# 중간발표 대본

안녕하세요, 7조 발표를 맡은 정하연입니다. 발표 시작하겠습니다.

먼저 original dataset을 어떤 모델에 어떻게 학습시켜봤는지, 그리고 결과는 어떠했는지 말씀드리겠습니다. 그리고 Handmade Dataset을 어떻게 Cleaning했는지 언급하고, 동일하게 모델 별 학습 결과를 말씀드리겠습니다. 마지막으로 추후 계획을 말씀드리겠습니다.

---

먼저 Original Dataset입니다. 저희는 Original Dataset을 이용한 모든 학습에서 testset의 비율을 20%로 설정하였습니다. 그리고, GridSearch로 최적의 파라미터를 찾는 과정을 거쳤습니다. 일반적으로 fold가 5 또는 10인 cross validation이 많이 활용되는데, 저희는 5로 설정하여 진행했습니다.

---

저희가 사용해본 모델은 크게 8가지 입니다.

softmax를 제외하고 모델 당 두 개 이상의 하이퍼파라미터를 사용하지 않았습니다.  하이퍼파라미터가 너무 많아질 경우 학습시간이 너무 길어져 교재에서도 중요하게 다뤄진 하이퍼파라미터들에 집중했습니다.

SVM은 10000개의 데이터를 한 번 학습시키는 데에도 30초가 걸려 너무 오래 걸려서 앞으로 사용하지 않기로 결정했습니다.

Decision Tree에서 불순도 측정에는 gini보다 entropy가 최적이었고, RandomForest에서 n_estimator는 시도한 개수 중 가장 많은 2000개였습니다.

Voting Classifier에는 svm을 제거한 총 6개의 베스트 모델들을 모두 넣었습니다.

---

왼쪽은 정확도이고, 오른쪽은 gridsearch에 걸린 시간인데, 각각 오름차순으로 정렬한 결과입니다.

random forest, extra tree, inn, voting이 정확도 95% 이상으로 상대적으로 좋은 성능을 보였으며, voting classifier를 soft로 진행할 때가 가장 좋은 정확도를 보였습니다. 하지만 voting classifier은 학습에 가장 오랜 시간이 걸렸습니다. soft voting classifier가 hard보다 성능이 더 좋았습니다.

정확도와 걸리는 시간 모두를 고려하면, MLP와 Extra Tree는 정확도도 상위권과 가깝고, search time도 비교적 작아서 original dataset에 대해서 가장 좋은 모델들이라고 말 할 수 있을 것 같습니다.

---

Handmade Dataset을 클리닝하고자 몇가지 아이디어를 내보았습니다.

첫째로, 숫자 10개, 기호 5개에 속하지 않는 라벨을 가진 데이터는 모두 삭제하는 방안입니다. 현재는 여기까지 클리닝된 데이터로 학습을 진행했습니다. 향후 추가적인 데이터 클리닝 과정을 거쳤을 때, 어떤 변화가 있는지 비교 분석 해보고자 합니다.

두번째로, 0에서 9, 기호 라벨 순서대로 데이터를 정렬한 후, 잘못 라벨링된 데이터를 육안으로 판별 후 삭제하거나 라벨을 수정해보고자 합니다. 예를 들어 +인데 이미지가 rotate 되어 있어 x로 잘못 라벨링된 경우, 라벨을 수정하는 것입니다. 이후 학습 시에 데이터를 다시 shuffle하는 과정을 거치고자 합니다.

세번째로 생각한 방안은, 이미지 해상도를 조정하여 각기 다른 숫자와 기호의 굵기를 동일하게 바꿔보는 것입니다. 데이터를 출력해보았을 때, 글자의 두께가 두꺼워 정확히 인식되지 못하는 경우도 있었기에 해당 방안을 떠올려보았습니다.

현재까지 도출된 아이디어는 크게 이 3가지 이며, 새롭고 참신한 아이디어를 계속해서 찾아나갈 예정입니다.

---

데이터 클리닝 진행과정은 다음과 같습니다.

데이터 클리닝 이전 데이터셋의 shape를 출력해보니 각각 약 15000개의 학습 데이터가 있었고, 2100개 가량의 테스트 데이터셋이 있었습니다.

저희는 숫자와 기호를 통합한 총 15개 클래스 분류기를 만들기 위해 데이터를 합쳤고, 15개 이외의 라벨을 가진 데이터는 삭제하였습니다. 학습 데이터셋에서 13.8%, 테스트 데이터셋에서 14.3%가 삭제되었습니다.

최종 데이터셋의 shape은 (26249, 784)와 (3730, 784)로 설정되었습니다.

---

handmade dataset을 앞과 동일하게 8개의 모델에 학습시킨 결과입니다.

Decision Tree에서 불순도 기준이 gini로 바뀌었습니다.

voting classifier는 두 가지로 나누어봤는데, Accuracy가 50% 이하인 모델을 포함한 경우와 포함하지 않은 경우로 나누었습니다. 50% 이하인 모델을 제거한 경우, soft voting이 더 좋았으며 모두 포함했을 때보다 정확도가 더 높습니다. 반대로, 6개의 모델을 모두 포함했을 경우엔 hard voting이 더 좋았으며, 정확도는 더 낮았습니다.

---

Grid Search 시간을 보시면, MLP와 SVM을 제외한 모든 모델들은 시간이 적게 걸렸습니다.

반면 각 모델의 베스트 모델들의 정확도를 확인한 결과, original dataset과 다르게 모든 모델의 정확도가 80%조차 못넘는 결과가 나왔습니다.

더 정교한 클리닝을 진행했을 때, 어떤 변화가 나타나는지 살펴볼 계획입니다.

---

추후 계획입니다.

말씀드린 데이터 클리닝 중 남은 데이터 클리닝 아이디어를 시도해보고자 합니다. 또한, 새롭고 다양한 클리닝 방안을 계속해서 모색해볼 예정입니다.

그리고 최종적으로 original과 클리닝 단계에 따른 결과를 비교하고자 합니다.

그리고 현재 Voting Classifier에는 svm을 제외한 6개의 모델이 모두 들어가있는데, Search Time과 Accuracy 둘 다 고려하여 좋은 순으로 조합하여 앙상블 학습을 시도해보려고 합니다.

이상 발표를 마치겠습니다. 감사합니다.

# 참고사항

- KNN
  - 주어진 데이터 포인트와 가장 가까운 3개의 이웃을 찾아 이 이웃들의 레이블을 살펴보고 다수결 방식으로 예측을 결정
- SVM
  -
- Extra Tree Classifier
  - `ExtraTreesClassifier`는 Extremely Randomized Trees의 약자로, Random Forest와 유사하지만 더 무작위성을 가지는 특징이 있습니다.
- Softmax (LogisticRegresison - multinomial 지정)
  -
- Decision Tree Classifier
  -
- Random Forest Classifier
  - 여러 개의 DT
- MLP Classifier
  -
- Voting Classifier