

01

파이썬 기반의 머신러닝과 생태계 이해

01. 머신러닝의 개념	1
머신러닝의 분류	2
데이터 전쟁	3
파이썬과 R 기반의 머신러닝 비교	4
02. 파이썬 머신러닝 생태계를 구성하는 주요 패키지	6
파이썬 머신러닝을 위한 S/W 설치	7
03. 넘파이	13
넘파이 ndarray 개요	14
ndarray의 데이터 타입	16
ndarray를 편리하게 생성하기 – arange, zeros, ones	19
ndarray의 차원과 크기를 변경하는 reshape()	20
넘파이의 ndarray의 데이터 세트 선택하기 – 인덱싱(Indexing)	23
행렬의 정렬 – sort()와 argsort()	33
선형대수 연산 – 행렬 내적과 전치 행렬 구하기	37
04. 데이터 핸들링 – 판다스	39
판다스 시작 – 파일을 DataFrame으로 로딩, 기본 API	40
DataFrame과 리스트, 딕셔너리, 넘파이 ndarray 상호 변환	49
DataFrame의 칼럼 데이터 세트 생성과 수정	52
DataFrame 데이터 삭제	54
Index 객체	57
데이터 셀렉션 및 필터링	61
정렬, Aggregation 함수, GroupBy 적용	75
결손 데이터 처리하기	79
apply lambda 식으로 데이터 가공	82
05. 정리	86

02

사이킷런으로 시작하는 머신러닝

01. 사이킷런 소개와 특징	87
02. 첫 번째 머신러닝 만들어 보기 - 붓꽃 품종 예측하기	88
03. 사이킷런의 기반 프레임워크 익히기	93
Estimator 이해 및 fit(), predict() 메서드	93
사이킷런의 주요 모듈	94
내장된 예제 데이터 세트	96
04. Model Selection 모듈 소개	100
학습/테스트 데이터 세트 분리 - train_test_split()	100
교차 검증	102
GridSearchCV - 교차 검증과 최적 하이퍼 파라미터 튜닝을 한 번에	113
05. 데이터 전처리	118
데이터 인코딩	118
피처 스케일링과 정규화	124
StandardScaler	125
MinMaxScaler	127
06. 사이킷런으로 수행하는 타이타닉 생존자 예측	128
07. 정리	142

03

평가

01. 정확도(Accuracy)	144
02. 오차 행렬	148
03. 정밀도와 재현율	152
정밀도/재현율 트레이드오프	155
정밀도와 재현율의 맹점	163

04

분류

04. F1 스코어	164
05. ROC 곡선과 AUC	166
06. 피마 인디언 당뇨병 예측	170
07. 정리	178
01. 분류(Classification)의 개요	179
02. 결정 트리	181
결정 트리 모델의 특징	183
결정 트리 파라미터	184
결정 트리 모델의 시각화	185
결정 트리 과적합(Overfitting)	195
결정 트리 실습 - 사용자 행동 인식 데이터 세트	197
03. 앙상블 학습	206
앙상블 학습 개요	206
보팅 유형 - 하드 보팅(Hard Voting)과 소프트 보팅(Soft Voting)	208
보팅 분류기(Voting Classifier)	209
04. 랜덤 포레스트	212
랜덤 포레스트의 개요 및 실습	212
랜덤 포레스트 하이퍼 파라미터 및 튜닝	214
05. GBM(Gradient Boosting Machine)	217
GBM의 개요 및 실습	217
GBM 하이퍼 파라미터 및 튜닝	220

06. XGBoost(eXtra Gradient Boost)	222
XGBoost 개요	222
XGBoost 설치하기	224
파이썬 래퍼 XGBoost 하이퍼 파라미터	225
파이썬 래퍼 XGBoost 적용 - 위스콘신 유방암 예측	229
사이킷런 래퍼 XGBoost의 개요 및 적용	235
07. LightGBM	239
LightGBM 설치	241
LightGBM 하이퍼 파라미터	242
하이퍼 파라미터 튜닝 방안	244
파이썬 래퍼 LightGBM과 사이킷런 래퍼 XGBoost,	244
LightGBM 하이퍼 파라미터 비교	244
LightGBM 적용 - 위스콘신 유방암 예측	245
08. 분류 실습 - 캐글 산탄데르 고객 만족 예측	247
데이터 전처리	248
XGBoost 모델 학습과 하이퍼 파라미터 튜닝	251
LightGBM 모델 학습과 하이퍼 파라미터 튜닝	255
09. 분류 실습 - 캐글 신용카드 사기 검출	257
언더 샘플링과 오버 샘플링의 이해	257
데이터 일차 가공 및 모델 학습/예측/평가	259
데이터 분포도 변환 후 모델 학습/예측/평가	263
이상치 데이터 제거 후 모델 학습/예측/평가	266
SMOTE 오버 샘플링 적용 후 모델 학습/예측/평가	270
10. 스택킹 앙상블	274
기본 스택킹 모델	275
CV 세트 기반의 스택킹	278
11. 정리	284

05

회귀

01. 회귀 소개	286
02. 단순 선형 회귀를 통한 회귀 이해	288
03. 비용 최소화하기 - 경사 하강법(Gradient Descent) 소개	290
04. 사이킷런 LinearRegression을 이용한 보스턴 주택 가격 예측	299
LinearRegression 클래스 - Ordinary Least Squares	299
회귀 평가 지표	300
LinearRegression을 이용해 보스턴 주택 가격 회귀 구현	301
05. 다항 회귀와 과(대)적합/과소적합 이해	307
다항 회귀 이해	307
다항 회귀를 이용한 과소적합 및 과적합 이해	310
편향-분산 트레이드오프(Bias-Variance Trade off)	314
06. 규제 선형 모델 - 릿지, 라쏘, 엘라스틱넷	315
규제 선형 모델의 개요	315
릿지 회귀	317
라쏘 회귀	320
엘라스틱넷 회귀	323
선형 회귀 모델을 위한 데이터 변환	325
07. 로지스틱 회귀	328
08. 회귀 트리	331
09. 회귀 실습 - 자전거 대여 수요 예측	338
데이터 클렌징 및 가공	339
로그 변환, 피처 인코딩과 모델 학습/예측/평가	342

06

차원 축소

10. 회귀 실습 - 캐글 주택 가격: 고급 회귀 기법	349
데이터 사전 처리(Preprocessing)	350
선형 회귀 모델 학습/예측/평가	354
회귀 트리 모델 학습/예측/평가	366
회귀 모델의 예측 결과 혼합을 통한 최종 예측	367
스태킹 앙상블 모델을 통한 회귀 예측	369
11. 정리	371
01. 차원 축소(Dimension Reduction) 개요	373
02. PCA(Principal Component Analysis)	375
PCA 개요	375
03. LDA(Linear Discriminant Analysis)	389
LDA 개요	389
붓꽃 데이터 세트에 LDA 적용하기	390
04. SVD(Singular Value Decomposition)	392
SVD 개요	392
사이킷런 TruncatedSVD 클래스를 이용한 변환	398
05. NMF(Non-Negative Matrix Factorization)	401
NMF 개요	401
06. 정리	403

07

군집화

01. K-평균 알고리즘 이해	405
사이킷런 KMeans 클래스 소개	406
K-평균을 이용한 붓꽃 데이터 세트 군집화	407
군집화 알고리즘 테스트를 위한 데이터 생성	411
02. 군집 평가(Cluster Evaluation)	415
실루엣 분석의 개요	416
붓꽃 데이터 세트를 이용한 군집 평가	417
군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법	419
03. 평균 이동	423
평균 이동(Mean Shift)의 개요	423
04. GMM(Gaussian Mixture Model)	428
GMM(Gaussian Mixture Model) 소개	428
GMM을 이용한 붓꽃 데이터 세트 군집화	430
GMM과 K-평균의 비교	432
05. DBSCAN	436
DBSCAN 개요	436
DBSCAN 적용하기 - 붓꽃 데이터 세트	440
DBSCAN 적용하기 - make_circles() 데이터 세트	443
06. 군집화 실습 - 고객 세그먼테이션	446
고객 세그먼테이션의 정의와 기법	446
데이터 세트 로딩과 데이터 클렌징	447
RFM 기반 데이터 가공	451
RFM 기반 고객 세그먼테이션	454
07. 정리	459

08

텍스트 분석

NLP이냐 텍스트 분석이냐?	460
01. 텍스트 분석 이해	461
텍스트 분석 수행 프로세스	462
파이썬 기반의 NLP, 텍스트 분석 패키지	462
02. 텍스트 사전 준비 작업(텍스트 전처리) – 텍스트 정규화	463
클렌징	464
텍스트 토큰화	466
스톱 워드 제거	468
Stemming과 Lemmatization	470
03. Bag of Words – BOW	471
BOW 피쳐 벡터화	471
사이킷런의 Count 및 TF-IDF 벡터화 구현: CountVectorizer, TfidfVectorizer	473
BOW 벡터화를 위한 희소 행렬	475
희소 행렬 – COO 형식	476
희소 행렬 – CSR 형식	477
04. 텍스트 분류 실습 – 20 뉴스그룹 분류	481
텍스트 정규화	481
피쳐 벡터화 변환과 머신러닝 모델 학습/예측/평가	484
사이킷런 파이프라인(Pipeline) 사용 및 GridSearchCV와의 결합	488
05. 감성 분석	491
감성 분석 소개	491
지도학습 기반 감성 분석 실습 – IMDB 영화평	491
비지도학습 기반 감성 분석 소개	495
SentiWordNet을 이용한 감성 분석	497
VADER를 이용한 감성 분석	504
06. 토픽 모델링(Topic Modeling) – 20 뉴스그룹	506

07. 문서 군집화 소개와 실습(Opinion Review 데이터 세트)	510
문서 군집화 개념	510
Opinion Review 데이터 세트를 이용한 문서 군집화 수행하기	510
군집별 핵심 단어 추출하기	519
08. 문서 유사도	522
문서 유사도 측정 방법 - 코사인 유사도	522
두 벡터 사잇각	523
Opinion Review 데이터 세트를 이용한 문서 유사도 측정	527
09. 한글 텍스트 처리 - 네이버 영화 평점 감성 분석	530
한글 NLP 처리의 어려움	530
KoNLPy 소개	531
데이터 로딩	534
10. 텍스트 분석 실습 - 캐글 Mercari Price Suggestion Challenge	538
데이터 전처리	540
피처 인코딩과 피처 벡터화	545
릿지 회귀 모델 구축 및 평가	551
LightGBM 회귀 모델 구축과 앙상블을 이용한 최종 예측 평가	553
11. 정리	554

09

추천 시스템

01. 추천 시스템의 개요와 배경	556
추천 시스템의 개요	556
온라인 스토어의 필수 요소, 추천 시스템	558
추천 시스템의 유형	559
02. 콘텐츠 기반 필터링 추천 시스템	560
03. 최근접 이웃 협업 필터링	561

04. 잠재 요인 협업 필터링	564
잠재 요인 협업 필터링의 이해	564
행렬 분해의 이해	567
확률적 경사 하강법을 이용한 행렬 분해	570
05. 콘텐츠 기반 필터링 실습 - TMDB 5000 영화 데이터 세트	573
장르 속성을 이용한 영화 콘텐츠 기반 필터링	574
데이터 로딩 및 가공	574
장르 콘텐츠 유사도 측정	577
장르 콘텐츠 필터링을 이용한 영화 추천	579
06. 아이템 기반 최근접 이웃 협업 필터링 실습	585
데이터 가공 및 변환	585
영화 간 유사도 산출	588
아이템 기반 최근접 이웃 협업 필터링으로 개인화된 영화 추천	591
07. 행렬 분해를 이용한 잠재 요인 협업 필터링 실습	597
08. 파이썬 추천 시스템 패키지 - Surprise	601
Surprise 패키지 소개	601
Surprise를 이용한 추천 시스템 구축	602
Surprise 주요 모듈 소개	606
Surprise 추천 알고리즘 클래스	610
베이스라인 평점	611
교차 검증과 하이퍼 파라미터 튜닝	612
Surprise를 이용한 개인화 영화 추천 시스템 구축	614
09. 정리	619