**Thuật toán Hạt Giống Nảy Mầm (Seed Germination)**

Giả sử hàm hồi quy*: Y = α1X1 + α2X2 +…. + αnXn*

Biến Y được gọi biến đáp ứng (response variable), các biến *Xi* được gọi là biến hồi quy (regressor). Biến đáp ứng từ dữ liệu thực tế được gọi *Y* thực tế (*Y-real*), biến đáp ứng tính từ hàm hồi quy gọi là *Y* ước lượng (*Y-esti*). Tương tự cho biến hồi quy *Xi-real* và *Xi-esti*.

Như vậy hàm hồi quy được đặc tả bằng các tham số *αi*

Bài toán đặt ra: dựa vào dữ liệu các biến Xi-real như: đường kính lưỡng đỉnh (ĐKLĐ), chu vi đầu (CVĐ), chiều dài xương tay (CDXT), chu vi bụng (CVB), chiều dài xương đùi (CDXĐ) để tính hàm hồi quy cho Y là trọng lượng thai (TLT) và tuổi thai (TT).

Quy ước

* + *α*(*r*) = {*αk, αk+1,…, αl*} là danh sách các tham số hồi quy của *r*
  + *Re*(*r*) là danh sách các biến hồi quy của *r*

**Giả định**

Một hàm hồi quy được cho là tối ưu khi thỏa hai điều kiện sau, được gọi **cặp điều kiện tối ưu**:

1. Độ tương quan giữa *Y-real* và *Y-esti* lớn nhất
2. Tổng residual (SS) nhỏ nhất

Từ nhận xét này, ta có 2 giả định cho thuật toán Hạt Giống Nảy Mầm (Seed Germination):

* Giả định thứ nhất: các biến hồi quy *Xi* trong hàm hồi quy tối ưu có xu hướng độc lập (từng đôi một) với nhau. Khái niệm độc lập được giảm nhẹ thành “các biến *Xi*, *Xj* từng đôi có hệ số tương quan gần bằng *0*, tức nhỏ hơn một ngưỡng ***δ***”. Đây là giả định Min.
* Giả định thứ hai: mỗi biến hồi quy *Xi* trong hàm hồi quy “đóng góp” nhiều cho hàm hồi quy. Khai niệm “đóng góp nhiều” được chuyển thành “hệ số tương quan giữa *Xi* và *Y* cao, tức lớn hơn một ngưỡng ***ε***”. Đây là giả định Max.

Thuật toán Hạt Giống Nảy Mầm cố gắng tìm những tổ hợp của các *Xi* sao cho thỏa mãn hai giả định trên. Suy ra, hàm hồi quy tối ưu là hàm mà các biến hồi quy thỏa hai điều kiện sau, được gọi là **cặp điều kiện nảy mầm**:

1. Từng đôi biến hồi quy *Xi, X­j* có hệ số tương quan *corr*(*X­i, Xj*) *< α* và nhỏ nhất
2. Mỗi biến *Xi* và *Y* có hệ số tương quan *corr*(*Xi, Y*) > β và lớn nhất

**Lưu đồ thuật toán Hạt Giống Nảy Mầm**

* Input*: L* làdanh sách tất cả các biến hồi quy *Xi* có thể, *L* = {*X1, X2,…, Xn*}
* Output: *seed* làhạt giống (hàm hồi quy tối ưu được trả về), được khởi tạo rỗng

Đặt danh sách *INPUT* là phần bù của *L* với *Re*(*seed*)

*INPUT* = *L* / *Re*(*seed*)

Đặt danh sách *GERM\_LIST* là danh sách các biến hồi quy thỏacặp điều kiện nảy mầm lấy từ *INPUT*.

Yes

*GERM\_LIST*  rỗng?

No

Duyệt từng biến hồi quy *X GERM\_LIST*, hội *X* với danh sách các biến hồi quy của *seed*. Ta có: *LIST = X Re*(*seed*)

Đặt *CANDIDATE* là danh sách các hàm hồi quy tối ưu mới

Tạo hàm hồi quy mới *r* từ *LIST*, nếu *r* thỏa cặp điều kiện tối ưu thì đưa *r* và *CANDIDATE: CANDIDATE = CANDIDATE*  {r}

NEW\_SEED = BEST\_SEED {S}

Gán *seed* vào *best: best = seed*

Gọi *best* là hàm hồi quy tốt nhất (thỏa cặp điều kiện tối ưu cao nhất) từ CANDIDATE.

*best* trùng với *seed* ?

Yes

No

**Thuật toán vét cạn (Brute-Force)**

Hạt Giống Nảy Mầm là thuật toán heuristic đệ quy dựa trên cặp giả định nảy mầm và chỉ cho kết quả một hàm khả dĩ tối ưu. Ngược lại, thuật toán vét cạn tìm hết cả cả các hàm hồi quy tối ưu nhưng chậm hơn Hạt Giống Nảy Mầm. Ý tưởng chính của vét cạn là duyệt tất cả các tổ hợp của tập các biến hồi quy và chọn tổ hợp nào thỏa cặp điều kiện tối ưu

**Lưu đồ thuật toán vét cạn**

* Input*: L* làdanh sách *n* biến hồi quy *Xi* có thể, *L* = {*X1, X2,…, Xn*}
* Output: *SEED\_LIST* làdanh sáchhạt giống (tập hàm hồi quy tối ưu được trả về). Số lượng tối đa của *SEED\_LIST* là *max*. SEED\_LIST được khởi tạo rỗng

Gọi *k* là chập của tổ hợp, nghĩa là số lượng biến hồi quy của hạt giống

*k* được khởi tạo là 1

No

*k n*

Yes

Tạo tổ hợp *COMB* chập *k* biến hồi quy từ *L*:

*COMB* = {*Xm+1, Xm+2,…, Xm+k*)

Xây dựng danh sách LIST các hàm hồi quy thỏa cặp điều kiện tối ưu. Đưa LIST và SEED\_LIST

*SEED\_LIST = SEED\_LIST LIST*

Chọn *max* hàm tối ưu nhất của SEED\_LIST

SEED\_LIST = {max hàm tối ưu nhất của SEED\_LIST}

*k = k + 1*