자료와 정보 57458 2024년 봄학기 탐구 과제 2
제출 마감: 2024년 6월 22일(토) 23:59까지

탐구 내용: 가우시안 분포를 띠는 어떤 데이터 세트에서 최대 우도 추정(MLE)에 의해 찾은 확률 분포와 파젠창과 k-NN을 이용하여 추정한 확률 분포의 정확도를 분석한다.

1. 자신의 키와 몸무게를 평균값으로, 자기 학번의 임의의 두 자릿수를 골라 각각 키와 몸무게의 분산(σ^2)으로 설정한 2차원 가우시안 함수를 정한다. 이 함수의 분포가 되도록 난수 발생기를 이용하여, 샘플의 수가 1000, 5000, 10,000개인 랜덤 변수의 샘플 집합을 각각 10개씩 모두 30개를 만든다.
2. 30개의 데이터 세트를 이용하여 최대 우도 추정 방법으로 각각의 세트에 대한 가우시안 분포의 파라미터를 구한다.
3. 30개의 데이터를 이용하여 파젠창의 방법으로 각각의 세트에 대한 확률 분포를 구한다. 파젠창의 크기는 1에서 설정한 표준편차(σ)의 1배, 2배, 3배의 크기를 사용한다.
4. 30개의 데이터를 이용하여 k-NN 방법으로 각각의 세트에 대한 확률 분포를 구한다. k의 크기는 3,5,7에 대하여 확률 분포를 구한다.
5. 2, 3, 4번 항목에서 구한 확률 분포와 1번 항목에서 설정한 확률 분포 사이에 발생하는 오차를 구하고, 30개의 데이터 세트를 이용하여 얻은 파라미터와 파젠창의 크기, k의 크기에 따라 어떻게 오차의 분포가 달라지는지를 분석한다.

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. 프로그램은 소스 코드, 데이터 파일을 읽어 분할표를 <학번>.csv 파일에 저장한 결과 파일. 설정 조건에 따라 다수의 분할표를 만든다면 <학번>-<순번>.csv로 생성 (올바른 파일 형식이 아닐 경우 0점)
4. 분석 결과를 도출하는 조건, 소프트웨어 구조와 실험 결과를 설명하는 보고서 (A4 일반 여백, 폰트 10, 행간 130%, 5장 내외).

평가 기준 및 방법: 다음의 항목을 기준으로 과제물을 평가합니다.

1. 탐구할 내용: 30% (탐구 과정의 완성도, 제출한 결과의 신뢰도)
2. 분석 방법: 30% (실험 방법, 오차 분석 방법은 각자 제출한 보고서에 기재된 프로그램 실행 방법을 기준으로 측정한다.)
3. 보고서 완성도: 40% (내용:20%, 형식:20%)

* 주의 사항
- **만일 두 사람의 과제 내용이나 데이터 세트의 구성이 일반적 수준 이상으로 같으면 두 학생의 성적 모두 낙제 처리합니다.**
- 오차를 분석할 때 분석 도구를 이용하면 50%만 인정합니다. 자신이 라이브러리등의 도움 없이 분석 코드를 직접 작성하면 100% 인정합니다. (단 데이터 생성용 난수 발생기는 라이브러리를 사용해도 좋습니다.)
- 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 다시 업로드하면 됩니다.)
- **과제 제출물을 제출하지 않으면 나머지 시험 점수와 관계없이 낙제(F)입니다. 단 과제 1이나 과제 2중 하나만 완료하면 됩니다. 둘다 제출하면 높은 점수를 인정합니다.**