SoSanhPhanCumNoiChung

May 30, 2024

0.0.1 1. K-Means Clustering

Ưu điểm:

1. **Đơn giản và dễ hiểu**: K-Means là thuật toán đơn giản và dễ triển khai, thường được sử dụng như một phương pháp phân cụm cơ bản.
2. **Tốc độ nhanh và hiệu quả**: Với thời gian tính toán là $O(nkdi)$ (trong đó n là số lượng điểm dữ liệu, k là số cụm, d là số chiều dữ liệu, i là số vòng lặp), K-Means rất nhanh và hiệu quả cho các tập dữ liệu lớn.
3. **Dễ dàng mở rộng**: K-Means có thể dễ dàng mở rộng cho dữ liệu lớn và được tối ưu hóa để sử dụng trong các môi trường tính toán phân tán.
4. **Phổ biến và có nhiều thư viện hỗ trợ**: Có sẵn trong nhiều thư viện phân tích dữ liệu như Scikit-learn (Python), WEKA (Java).

Nhược điểm:

1. **Yêu cầu xác định số cụm trước**: Người dùng phải xác định trước số lượng cụm k, điều này có thể khó khăn nếu không biết trước cấu trúc dữ liệu.
2. **Nhạy cảm với giá trị ban đầu**: Kết quả của K-Means có thể bị ảnh hưởng bởi giá trị khởi tạo ban đầu. Sử dụng phương pháp khởi tạo thông minh như K-Means++ có thể cải thiện điều này.
3. **Giả định cụm hình cầu và cùng kích thước**: K-Means hoạt động tốt với các cụm có hình cầu và cùng kích thước, nhưng không hiệu quả với các cụm có hình dạng phức tạp hoặc kích thước không đều.
4. **Không xử lý tốt dữ liệu nhiễu**: K-Means không xử lý tốt các điểm nhiễu và ngoại lệ.

0.0.2 2. Gaussian Mixture Model (GMM)

Ưu điểm:

1. **Linh hoạt hơn K-Means**: GMM không giả định các cụm có hình dạng cầu hoặc cùng kích thước, thay vào đó mô hình các cụm là các phân phối Gaussian, cho phép xử lý các cụm có hình dạng phức tạp.
2. **Xác suất gán cụm**: GMM gán điểm dữ liệu cho cụm dựa trên xác suất, điều này cho phép mô hình xử lý tốt hơn các trường hợp ranh giới giữa các cụm không rõ ràng.
3. **Xử lý dữ liệu nhiễu tốt hơn**: GMM có thể xử lý tốt hơn các điểm nhiễu so với K-Means nhờ khả năng gán xác suất.

Nhược điểm:

1. **Phức tạp hơn và yêu cầu nhiều tài nguyên tính toán hơn**: GMM phức tạp hơn K-Means và yêu cầu nhiều tài nguyên tính toán hơn, đặc biệt là khi số chiều của dữ liệu lớn.
2. **Cần xác định số cụm trước**: Tương tự như K-Means, GMM cũng yêu cầu xác định trước số lượng cụm.
3. **Nhạy cảm với giá trị ban đầu**: GMM cũng bị ảnh hưởng bởi giá trị khởi tạo ban đầu, và việc khởi tạo không tốt có thể dẫn đến hội tụ kém.

0.0.3 3. Agglomerative Hierarchical Clustering

Ưu điểm:

1. **Không yêu cầu xác định số cụm trước**: Khác với K-Means và GMM, Agglomerative Hierarchical Clustering không yêu cầu xác định trước số cụm. Số cụm có thể được xác định bằng cách cắt dendrogram tại một mức độ nhất định.
2. **Tạo ra cấu trúc phân cấp**: Phương pháp này tạo ra một cây phân cấp (dendrogram), giúp hiểu rõ hơn về cấu trúc phân cụm của dữ liệu.
3. **Đa dạng về phương pháp hợp nhất cụm**: Có nhiều phương pháp hợp nhất cụm (single linkage, complete linkage, average linkage, Ward's method) giúp tùy chỉnh quá trình phân cụm cho từng loại dữ liệu cụ thể.

Nhược điểm:

1. **Chi phí tính toán cao**: Thời gian tính toán là $O(n^3)$ và yêu cầu bộ nhớ $O(n^2)$ đối với dữ liệu lớn, do đó khó áp dụng cho các tập dữ liệu rất lớn.
2. **Khó khăn trong việc mở rộng**: Không dễ mở rộng cho dữ liệu lớn và phân tán.
3. **Nhạy cảm với dữ liệu nhiễu**: Có thể bị ảnh hưởng bởi dữ liệu nhiễu và các điểm ngoại lệ.

0.0.4 Tổng kết

Thuật toán	Ưu điểm	Nhược điểm
K-Means	Đơn giản, nhanh, dễ mở rộng, phổ biến	Cần xác định số cụm trước, nhạy cảm với giá trị ban đầu, giả định cụm hình cầu, không xử lý tốt dữ liệu nhiễu
GMM	Linh hoạt hơn, xác suất gán cụm, xử lý dữ liệu nhiễu tốt hơn	Phức tạp, yêu cầu nhiều tài nguyên, cần xác định số cụm trước, nhạy cảm với giá trị ban đầu
Agglomerative Clustering	Không yêu cầu xác định số cụm trước, tạo ra cấu trúc phân cấp, đa dạng phương pháp hợp nhất cụm	Chi phí tính toán cao, khó mở rộng, nhạy cảm với dữ liệu nhiễu

Silhouette Score là một thước đo để đánh giá chất lượng của các cụm trong phân cụm. Nó tính toán mức độ tương đồng của một điểm dữ liệu với cụm của chính nó (độ kết nối) so với các cụm khác (độ tách biệt).

0.0.5 Công thức tính Silhouette Score

Đối với mỗi điểm dữ liệu i, Silhouette Coefficient $s(i)$ được tính theo các bước sau:

1. **Tính** $a(i)$: Là khoảng cách trung bình từ điểm dữ liệu i đến tất cả các điểm khác trong cùng cụm với nó. Điều này đo lường mức độ tương đồng của i với cụm của nó. $a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, j \neq i} d(i,j)$ Trong đó:
   - C_i là cụm chứa điểm i.
   - $d(i,j)$ là khoảng cách giữa điểm i và điểm j.
2. **Tính** $b(i)$: Là khoảng cách trung bình từ điểm dữ liệu i đến tất cả các điểm trong cụm gần nhất không chứa i. Điều này đo lường mức độ tách biệt của i với cụm gần nhất không chứa nó. $b(i) = \min_{C_k \neq C_i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$ Trong đó:
   - C_k là cụm khác với C_i.
3. **Tính Silhouette Coefficient** $s(i)$: Công thức tính toán cho mỗi điểm dữ liệu i: $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$ Silhouette Coefficient $s(i)$ nằm trong khoảng [-1, 1]. Giá trị $s(i)$ càng gần 1 thì điểm i càng phù hợp với cụm của nó và không phù hợp với cụm lân cận. Giá trị $s(i)$ gần 0 cho thấy điểm i nằm giữa hai cụm, và giá trị $s(i)$ gần -1 cho thấy điểm i có thể đã được phân cụm sai.
4. **Tính Silhouette Score cho toàn bộ tập dữ liệu**: Trung bình Silhouette Coefficient của tất cả các điểm dữ liệu: Silhouette Score $= \frac{1}{n} \sum_{i=1}^{n} s(i)$ Trong đó:
   - n là tổng số điểm dữ liệu.

0.0.6 Ý nghĩa của Silhouette Score

- **Silhouette Score** gần 1: Các cụm được phân biệt rõ ràng và các điểm dữ liệu được gán đúng cụm.
- **Silhouette Score** gần 0: Các điểm dữ liệu nằm ở ranh giới giữa hai cụm, khó xác định cụm chính xác.
- **Silhouette Score** gần -1: Các điểm dữ liệu có thể đã bị phân cụm sai, nên được xem xét lại.

0.0.7 Ví dụ tính toán

Giả sử ta có ba cụm C_1, C_2, C_3 và một điểm dữ liệu i thuộc cụm C_1: 1. **Tính** $a(i)$: Khoảng cách trung bình từ i đến các điểm khác trong C_1. 2. **Tính** $b(i)$: Khoảng cách trung bình từ i đến các điểm trong cụm lân cận gần nhất (giả sử là C_2). 3. **Tính** $s(i)$: Sử dụng công thức $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$.

Tóm lại, Silhouette Score cung cấp một cách trực quan và dễ hiểu để đánh giá chất lượng phân cụm, giúp hiểu rõ hơn về mức độ gắn kết và tách biệt của các cụm trong dữ liệu.