기계학습 알고리즘의 실험적 분석과 결과

**초록**

실생활에서 이미지 처리를 위해 사용되는 알고리즘은 과거에 비해 다양해졌다. 다양한 알고리즘의 출현은 처한 상황에 적합한 알고리즘을 선택하는 문제로 귀결된다. 본 논문에서는 각 알고리즘들의 분류 속도, 학습 속도, 정확도를 비교해보고 상황에 따른 알고리즘이 무엇인지 분석하고자 한다.

1. **문제 정의**

이미지 분류에서 기계학습 알고리즘을 선택하는 방법은 다양하다. 더 좋은 알고리즘은 더 많은 자원 활용을 요구하기도 하며, 결과적으로 불필요하게 많은 자원을 이용해 낭비하는 경우가 생긴다. 본 논문에서는 이미지 분류 알고리즘에 대해 모델링하고 수치적으로 분석하고자 한다.

* 1. **실험 조건**

실험으로 사용하는 데이터는 MNIST의 손글씨 8 x 8 크기의 이미지를 사용한다. 1797개의 이미지 갯수를 사용하고, 이미지들은 0부터 9까지 해당되는 숫자로 라벨링 되어있다. 데이터는 훈련 데이터와 테스트 데이터를 7대 3의 비율로 나눈다. 이미지는 회색조로 표현된다.

자연이(가) 표시된 사진

자동 생성된 설명

그림 1: MNIST 손글씨

* 1. **분석 지표**

알고리즘은 분류 속도, 학습 속도, 그리고, 정확도를 이용하여 분석한다. 분류 속도는 학습이 완료된 모델에서 테스트 데이터의 답을 추론하는데까지 걸리는 시간을 의미한다. 분류 속도가 느리면 이미지 분류를 위한 입력이 주어졌을때 정답을 추론하는 시간이 길어진다. 반대로 분류 속도가 빨라지면 이미지 분류를 위해 입력이 주어질때 정답을 추론하는 시간이 짧아진다. 분류 속도의 정확한 측정을 위해 필요한 클록 수를 기준으로 평가한다.

학습 속도는 분류를 위해 정답을 추론하는 모델을 학습시키는데 걸리는 시간을 의미한다. 학습 속도가 짧으면 모델이 모든 데이터를 탐색하면서 모델을 학습시키는 시간이 작다는 것을 말한다. 반대로 학습 속도가 길면 모델이 모든 데이터를 탐색하면서 모델을 학습시키는 시간이 길어지는 것을 말한다. 학습 속도의 정확한 측정을 위해 필요한 클록 수를 기준으로 평가한다.

정확도는 알고리즘이 정답을 추론했을때 실제 정답을 얼마나 맞추었는지를 말한다. 정확도가 높을수록 알고리즘이 정답을 맞출 확률이 증가하는 것을 의미한다.

* 1. **알고리즘**

실험에 사용되는 알고리즘은 로지스틱 회귀, 서포트 벡터 머신, 선형 판별 분석, KNN, 심층신경망, 합성신경망이다.

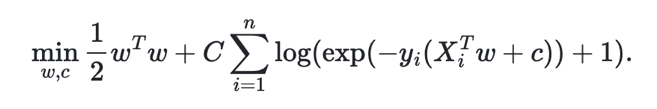
로지스틱 회귀 알고리즘은 분류를 위한 선형 모델이다. 최대 엔트로피 분류 혹은 로그 선형 분류라고도 부른다. 로지스틱 회귀 알고리즘은 이진 분류로도 사용할 수 있다. 본 논문에서는 싸이킷런의 이진 분류 방법을 이용하여 실험한다. 여기서 사용되는 공식은 그림 2와 같다.

그림 2: 로지스틱 회귀 목표 함수

서포트 벡터 머신은 패턴 인식, 자료 분석을 위한 모델이며 주로 분류와 회귀 분석을 위해 사용한다. 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단한다.

선형 판별 분석은 데이터 분포를 학습해 결정경계를 만들어 데이터를 분류하는 모델이다. 데이터를 결정경계에 해당하는 벡터에 사영시킨 결과를 이용한다. 베이즈 룰을 이용한 방법을 통해서 선형 판별 분석을 이해할 수 있다.

알고리즘의 구체적인 구현 방법은 실험에 후술한다.

1. **실험**
   1. 로지스틱 회귀
   2. 서포트 벡터 머신
   3. 선형판별분석
   4. KNN
   5. 심층신경망
   6. 합성신경망
2. 결과
3. 결론

사용하는 알고리즘

1. Logistic regression
2. Svm
3. FDA
4. Knn
5. DNN
6. CNN