2장, 사이킷런으로 시작하는 머신러닝

신림프로그래머 최범균

프로세스 기초

- 프로세스
  - 1. 피처 처리
  - 2. 데이터 세트 분리
  - 3. 모델 학습
  - 4. 예측 수행, 평가

단순 예

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
5.1	3.5	1.4	0.2	0
4.9	3.0	1.4	0.2	0
4.7	3.2	1.3	0.2	0

```
iris = load_iris()
iris_data = iris.data # ndarray: (150, 4)
iris_label = iris.target # ndarray: (150, )

X_train, X_test, y_train, y_test = train_test_split(iris_data, iris_label,
                                                    test_size=0.2, random_state=11)
# X_train: ndarray: (120, 4), y_train: ndarray: (120, )
# X_test: ndarray: (30, 4), y_test: ndarray: (30, )

dt_clf = DecisionTreeClassifier(random_state=11)

dt_clf.fit(X_train, y_train) # 학습 수행

pred = dt_clf.predict(X_test) # pred: ndarray: (30, )

print('예측 정확도: {0:.4f}'.format(accuracy_score(y_test, pred)))
```

사이킷 프레임워크

- Estimator
  - 지도 학습(supervised learning)
    - 분류, 회귀
  - 비지도 학습
    - 차원 축소, 클러스터링, 피처 추출 등
  - 주요 메서드
    - 학습: fit()
    - 예측: predict()
- 하이퍼 파라미터 튜닝
- 데이터 셋

검증

- 과적합 방지 위해 교차 검증 사용

K Fold, Stratified Fold

- K Fold : 단순 분할
- Stratified Fold : 레이블(타겟) 분포를 고려한 분할
    - 레이블

cross_val_score()

```
scores = cross_val_score(dt_clf,
       data, label,
       scoring='accuracy',
       cv=5)

print('교차 검증별 정확도:',
        np.round(scores, 4))
print('평균 검증 정확도:',
        np.round(np.mean(scores), 4))
```

```
cv_results = cross_validate(dt_clf2,
        data, label,
        scoring='accuracy',
        cv=5,
        return_estimator=True)

print('교차 검증별 정확도:',
       np.round(cv_results['test_score'], 4))
print('평균 검증 정확도:',
       np.round(np.mean(cv_results['test_score']), 4))

pred = cv_results['estimator'][0].predict(data)
```

https://scikit-learn.org/stable/modules/model_evaluation.html

데이터 전처리

- 예측을 계산할 수 있도록 원본 데이터 변경
  - 데이터 인코딩
  - null 처리 : 피처의 평균 등으로 대체
  - 제거 : 결과에 영향이 없는 피처 삭제
- 문자열 → 숫자
  - 레이블 인코딩 : 문자열 값 → 숫자
  - 원-핫 인코딩 : 문자열 값마다 피처 생성
    - 회귀 등 숫자에 따라 중요도가 결정되는 ML 알고리즘에 사용

분류	가격
TV	500,000
냉장고	3,000,000
TV	2,500,000

분류	가격
1	500,000
2	3,000,000
1	2,500,000

TV	냉장고	가격
1	0	500,000
0	1	3,000,000
1	0	2,500,000

피처 스케일링과 정규화

- 피처마다 값의 범위가 차이나면 ML 알고리즘 성능에 영향
  - 예: y = w0·x0 + w1·x1 + w2·x2 + … + wn·xn
  - x0은 0~1사이 값, x2는 0~100, x3은 -1000~1000
- 표준화
  - 데이터 피처 각각이 평균이 0이고 분산이 1인 가우시안 정규 분포를 가진 값으로 변환
- 정규화
  - 서로 다른 피처의 크기를 통일하기 위해 크기를 변환
  - 즉, 개별 데이터의 크기를 모두 똑같은 단위로 변경
- 사이키런의 대표적인 피처 스케일링
  - StandardScaler : 표준화
  - MinMaxScaler : 정규화 (0~1)

타이타닉 예

- 데이터 탐색
- 피처 전처리
- 학습
- 평가
- 예측