k-NN 보고서

이름:곡우환 학번:20194392

1. 프로젝트 소개

비디오 콘텐츠 만족도 조가에 대한 20000개의 설문 응답 데이터 세트가 있다. 이 데이터 세트가 사용자 6가지의 특성이 있고 사용자가 콘텐츠의 좋음과 나쁨중 하나를 선택하여 응답하게 되어 있다. 이 데이터와 k-NN알고리즘을 이용하여 새로운 사용자에 대해 콘텐츠의 만족과 불만족을 분류한다. 즉 새로운 사용자에게 해당 콘텐프의 만족도를 추정한다.

1. 사용하는 프로그램 언어: 파이썬, Jupyter Notebook
2. 프로그램 실행 방법:
3. 훈련 데이터와 데스트 데이터를 만든다.

훈련 데이터 세트와 데스트 데이터 세트를 렌텀으로 9:1로 나눈다. 렌텀으로 10번씩 나눠서 진행한다. 즉 훈련 데이터 세트 10개와 데스트 데이터 세트10개가 있다. 훈련 데이터 세트와 데스트 데이터 세트를 나누는 알고리즘(함수)를 직접 만든다.

文本

描述已自动生成

1. k-NN 알고리즘을 구성한다.

Step1: 유사 판정 기준을 정의한다. 데이터 포인트 사이의 Euclidean 거리 크기를 통해 유사를 정의한다. Euclidean 거리는 유클리드 공간에서 두 점 사이의 거리다.



文本

描述已自动生成

Step2: 사용자에 대해 콘텐츠의 만족과 불만족을 분류하는 알고리즘(k-NN)을 구현한다. 모든 데이터와 목표 데이터 지점 사이의 Euclidean 거리를 계산한다. Euclidean 거리가 가장 작은 k개 데이터의 인덱스를 반환한다. 가장 유사한 k개 데이터 포인트중에서 만족하는 데이터 개수와 불만족하는 데이터 개수를 통계한다. 두 클래스의 데이터 포인트가 어느 클래스 데이터 표인트 개수 더 많은지를 비교한다. 최종 목표 데이터 클래스는 더 많은 데이터 포인트가 있는 클래스다.

文本

描述已自动生成

1. k-NN알고리즘을 평가방법을 구현

분류기 성능을 평가한다. 먼저 불만족을 0, 만족을 1로 표시한다.

분류기 성능을 평가방법을 구현한다. 분류결과와 맞는 데이터 개수를 통계한다. 분류 정확률=(분류결과와 맞는 데이터 개수)/(총 분류 데이터 개수)

文本

描述已自动生成

表格

中度可信度描述已自动生成

1. 새로운 사용자의6가지 특징을 입력 및 분류.

文本

描述已自动生成

1. 실험(분류) 결과

图表, 折线图

描述已自动生成

1. 결론 분석

결과 그림에 따라 데스트 데이터 세트는 k값(3,5,7,9,11,13,15,17,19,21,23,25,27,29) 이 구간에서 클수록 분류의 정확도는 떨어진다. 또한 k값이 클수록 프로그램 실행시간도 느려진다. 훈련 데이터 세트는 데이터 양이 많아서 프로그램 실행시간 아주 길다.