Problema 1.3: el coeficient de Gini

Marc Valls Camps

Enunciat

El coeficient de Gini és una mesura de la desigualtat ideada per l'estadístic italià Corrado Gini. Normalment s'utilitza per mesurar la desigualtat en els ingressos, dins d'un país, però pot utilitzar-se per a mesurar qualsevol forma de distribució desigual. El coeficient de Gini és un nombre entre 0 i 1, on 0 es correspon amb la perfecta igualtat (tots tenen els mateixos ingressos) i on el valor 1 es correspon amb la perfecta desigualtat (una persona té tots els ingressos i els altres cap).

Formalment, si $r = (r_1, \ldots, r_n)$, amb $n > 1$, és un vector de valors no negatius, el *coeficient de Gini* es defineix com:

$$G(r) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |r_i - r_j|}{2(n-1) \sum_{i=1}^{n} r_i}$$

Proporcioneu un algorisme eficient per calcular el coeficient de Gini donat el vector r.

Solució

Definició alternativa del coeficient de Gini

Calcular un sumatori dins d'un altre sumatori tindria cost $O(n^2)$. Necessitarem una altra definició per millorar-ho. Ho aconseguirem analitzant quantes vegades contribueix a la suma total cada element. Sigui x el vector r ordenat de gran a petit:

1. $|r_i - r_j| = max(r_i - r_j, r_j - r_i) = max(r_i, r_j) - min(r_i, r_j)$
2. L'element més gran de tot el vector, x_1, sempre contribuirà en positiu, excepte quan $r_i = r_j = x_1$
3. El segon més gran, x_2, sempre contribuirà en positiu, excepte quan $r_i = r_j = x_1$, i en negatiu només quan $r_i = x_1, r_j = x_2$ o $r_i = x_2, r_j = x_1$
4. ...

Les taules següents ensenyen a que em refereixo:

i\j	1	2	3	...	n
1	0	+	+		+
2	+				
3	+				
...	...				
n	+				

Contribucions de x_1 a la suma total

i\j	1	2	3	...	n
1		-			
2	-	0	+	...	+
3		+			
...		...			
n		+			

Contribucions de x_2 a la suma total

i\j	1	2	3	...	n
1					-
2					-
3					-
...					...
n	-	-	-	...	0

Contribucions de x_n a la suma total

D'aquestes taules en deduïm que en realitat que l'element x_i contribueix sumant $2n - 2i$ vegades, i restant $2i - 2$ vegades. Independent de k, sempre passarà una vegada que $r_i = r_j = x_k$ i no es contribuirà ni sumant ni restant.

Si ho sumem, ens surt que cada element x_i contribuirà $2n - 2i - (2i - 2) = 2n - 4i + 2$ al total. Per tant, una definició alternativa del coeficient de Gini per a un vector ordenat x seria:

$$G(x) = \frac{\sum_{i=1}^{n} (2n - 4i + 2) \cdot x_i}{2(n-1) \sum_{i=1}^{n} x_i}$$

Codi en C++

En el codi s'ha adaptat la fórmula per tal de tenir en compte que treballem amb vectors indexats a 0.

```
#include<vector>
#include<algorithm>
using namespace std;

double calculate_gini(vector<double>& r){
    int n = r.size();

    sort(r.begin(), r.end(), greater<double>());

    double num = 0;
    double den = 0;
    for (int i = 0; i < n; ++i){
        num += (2*n -2 -4*i) * r[i];
        den += r[i];
    }

    den *= 2 * (n - 1);

    return num/den;
}
```

Correcció i terminació

La demostració de correcció i terminació és trivial, només s'està traduïnt a codi una fórmula matemàtica alternativa que hem trobat per al coeficient de Gini.

Cost de l'algorisme

Després d'ordenar decreixentment amb cost $O(n \log n)$, els càlculs es fan amb una sola passada, amb cost $O(n)$.

Cost: $O(n \log n)$