Algoritmi genetici - Matematica

Alessio Marchetti

1 Introduzione

Gli algoritmi genetici (GA) sono algoritmi di ricerca basati sulle meccaniche della selezione naturale e della genetica. Essi si basano sulla manipolazione di una popolazione di individui, ciascuno identificato da una propria stringa, che ne definisce il comportamento. Tale stringa può essere assimilata al DNA. Sulla popolazione vengono essenzialmente eseguiti tre tipi di azioni:

1. *riproduzione*: Da una generazione (ovvero una popolazione in un certo istante) si selezionano gli individui più adatti alla sopravvivenza e si portano alla generazione successiva.

2. *crossover*: Gli individui di una popolazione scambiano fra di loro porzioni di DNA.

3. *mutazione*: Sezioni di DNA variano casualmente, con una frequenza fissata generalmente molto bassa.

2 Simulazione "a mano"

Per avere un'idea del funzionamento di un GA mostro un esempio di funzionamento molto semplice e di cui è ben nota una soluzione. Si tratta di trovare il massimo e il punto di massimo della funzione $f(x) = x^2$ nell'intervallo $[0, 31]$.

Come prima cosa è necessario definire un vocabolario che andrà a comporre le stringhe del DNA. Lo faccio nella maniera più semplice possibile, ovvero

$$V = \{0, 1\}$$

Inoltre definisco un individuo come un elemento dell'insieme V^5. Per esempio è un individuo `01001`. Ogni individuo codifica un certo valore di x, che per comodità scelgo essere la sua rappresentazione in base decimale. Il valore corrispondente a quello scelto in precedenza sarà dunque 9.

Per l'implementazione di un GA è inoltre necessaria una funzione detta di *fitness*, tale che, dato un individuo, ne indichi la sua idoneità. Nell'esempio

preso in esame è spontaneo scegliere f stessa. Quindi `01001` ha un fitness di $f(9) = 9^2 = 81$.

In questo esempio scelgo di avere una popolazione composta da $N = 4$ individui. In genere questo numero è molto più grande per avere un algoritmo efficiente. La popolazione iniziale è generata casualmente, ovvero ogni lettera di ogni stringa è il risultato di un lancio di moneta. Ottengo la seguente popolazione, con relativo fitness.

k	Stringa	Valore x	fitness
1	`01101`	13	169
2	`11000`	24	576
3	`01000`	8	64
4	`10011`	19	361
	totale		1170

2.1 Riproduzione

A questo punto voglio generare una nuova generazione a partire dagli individui con fitness maggiore. Per fare ciò ad ogni individuo i_k assegno la probabilità di riproduzione

$$p_k = \frac{f(i_k)}{\sum\limits_j f(i_j)}$$

dove al numeratore compare il fitness di tale individuo e a denominatore la somma di tutti i fitness. Risulta chiaro che a maggiore fitness corrisponde maggiore probabilità di riproduzione e che la somma di tutti i p_k valga 1. La tabella risulta dunque essere

k	Stringa	Valore x	fitness	p_k
1	`01101`	13	169	0.14
2	`11000`	24	576	0.49
3	`01000`	8	64	0.05
4	`10011`	19	361	0.31
	totale		1170	1.00

Una volta determinate le probabilità, scelgo casualmente gli individui per la nuova popolazione. Ciascuno di essi ha probabilità p_k di essere uguale all'individuo i_k. Su una popolazione di N individui ci aspettiamo quindi di avere Np_k individui uguali a i_k. Se si definisce il fitness medio

$$\overline{f} = \frac{\sum\limits_j f(i_j)}{N}$$

si ha che

$$Np_k = N\frac{f(i_k)}{\sum\limits_j f(i_j)} = \frac{f(i_k)}{\overline{f}}$$

Questo significa che se un certo individuo ha un fitness superiore alla media, ovvero $f > \overline{f}$, avrà $f/\overline{f} > 1$, cioè tenderà ad aumentare il suo numero di copie nella generazione successiva. Analogamente un individuo con un fitness inferiore alla media tenderà a diminuire il proprio numero di copie.

Vado dunque a eseguire la riproduzione sulla popolazione in esame, ottenendo i seguenti risultati.

k	Stringa	Valore x	fitness	p_k	$f/\overline{f}$	numero di individui nella nuova generazione
1	`01101`	13	169	0.14	0.58	1
2	`11000`	24	576	0.49	1.96	2
3	`01000`	8	64	0.05	0.21	0
4	`10011`	19	361	0.31	1.23	1
totale			1170	1.00	4.00	
media			229.5	0.25	1.00	
massimo			576	0.49	1.96	

2.2 Crossover

Supponiamo in una certa popolazione di avere individui con stringhe composte da un numero l di lettere tratte dal vocabolario V. Nel nostro esempio $l = 5$. Indico con i_k^j la j-esima lettera della stringa dell'individuo i_k. Vediamo cosa succede eseguendo un crossover su due individui i_1 e i_2. Scelgo casualmente un un intero c compreso tra 1 e $l-1$ detto sito di crossover. Il risultato del crossover è un individuo i'_1 tale che $i'^n_1 = i^n_1$ se $n \leq c$ e $i'^n_1 = i^n_2$ altrimenti. Allo stesso modo viene prodotto un individuo i'_2 tale che $i'^n_1 = i^n_2$ se $n \leq c$ e $i'^n_1 = i^n_1$ altrimenti. In altri termini da un certo punto in poi le due stringhe si scambiano tra di loro. Per eseguire un crossover su un'intera popolazione, si devono accoppiare casualmente gli individui, e poi eseguire il crossover su ciascuna coppia.

Partiamo dalla popolazione derivante dalla riproduzione e accoppiamo casualmente gli individui. Scegliamo inoltre casualmente il sito di crossover.

k	Stringa	Compagno	Sito di crossover
1	`01101`	2	4
2	`11000`	1	4
3	`11000`	4	2
4	`10011`	3	2

Andiamo a studiare la prima coppia. La prima stringa si spezza come `0110-1`, mentre la seconda come `1100-0`. Si otterranno dunque i nuovi individui `01100` e `11001`. Si procede in modo analogo sulla seconda coppia per ottenere `11011` e `10000`.

2.3 Mutazione

Durante la mutazione ciascuna lettera di ogni stringa ha un probabilità di $\varepsilon = 0.001$ di diventare un'altra lettera del vocabolario. In questo caso ci aspettiamo $Nl\varepsilon = 4 \cdot 5 \cdot 0.001 = 0.02$ mutazioni. Di fatto simulando non ne troviamo alcuna.

2.4 Analisi dei risultati

Andiamo infine ad analizzare quali sono stati i risultati delle tre operazioni. Nella tabella seguente sono presenti alcuni dati sulla seconda generazione comparati con la prima.

k	Stringa	Valore x	fitness gen. 2	fitness gen. 1
1	`01100`	12	144	169
2	`11001`	25	625	576
3	`11011`	27	729	64
4	`10000`	16	256	361
media			438.5	229.5
massimo			625	576

Risulta dunque evidente che si ha avuto un incremento del fitness sia nel valore massimo che nel valore medio.

3 Fondamenti matematici

Una grossa domanda che è necessario porsi per comprendere il perché del funzionamento dei GA è la seguente: quali informazioni porta una certa popolazione? Ovviamente una risposta è data dagli individui stessi, il cui fitness cresce con l'avanzare delle generazioni. Un secondo punto di vista si può trovare considerando le similitudini tra le stringhe della popolazione. In altre parole il miglioramento generale della popolazione tra una generazione e l'altra è parallelo all'accentuarsi di alcune similitudini. Per studiare questo processo si fa uso della nozione di schema.

3.1 Schemi

Consideriamo sempre una poplazione di N individui caratterizzati da stringhe di lunghezza l composte da lettere del vocabolario V. Scelgo un carattere jolly, che sarà identificato dal simbolo $*$. A questo si può costruire

un vocabolario esteso $V_+ = V \cup \{*\}$. Uno schema h è un elemento dell'insieme V_+^l. Proseguendo nell'esempio di prima, è uno schema `10*1*`. Si può stabilire una corrispondenza tra uno schema e un insieme di possibili stringhe, tale che ogni lettera dello schema diversa da `*` sia uguale a quella corrispondente nella stringa. Dunque allo schema `10*1*` saranno associate le stringhe `10010`, `10011`, `10110` e `10111`. Queste stringhe prendono il nome di rappresentanti dello schema.

Definisco ora due funzioni che saranno utili in seguito. La prima è l'ordine, che si denota con $o(h)$ ed è il numero di caratteri fissi (cioè diversi dal carattere jolly) nello schema. Quindi $o(\texttt{110*1**1}) = 5$. La seconda funzione è detta lunghezza caratteristica di uno schema ed è la distanza tra la prima lettera fissa e l'ultima. Si scrive come $\delta(h)$. Per esempio $\delta(\texttt{10*1**}) = 3$. Infatti il primo carattere fisso si trova in posizione 1 e l'ultimo in posizione 4. Dunque il valore che cerchiamo è $4 - 1 = 3$.

Inoltre possiamo estendere la funzione di fitness agli schemi. Sia h uno schema e $i_1, i_2, \ldots, i_n$ suoi rappresentanti. allora il fitness medio di h vale

$$f(h) = \frac{\sum_{j=1}^{n} f(i_j)}{n}$$

3.2 Schemi e riproduzione

Chiamiamo $m(h,t)$ il numero di rappresentanti di uno schema h nella generazione t. Parlando del fitness medio rispetto ai rappresentanti presenti nella popolazione, chiamati $r_1, r_2, \ldots, r_n$ con $n = m(h,t)$, ci aspettiamo che il numero di copie di r_j sia

$$\frac{f(r_j)}{\overline{f}}$$

e che dunque

$$m(h, t+1) = \frac{f(r_1)}{\overline{f}} + \frac{f(r_2)}{\overline{f}} + \cdots + \frac{f(r_n)}{\overline{f}} =$$

$$\frac{\sum_{j=1}^{n} f(r_j)}{\overline{f}} = \frac{n}{\overline{f}} \frac{\sum_{j=1}^{n} f(r_j)}{n} = n\frac{f(h)}{\overline{f}} = m(h,t)\frac{f(h)}{\overline{f}}$$

Possiamo scegliere una certa λ tale che $f(h) = (1+\lambda)\overline{f}$. Tale valore indica quanto il fitness di uno schema si discosta dalla media, e sarà positivo se h è un buono schema e negativo altrimenti. L'equazione precedente diventa dunque

$$m(h, t+1) = m(h,t)\frac{(1+\lambda)\overline{f}}{\overline{f}} = (1+\lambda)m(h,t)$$

Assumiamo che λ non vari significativamente, ovvero consideriamo uno schema che rimane sempre "buono" oppure sempre "cattivo". In tal caso si

avrà

$$m(h,t) = (1+\lambda)^t m(h,0)$$

Ovvero il numero di rappresentanti di uno schema crescerà o decrescerà con regime esponenziale.

3.3 Schemi e crossover

Occupiamoci ora di come il crossover influisca sul numero di rappresentanti di un certo schema. Prendiamo come esempio gli schemi

$$h_1 = \texttt{*1****0}$$
$$h_2 = \texttt{***10**}$$

di cui l'individuo $i = \texttt{0111000}$ è un rappresentante. Risulta evidente che è molto più facile che un crossover distrugga h_1 piuttosto che h_2. Infatti il primo schema viene distrutto dai crossover su i con sito $c = 2,3,4,5,6$. Per il secondo si può avere solo per $c = 4$. Risulta evidente che il numero di crossover "distruttivi" è $\delta(h)$. Il numero dei possibili è ovviamente $l-1$ con l lunghezza della stringa. Dunque la probabilità che uno schema venga salvato da un crossover è

$$p_\mathrm{s} = 1 - \frac{\delta(h)}{l-1}$$

Nel caso in cui un crossover distruttivo avvenisse tra due rappresentanti dello schema, lo schema rimarrebbe comunque intatto. Ne consegue che quello trovato prima è solo un limite inferiore. Si ha dunque che il contributo relativo al crossover al numero di rappresentanti di h è

$$m(h,t+1) \geq p_\mathrm{s} m(h,t) = m(h,t)\left(1 - \frac{\delta(h)}{l-1}\right)$$

3.4 Schemi e mutazione

Studiamo il comportamento degli schemi con una mutazione con frequenza ε, ovvero ogni carattere ha la probailità di modificarsi pari a ε. Risulta dunque evidente che la probabilità di uno schema di resistere ad una mutazione è pari a

$$p_\mathrm{r} = (1-\varepsilon)^{o(h)}$$

Siccome stiamo trattando valori di $\varepsilon \ll 1$, possiamo sviluppare al primo ordine per ottenere

$$p_\mathrm{r} = 1 - \varepsilon o(h)$$

e dunque considerando solo il contributo della mutazione si ottiene

$$m(h,t+1) = m(h,t)\left(1 - \varepsilon o(h)\right)$$

3.5 Teorema degli schemi

Non rimane che combinare le equazioni per ottenere che

$$m(h, t+1) \geq m(h, t)\,(1+\lambda)\left(1 - \frac{\delta(h)}{l-1}\right)(1 - \varepsilon o(h))$$

Ciò significa che schemi corti, con piccola lunghezza caratteristica e con un fitness sempre superiore alla media, tenderanno ad aumentare il numero di propri rappresentanti con andamento esponenziale. Questo risultato è noto come teorema degli schemi o teorema fondamentale degli algoritmi genetici. Infatti la vera potenza dei GA sta nel riuscire a processare parallelamente un grande numero di schemi.

4 Importanza della diversità tra individui

Consideriamo due generazioni consecutive t e $t+1$. Partendo da condizioni casuali (come nell'esempio) e procedendo a svolgere le varie operazioni, chiamiamo $p(i,t)$ la probabilità di trovare l'individuo i nella generazione t. Ovviamente su una popolazione di N individui, ci aspettiamo $Np(i,t)$ individui uguali a i. Per la legge della riproduzione

$$p(i, t+1) = p(i,t)\frac{f(i)}{\overline{f}(t)}$$

con $\overline{f}(t)$ fitness medio nella generazione t. Inoltre possiamo risrivere il fitness medio di una popolazione nei termini dei $p(i_j, t)$. Si ha infatti

$$\overline{f}(t) = \sum_j p(i_j) f(i_j)$$

Dimostriamo per induzione che $\sum_j p(i_j, t) = 1$. Ovviamente il caso base è facile e si ha

$$p(i_j, 0) = \frac{1}{|V|^l}$$

e siccome i possibili i_j sono in tutto $|V|^l$ la somma vale 1. Supponiamo allora che la somma valga 1 per una certa generazione t. Allora

$$\sum_j p(i_j, t+1) = \sum_j p(i_j)\frac{f(i_j)}{\overline{f}(t)} = \frac{\sum_j p(i_j) f(i_j)}{\overline{f}(t)} = \frac{\overline{f}(t)}{\overline{f}(t)} = 1$$

Definiamo la varianza come

$$\sigma^2(t) = \sum_j p(i_j, t)\left[f(i_j) - \overline{f}(t)\right]^2$$

dove la sommatoria è intesa su tutti i possibili individui. Tanto maggiore è la varianza tanto è maggiore la diversità tra gli individui nella popolazione. Si ha che

$$\begin{aligned}\sigma^2(t) = &\sum_j p(i_j,t)\left[f^2(i_j) - 2f(i_j)\overline{f}(t) + \overline{f}^2(t)\right] = \\ &\dots \\ &\texttt{...Passaggi algebrici...} \\ &\dots \\ &\sum_j p(i_j,t)\left[f^2(i_j) - f(i_j)\overline{f}(t)\right]\end{aligned}$$

Definiamo inoltre la risposta alla riproduzione

$$R(t) = \overline{f}(t+1) - \overline{f}(t)$$

che indica di quanto la riproduzione riesca a far crescere il fitness medio. Dunque

$$\begin{aligned}R(t) &= \sum_j \left[p(i_j,t+1) - p(i_j,t)\right] f(i_j) = \\ &\qquad\qquad\qquad\qquad\qquad\qquad \texttt{passggi algebrici} \\ &= \frac{1}{\overline{f}(t)} \sum_j p(i_j,t)\left[f^2(i_j) - f^2(i_j)\overline{f}(t)\right] = \\ &\qquad\qquad\qquad\qquad\qquad\qquad \frac{\sigma^2(t)}{\overline{f}(t)}\end{aligned}$$

Ne consegue che si ha un aumento maggiore nel fitness di una popolazione quando la varianza è grande, e cioè quando gli individui sono diversi tra di loro. A questo fine sono infatti volte le operazioni di crossover e mutazione.