예습

Sklearn model KFold
Test set에 대한 성능 평가 신뢰성 높이고, 모든 데이터가 최소 한 번은 테스트셋으로 쓰이도록 합니다.
https://3months.tistory.com/321

1-9 (train)10(test) - 50 바퀴 돌려 학습
Rest train (train) 9(test) - 50 바퀴 돌려 학습
Rest train (train)8(test)

Epoch:
모델한테 몇번을 학습시켜 통과시킬지
Default 50 :한번학습시킬때 50번 보여줌

Batch size : 몇번으로 쪼개서 보여줄 것인가 Default 50- 데이터를 한번 넣을 때 50개씩 넣는다 몇개씩 나눠서 넣을 지
9900개의 train set을 몇개씩 나눠 넣을지
Epoch 이 50 이기에 총 50번 보는데, 9900개를 한번에 모델에 보여줄 수 없기에 batch size로 나눠서 보여줌
9900/50 = 50개씩 넣은순간에 모델 업데이트, weight update (optimal 하게) 198번 넣어야함
1epoch 돌란다는 의미:198 iter 돌린다
1 iter당 weight update
11000개 총 데이터 = 1 fold 당 1100개 = 9900개 train
Iteration - weight 이 몇번 변하는가

- 모델 성능을 정한다
- Parameter: weight 와 같이 모델의 숫자값으로 모델이 optimize
- Hyper parameter: 사람이 조정

Learning rate
Weight 업데이트 과정에서 미분하고 거기다가 learning rate을 곱해서 하게됨
얼마나 빨리갈지에 대한 이야기
학습의 속도 결정 : 작아야함

Middle dimension- latent layer, latent space

Binary output - cross entropy loss
Binary output 이면
데이터의 형태가 무조건 (0,1)이면 loss cross entropy
아니면 -MSE model 을 쓴다

칼람들은 fid아닌거 빼고 다 쓴다

Combine 가져와서 x라는 데이터프래임 만듬
Y 는 nih토탈사용해서 데이터 프래임 만듬
학습을 해서 예측
예측을 하기 위해서는 train - test 나눠서
스스로 평가해야함

Classification : 남자/ 여자 , 불량품/ 아닌것
Regressor: 숫자 맞춤 , 집값예측하는 모델

오늘자 한일

- Day4: 커피 같이 마시기, 인사, 논문 하나 읽기, 점심, random forest model 이해
- Day5: 논문읽기, Friday meeting with 교수님, 점심, 오후 paper meeting

논문읽기

-multi-trait genomic methods were effective in boosting predictive power.
-psr 에 따라 예측률 다르지만
however results were similar for different multi-trait and polygenic score methods.

education attainment, intelligence 예측
-large sample size available for EA GWA studies
- genetic correlation between EA and intelligence,
-> EA GPS predicted as much or more variance in intelligence than did GPS derived from GWAS of the target trait of intelligence itself
-EA(education attainment) GPS predict educational achievement(점수) and intelligence better than---- do GWA of the target traits themselves suggests the usefulness of multivariate approaches

-aim: estimate how much variance in intelligence and educational achievement can be predicted by applying multi-trait genomic approaches and leveraging highly powered GWA summary statistics

1)compare three polygenic score methods ((PRSice [22], LDpred [23], and Lassosum)
2)test how much variance the new IQ3 and EA3 GPS maximally predict with three highly (genetically) correlated traits (‘Income’ [25], ‘Age when completed full time education’ [26], and ‘Time spent using computer’ [26]) to boost predictive power
3)compare the performance of three multi-trait methods(Genomic SEM [27], MTAG [4] and SMTpred

Genotypes for 10,346 individuals > SNP imputation using the Haplotype Reference Consortium reference panels
Following imputation, we excluded variants with minor allele frequency and selected variants with an info score of 1, resulting in 515,000 SNPs used for analysis

outcome variable: intelligence(verbal and nonverbal web-based test score) and educational achievement at ages 12 and 16(기본과목의 평균점수)
GWASs IQ3 and EA3 summary statistics to construct genome-wide polygenic scores (GPS) PSR

오늘 깨달은점: genotype data 는 snp데이터이고
중간은 snp 중 가장 independent 하면서 잘 나타내는 것이고
11000명의 데이터이지만 4500정도로 추려지는 이유가 European 만 해당하기에

-Multivariate/multitrait genetic method = psr 해서 예측한 것
- psr 측정도구가 여러개 있는데, 우리는 lassosum 사용

random forest model 이해

if args.binary_output:
    rf_model = RandomForestClassifier(n_estimators=100, max_depth=20, random_state=2)
Else: (continuous output)
    rf_model = RandomForestRegressor(n_estimators=100, max_depth=20, random_state=2)

Snp 넣은것과 psr concatenate한 것이 Binary output 이 되는 기준이 무엇인가...궁금
일단 random forest model에 돌리는 것은
Snp 넣은것과 psr concatenate 한 것일텐데..data를 봐야 알 수 있나?

유전자 변이들의 질병에 영향끼치는 것은 알지만 Clinic benefit 까지 가기에 GWAS의 개개인별 loci 전반적 위험성의 아주 작은 부분 설명, ㅏ ㅑ 우됴

Gps 활용하는 방법
-Gwas 자료 기준으로 후보 predictor 고르고
-uk biobank 큰 데이터셋활용해서 각 질병들에 대해 최적의 예측 방법찾아냄
-Test set 활용해 Performance 평가

-Gps 활용하기 위해 고려해야할 상황
1) 질병의 유전적 아키텍처 이해
2) (SNP)-based heritability 정도 알아야함 (that is, the genetic risk component that can be captured by SNP arrays);
- 처음에 사용하는 gwas 자료들이 predictor 고르기 충분한지

- Polygenice predictor를 최적화하고 검증하는데 다른 적절한 데이터가 있을 수 있는가
- Polygenic risk 해석에다른 유전자적이나 non-gentic factor 고려되어야 하는가
- Risk score 의 퍼포먼스가 타겟 인구에 대해 일반화 가능한가

-chronic kidney disease (CKD) 위험 예측
Ckd- a significant SNP based heritability 보이는 반면
but rather a highly heterogenous group of primary and secondary pathogenic processes.
> SNP-based heritability of CKD 파악 어려움 , 진행단계 다양
-gwas study 더 이뤄져야하는데-specific kidney disease 에대해서 데이터 부족
> investment in more-powerful GWAS for primary kidney disorders

- lack of external validation data sets
- kidney disorders cannot be reliably defined using International Classification of Disease codes, and require more complex phenotyping algorithms
-GPS approach can only estimate a relative risk of disease, whereas the absolute risk for an individual will vary with age, lifestyle and other non-genetic factors.
- GWASbased risk models are inherently population-specific, due to differences in SNP allelic frequencies, linkage disequilibrium patterns, and allelic effects between populations.