

# Midterm Exam

2019.10.18 (Fri)

Yukyung Choi

시험에 앞서...

여러분의 실력을 검증하기 위한 문항 수와 시간은 충분합니다.

“전략”을 잘 세우고, 문제 풀이에 임하시기 바랍니다.

문제의 난이도와 점수를 잘 살펴 전략을 세우세요.

시험에 필요한 **오픈 라이브러리 메뉴얼**은 얼마든지 인터넷 접근이 가능합니다.

하지만, 사용법 혹은 이론 설명이 작성된 블로그, 이를 이용한 문제 풀이 사이트의 접근은 제한합니다.

판단하기 힘들 때는 감독관에게 질문하세요.

**이와 관련된 부정행위 적발 시 “F” 처리 합니다.**

**[T1번][5점]**

Unsupervised Learning 과 Supervised Learning 을 설명하시오.

**[T2번][5점]**

Local Minima와 Global Minima에 대해 설명하시오.

**[T3번][5점]**

Feature Vector 란 무엇인지 설명하시오.

**[T4번][5점]**

Clustering 이란 무엇인지 설명하시오.

**[T5번][5점]**

SVM의 Support Vector란 무엇인지 설명하시오.

**[T6번][5점]**

SVM의 중요 Parameter는 무엇이며, 각각이 가지는 의미를 설명하시오.

**[T7번][5점]**

SVM를 활용한 다중 분류 방법을 설명하시오.

**[T8번][5점]**

Kernel Trick에 대해서 설명하시오.

[P1번][레벨1][10점]

입력으로 주어지는 합성 데이터의 정보는 아래와 같다. 군집화 알고리즘인 Kmeans의 다양한 알고리즘을 비교 실험하고자 한다. 주어진 3개의 Kmeans 알고리즘에 대하여 정확도와 시간복잡도를 구하여라. 그리고 본 실험을 바탕으로 결론을 도출하라.

[제공 코드] <https://colab.research.google.com/drive/19iWQVTJeKTc3SOUonRcgabJ-hlxVxz1k>

[입력] 데이터의 수: 1000, 클래스의 수: 10

[출력]

	Accuracy (Purity Check)	Time Complexity
Lloyd Kmeans		
Kmeans++		
Elkan Kmeans		

\* Purity Check 함수는 제공된 함수를 참고하라. (Colab 참고)

\* 시간 측정을 위해 시간 함수를 사용하라. (아래 참고)

```
import time

# 기본 단위 초
start_time = time.time()
end_time = time.time()
elapsed = end_time - start_time

print("training Runtime: %0.20f Seconds"%(elapsed))
```

[주의] 시간 측정은 초기화부터 군집화까지 전 과정 측정

[주의] Random Seed 는 (42) 로 모두 고정

[결론] 위의 실험 결과를 바탕으로 kmeans 알고리즘에 따른 장단점을 10줄 내외로 작성하시오.

[P2번] [레벨 1] [20점]

입력으로 주어지는 합성 데이터의 정보는 아래와 같다. SVM 알고리즘에 따른 비교 실험을 진행 하고자 한다. 주어진 3개의 SVM 알고리즘에 대하여 정확도 시간복잡도를 구하여라. 그리고 본 실험을 바탕으로 결론을 도출 하라.

[제공 코드] <https://colab.research.google.com/drive/1TRZO1hWgBcl0XHslQyHM4QH33vc3Zxqd>

[입력 1] 데이터의 수: 100, 클래스의 수: 2

[입력 2] 데이터의 수: 100, 클래스의 수: 2

[출력]

	DATA1-Accuracy (Average Precision)	DATA2-Accuracy (Average Precision)	Time Complexity
Linear-SVM			
Poly-SVM			
RBF-SVM			

\* Precision측정을 위해 sklearn의 classification\_report를 참고하라

\* 시간 측정을 위해 시간 함수를 사용하라.(아래 참고)

```
import time

# 기본 단위 초
start_time = time.time()
end_time = time.time()
elapsed = end_time-start_time

print("training Runtime: %0.20f Seconds"%(elapsed))
```

[주의] 시간 측정은 초기화부터 학습까지 전 과정 측정

[주의] Random\_Seed 는 (42) 로 모두 고정

[결론] 위의 실험 결과를 바탕으로 SVM알고리즘에 따른 장단점을 10줄 내외로 작성하시오.

[P3번][레벨1] [20점]

입력으로 주어지는 MNIST 데이터를 이용하여 Supervised Classification과 Unsupervised Classification의 대표 알고리즘으로 필기체 숫자 분류 문제를 해결 하고자 한다. 각각의 알고리즘의 정확도를 구하여라. 그리고 본 실험을 바탕으로 결론을 도출하라.

[제공 코드] <https://colab.research.google.com/drive/1YtsaPAq15nNPB8KKqG44GTEyU9rpp4ns>

[입력] 클래스의 수: 10, MNIST: 필기체 숫자 (0~9)\

[전처리] 데이터 전처리는 StandardScaler() 만을 사용하라.

\* from sklearn.preprocessing import StandardScaler 매뉴얼을 참고

[출력]

	Accuracy
Supervised	(Average Precision)
Unsupervised	(Purity)

[주의] Random\_Seed 는 (42) 로 모두 고정

[결론] 분류 방법론에 따른 실험 결과를 10줄 내외로 작성하시오.

[P4번] [레벨2] [35점]

Kaggle Leaderboard: 상위 Rank 3 명은 가산점 10점 부여 (단, 가산점은 만점을 넘어갈 수 없음)

[주의사항] <https://github.com/sejongresearch/2019.Fall.PatternRecognition/issues/20#issuecomment-543639828>

입력으로 주어지는 Cats & Dogs 데이터를 이용하여 영상 분류 문제를 해결하고자 한다. 패턴인식 수업시간에 배운 전통적인 머신러닝 기법을 이용하여 문제를 해결하라. 결과 파일은 리더보드에 제출한다.

[제공 코드] <https://colab.research.google.com/drive/1Z9L6H9CSoso02Pv9o-5qCLs1qur2bcrd>

[입력] 클래스의 수: 2 (개 vs 고양이)

[전처리] 영상 크기를 32x32로 resize 하여 feature vector 로 사용한다.

[출력] result.csv 로 아래와 같이 만든다, Label = 0 은 고양이, Label = 1 은 개이다.

id	Label
1	0
2	1

[제출] result.csv 파일은 Kaggle 에 Submit 하여 등수를 확인 가능하다.

“unizard-test” 와 같은 로그 메시지는 [이름+제출시간]으로 설정해야 한다.

Ex) ykchoi-2010 → 오후 8시 10분 버전

```
[17] 1 | kaggle competitions submit -c 2019-fall-pr-project -f result_yk.csv -m "unizard-test"
100% 38.0k/38.0k [00:03<00:00, 11.3kB/s]
Successfully submitted to InClass Competition at Sejong University
```

[주의] Random Seed 는 (42) 로 모두 고정

[주의] 인공지능망 모델 사용은 금지한다.

[주의] LeaderBoard에 제출 후 결과 확인 은 5분 후에 가능하다.

[P5번][레벨2][35점]

Kaggle Leaderboard : 상위 Rank 3 명은 가산점 10점 부여 (단, 가산점은 만점을 넘어갈 수 없음)

입력으로 주어지는 Music 데이터를 이용하여 음악 장르 분류 문제를 해결하고자 한다. 패턴인식 수업시간에 배운 전통적인 머신러닝 기법을 이용하여 문제를 해결하라. 결과 파일은 리더보드에 제출한다.

[제공 코드] [https://colab.research.google.com/drive/1E3GQyKCXxdyiW2Qz6dT7U\\_sTdelimb5A](https://colab.research.google.com/drive/1E3GQyKCXxdyiW2Qz6dT7U_sTdelimb5A)

[입력] 클래스의 수: 10 (blues, country, rock, jazz, reggae, hiphop, classical, disco, pop, metal)

음악 데이터의 Feature Vector는 총 28가지로 구성되어 있으며, 정보는 다음과 같다.

tempo	beats	chroma_stft	rmse	spectral_centroid	spectral_bandwidth	rolloff	zero_crossing_rate					
mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	mfcc9	mfcc10	mfcc11	mfcc12	mfcc13
mfcc14	mfcc15	mfcc16	mfcc17	mfcc18	mfcc19	mfcc20						

[전처리] 데이터를 scale 을 이용하여 정규화 하라.

**\*from sklearn.preprocessing import scale**

[출력] result.csv 로 아래와 같이 만든다,

Blues(0), country(1), rock(2), jazz(3), reggae(4), hiphop(5), classical(6), disco(7), pop(8), metal(9)

id	Label
1	0
2	1

[제출] result.csv 파일은 Kaggle 에 Submit 하여 등수를 확인 가능하다.

“unizard-test” 와 같은 로그 메시지는 [이름+제출시간]으로 설정해야 한다.

Ex) ykchoi-2010 ➔ 오후 8시 10분 버전

```
[17] 1 ! kaggle competitions submit -c 2019-fall-pr-project -f result_yk.csv -m "unizard-test"
100% 38.0k/38.0k [00:03<00:00, 11.3kB/s]
Successfully submitted to InClass Competition at Sejong University
```

[주의] Random Seed 는 (42) 로 모두 고정

[주의] 인공지능경망 모델 사용은 금지한다.

[주의] LeaderBoard에 제출 후 결과 확인 은 5분 후에 가능하다.