# CHƯƠNG II: Bài toán học máy

## 1. Bài toán phân cụm hạt giống lúa mì

### 1.1 Mô tả bộ dữ liệu

**Nguồn gốc:** Piotr Kulczycki, Piotr A. Kowalski, Szymon Lukasik,  
Khoa Công nghệ Thông tin và Điều khiển Tự động Slawomir Zak ,  
Đại học Công nghệ Cracow, Warszawska 24, PL 31-155 Cracow, Ba Lan  
và  
Viện Nghiên cứu Hệ thống, Học viện Khoa học Ba Lan, Newelska 6,  
PL 01 -447 Warsaw, Ba Lan  
e-mail: {kulczycki, pakowal, slukasik, slzak} @ ibspan.waw.pl  
  
Hồ sơ về nấm được rút ra từ Hướng dẫn thực địa về nấm ở Bắc Mỹ của Hiệp hội Audubon (1981). GH Lincoff (Pres.), New York: Alfred A. Knopf  
  
**Nhà tài trợ:**  
  
Jeff Schlimmer ( Jeffrey.Schlimmer **'@'** a.gp.cs.cmu.edu )

**Thông tin tập dữ liệu:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Đặc điểm của Tập dữ liệu:** | Đa biến | **Số phiên bản:** | 210 | **Khu vực:** | Đời sống |
| **Đặc điểm thuộc tính:** | Thực tế | **Số thuộc tính:** | 7 | **Ngày được tặng** | 2012-09-29 |
| **Nhiệm vụ liên quan:** | Phân loại, phân cụm | **Giá trị bị mất?** | N/A | **Số lượt truy cập web:** | 332411 |

Nhóm được kiểm tra bao gồm các hạt nhân thuộc ba giống lúa mì khác nhau: Kama, Rosa và Canadian, mỗi loại 70 nguyên tố, được chọn ngẫu nhiên cho  
thí nghiệm. Hình ảnh chất lượng cao của cấu trúc nhân bên trong đã được phát hiện bằng kỹ thuật tia X mềm. Nó không phá hủy và rẻ hơn đáng kể so với các kỹ thuật hình ảnh phức tạp khác như kính hiển vi quét hoặc công nghệ laser. Hình ảnh được ghi lại trên tấm KODAK X-quang 13x18 cm. Các nghiên cứu được thực hiện bằng cách sử dụng kết hợp hạt lúa mì thu hoạch có nguồn gốc từ các cánh đồng thí nghiệm, được khám phá tại Viện Vật lý Nông học của Viện Hàn lâm Khoa học Ba Lan ở Lublin.  
  
Tập dữ liệu có thể được sử dụng cho các nhiệm vụ phân loại và phân tích cụm.

**Thông tin thuộc tính:**

Để xây dựng dữ liệu, người ta đo bảy thông số hình học của hạt lúa mì:  
1. diện tích A,  
2. chu vi P,  
3. độ nén C = 4 \* pi \* A / P ^ 2,  
4. chiều dài của nhân,  
5. chiều rộng của nhân ,  
6. hệ số không đối xứng  
7. chiều dài của rãnh nhân.  
Tất cả các tham số này đều có giá trị thực liên tục.

### 1.2 Mô tả bài toán

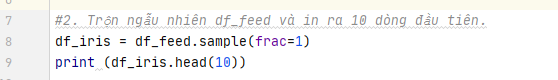
Bài toán sử dụng bộ dữ liệu các thuộc tính vật lý của hạt giống lúa mì. Áp dụng ký thuật K-Means Claustering

1.3 Mô tả chi tiết về cách xử lý dữ liệu

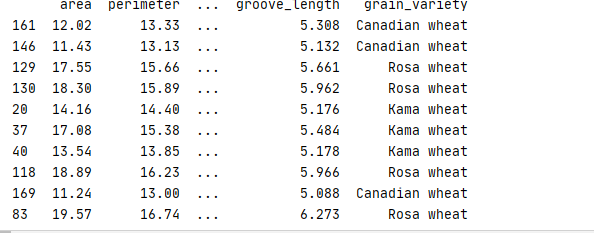
Đọc file dữ liệu seeds.csv



Trộn bộ df\_feed và in ra 10 dòng đầu tiên :



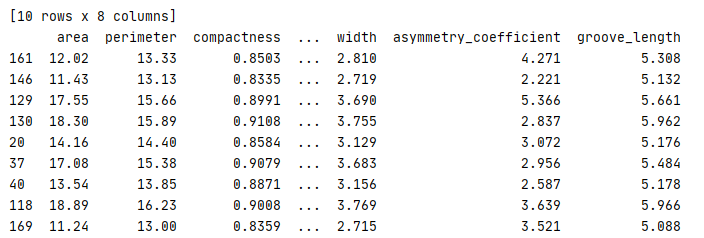
Kết quả



Gán X là 7 cột (area ,perimeter, compactness, length, width, asymmetry\_coefficent, groove\_length) và hiển thị 10 hàng đầu:



Kết quả:



1.4 Mô tả về KMeans

***Thuật toán K-Means được mô tả như sau:***





















 Hình 1. Thuật toán K-means

* ***Thuật toán K-Means thực hiện qua các bước chính sau:***

1.    Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

2.    Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)

3.    Nhóm các đối tượng vào nhóm gần nhất

4.    Xác định lại tâm mới cho các nhóm

5.    Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

Thuật toán k-means có thể được chia thành các bước như sau:

Bước 1: Tạo các trung tâm ngẫu nhiên

https://upload.wikimedia.org/wikipedia/commons/thumb/c/c6/Random_centers.png/220px-Random_centers.png

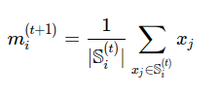
Bước 2: Gán các điểm dữ liệu vào các cụm

Với mỗi điểm dữ liệu, ta sẽ tính khoảng cách của nó tới các trung tâm (bằng [Khoảng cách Euclid](https://vi.wikipedia.org/wiki/Kho%E1%BA%A3ng_c%C3%A1ch_Euclid)). Ta sẽ gán chúng vào trung tâm gần nhất. Tập hợp các điểm được gán vào cùng 1 trung tâm sẽ tạo thành cụm.

https://upload.wikimedia.org/wikipedia/commons/thumb/a/a3/Clustering_point.png/414px-Clustering_point.png

Bước 3:Cập nhật trung tâm

Với mỗi cụm đã tìm được ở bước 2, trung tâm mới sẽ là [trung bình cộng](https://vi.wikipedia.org/wiki/Trung_b%C3%ACnh_c%E1%BB%99ng_%C4%91%C6%A1n_gi%E1%BA%A3n) của các điểm dữ liệu trong cụm đó.

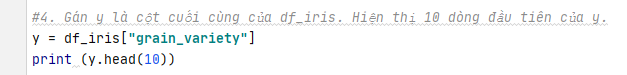


Thuật toán sẽ lặp lại các bước trên cho tới khi đạt được kết quả chấp nhận được.

Ngắn gọn phương pháp k-means là 1 trong số những thuật toán phân cụm (clustering). Đầu vào tập dữ liệu cần phân cụm và số cụm (cluster), đầu ra chúng ta sẽ được kết quả dữ liệu đã được phân về các cluster.

Mục đích là phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

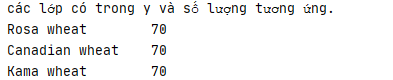
Gán y là cột cuối cùng của df\_iris.



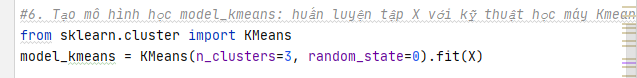
Hiện thị các lớp có trong y và số lượng tương ứng



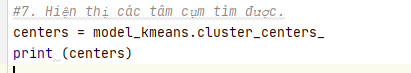
Kết quả:



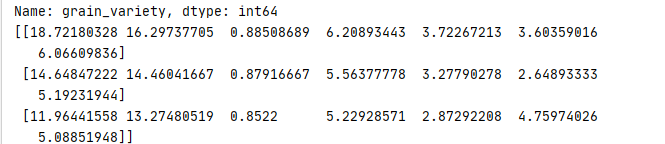
Tạo mô hình học model\_kmeans: huấn luyện tập X với kỹ thuật học máy Kmeans (số cụm = 3).



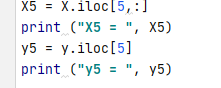
Hiện thị các tâm cụm tìm được.



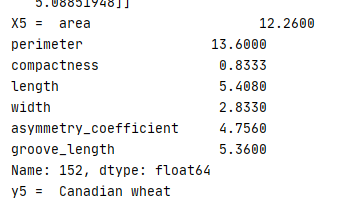
Kết quả:



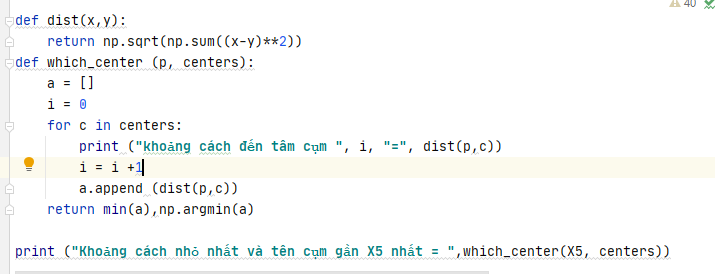
Hiện thị điểm dữ liệu X5 (dòng số 5 của X) và nhãn y5 tương ứng.



Kết quả:



Tìm cụm gần điểm X5 nhất.



Kết quả:

