머신러닝

학번: 2016722074

학과: 컴퓨터정보공학과

이름: 김영태

과제 제목 및 목적

Naïve Bayesian Classifier을 직접 디자인하고 flower dataset을 이용해 직접 데이터를 분류해 보는 실습을 진행한다. 해당과제를 진행하면서 전체적인 classifier의 구조를 이해하고 classifier의 동작 과정을 이해할 수 있다. 해당 과제의 데이터셋에는 총 4개의 특성이 있다. 4개의 특성이 있는 데이터셋을 사용해서 데이터들의 Likelihood와 Prior를 구하고 이 값들을 베이즈 정리를 사용해서 Posterior 값을 구한다.

소스코드 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

먼저 feature\_normalization 함수이다.

데이터를 인자로 전달받고 이 데이터의 각 feature마다 분산과 평균을 구한다. 또 이때 구한 평균과 분산으로 이 데이터의 normalization을 진행한다. 이때 이 값은 데이터에서 평균을 빼고 이 값을 분산으로 나눈 값이다. 최종적으로 원본데이터를 변환한 normal\_feature을 리턴한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 get\_normal\_parameter 함수이다. 함수의 인자로 데이터와 레이블값, 레이블의 개수를 전달받는다. 레이블된 데이터를 저장해 함수 내에서 사용하기 위해 3차원 배열 데이터 labeled\_data를 선언하고 0으로 초기화했다. 이 배열은 함수의 인자로 전달받은 label의값에 해당하는 값에 따라 데이터를 레이블링해 새로운 데이터(레이블링된)가 만들어진다. 이제 이 배열을 data\_point 축을 따라 평균과 분산 값을 구하고 이 두 값을 반환했다.

텍스트, 스크린샷, 검은색, 화면이(가) 표시된 사진

자동 생성된 설명

다음은 Prior를 구하는 함수이다. 이 함수는 레이블 값을 전달받아 사전지식, 즉 prior를 계산한다. 여기서 prior란 각 클래스에 속할 확률을 의미하는데, 이 과제를 예시로 들면 총 3개의 클래스가 있으므로 이 prior배열의 길이는 3이고 이 배열의 모든 값을 더하면 1이된다. prior의 각 클래스에 해당하는 값은 각 클래스에 데이터가 속할 확률을 의미한다. 따라서 코드를 보면 천체 레이블의 길이를 구하고, 각 클래스마다 몇 개의 데이터가 있는지 숫자를 세서 확률을 구했다.

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 pdf의 값을 구하는 함수이다. 첫번째 함수는 가우시안 함수인데 이 함수는 평균과 분산만 있다면 pdf를 구할 수 있다. 따라서 pdf를 구하기 위한 평균과 표준편차값을 함수의 인자로 전달받는다. 또 데이터의 Likelihood값을 구하는 것이 목적이기 때문에 데이터 x역시 주어진다. 가우시안 함수는 e의 –(x-평균)^(2/2\*분산)을 루트 2\*pi\*분산 으로 나눈 값 이기 때문에 numpy내장함수를 사용해서 코드를 구현했다. 밑의 가우시안 로그 pdf의 경우는 앞의 가우시안 pdf의 값에 로그를 취한것이기 때문에 numpy 내장함수를 사용해서 로그를 취한 뒤 값을 리턴했다.

텍스트이(가) 표시된 사진

자동 생성된 설명

마지막으로 지금까지 구한 Likelihood와 Prior값을 이용해서 Posterior 값을 구하고 각 데이터 포인트에서 어떤 레이블, 즉 어떤 클래스에서의 값이 제일 높은 지 비교하여 어떤 클래스에 속할지 예측한다.

코드의 구현을 보면, 각 레이블은 각 클래스를 의미한다. 각 데이터 포인터에서의 레이블마다의 Likelihood를 구한다. 이때 코드에서 사용한 함수는 Gaussian Log PDF이므로 각 데이터 포인트에서의 pdf값은 로그를 취했으므로 곱셈이 아니라 덧셈을 수행한다. 각 레이블과 데이터 포인트에서의 Likelihood를 구했으므로 이 값에 Prior를 추가해 Posterior값을 계산한다. 계산에 로그 사용했기 때문에 Prior값을 곱하지 않고 더한다. 각 데이터 포인트마다 각 클래스에 속할 확률을 구했으므로 최종적으로 이 Posterior의 값을 비교하여 어떤 클래스에 속할지 결정하게 된다.

실행결과

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

참고문헌

<https://numpy.org/doc/stable/index.html>

<https://ko.wikipedia.org/wiki/%EB%82%98%EC%9D%B4%EB%B8%8C_%EB%B2%A0%EC%9D%B4%EC%A6%88_%EB%B6%84%EB%A5%98>

<http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf>