패턴 인식 49147 2022년 봄학기 프로젝트 과제
제출 마감: 2022년 5월 7일(토) 24:00까지

데이터 세트: 비디오 콘텐츠 만족도 조사에 대한 20,000개의 설문 응답 데이터 세트가 있다. 이 만족도 조사에서는 서비스 공급자는 응답자(사용자)의 콘텐츠 이용 특성 가운데에서 6가지의 특성에 따라 사용자의 특성을 구별하고, 사용자가 콘텐츠의 좋음과 나쁨 중 하나를 선택하여 응답하게 되어 있다. 이 데이터를 활용하여 새로운 사용자에게 해당 콘텐츠의 만족도를 추정하고자 한다. 사용자의 특성 정보는 검증을 거친 유효한 정보이며 사용자의 특성 정보명은 공개되지 않는다.

프로젝트 내용: 새로운 사용자의 6가지 특징이 입력되었을 때 이 사용자의 만족도를 예측합니다. 20,000개의 설문 응답 데이터와 k-NN 알고리즘을 이용하여 새로운 사용자에 대해 콘텐츠의 만족과 불만족을 분류합니다. 이 예측 결과는 신규 사용자의 만족도 조사 추정에 사용되며 해당 사용자들의 실제 응답을 토대로 모델의 정확성을 판단합니다.

1. 데이터 세트를 9:1로 나누어 실제 알고리즘에 사용할 훈련 데이터와 성능 확인을 위한 테스트 데이터 세트로 사용한다. 훈련 데이터와 테스트 데이터는 배타적으로 나누어야 하며 서로 다르게 나눈 훈련/테스터 데이터로 나누어진 세트를 10개 이상 구성하여 분류 결과를 도출해야 한다. 즉 2만개의 데이터를 18000개와 2000개로 나누는 세트를 10개 이상 서로 다르게 구성해서 사용한다. (랜덤 선택된 세트를 10개 구성)
2 k-NN 알고리즘을 적용하여 예측 시스템을 구성하고 프로그램 코드로 구현하고 테스트 데이터 세트에 대한 분류 결과를 측정한다. 분류기의 성능은 훈련/테스트 데이터 세트의 조합을 각각 이용하여 측정한 후 합산하여 도출한다.

제출할 내용: 다음의 내용을 자신의 학번으로 파일명으로 하는 하나의 압축 파일을 만들고 과제란에 마감 이전에 제출해야 합니다.

1. 실행 파일 및 소스 코드 (윈도우 또는 리눅스에서 실행 가능한 파일과 실행 파일을 생성 가능한 보조 파일도 함께 제출해야 합니다.)
2. 자신이 생성한 훈련/테스트 데이터 세트
3. 프로그램은 소스 코드와 데이터 파일을 읽어 응답자별 추정 만족도를 <학번>.csv 파일에 저장한 결과 파일. (올바른 파일 형식이 아닐 경우 0점)
4. 분류기의 설계, 소프트웨어 구조와 실험 결과를 설명하는 보고서 (A4 일반 여백, 폰트 10, 행간 130%, 10장 내외) 보고서에는 사용한 언어와 프로그램 실행 방법을 반드시 기재하며 알고리즘, 구현 방법, 결과 등을 상세히 기술한다.

평가 기분 및 방법: 다음의 항목을 기준으로 과제물을 평가합니다.

1. 구현 내용 30% (소스 코드의 완성도, 제출한 결과의 신뢰도)
2. 분류기 성능: 30% (미공개 평가용 데이터 파일을 이용하여 분류 정확도와, 실행 시간을 측정하여 평가함. 프로그램 실행 시간은 각자 제출한 보고서에 기재된 프로그램 실행 방법을 기준으로 측정한다.)
3. 보고서: 40% (완성도15%, 내용15%, 형식:10%)

* 주의 사항
  - **만일 두 사람의 과제 내용이나 데이터 세트의 구성이 일반적 수준 이상으로 같으면 두 학생의 성적 모두 낙제 처리합니다.**
  - 제한된 마감 시간 이후에 제출되는 어떠한 형태의 것도 평가에서 제외됩니다. 반드시 제한 시간 이전에 업로드 완료해야 합니다. 마감에 인접하여 통신장애로 인해 업로드하지 못한 것도 인정되지 않습니다. 미리미리 업로드하기 바랍니다. (수정이 필요하다면 재업로드 하면 됩니다.)
  - **과제 제출물을 제출하지 않으면 나머지 시험 점수와 관계없이 낙제(F)입니다.**