1 Introduzione alla previsione

In sintesi, l'arte di prevedere è in realtà l'arte di riconoscere *elementi ripetitivi* nei dati del passato. Solo grazie ad essi abbiamo il suggerimento di cosa potrebbe accadere nel futuro.

I due elementi ripetitivi più evidenti sono il *trend* ed i *pattern* ciclici (o periodici). Anche il trend, in senso lato, è un elemento ripetitivo.

Il metodo ripetitivo più ingenuo è la pura ripetizione. Se indichiamo con $d(n)$ il dato al tempo n (tempo discreto, nei nostri esempi) e con $p(n+1)$ il valore che prevediamo al tempo $n+1$, il metodo di ripetizione più ingenuo stabilisce:

$$p(n+1) = d(n).$$

Ovviamente nessuno si aspetta grosse prestazioni da un tale metodo. Però esso aiuta a focalizzare alcune domande. Come mai riteniamo poco furbo questo metodo? Essenzialmente per il fatto che non utilizza ciò che è avventuo in precedenza. Per migliorare le cose potremmo prendere una funzione degli ultimi k dati:

$$p(n+1) = f\left(d(n), d(n-1), ..., d(n-k+1)\right).$$

Ma quale f? In assenza di idee più furbe, si può tentare con una f lineare (affine). Questo conduce al metodo di previsione basato sulla *regressione lineare multipla*:

$$p(n+1) = b + a_n d(n) + a_{n-1} d(n-1) + ... + a_{n-k+1} d(n-k+1).$$

I coefficienti vengono trovati col metodo dei minimi quadrati, sui dati noti: si cercano i coefficienti che meglio avrebbero predetto i dati noti. Lo esemplificheremo nel seguito.

Altrettanto naturale è il tentativo di specificare *a priori* la dipendenza dagli ultimi dati, invece che cercarla coi minimi quadrati. In un certo senso, invece che subire la peculiarità dei dati sperimentali, si decide a priori che i dati precedenti devono influire in un certo modo. Nascono in questa logica due indirizzi principali:

- il metodo di *media mobile* con finestra di lunghezza k:

$$p(n+1) = \frac{1}{k}\left(d(n) + d(n-1) + ... + d(n-k+1)\right)$$

- il metodo di *smorzamento esponenziale* di parametro $\alpha \in (0,1)$:

  $$p(n+1) = \alpha d(n) + \alpha (1-\alpha) d(n-1) + \alpha (1-\alpha)^2 d(n-2) + ... + r(1)$$

  dove la somma si intende estesa a tutti i dati disponibili (quindi la finestra di dati utilizzati non è fissata), ed $r(1)$ è un valore di aggiustamento iniziale. La strana forma dei coefficienti è dovuta alla relazione

  $$\alpha \left(1 + (1-\alpha) + (1-\alpha)^2 + ...\right) = 1$$

  quindi la logica è di eseguire una media pesata (con somma dei pesi pari ad uno) dei valori precedenti, ma scalati esponenzialmente.

Quest'ultimo metodo si può riformulare in modo iterativo, senza l'uso della somma esplicita su tutti i valori. Questa riformulazione offre lo spunto per generalizzazioni importantissime che ora discuteremo, ovvero lo smorzamento esponenziale con trend e quello con trend e stagionalità (Holt Winters).

Questi ultimi metodi, come vedremo tra un attimo, enfatizzano l'importanza del concetto di *modello*. Questa è un'altra delle idee portanti in teoria delle previsione. Sempre nell'ottica di ripetere ciò che si è già visto, si cerca un modello di ciò che si è visto e lo si utilizza per la previsione futura. A suo modo, anche la ricerca di trend e ciclicità è una ricerca di un modello, ma l'accezione usale del termine "modello" è quella di *relazione funzionale o ricorsiva* (o differenziale, nel continuo). Ad esempio, scoprire che per i dati passati vale la proprietà che il dato al tempo $n+1$ è legato ad alcuni dati precedenti ed alcune altri fattori F_j da una relazione funzionale, a meno di errore, del tipo

$$d(n+1) = f\left(d(n), d(n-1), ..., F_1, F_2, ...\right) + \varepsilon(n)$$

significa aver scoperto un modello del problema e questo si userà per la previsione nel modo ovvio

$$p(n+1) = f\left(d(n), d(n-1), ..., F_1, F_2, ...\right)$$

se l'errore ha media nulla.

2 Metodi di smorzamento esponenziale

2.1 Smorzamento esponenziale (semplice)

Lo smorzamento esponenziale nella sua versione più semplice, quella già indicata sopra, si può riformulare nel seguente modo:

$$p(n+1) = \alpha d(n) + (1-\alpha)\, p\,(n)\,.$$

Questo significa che al tempo n abbiamo sia il dato $d(n)$ sia una previsione $p\,(n)$ ottenuta precedentemente e calcoliamo la previsione successiva usando entrambi, in modo pesato: se abbiamo scelto α grande, diamo più peso al dato sperimentale attuale, mentre se abbiamo scelto α piccolo, tendiamo a fidarci di più della previsione già fatta, modificandola solo di poco col nuovo dato sperimentale.

Siccome serve la previsione al tempo precedente, nasce il problema dell'*inizializzazione*: dobbiamo cominciare ad effettuare una previsione ad un certo istante, e solo da lì in poi possiamo usare il metodo. La prima previsione deve essere fatta con un'altro metodo. Spesso si fa la scelta banale: se $d(1)$ è il primo dato sperimentale, prevediamo il valore al tempo $n = 2$ con la formula $p(2) = d(1)$. Arrivati al tempo $n = 2$, abbiamo in nostro possesso sia il dato $d(2)$ sia la previsione (per quanto rozza) $p(2)$ e quindi possiamo innescare lo smorzamento esponenziale, prevedendo $p(3)$.

Però si può effettuare una inizializzazione meno banale, ad esempio aspettando un po' di tempo ed usando una regressione o una media mobile. L'esperienza però non conferma una grande utilità di inizializzazioni più complesse, in quanto lo smorzamento esponenziale si auto-aggiusta piuttosto velocemente e quindi nel tempo richiesto da inizializzazioni complesse ha già fatto un buon lavoro.

Mostriamo che le due formulazioni del metodo coincidono. Vale, iterativamente,

$$\begin{aligned}
p(n+1) &= \alpha d(n) + (1-\alpha)\left(\alpha d(n-1) + (1-\alpha)\, p\,(n-1)\right) \\
&= \alpha d(n) + \alpha\,(1-\alpha)\, d(n-1) + (1-\alpha)^2\, p\,(n-1) \\
&= \alpha d(n) + \alpha\,(1-\alpha)\, d(n-1) + (1-\alpha)^2\left(\alpha d(n-2) + (1-\alpha)\, p\,(n-2)\right) \\
&= \alpha d(n) + \alpha\,(1-\alpha)\, d(n-1) + \alpha\,(1-\alpha)^2\, d(n-2) + (1-\alpha)^3\, p\,(n-2)
\end{aligned}$$

e quindi si vede che vale la relazione descritta nell'introduzione, con $r(1) = (1-\alpha)^{n-1}\, p\,(2)$.

Ecco alcune osservazioni.

- Si capisce che se vogliamo dare più rilievo agli ultimi dati, dobbiamo usare α grande e viceversa un α piccolo se vogliamo dare circa lo stesso rilievo ai dati di una più ampia finestra di tempo. Quando operare l'una o l'altra scelta? Se la serie di dati è stazionaria e somiglia ad un rumore bianco (valori indipendenti), conviene un α piccolo: essendo impossibile prevedere le fluttuazioni specifiche in quanto sono indipendenti, conviene almeno assestarsi sul valor medio il prima possibile e senza subire le fluttuazioni. Quando invece la serie ha dell variazioni strutturali, sia a breve termine, sia di tipo trend o stagionalità, conviene inseguire di più gli ultimi valori in quanto essi sono più rappresentativi della variazione in atto; per far questo conviene prendere un α grande.

- C'è però un modo automatico di scegliere α. Si finge di applicare il metodo ai dati noti come se non lo fossero. Si fissa un α, si usa il metodo sui dati noti, prevedendoli passo per passo e si calcolano i residui
$$e(n) = p(n) - d(n).$$
Di questi residui si calcola poi un qualche riassunto che ci sembri significativo. Il più tradizionale è lo scarto quadratico medio (la *radice quadrata* del cosidetto *MSE*, mean square error):
$$\widehat{\sigma}_e := \sqrt{\frac{1}{N - N_0} \sum_{n=N_0+1}^{N} e^2(n)}$$
dove si esclude una finestra iniziale di valori, sia per il fatto che non li abbiamo a causa dell'inizializzazione, sia per evitare magari di dar peso ad una fase iniziale di assestamento (la cui scelta però è soggettiva). Uno meno sensibile a valori eccessivamente alti e magari anomali, è il *MAPE* (mean absolute percentage error), espresso in percentuale:
$$MAPE = \frac{100}{N - N_0} \sum_{n=N_0+1}^{N} \frac{|e(n)|}{|d(n)|}.$$
A questo punto, si cerca il valore di α che produce il valore minimo dell'indicatore scelto. Per semplicità di solito si testano solo alcuni valori di α (es. 0.1, 0.2, ... , 0.9) e magari si raffina un po' la ricerca nelle vicinanze del valore migliore. Il software R utilizza $\widehat{\sigma}_e$.

- Resta il fatto che il nostro intuito potrebbe indurci a modificare un poco tale scelta automatica.

A questo punto conviene familiarizzare con R. Prendiamo un white noise, cioè una sequenza di numeri gaussiani indipendenti:

x <- 1:30
y.stat <- rnorm(x)
ts.plot(y.stat)

e di questi effettuiamo la previsione con smorzamento esponenziale semplice:

p.stat <- HoltWinters(y.stat, gamma=0, beta=0)
plot(p.stat)
p.stat

Vediamo che la funzione HoltWinters ha selezionato un α piuttosto piccolo. Se la serie del white noise fosse stata più lunga, avrebbe preso un α ancora più piccolo. Si può completare questa indagine controllando cosa succede se si impone ad es. $\alpha = 0.7$: la previsione può sembrare ad occhio migliore ma questo è un caso in cui l'occhio inganna se non è supportato dai ragionamenti giusti. Infatti, il grafico della previsione e quello dei dati sono molto simili (uguali nel caso limite $\alpha = 1$) ma traslati di un'unità. L'occhio non istruito riconosce la somiglianza e pensa che la previsione sia buona. Invece gli scarti in verticale (che sono quelli che contano) sono più elevati.

Un secondo esempio è invece una retta più white noise:

y.trend <- x+y.stat
ts.plot(y.trend)

Qui usando il solito comando:

p.trend <- HoltWinters(y.trend, gamma=0, beta=0)
plot(p.trend)
p.trend

si trova che $\alpha = 1$ è il migliore. Bisogna inseguire il trend. Basta verificare cosa accade imponendo ad es. $\alpha = 0.2$ per non avere dubbi sull'opportunità del valore $\alpha = 1$.

2.2 Sulla media mobile

Sul metodo di media mobile c'è poco da dire, a meno di non esaminare le sue generalizzazioni a processi ARIMA o a regressione multipla (che dicuteremo dopo). Dal punto di vista pratico, ecco alcune osservazioni.

- L'esperienza mostra che usualmente è un po' meno accurato dello smorzamento esponenziale, quindi nettamente meno accurato di Holt Winters quando ci sono evidenti trend e stagionalità. Senza trend e stagionalità si comporta quasi come lo smorzamento esponenziale, abbiamo detto. Questa somiglianza si può quantificare: c'è una relazione tra k ed α sotto la quale i due metodi sono quasi uguali:

$$\frac{\alpha}{2-\alpha} = \frac{1}{k}$$

  che però non dimostriamo (inoltre richiede opportune ipotesi di stazionarietà della serie).

- Per quanto rozzo, il metodo di media mobile viene usato spesso nelle applicazioni. Anche in finanza, si può osservare sulle pagine web già utilizzate che esse offrono la possibilità di sovrapporre al grafico dei dati la media mobile, con k a scelta. Pare che alcuni operatori usino effettivamente la media mobile ed agiscano quando il grafico dei dati taglia quello della media mobile.

- Ripetiamo simmetricamente l'osservazione fatta sopra per lo smorzamento esponenziale. Se vogliamo dare più rilievo agli ultimi dati, dobbiamo usare k piccolo e viceversa un k grande se vogliamo dare lo stesso rilievo ai dati di una più ampia finestra di tempo. Quando operare l'una o l'altra scelta? Se la serie di dati è stazionaria e somiglia ad un rumore bianco (valori indipendenti), conviene un k grande: essendo impossibile prevedere le fluttuazioni specifiche in quanto sono indipendenti, conviene almeno assestarsi sul valor medio il prima possibile e senza subire le fluttuazioni. Quando invece la serie ha dell variazioni strutturali, sia a breve termine, sia di tipo trend o stagionalità, conviene inseguire di più gli ultimi valori in quanto essi sono più rappresentativi della variazione in atto; per far questo conviene prendere un k piccolo.

- In ogni caso ha senso scegliere il k che minimizza l'errore che il metodo commette sui dati noti. Però il nostro intuito potrebbe indurci a modificare un poco tale scelta automatica, alla luce dell'osservazione precedente, se con l'intuito fossimo sicuri di un comportamento stazionario oppure con variazioni strutturali.

2.3 Smorzamento esponenziale con trend

Una versione fatta in casa, a mano, di questa filosofia, è quella di estrarre un trend dai dati (tipicamente con una regressione, lineare o meno, globale o meno), sottrarre il trend a i dati, ed applicare lo smorzamento esponenziale (semplice) alla serie modificata.

C'è però una variante automatica dello smorzamento esponenziale semplice che è in grado di calcolare un trend. Lo fa non globalmente o localmente, ma *iterativamente*, sulla falsariga dello smorzamento stesso. Ci sono molti modi di capire questa variante; seguiamo l'idea di *modello*. Supponiamo di ipotizzare che i dati seguano il semplice modello lineare

$$d(t) = (at + b) + \varepsilon(t).$$

In realtà, se pensassimo che questo modello è globalmente accurato, basterebbe usare la regressione. Quindi quello che pensiamo veramente è che un tale modello lineare sia una sorta di canovaccio, ma che col passare del tempo i valori dei coefficienti della retta debbano essere aggiornati. Per fare un parallelo, è come se nel paragrafo precedente avessimo usato il modello

$$d(t) = b + \varepsilon(t)$$

con $\varepsilon(t)$ white noise. Se credevamo completamente in un tale modello, bastava calcolare la media b dei dati. Invece pensavamo che al variare del tempo potesse essere utile aggiornare un po' le cose, e quindi abbiamo trovato un modo per dare un po' più di peso agli ultimi valori sperimentali.

Ora stiamo ipotizzando il modello lineare ma con la propensione ad aggiornare iterativamente la nostra stima della retta. Introduciamo quindi dei valori dipendenti dal tempo per i coefficienti della retta: li cambieremo ad ogni passo temporale. Indichiamoli con $s(n)$ ed $m(n)$ com'è uso. Il primo corrisponde all'intercetta, ma non quella calcolata rispetto agli assi originari, bensì quella rispetto ad un asse delle ordinate posizionato al tempo n. Invece $m(n)$ è il coefficiente angolare della retta.

Immaginiamo di essere al tempo n. Conosciamo il dato $d(n)$. Inoltre, al passo precedente eravamo arrivati ad una stima di s ed m, che indichiamo con $s(n-1)$ ed $m(n-1)$. Il problema allora è, sulla base di $d(n)$, $s(n-1)$, $m(n-1)$, stimare i nuovi valori di $s(n)$ ed $m(n)$. Se ci riusciamo, prevediamo il valore futuro al tempo $n+1$ con la formula

$$p(n+1) = s(n) + m(n)$$

e più in generale prevediamo il valore dopo k unità di tempo con la formula

$$p(n+k) = s(n) + m(n) \cdot k.$$

Infatti, nel sistema di coordinate in cui l'ordinata è posizionata al tempo n, l'intercetta è $s(n)$, quindi la retta si legge

$$r(t) = s(n) + m(n) \cdot t$$

e qui t sta ad indicare il tempo trascorso dopo l'istante n.

L'aggiornamento dei coefficienti è dato dalle formula

$$\begin{aligned} s(n) &= \alpha d\,(n) + (1-\alpha)\,(s(n-1) + m(n-1)) \\ m\,(n) &= \beta\,(s(n) - s(n-1)) + (1-\beta)\,m(n-1). \end{aligned}$$

Queste vanno eseguite nell'ordine scritto, in quanto si usa $s(n)$ nella seconda. La logica dietro questa scelta di iterazioni è, per quanto riguarda la prima equazione, che l'intercetta nel sistema di assi al tempo n è data da $s(n-1) + m(n-1)$, sulla base delle vecchie stime; questa la aggiorniamo dando un po' di peso a $d\,(n)$. Per la seconda equazione, aggiorniamo la vecchia stima $m(n-1)$ usando un "coefficiente angolare" calcolato sulla base degli ultimi dati; avremmo dovuto scrivere $d(n) - d(n-1)$, ma questo è troppo semsibile alle oscillazioni casuali dei dati, così si prende $s(n) - s(n-1)$ che gli somiglia ma è più stabile, sente l'influsso dei dati passati secondo la logica dello smorzamento.

Questo modello è un esempio delle cosidette *variabili nascoste.* In anni recenti hanno preso piede i modelli denominati *hidden Markov chains.* Questi come altri non markoviani sono basati sull'idea che dietro i dati sperimentali ci siano variabili magari non direttamente osservabili che però sono suscettibili di più logiche relazioni ricorsive o funzionali e che i dati sperimentali siano solo un sottoprodotto, una marginale, del modello nascosto.

Circa l'inizializzazione, si può seguire una strategia banale, ovvero prendere $s(2) = d(1)$ come nello smorzamento semplice e $m(2) = 0$. Il software R migliora un po' questa scelta aspettando un altro passo temporale e inizializzando

$$s(3) = d(2), \quad m(3) = d(2) - d(1).$$

Queste scelte sono però davvero povere in certi casi. Quindi può convenire l'attesa di diverse unità temporali (es. $n_0 = 4$), eseguire una regressione lineare e trovata la retta $at + b$, prendere

$$s(n_0) = an_0 + b, \quad m(n_0) = a.$$

La scelta dei parametri α, e β viene effettuata per ottimizzazione degli scarti, come sopra.

Tornando all'esempio della serie data da una retta più white noise, dai comandi:

p.vero.trend <- HoltWinters(y.trend, gamma=0)
plot(p.vero.trend)
p.vero.trend

otteniamo una previsione molto migliore. Si osservi che il valore di β è basso: questo non significa che conti poco, ma al contrario che la tendenza è abbastanza costante, quindi conviene modificarla poco in base agli ultimi valori.

2.4 Smorzamento esponenziale con trend e stagionalità (Winters)

Di questo metodo esiste una versione per stagionalità additiva e duna per quella moltiplicativa; descriviamo solo quest'ultima, essendo l'altra del tutto simile e forse più elementare. Si ipotizza il modello

$$d(t) = (at + b)\, F(t) + \varepsilon\,(t)$$

con $F(t)$ funzione periodica di periodo P. Per semplificare l'esposizione, fingiamo di non avere il rumore $\varepsilon\,(t)$, quindi di lavorare sull'equazione

$$d(t) = (at + b)\, F(t).$$

Idealmente, si introduce la grandezza ausiliaria $y(t) = \frac{d(t)}{F(t)}$ che soddisfa

$$y(t) = at + b.$$

A questa possiamo applicare lo smorzamento con trend. Si trova

$$p_y(n + k) = s_y(n) + m_y(n) \cdot k$$

dove $s(n)$ e $m(n)$ soddisfano

$$\begin{aligned} s_y(n) &= \alpha y\,(n) + (1 - \alpha)\,(s_y(n - 1) + m_y(n - 1)) \\ m_y\,(n) &= \beta\,(s_y(n) - s_y(n - 1)) + (1 - \beta)\, m_y(n - 1). \end{aligned}$$

Il problema è che per innescare questo sistema bisogna conoscere $y(n)$ e per questo bisognerebbe conoscere $\frac{d(n)}{F(n)}$, mentre $F(n)$ per ora è incognita. L'idea è di stimare anche la periodicità in modo iterativo, così da aggiustarla se è il caso. Allora al posto di $y(n)$ si mette $\frac{d(n)}{F(n-P)}$, immaginando che nella struttura iterativa che torveremo alla fine il valore $F(n-P)$ sia noto (e riteniamo sia una buona approssimazione di $F(n)$ in quanto cerchiamo una F periodica).

Poi bisogna creare un'equazione iterativa per F. Un'idea ispirata alla filosofia dello smorzamento esponenziale è

$$F(n) = \gamma \frac{d(n)}{y(n)} + (1-\gamma) F(n-P)$$

(si ricordi la definizione $y(t) = \frac{d(t)}{F(t)}$). Però non conosciamo $y(n)$. Noti però $s_y(n)$ ed $m_y(n)$, $s_y(n)$ è una stima di $y(n)$. In definitiva, si arriva al sistema:

$$\begin{aligned} s(n) &= \alpha \frac{d(n)}{F(n-P)} + (1-\alpha)(s(n-1) + m(n-1)) \\ m(n) &= \beta (s(n) - s(n-1)) + (1-\beta) m(n-1) \\ F(n) &= \gamma \frac{d(n)}{s(n)} + (1-\gamma) F(n-P). \end{aligned}$$

L'inizializzazione qui è più complessa. Serve F su un intero periodo per innescare l'iterazione. Allora si sacrifica il primo periodo (a volte più di uno), su quello si trova una retta di regressione

$$z(t) = at + b$$

e la si usa come se fosse una stima di $y(t)$. Quindi dalla definizione $y(t) = \frac{d(t)}{F(t)}$ si stima

$$F(n) = \frac{d(n)}{an+b}.$$

In definitiva, per $n = 1, 2, ..., P$ si prendono questi valori di F e poi si pone $s(P) = aP + b$, $m(P) = a$. Si comincia quindi a prevedere il valore al tempo $P+1$.

3 Regressione lineare multipla

Il metodo basato sulla regressione lineare multipla ha tanti pregi ed appigli teorici:

- con fantasia ed abilità, permette di considerare anche funzioni non lineari, tramite opportune trasformazioni dei dati, quindi punta verso il modello generale

$$p(n+1) = f\left(d(n), d(n-1), ..., d(n-k+1)\right)$$

- permette di inserire predittori di natura diversa all'interno di uno schema iterativo: nulla vieta di aggiungere tra i predittori, oltre ai valori ai tempi precedenti, altri fattori F_j che sembrino coinvolti nel problema:

$$p(n+1) = b + \sum_{i=0}^{k-1} a_{n-i} d(n-i) + \sum_{j=1}^{p} c_j F_j$$

- è un caso particolare dei modelli ARIMA (i cosidetti AR).