Distance Functions of vectors

거리 측정 함수 를 알아보자

왜 쓰는가?

데이터 혹은 풀고자 하는 문제 마다 적합한 거리 함수를 써야함

알고리즘은 그대로 인데 거리 함수 에 따라 성능이 완전 달라짐

각 거리 함수의 특징을 마스터 해야함!!

-유클리디안 거리 측정

피타고라스 정리에 의해 만들어진 최소 직선 거리 이다.

차원의수가 작을때 백터의 크기를 측정 하는게 중요하면

유클리드 쓰는게 제일 좋다(KNN 이웃하는 거리를 판별하기 위해 사용)

단점

정규화가 없다고 가정 하에

차원의 영향력을 많이 받는다

차원이 커짐에 따라서 데이터가 더 필요해 지고 거리의 개념이 많이 없어짐

차원이 많아지면 데이터가 많아짐 = 거리가 더 멀고 1의 값이랑 비슷해짐 결국 거리가 의미가 없어짐

상대적으로 데이터간 거리가 줄어든다.

해결책

정규화 를 한다

다리가 있으면 1 없으면 0 =binary dtm 이 필요 할수 있다.

파충류 vs 양서류

개구리 vs 뱀 악어

뱀 악어의 거리가 악어 와 개구리의 거리 보다 작아야함 양서류니까

그러나 다리가 4개인 악어와 개구리는 유사성 이 존재 한다.

그래서 유클리드 값이 개구리 악어가 더 작아져서

더 유사하게 나타난다 파충류 와 양서류 인데

-맨해튼 거리

A - B 사이의 거리를 x 축이나 y축을 직선으로만 가서 조달하는 거리

D = sum(|x -y|)

차원의 저주를 받는다

Outlier 에 강건하다는 점이 문제 일수도 있다

Minkowski = 유클리디와 맨해튼의 일반화 공식

거리가 백터의 차원에 비례한다

-코사인 거리

차원의 비례한 거리를 방지하는 방식

벡터의 크기만을 고려하지 않고 각도로 판단하는방식

문서의 길이의 차가 큰경우 = 편차 가 큰경우다?

뉴스에 달린 댓글 유사도 = 블로그 포스팅 유사도 = 단어의 개수 차이가 크다

유클리드와 맨해튼으로 구하면 카운팅 만으로고 거리를 나눌수 없다.