**Machine Learning**

**NAIVE BAYESIAN CLASSIFIER DESIGN**

**담당교수 : 박철수**

**제출일 : 2022. 04. 17.**

**학과 : 컴퓨터정보공학부**

**학번 : 2017202088**

**이름 : 신해담**

1. **과제 목적**
   * 3종의 꽃과 4 feature로 구성된 주어진 data를 가지고 training해서, feature를 통해 꽃의 종류를 분류하는 NAÏVE BAYESIAN CLASSIFIER를 디자인한다. 각 feature는 가우시안 분포를 따른다고 가정하며, bayesian theorem을 이용해서 확률을 계산한다. 구현한 classifier는 무작위 test data에 대해 90% 이상의 정확도를 보여야 한다.
2. **코드 설명**
   * **feature\_normalization**  
     입력 받은 data를 각 feature에 대해 정규화해서 반환하는 함수이다. 각 feature data 값에 평균을 빼고 표준편차로 나누어서, 해당 feature에 대한 data가 평균이 0이고 표준편차가 1인 표준 정규분포를 따르도록 할 수 있다.  
     텍스트이(가) 표시된 사진

     자동 생성된 설명
   * **split\_data**  
     입력 받은 data와 label을 split\_factor를 기준으로 분할해서 반환하는 함수이다. training을 통해 classifier를 디자인한 다음에는 구현한 classifier가 얼마나 정확한지 test가 필요하므로, 미리 data를 어디에 사용할 것인지 구분해 놓을 필요가 있다.  
     
   * **get\_normal\_parameter**  
     data와 label을 입력 받아서 각 label에 해당하는 data의 평균과 분산을 반환하는 함수이다. data를 classify하기 위해서는 각 label에 해당하는 data의 분포가 어떤 형태를 띄고 있는지 알아야 한다. data의 분포를 통해 새로운 data가 어느 label에 해당할 확률이 높은 지 계산할 수 있기 때문이다.  
     텍스트이(가) 표시된 사진

     자동 생성된 설명
   * **get\_prior\_probability**label을 입력 받아서 prior를 계산해서 반환하는 함수이다. prior는 각 label이 나올 확률이다.  
     텍스트이(가) 표시된 사진

     자동 생성된 설명
   * **Gaussian\_PDF**  
     x, mu, sigma를 입력 받아 mu, sigma를 parameter로 하는 gaussian pdf에 x를 대입했을 때의 결과값을 반환하는 함수이다.  
     텍스트, 화면, 스크린샷이(가) 표시된 사진

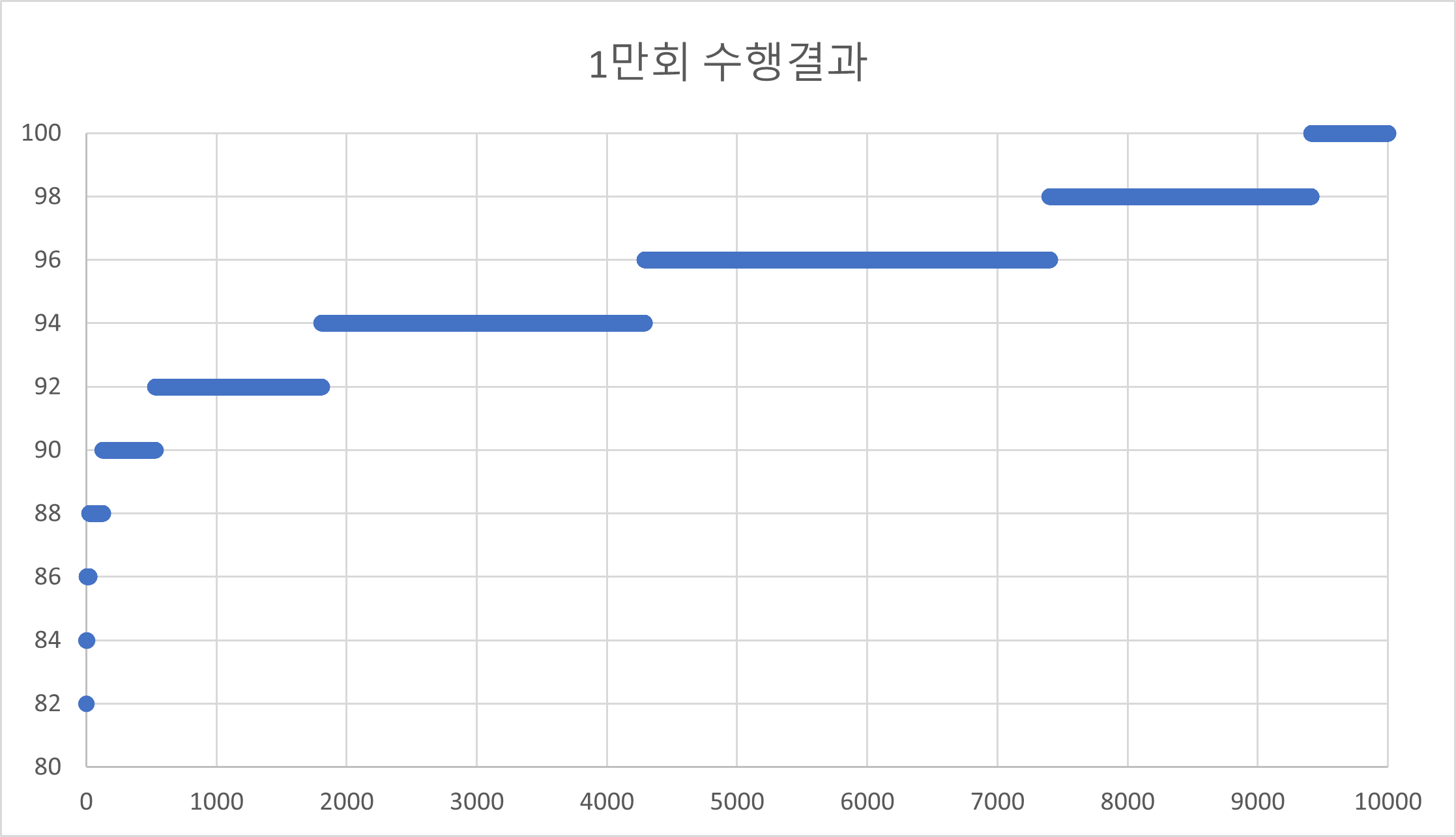
     자동 생성된 설명
   * **Gaussian\_Log\_PDF**  
     x, mu, sigma를 입력받아 mu, sigma를 parameter로 하는 gaussian pdf에 x를 대입했을 때의 결과값을 log를 취해 반환하는 함수이다.  
     텍스트이(가) 표시된 사진

     자동 생성된 설명
   * **Gaussian\_NB**  
     입력 받은 mu, sigma, prior, data를 가지고 posterior를 계산해서 반환하는 함수이다. 먼저 Gaussian\_PDF 함수 또는 Gaussian\_Log\_PDF 함수를 이용해서 특정 data의 feature값들의 각 label에 대한 유사도 likelihood를 계산한다. 그리고 likelihood 값에 prior을 추가해서 posterior를 계산한다. Bayesian theorem에서는 p(x)를 나누어주지만, p(x)가 모든 data에서 동일하여 classifier에 영향을 끼치지 않는다. GaussianPDF 함수를 사용하면 각 확률값을 곱하고, Gaussian\_Log\_PDF 함수를 사용하면 각 확률값에 log를 취해서 더해준다.  
     텍스트이(가) 표시된 사진

     자동 생성된 설명
   * **classifier**  
     data별 posterior의 값을 비교해서 posterior가 최대가 되는 위치를 반환하는 함수이다. data의 label별로 posterior값을 계산했으므로, posterior가 가장 큰 위치의 label에 속한다고 판단할 수 있다.  
     텍스트이(가) 표시된 사진

     자동 생성된 설명
   * **accuracy**  
     pred(예측값), gnd(실제값)을 비교해서 정확도를 측정하는 함수이다.  
     텍스트이(가) 표시된 사진

     자동 생성된 설명
3. **실행 결과**
   * **콘솔 결과**텍스트이(가) 표시된 사진

     자동 생성된 설명
   * **1만번 실행**1만번 실행한 결과 98.75%가 90% 이상의 정확도를 보였다.
4. **Reference**
   * Normal distribution.<https://en.wikipedia.org/wiki/Normal_distribution>
   * 박철수. Basic of Machine Learning -Likelihood Ratio Test-