)$ noto (x_i e y_i possono essere vettoriali). Si vogliono calcolare alcune proprietà di Y quali il valore atteso, la varianza, la deviazione standard, i percentili e così via ma f è troppo complessa per essere trattata analiticamente.

Il metodo Monte Carlo permette di stimare queste grandezze simulando un numero elevato di osservazioni x_i^* ($*$ sta per "simulato"), calcolando f per ognuno di essi e quindi delle statistiche complessive.

Applicabilità

Ogniqualvolta f è computabile (ad esempio un modello computazionale), ma troppo complessa per essere studiata analiticamente. E' tanto più facile applicare il metodo ed ottenere risultati precisi quanto più f è efficientemente calcolabile e quanta più potenza di calcolo si ha a disposizione.

x è una variabile casuale e deve essere possibile simularla, quindi il suo modello deve essere noto. Anche qui valgono le stesse considerazioni di efficienza computazionale.

Un caso particolare di uso di questa tecnica è la stima parametrica, ossia quando f è uno stimatore del parametro θ^0 del modello di x, quindi y coincide con $\hat{\theta}$. Il metodo può essere applicato per ottenere informazioni aggiuntive sullo stimatore quale, ad esempio, la varianza.

Validità

Il metodo Monte Carlo può fornire intervalli di confidenza sui risultati ottenuti e generalmente è possibile aumentarne arbitrariamente la precisione (a patto di avere sufficiente potenza o tempo di elaborazione).

Implementazione

1. Si simulano n variabili x_i^* ;
2. Per ciascuna delle n variabili, si calcola il valore $y_t^* = f(x_t^*)$.

3. A questo punto $E[Y_t] \approx \frac{1}{n}\sum_{i=1}^{n} y_t^*$, $V[Y_t] \approx \frac{1}{n}\sum_{i=1}^{n} (y_t^* - E[Y_t])^2$,

   $F(y) \approx \frac{1}{n}\sum_{i=1}^{n} I(y_i \le y)$ (funzione di ripartizione), e così via;

4. Dato che le y_t sono indipendenti e identicamente distribuite per m sufficientemente grande vale il teorema centrale del limite, quindi si possono creare intervalli di confidenza e test d'ipotesi sui valori stimati;
5. Si possono costruire anche degli intervalli di confidenza empirici: ordinando i risultati in ordine crescente, si ottiene il vettore $y_1^* \le y_2^* \le ... y_n^*$ che può essere usato per stimare i percentili (ad esempio se n=1000 e si vuole costruire un intervallo bilaterale centrato in 0 per $\alpha = 95$ si ha $\frac{1-\alpha}{2} = 2{,}5\%$ quindi l'intervallo sarà tra y_{25}^* e y_{975}^*).

Bootstrap parametrico

Idea

Sia y=f(x) una variabile generata dall'applicazione di una funzione f. Sia la distribuzione di X nota nella forma ma non nei parametri, ad esempio sia la funzione di ripartizione pari a $F(x|\theta^0)$. Si vogliono calcolare alcune proprietà di Y quali il valore atteso, la varianza, la deviazione standard come nel caso dei metodi Monte Carlo.

L'idea è di stimare θ^0 con un'altra delle tecniche descritte, e poi applicare il metodo Monte Carlo su $F(x|\hat{\theta})$.

Applicabilità e validità

Valgono considerazioni analoghe a quelle del metodo Monte Carlo.

Bootstrap non parametrico

Idea

Siano $x_1, x_2, ... x_n$ n osservazioni di processi casuali (ogni x_i è un vettore di k elementi $x_{i,1}, x_{i,2}, ... x_{i,t}, ... x_{i,k}$) provenienti da uno stesso modello $M(\theta^0)$ ignoto nella forma e nel parametro. Sia inoltre f una funzione nota e computabile e siano y_i le variabili generate dall'applicazione di tale funzione alle osservazioni, ossia $y_i = f(x_i)$.

Si vogliono calcolare alcune proprietà di Y quali il valore atteso, la varianza, la deviazione standard e così via come nei casi precedenti. L'idea è di "simulare" m nuovi campioni x_i^* estraendoli a caso dagli n x_i noti. A questo punto si possono costruire le corrispondenti y_i^* applicando f, e queste saranno informative di Y (ossia saranno iid Y). Da queste si potrà stimare media, varianza, percentili e così via con le stesse considerazioni viste per il metodo Monte Carlo.

Applicabilità e validità

Valgono considerazioni analoghe a quelle del metodo Monte Carlo, a parte i seguenti fatti:

- m ed n dovrebbero essere grandi, teoricamente tendenti a infinito;
- non è necessario conoscere la distribuzione di X né quella di Y a priori.

Implementazione

L'unica questione non affrontata nei precedenti paragrafi è quella dell'estrazione delle variabili simulate dall'insieme originale, con reintroduzione. Il problema si risolve generando un indice casuale tra 1 ed n, il che può essere ricondotto al calcolo: $i = nu$ dove u è una realizzazione di una variabile casuale uniforme tra 0 e 1, che è relativamente facile da implementare su un calcolatore.

Metodo Monte Carlo stratificato

Idea

Nelle stesse ipotesi del metodo Monte Carlo e considerando il caso con x_i monodimensionali, si vuole creare uno stimatore del valore atteso di Y più efficiente del metodo originale, ossia a varianza inferiore. Il risultato pratico è che occorreranno meno simulazioni, quindi meno tempo di elaborazione, per avere una buona stima della media.

Per far questo si agisce sul **campionamento**, ossia il procedimento di estrazione delle x_i^* , in modo che "contengano più informazione" sulla media, che verrà quindi stimata meglio.

Laddove il metodo Monte Carlo adotta un **campionamento casuale semplice**, ossia estrae i valori di x_i^* da tutto il dominio, questa variante implementa un **campionamento stratificato**, ossia prende i valori da una partizione in intervalli equiprobabili del dominio avendo cura che ogni intervallo venga utilizzato un numero uguale di volte. In questo modo si "ha maggiore confidenza" che anche le parti meno probabili (come le code) della distribuzione siano equamente rappresentate nel campione simulato.

Applicabilità

Valgono le stesse considerazioni fatte nel metodo Monte Carlo, ma:

- si vuole stimare solo $E[Y]$ e
- le x_i sono monodimensionali.

Validità

La stima del valore atteso calcolata da questo metodo, $\hat{E}[Y]_{MCS}$, è corretta ed ha varianza inferiore rispetto a quella del metodo Monte Carlo, $\hat{E}[Y]_{MC}$. Infatti si può dimostrare che:

$$V(\hat{E}[Y]_{MC}) = V(\hat{E}[Y]_{MCS}) \ddot{u} \frac{1}{n} V_C[E[f(x)|C_i]] \quad .$$

Implementazione

Per semplicità si espone il caso in cui le x_i siano monodimensionali, anche se questo non è limitante per l'algoritmo.

Inizialmente si sceglie il numero s di sotto-intervalli in cui dividere il dominio, e si calcolano gli intervalli stessi. In particolare, detta F la funzione di ripartizione delle x_i , si calcolano i valori:

$$l_{j,d} = F^{-1}\left(\frac{j-1}{s}\right)$$

$$l_{j,s} = F^{-1}\left(\frac{j}{s}\right)$$

per $1 < j < s,\ j \in \mathbb{N}$

che rappresentano rispettivamente i limiti destro e sinistro del j-esimo sotto-intervallo. Le formule sono giustificate dal fatto che questi intervalli devono essere equiprobabili, ossia deve valere la relazione $F(l_{j,d}) - F(l_{j,s}) = \frac{1}{n}$ per ogni j. Chiamiamo questi intervalli C_j, ossia $C_j = [l_{j,s}, l_{j,d}]$.

Il campione viene generato iterativamente: per ogni intervallo (j che va da 1 ad s) si estrae casualmente un valore x_i, con distribuzione $f(x_i|C_j)$ (il condizionamento è dovuto al fatto che si è scelto a priori un intervallo); raggiunto l'intervallo s-esimo, si ricomincia da 1. In particolare, x_i si può calcolare come:

$$x_i = F^{-1}\left(\frac{j-1}{s} + \frac{u}{s}\right)$$

dove u è la realizzazione di una variabile casuale uniforme tra 0 e 1.

A questo punto la stima di $E[Y]$, che si denota con $\hat{E}[Y]_{MCS}$ si può calcolare come:

$$\hat{E}[Y]_{MCS} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Ipercubi latini

Idea

Questo metodo è un'altra variante del metodo Monte Carlo e, come nel caso precedente, agisce variando il procedimento di campionamento delle x_i^*. In questo caso si opera un **campionamento a ipercubi latini**, che è un'estensione a k dimensioni del concetto di stratificazione.

Laddove il campionamento stratificato estrae un valore da ognuno degli s intervalli in cui il dominio monodimensionale è suddiviso, nel caso del **quadrato latino** vengono estratti s valori in modo tale che ogni riga ed ogni colonna della matrice $s \times s$ in cui il dominio bidimensionale è suddiviso abbia uno e un solo valore.

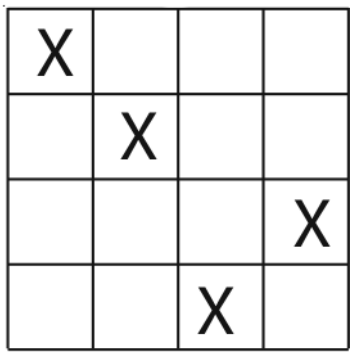

Figura 1: quadrato latino. Il dominio è suddiviso in intervalli bidimensionali equiprobabili e viene preso un campione per ogni riga e per ogni colonna.

Analogamente, si può suddividere un dominio a k dimensioni in s^k piccoli ipercubi e campionare in modo tale che sia estratto uno e un solo valore ognuno per ognuno dei k "assi" (sottospazi) realizzando appunto un **ipercubo latino**.

Applicabilità

Valgono le stesse considerazioni fatte nel metodo Monte Carlo, ma si vuole stimare solo $E[Y]$.

Validità

Si dimostra che $V[\hat{E}_{LHS}[Y]] \leq V[\hat{E}_{MC}[Y]]$, dove LHS sta per ipercubi latini, e in particolare se f è monotona in almeno k-1 dimensioni vale la disuguaglianza stretta.

Implementazione

L'implementazione è analoga al caso stratificato, con l'eccezione dell'estrazione delle variabili che è svolta prendendo una permutazione di s elementi per ogni dimensione. Questo assicura che non esistano due campioni su una stessa dimensione e che tutte le dimensioni siano correttamente rappresentate.

Validazione

Problema generale

Si supponga di aver stimato il parametro $\hat{\theta}$ del modello di un modello $M(\theta^0)$ che ha generato le osservazioni $y_1, y_2, \dots y_n$. Si vogliono definire dei test per assicurare che il modello e il parametro siano buoni.

Metodologie

Verifica del rumore bianco (o test di bianchezza)

Se una serie di dati proviene da un rumore bianco, il correlogramma dovrebbe avere forma di un impulso: $\rho(0)=1$ mentre $\rho(h)=0$ per $h>0$. Posso calcolare l'autocorrelazione campionaria:

$r(h)=\frac{\sum_{t=1}^{n-h}(x_t-\bar{x})(x_{t+h}-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}$ (per grandi campioni). Se n è elevato e i dati sono iid,

$r(h) \sim N(0, \frac{1}{n-h})$, quindi posso pensare di costruire un intervallo di confidenza attorno a ogni valore di r(h): riunendoli in un grafico trovo una "banda di accettazione" che si allarga con h.

Volendo invece fare una prova unica, posso usare il test di Ljung-Box che usa la seguente statistica Q: $Q=\sum_{h=1}^{H}(n-h)r(h)^2 \sim X^2_H$

Intervalli di confidenza sui parametri della regressione lineare

Intervalli di confidenza sui parametri: si possono costruire nota la matrice delle varianze $V[\hat{\beta}]=\sigma_\epsilon^2(X^T X)^{-1}$, dove σ_ϵ^2 è la varianza dell'errore del modello, che si può stimare con:

$$\sigma_\epsilon^2 \approx s_\epsilon^2 = \frac{1}{n(k+1)} \sum_i e_i^2$$

Tabella ANOVA: analisi della composizione delle varianze (devianze).

$$D_{tot} = \sum_{i=1}^{k} (y_i - \bar{y})^2$$

$$D_{res} = \sum_{i=1}^{k} (y_i - \hat{\beta} x_i)^2$$ (nota: la regressione tenta di minimizzare questo parametro)

$$D_{sp} = \sum_{i=1}^{k} (\hat{\beta} x_i - \bar{y})^2$$

$D_{tot} = D_{sp} + D_{res}$ (quando è valida)

Da cui il parametro di valutazione $R^2 = \frac{D_{sp}}{D_{tot}} = 1 - \frac{D_{res}}{D_{tot}}$ (per grandi campioni). $R \approx 1$ indica che il modello è buono se $n \gg k$; se $n \approx k$ rischio di fare "overfitting" (interpolazione esatta dei punti).

Supponendo che il modello sia corretto, ci si chiede se le variabili esplicative x_i scelte siano adeguate. Se non lo fossero, i corrispettivi coefficienti β_i risulterebbero vicini a zero, è possibile a questo punto eseguire un test d'ipotesi sulla loro nullità tramite la statistica t. I software lo fanno automaticamente, e possono produrre due tipi di output: una tabella con i valori del test t oppure un'equazione del tipo:

$$y = \beta_1(\pm\sigma_1)x_1 + \beta_2(\pm\sigma_2)x_2 + ...\epsilon(\pm\sigma_{res})$$

Se una σ_i è più della metà della corrispondente β_i la variabile è poco significativa.

Analisi di sensitività

Problema generale

Si supponga di avere elaborato un modello di un fenomeno $Y = f(X)$, con f funzione computabile, X ed Y variabili casuali eventualmente vettoriali. Si suppone inoltre che f sia troppo complessa per essere studiata analiticamente, e quindi occorrano metodi statistici. Ci si pone il problema di capire quanto Y è sensibile rispetto alle variazioni di X, ad esempio comprendendo quali sono le variabili rispetto a cui Y è più sensibile, negli intorni di quali punti, quali sono le fonti di incertezza del modello e così via.

Metodologie

Verifica dell'incertezza complessiva del modello

Il modo più grossolano per valutare la sensitività di f è di calcolare la varianza di Y. La varianza infatti dà una stima complessiva dell'incertezza anche se non permette di capirne le cause. Il calcolo è quello usuale:

$$V(Y) = \int_{\Omega} (f(x) - E[f(X)])^2 dx$$

dove Ω è il dominio di X.

Verifica della variabilità locale

Un altro modo per valutare la variabilità di f, in particolare nell'intorno di un punto x_0, è di calcolare la derivata parziale di f rispetto a x in quel punto ($\frac{\partial f(x)}{\partial x}|x_0$). Il risultato è la "pendenza" della retta tangente la funzione in quel punto, e quindi dà un'idea di quanto velocemente può cambiare per piccoli scostamenti di x.

Verifica della variabilità globale

Analogamente al paragrafo precedente, si può calcolare la derivata e studiarne l'andamento su tutto il dominio ($\frac{\partial f(x)}{\partial x}$), per farsi un'idea complessiva della sensitività della funzione.

Indice di sensitività

Un modo un po' più raffinato per capire le cause di un certo livello di variabilità si può ricavare dalla legge di scomposizione della varianza, la quale afferma che per ogni coppia di variabili casuali X ed Y vale:

$$V[Y]=V[E[Y|X]]+E[V[Y|X]]$$

Nella formula, $E[Y|X]$ e $V[Y|X]$ sono variabili casuali (dipendono dal valore assunto da X) e indicano rispettivamente il valore atteso di Y fissato un valore per X e la sua varianza. In particolare quindi, il termine $V[E[Y|X]]$ assume il significato di variabilità di Y cambiando tutti i possibili valori di X. Quindi si può dire che questo termine sia *la parte della variabilità di Y che dipende dal valore assunto da X*, e per questa ragione prende il nome di **varianza spiegata**. L'altra parte, di contro, si chiama varianza **non spiegata** o **residua**.

Supponendo di aver calcolato o stimato il valore dei tre oggetti, si può definire l'**indice di Pearson**:

$$S_j=\frac{V[E[Y|X]]}{V[Y]}$$

che è la percentuale della varianza di Y che dipende da X e può essere usato per quantificare la sensitività di f.

Analisi di un emulatore

E' anche possibile, qualora sia troppo difficile studiare f, ricorrere a un emulatore g, ossia un'altra funzione più semplice (ad esempio lineare) che la approssima. Chiaramente questo aggiunge incertezza alla stima della sensitività (poiché g non è esattamente f e, magari, va anch'essa stimata) ma può essere utile per ridurre il carico computazionale.

Nel caso particolare di g lineare in x e in un parametro β con errore rappresentato da rumore gaussiano bianco, si ha:

$$y\approx x\beta+\beta_0$$

che è esattamente il caso della regressione lineare. Quindi, si può utilizzare lo stesso coefficiente $R^2=\frac{D_{sp}}{D_{tot}}=1-\frac{D_{res}}{D_{tot}}$ per valutare quanto y è spiegata da x (vedi "Minimi quadrati lineari" e "Intervalli di confidenza sui parametri della regressione lineare").

Inferenza sui modelli con variabili non osservabili

Problema generale

Si suppone di avere a disposizione delle osservazioni casuali $y_{1,} y_{2,} \cdots$, provenienti da un modello $M(\theta^0)$ noto di parametro θ^0 anch'esso noto; ad ogni istante di tempo si abbia a disposizione una nuova osservazione. Per semplicità di notazione, si indicano la sequenza di variabili consecutive dall'istante 1 all'istante corrente t con $y_{1:t}$. Si suppone inoltre che le osservazioni y_i dipendano stocasticamente dal valore delle variabili $x_{0,} x_{1,} \ldots$ le quali, però, non sono osservabili.
Si vogliono effettuare le seguenti operazioni:

- il **filtraggio**, ossia la stima dello stato corrente x_t date tutte le osservazioni fatte fino a quel momento $y_{1:t}$. Il problema può anche essere ricondotto al calcolo di $P(X_t = x_t | Y_{1:t} = y_{1:t})$: note le probabilità si stima x_t per massimizzazione. Per semplicità di notazione, si indicano le probabilità senza indicare esplicitamente le variabili casuali se non ci sono ambiguità, in questo caso $P(X_t = x_t | Y_{1:t} = y_{1:t})$ si denota con $P(x_t | y_{1:t})$;
- la **predizione**, ossia la stima dello stato futuro x_{t+k} date tutte le osservazioni fatte fino a quel momento $y_{1:t}$. Come sopra, il problema può essere formulato nei termini del calcolo di $P(x_{t+1} | y_{1:t})$;
- lo **smoothing**, ossia la stima di uno stato passato x_{t-k} date tutte le osservazioni fatte fino a quel momento $y_{1:t}$. Anche qui il problema può essere formulato nei termini del calcolo di $P(x_{t-k} | y_{1:t})$;
- il calcolo della **verosimiglianza delle osservazioni** ottenute, ossia $P(y_{1:t})$;
- la stima della **sequenza di stati a massima verosimiglianza**, ossia la sequenza di stati che più probabilmente ha portato alla sequenza di osservazioni ottenute $\underset{x_{1:t}}{argmax}\, P(y_{1:t} | x_{1:t})$. Come sopra il problema è riconducibile al calcolo della probabilità $\underset{x_{1:t}}{max}\, P(y_{1:t} | x_{1:t})$;

Metodologie

Modelli markoviani latenti

Idea

Un modello si dice **markoviano** (o **catena markoviana**) se rispetta le seguenti ipotesi:

1. le realizzazioni (uscite) del modello appartengono a un insieme finito e noto S. n_S è la sua cardinalità;
2. la probabilità che un uscita assuma un certo valore dipende solo dall'uscita immediatamente precedente (ipotesi di **markovianità di ordine 1**).

Un modello markoviano si dice **omogeneo** se le probabilità di cui al punto 2 non cambiano nel tempo (questa ipotesi si considera sempre verificata).

Si dice che un modello è **markoviano latente** se le sue uscite dipendono dalle uscite di una catena markoviana che non è osservabile. Si denotano con x_t le uscite non osservabili della catena, dette **stati** del sistema, e con y_t le uscite osservabili del modello markoviano latente. La definizione può essere precisata con le seguenti ipotesi:

1. in ogni istante il modello si trova in uno stato x_t non osservabile generato da una catena

markoviana, ossia il numero di stati è finito e la probabilità di trovarsi in uno stato x_t dipende solo dallo stato immediatamente precedente: $P(x_t|x_{t-1}, x_{t-2}, \dots x_0) = P(x_t|x_{t-1})$.

2. la probabilità di ottenere una certa uscita y_t dipende esclusivamente dallo stato in quell'istante $P(y_t|x_t, x_{t-1}, \dots x_0, y_{t-1}, y_{t-1}, \dots y_0) = P(y_t|x_t)$.

Il modello è completamente definito dalla probabilità congiunta totale $P(x_{0,} x_{1,} \dots y_{1,} y_{2,} \dots)$. Si dimostra che nelle ipotesi di cui sopra si può ricavare la congiunta totale avendo a disposizione:

4. le **probabilità iniziali** di trovarsi in ogni stato $P(x_0)$;
5. le **probabilità di transizione** per passare da uno stato all'altro $P(x_{t+1}|x_t)$;
6. le probabilità dette **modello dei sensori** (o **confounding probabilities**) $P(y_t|x_t)$ che indicano quanto lo stato reale influenza le misure osservate.

Queste tre quantità, insieme al numero di stati n_S implicitamente contenuto, rappresentano una parametrizzazione completa θ^0 per un qualunque modello markoviano latente.

Applicabilità

Tutti i casi in cui le ipotesi di cui sopra sono (almeno approssimativamente) vere. La più importante è che lo stato dipenda solo dal precedente, se fosse troppo limitante è possibile:

1. aumentare l'ordine della catena markoviana (x_t dipende anche da x_{t-1} , x_{t-2} e così via);
2. aumentare il numero degli stati e associare ad ogni stato informazioni sul precedente. Ad esempio, si può raddoppiare il numero degli stati e modificare le probabilità per fare in modo che ogni stato "contenga" l'informazione dello stato precedente.

L'altra assunzione importante è quella che riguarda la finitezza del numero di stati possibili, in caso di insiemi di stati infiniti è più opportuno il filtro di Kalman.

Dal punto di vista delle prestazioni di elaborazione, si può dimostrare che:

- il filtraggio e la predizione hanno complessità $O(1)$;
- lo smoothing ed il calcolo della verosimiglianza delle osservazioni hanno complessità $O(t)$;
- il calcolo della sequenza di stati a massima verosimiglianza ha complessità $O(n_S^2 t)$;

Implementazione

Filtraggio

Si può derivare un algoritmo ricorsivo che calcola lo stato dal filtraggio precedente, ossia una opportuna funzione f tale che $P(x_{t+1}|y_{1:t+1}) = f(P(x_t|y_{1:t}), y_{t+1})$. Infatti:

$$P(x_{t+1}|y_{1:t+1}) = P(x_{t+1}|y_{t+1}, y_{1:t}) = P(y_{t+1}|x_{t+1}, y_{1:t}) \frac{P(x_{t+1}, y_{1:t})}{P(y_{1:t+1})}$$

(per la regola di Bayes estesa: $P(A|B,C) = P(B|A,C) \frac{P(A,C)}{P(B,C)}$)

quindi, per l'ipotesi 3:

$$P(x_{t+1}|y_{1:t+1}) = P(y_{t+1}|x_{t+1}) \frac{P(x_{t+1}, y_{1:t})}{P(y_{1:t+1})}$$

l'ultima probabilità può essere riscritta in forma condizionata:

$$P(x_{t+1}|y_{1:t+1}) = P(y_{t+1}|x_{t+1}) \frac{P(x_{t+1}|y_{1:t}) P(y_{1:t})}{P(y_{1:t+1})}$$

e come somma su tutte le possibili x_t :

$$P(x_{t+1}|y_{1:t+1}) = \frac{P(y_{1:t})}{P(y_{1:t+1})} P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}, x_t|y_{1:t})$$

quindi, per Bayes:

$$P(x_{t+1}|y_{1:t+1}) = \frac{P(y_{1:t})}{P(y_{1:t+1})} P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t, y_{1:t}) \frac{P(x_t, y_{1:t})}{P(y_{1:t})}$$

semplificando e applicando l'ipotesi di markovianità:

$$P(x_{t+1}|y_{1:t+1}) = \frac{1}{P(y_{1:t+1})} P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t) P(x_t, y_{1:t})$$

applicando nuovamente Bayes:

$$P(x_{t+1}|y_{1:t+1}) = \frac{1}{P(y_{1:t+1})} P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t) P(x_t|y_{1:t}) P(y_{1:t})$$

quindi la formula finale:

$$P(x_{t+1}|y_{1:t+1}) = \frac{P(y_{1:t})}{P(y_{1:t+1})} P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t) P(x_t|y_{1:t})$$

Quest'ultima forma dipende:

- dalla verosimiglianza del campione al tempo t e al tempo t+1, rispettivamente indicate con $P(y_{1:t})$ e $P(y_{1:t+1})$, calcolabili con gli algoritmi descritti in seguito;
- dal modello dei sensori $P(y_{t+1}|x_{t+1})$;
- dalle probabilità di transizione $P(x_{t+1}|x_t)$;
- dal valore del filtro all'istante precedente $P(x_t|y_{1:t})$, che definisce la ricorsività della formula

che sono tutte quantità note.

Predizione

La predizione a k passi può essere calcolata tramite la formula ricorsiva seguente:

$$P(x_{t+k+1}|y_{1:t}) = \sum_{x_{t+k}} P(x_{t+k+1}|x_{t+k}) P(x_{t+k}|y_{1:t})$$

infatti:

$$P(x_{t+k+1}|y_{1:t}) = \sum_{x_{t+k}} P(x_{t+k+1}|x_{t+k}) \frac{P(x_{t+k}, y_{1:t})}{P(y_{1:t})}$$

$$P(x_{t+k+1}|y_{1:t}) = \sum_{x_{t+k}} \frac{P(x_{t+k+1}, x_{t+k}, y_{1:t})}{P(y_{1:t})}$$

$$P(x_{t+k+1}|y_{1:t}) = \frac{P(x_{t+k+1}, y_{1:t})}{P(y_{1:t})} \quad .$$

Smoothing

Il problema dello smoothing si riconduce al calcolo di $P(x_{t-k}|y_{1:t})$. Per semplicità, si definisce t_1 l'istante $t-k$ che si intende stimare, e si calcola la probabilità come segue:

$P(x_{t_1}|y_{1:t}) = P(x_{t_1}|y_{1:t_1}, y_{t_1+1:t})$ (scomposizione delle y)

$P(x_{t_1}|y_{1:t}) = P(y_{t_1+1:t}|x_{t_1}, y_{1:t_1}) \frac{P(x_{t_1}, y_{1:t_1})}{P(y_{1:t})}$ (regola di Bayes)

$P(x_{t_1}|y_{1:t}) = P(y_{t_1+1:t}|x_{t_1}) \frac{P(x_{t_1}, y_{1:t_1})}{P(y_{1:t})}$ (ipotesi 2)

$P(x_{t_1}|y_{1:t}) = P(y_{t_1+1:t}|x_{t_1}) \frac{P(x_{t_1}|y_{1:t_1})}{P(y_{1:t}) P(y_{1:t_1})}$ (definizione di probabilità condizionata).

L'ultima forma dipende:

- dalla verosimiglianza delle osservazioni fino al tempo t e fino al tempo t_1 , calcolabili con gli algoritmi descritti in seguito;
- dall'espressione del filtraggio fino al tempo t_1 ($P(x_{t_1}|y_{1:t_1})$);
- dalla probabilità $P(y_{t_1+1:t}|x_{t_1})$ detta **backward**, ed esprime la probabilità di una sequenza di stati successiva a t_1 noto lo stato in quell'istante.

Tutte le quantità sono note, infatti la probabilità backward si può calcolare come segue:

$P(y_{t_1+1:t}|x_{t_1}) = \sum_{x_{t_1+1}} P(y_{t_1+1:t}, x_{t_1+1}|x_{t_1})$ (legge delle probabilità totali)

$P(y_{t_1+1:t}|x_{t_1}) = \sum_{x_{t_1+1}} P(y_{t_1+1:t}|x_{t_1+1}, x_{t_1}) \frac{P(x_{t_1+1}, x_{t_1})}{P(x_{t_1})}$ (regola di Bayes)

$P(y_{t_1+1:t}|x_{t_1}) = \sum_{x_{t_1+1}} P(y_{t_1+1:t}|x_{t_1+1}, x_{t_1}) P(x_{t_1+1}|x_{t_1})$ (definizione di probabilità condizionata)

$P(y_{t_1+1:t}|x_{t_1}) = \sum_{x_{t_1+1}} P(y_{t_1+1:t}|x_{t_1+1}) P(x_{t_1+1}|x_{t_1})$ (ipotesi 2)

$P(y_{t_1+1:t}|x_{t_1}) = \sum_{x_{t_1+1}} P(y_{t_1+1}|x_{t_1+1}) P(y_{t_1+2:t}|x_{t_1+1}) P(x_{t_1+1}|x_{t_1})$ (divisione di probabilità condizionatamente indipendenti)

L'ultima espressione dipende:

- dal modello dei sensori $P(y_{t_1+1}|x_{t_1+1})$;
- dalle probabilità di transizione $P(x_{t_1+1}|x_{t_1})$;
- dal valore "successivo" della probabilità stessa $P(y_{t_1+2:t}|x_{t_1+1})$, che ne denota la forma ricorsiva.

Per come è definita, la probabilità backward si calcola a ritroso nel tempo, da questo fatto il nome.

Verosimiglianza delle osservazioni

Esistono più modi per calcolare la verosimiglianza.

Un modo è usare la probabilità **forward** che è definita come la probabilità che si verifichi una sequenza di stati precedenti a un tempo t, $y_{1:t}$, e che lo stato sia x_t : $P(x_t, y_{1:t})$.

Similmente alla probabilità backward, $P(x_t, y_{1:t})$ può essere calcolata in modo ricorsivo:

$P(x_{t+1}, y_{1:t+1}) = P(x_{t+1}, y_{t+1}, y_{1:t})$ (scomposizione delle osservazioni)

$P(x_{t+1}, y_{1:t+1}) = P(y_{t+1}|x_{t+1}, y_{1:t}) P(x_{t+1}, y_{1:t})$ (definizione di probabilità condizionata)

$P(x_{t+1}, y_{1:t+1}) = P(y_{t+1}|x_{t+1}) P(x_{t+1}, y_{1:t})$ (ipotesi 2)

$P(x_{t+1}, y_{1:t+1}) = P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}, x_t, y_{1:t})$ (teorema delle probabilità totali)

$P(x_{t+1}, y_{1:t+1}) = P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t, y_{1:t}) P(x_t, y_{1:t})$ (definizione di probabilità condizionata)

$P(x_{t+1}, y_{1:t+1}) = P(y_{t+1}|x_{t+1}) \sum_{x_t} P(x_{t+1}|x_t) P(x_t, y_{1:t})$ (ipotesi 1).

Dalla probabilità forward si può ricavare la verosimiglianza sommando su tutte le x_t :
$\sum_{x_t} P(x_t, y_{1:t}) = P(y_{1:t})$.

Un altro modo è di moltiplicare la probabilità forward per la probabilità backward. Si ottiene, previo cambio di variabili:

$P(y_{t_1+1:t}|x_{t_1}) P(x_{t_1}, y_{1:t_1})$ ossia

$P(y_{t_1+1:t}, x_{t_1}, y_{1:t_1}) = P(y_{1:t}, x_{t_1})$

Anche qui si può ottenere la verosimiglianza sommando su tutte le possibili x_{t_1} .

Sequenza di stati a massima verosimiglianza (algoritmo di Viterbi)

Prendiamo inizialmente in considerazione il problema del calcolo della probabilità della sequenza più probabile di stati che hanno portato a una certa sequenza di osservazioni $\max_{x_{1:t}} P(y_{1:t}|x_{1:t})$.
Questa probabilità dipende dal tempo in cui è calcolata, e quindi può essere pensata come funzione dell'ultimo stato della sequenza x_t ; si pone per definizione

$V(x_t) = \max_{x_{1:t}} P(y_{1:t}|x_{1:t})$

E' possibile calcolare la probabilità da massimizzare in modo relativamente semplice ragionando per ricorsione. Infatti:

$P(y_{1:t}|x_{1:t}) = P(y_{1:t-1}|x_{1:t-1}) P(x_{t-1}|x_t) P(y_t|x_t)$

ossia: la probabilità che la sequenza di osservazioni $y_{1:t}$ sia stata osservata dalla sequenza di stati $x_{1:t}$ è pari alla probabilità:

- che la sequenza troncata al passo precedente sia la stessa ($y_{1:t-1}$ dato $x_{1:t-1}$)
- e che si sia verificata la transizione da x_{t-1} a x_t
- e che si sia osservato il dato y_t dato che lo stato era x_t .

Essendo i tre eventi indipendenti (per l'ipotesi di markovianità e l'ipotesi 3), la probabilità voluta può essere calcolata semplicemente come prodotto delle tre.

A questo punto per calcolare $V(t)$ occorre massimizzare la formula ottenuta su tutte le possibili sequenze $x_{1:t}$:

$$V(x_t) = \max_{x_{1:t}} P(y_{1:t}|x_{1:t})$$

$$V(x_t) = \max_{x_{1:t}} [P(y_{1:t-1}|x_{1:t-1}) P(x_{t-1}|x_t) P(y_t|x_t)]$$

Il massimo del prodotto è uguale al prodotto dei massimi:

$$V(x_t) = \max_{x_{1:t}} [P(y_{1:t-1}|x_{1:t-1})] \max_{x_{1:t}} [P(x_{t-1}|x_t)] \max_{x_{1:t}} [P(y_t|x_t)]$$

il primo termine dipende solo dagli stati da x_1 a x_{t-1}, il secondo da x_t e x_{t-1} e il terzo solo da x_t:

$$V(x_t) = \max_{x_{1:t-1}} [P(y_{1:t-1}|x_{1:t-1})] \max_{x_{t-1}, x_t} [P(x_{t-1}|x_t)] \max_{x_t} [P(y_t|x_t)]$$

Inoltre, dal calcolo del primo termine sarà fissato un certo valore per x_{t-1}, quindi il secondo termine dipende effettivamente solo da x_t:

$$V(x_t) = \max_{x_{1:t-1}} [P(y_{1:t-1}|x_{1:t-1})] \max_{x_t} [P(x_{t-1}|x_t)] \max_{x_t} [P(y_t|x_t)]$$

Analogamente fissati i primi due termini sarà fissato anche x_t, quindi il terzo termine è unico e non deve essere massimizzato del tutto:

$$V(x_t) = \max_{x_{1:t-1}} [P(y_{1:t-1}|x_{1:t-1})] \max_{x_t} [P(x_{t-1}|x_t)] P(y_t|x_t)$$

A questo punto si può osservare che il primo termine è pari a $V(x_{t-1})$ e quindi riscrivere $V(x_t)$ nella forma:

$$V(x_t) = \max_{x_t} [V(x_{t-1}) P(x_{t-1}|x_t)] P(y_t|x_t)$$

Questa formula calcola correttamente la probabilità, ma se implementata direttamente in un calcolatore elettronico può dar luogo a problemi di precisione numerica poiché le probabilità coinvolte possono diventare molto piccole. Per ovviare a questi problemi si può calcolare la log-probabilità $v(x_t)$:

$$v(x_t) = \log(V(x_t)) = \max_{x_t} [v(x_{t-1}) + \log(P(x_{t-1}|x_t))] + \log(P(y_t|x_t))$$

La formula ricorsiva necessita di un valore iniziale per v, che si suppone noto (dipende dal contesto di applicazione).

A questo punto si può affrontare il problema del calcolo della sequenza ottima, che può essere espresso come:

$$\Psi(x_t) = \operatorname*{argmax}_{x_{1:t}} P(y_{1:t}|x_{1:t})$$

per la definizione di V data, si può scrivere:

$$\Psi(x_t) = \operatorname*{argmax}_{x_{1:t}} [V(x_{t-1}) P(x_{t-1}|x_t) P(y_t|x_t)]$$

Si possono fare, a questo punto, considerazioni analoghe alle precedenti su quali stati influenzano quali probabilità. La conclusione, in questo caso, è che $P(y_t|x_t)$ è costante e non influenza il calcolo poiché x_t è già fissato dai termini precedenti. Quindi:

$$\Psi(x_t) = \operatorname*{argmax}_{x_t} [V(x_{t-1}) P(x_{t-1}|x_t)]$$

Anche in questo caso si può calcolare una variante logaritmica, $\psi(t)$:

$$\psi(t) = \operatorname*{argmax}_{x_t} [v(x_{t-1}) + \log(P(x_{t-1}|x_t))]$$

Altri argomenti al momento non meglio collocati

Stima della media

Nel caso in cui le osservazioni siano IID, vale lo stimatore corretto e ottimo: $\hat{x} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

varianza $V[\bar{X}] = \frac{\sigma^2}{n}$. La sua deviazione standard è inversamente proporzionale a $\sqrt{n}$ quindi lo stimatore si dice $\sqrt{n}$ **-convergente**.
Il teorema centrale del limite, inoltre, afferma che
$\bar{X} \sim N(\mu, \sigma^2)$ per $n \to \infty$

dove $\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$, $X_i iid\, F(\mu, \sigma^2)$ con F qualunque, μ è il valore atteso finito della distribuzione e σ^2 la sua varianza finita. Quindi, a patto di avere sufficienti osservazioni e a patto che media e varianza esistano finite, il teorema assicura che $\bar{X}$ ha distribuzione normale.
Nota che è possibile che la media sia infinita, poiché l'integrale $E[X] = \int_{-\infty}^{+\infty} x f(x) dx$ potrebbe divergere nonostante $\int_{-\infty}^{+\infty} f(x) dx = 1$.

Nel caso in cui i dati siano non indipendenti, invece, risulta: $V[\bar{X}] = \frac{\sigma^2}{n} + \frac{1}{n^2}\sum_{t_1 \neq t_2} E(y_{t_1}, y_{t_2})$

quindi possono verificarsi diversi casi:

1. la somma tende a zero per $n \to \infty$, non ci sono particolari problemi;
2. la somma tende a una costante, quindi avrò intervalli di confidenza più ampi (es. modelli ARMA a memoria corta);
3. la somma tende ad infinito, lo stimatore potrebbe perdere anche la correttezza (es. modelli a memoria lunga).

Intervalli di confidenza sulla media

$\bar{x} \pm \frac{\sigma}{\sqrt{(n)}} z_{\frac{\alpha}{2}}$ dove $z_{\frac{\alpha}{2}}$ è il percentile di ordine $\frac{\alpha}{2}$ della normale standard.

Se σ non è noto e i dati non sono sufficientemente numerosi (la media non è distribuita come una normale): $\bar{x} \pm \frac{S}{\sqrt{(n)}} t_{\frac{\alpha}{2}, n-1}$ dove S è la varianza campionaria e $t_{\frac{\alpha}{2}, n-1}$ è il percentile di ordine $\frac{\alpha}{2}$ della t di Student con $n-1$ gradi di libertà.

Stima dei parametri di un modello AR

M: $y_t = \alpha y_{t-1} + e(t)$, $e(t) \sim N(\mu, \sigma^2)$, processo depolarizzato.

Stimatore LS per α : $\hat{\alpha_{LS}} = \frac{COV[y_t, y_{t-1}]}{V[y_t]} = \frac{\sum_t y_t y_{t-1}}{\sum_t y_t^2}$, varianza $V[\hat{\alpha}] \propto \frac{1}{n}$, lo stimatore è $\sqrt{n}$ **-convergente**, non ottimo.

Test d'ipotesi

$X_{1,} X_{2,} \ldots X_n iid\ N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$

$H_1: \mu > \mu_0$

$x_C = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{(n)}}$, accetto H_0 se $\bar{x} < x_C$ fissato α .

Rischio del primo tipo (probabilità α): rifiutare H_0 se è vera.

Rischio del secondo tipo (probabilità β): accettare H_0 se è vera H_1 .

Nota: per $n \to \infty$ $\bar{x} \to \mu$, ossia la varianza della media campionaria diventa sempre più piccola (la campana si "stringe" attorno al valore vero μ). Quindi α andrebbe scelto a seconda di n: se n è elevato il canonico $\alpha = 5\%$ porterebbe probabilmente a rifiutare molti casi in cui in realtà non c'è indecisione (la campana relativa a H_1 è molto "lontana" sull'asse delle x).

Per i test bilaterali, $H_1: \mu \neq \mu_0$, si può anche utilizzare la statistica seguente per ricondursi a un test monolaterale:

$\left(\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}\right)^2 \sim X_1^2$ che ha distribuzione chi-quadro con un grado di libertà.

Test d'ipotesi e intervalli di confidenza sono legati fra loro. Fissato il livello di significatività, infatti:

$IC_\alpha: \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ intervallo per μ ,

$A_\alpha: \mu_0 \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ intervallo di accettazione dell'ipotesi nulla per $\bar{x}$.