

3.3.3. Khoảng cách giữa các đa giác

Đa giác được định nghĩa là một mặt phẳng, được giới hạn bởi hữu hạn các đoạn thẳng nối với nhau để tạo thành một đa giác kín. Đa giác lồi là đa giác không có đoạn thẳng nối giữa hai điểm trên đường biên đi ra ngoài đa giác. Khoảng cách giữa các đa giác lồi có thể là (hình 3.6):

- khoảng cách giữa hai đỉnh
- khoảng cách giữa một cạnh và một đỉnh
- khoảng cách giữa hai cạnh
- 0 khi hai đa giác giao nhau/chồng lên nhau (có thể xác định bằng thuật toán Separating Axis [19])

Trong ba trường hợp đầu tiên, ta có thể xử lý bằng cách xem mỗi cặp cạnh-đỉnh là một đoạn thẳng và liệt kê khoảng cách giữa tất cả các cặp đoạn thẳng. Khoảng cách từ một điểm tới một cạnh có thể tính như sau [33]:
Cho điểm P là hình chiếu vuông góc của điểm C trên cạnh AB .

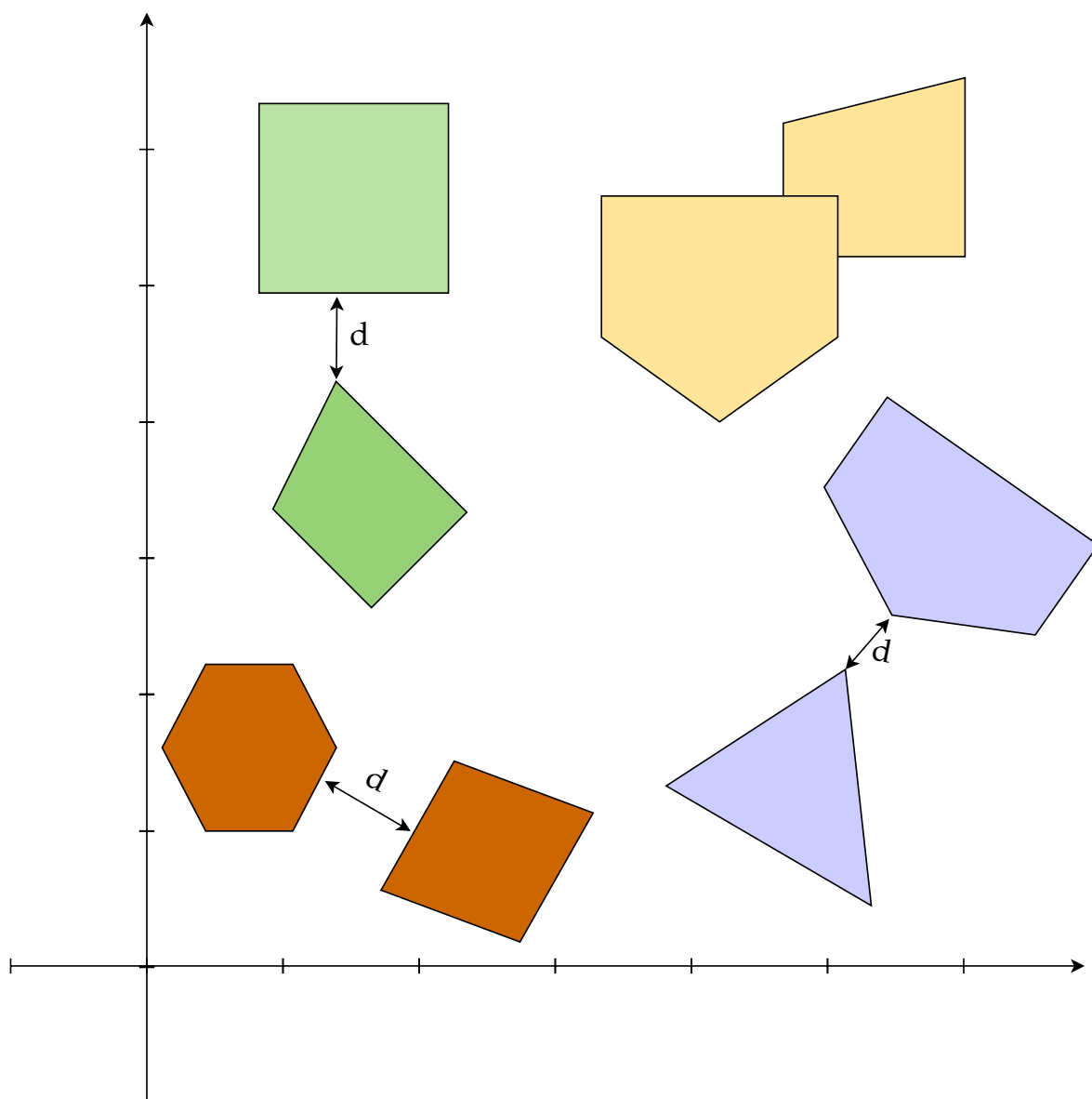
$$r = \frac{\vec{AB} \cdot \vec{AC}}{|\vec{AB}|^2} \quad (3.2)$$

Ta có:

- $P \equiv A$, nếu $r = 0$
- $P \equiv B$, nếu $r = 1$
- $P \in AB$, nếu $0 < 1 < r$
- P nằm trên phần kéo dài của AB

Khi đó,

$$x_P = x_A + r(x_B - x_A), y_P = y_A + r(y_B - y_A)$$



Hình 3.6: Các trường hợp có thể khi xác định khoảng cách giữa các đa giác

Tính giá trị s như sau:

$$s = \frac{(y_A - y_C)(x_B - x_A) - (x_A - x_C)(y_B - y_A)}{(x_B - x_A)^2 + (y_B - y_A)^2}$$

Khoảng cách C tới P , hay khoảng cách từ C tới AB là

$$d(C, AB) = |s| \times |\overrightarrow{AB}|$$

3.3.4. Giá trị ngoại lai

Giá trị ngoại lai là điểm giá trị có sự khác biệt hoàn toàn so với phần còn lại của tập dữ liệu. Một giá trị ngoại lệ có thể xảy ra trong bất kỳ phân phối nào và gây ra ảnh hưởng lớn tới trong phân tích thống kê. Có thể phát hiện giá trị ngoại lệ bằng cách trực quan hóa hoặc dựa trên các phương pháp giả định điểm thử, như:

- Tiêu chuẩn Chauvenet: phương pháp này đưa ra giả định khi xem xét việc loại bỏ dữ liệu và yêu cầu về tính chuẩn hóa của dữ liệu
- Tiêu chuẩn Peirce: áp dụng cho phân phối chuẩn, có khả năng tìm và loại bỏ nhiều hơn một giá trị ngoại lệ vì không đưa ra giả định như tiêu chuẩn Chauvenet
- Kiểm thử Grubbs: còn gọi là thử nghiệm dư tối đa chuẩn hóa. Tại một thời điểm, nó phát hiện một ngoại lệ và các điểm được kiểm tra này luôn là giá trị lớn nhất hoặc nhỏ nhất của tập dữ liệu. Phương pháp này không nên áp dụng với tập dữ liệu có kích thước nhỏ hơn 6.
- Kiểm thử Q-test của Dixon: tiêu biểu là phương pháp dựa trên Interquartile range và phương pháp này chỉ nên sử dụng một lần cho tập dữ liệu.
- Z-score: là một phương pháp đơn giản dựa trên giá trị trung bình và độ lệch chuẩn, áp dụng cho dữ liệu có phân phối chuẩn.