숭실대학교 Soongsil University

데이터 편향이 예측에 미치는 영향

➢ 타이타닉호 생존자 예측 자세히 들여다 보기

숭실대학교 베어드교양대학
서유화 교수
yhsuh@ssu.ac.kr

목차

- 지도학습(SVM)을 이용한 머신러닝 과정 복습
- 데이터 편향에 따른 머신러닝 예측 결과 비교
- 데이터 왜곡에 따른 예측 결과 비교

학습 목표

- 사이킷런을 이용한 머신러닝의 지도학습(SVM) 모델을 만들 수 있다.
- 데이터 편향으로 발생할 수 있는 문제점들을 설명할 수 있다.
- 데이터 왜곡으로 발생할 수 있는 문제점들을 설명할 수 있다.
- 데이터 윤리가 필요한 이유를 설명할 수 있다.
- 인공지능을 사용하는 우리의 시각과 자세에 대해 고찰한다.

지난시간 배운 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성,) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티처블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘 기반 의사결정 시스템의 한계 2. 윤리가 적용된 인공지능 알고리즘	1. 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	

지도학습(SVM)을 이용한 머신러닝 과정(온라인)

1. 사용 패키지와 모듈 임포트

In [1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn import svm   #SVM 모델
from sklearn import metrics  #정확도 비교
```

2. 데이터 가져오기

In [2]:
```
data_df = pd.read_csv('./titanic.csv')
data_df.head()
```

3. 데이터셋 나누기 : 학습용과 테스트용
   - train_test_split(데이터, 분할비율)

In [5]:
```
train, test = train_test_split(data_df, test_size= 0.2)
print("train data", train.shape)
print("test data",test.shape)
```

4. 학습에 사용할 변수 선택하기

In [6]:
```
data_df.info()
```

In [7]:
```
data_df.corr()
```

5. 학습용 데이터셋 : 학습데이터와 레이블(정답) 나누기

In [8]:
```
train_data_df = train[['pclass','sex', 'age', 'parch','fare']]
train_data_df
```

In [9]:
```
train_label_df = train[['survived']]
train_label_df
```


지도학습(SVM)을 이용한 머신러닝 과정(온라인)

6. 테스트 데이터셋 : 테스트데이터와 레이블(정답) 나누기

In [11]:
```
test_data_df = test[['pclass','sex', 'age', 'parch','fare']]
test_data_df
```

In [12]:
```
test_label_df = test[['survived']]
test_label_df
```

7. 모델 학습하기(SVM)
   - svm.SVC(비용값, 감마값) : svm학습모델생성
   - fit(학습데이터, 학습데이터정답(레이블))

In [15]:
```
clf = svm.SVC(C = 1, gamma = 0.1)
clf.fit(train_data, train_label)
```

8. 테스트 데이터로 예측하기(SVM)
   - predict(테스트데이터)

In [16]:
```
pred_svm = clf.predict(test_data)
pred_svm  #svm이 예측한 생존 여부 값
```

9. 모델 예측 정확도 확인
   - metrics.accuracy_score(테스트데이터정답, 예측값)

In [18]:
```
ac_score = metrics.accuracy_score(test_label, pred_svm)
print('accuracy : ', ac_score)
```


타이타닉생존자 예측 자세히 들여다 보기

- 오프라인 실습파일
  - 4주_5강_데이터편향이 예측에 미치는 영향_타이타닉호 생존자 예측하기.ipynb
- 지도학습을 이용한 타이타닉호 생존자 예측하기 (온라인)에서 추가된 코드

1. 사용 패키지와 모듈 임포트
   - import seaborn as sns
     - Matplotlib을 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 패키지 (https://seaborn.pydata.org/)

2. 데이터 가져오기
   - age 데이터를 범주화하기 위해서 10으로 나눈 몫에서 10을 곱해줌
   - 생존자(1)와 사망자(0) 막대 그래프로 그리기
     - 열이름.value_counts() : 해당열에 대해 존재하는 그룹별로 개수를 세줌
     - 열이름.value_counts().plot(kind='bar') : 그룹별로 센 개수를 막대 그래프로 표현

3. 데이터셋 나누기 (동일)

타이타닉생존자 예측 자세히 들여다 보기

- 지도학습을 이용한 타이타닉호 생존자 예측하기(온라인)에서 추가된 코드

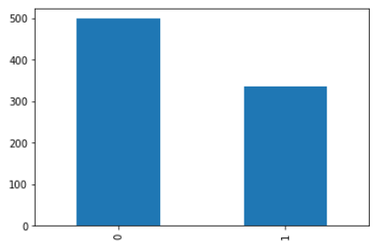
4. 학습에 사용할 변수(특징, Feature) 선택하기
   - 학습데이터에서 각 피쳐(Feature)가 생존자 예측 분류에 미치는 영향을 상세히 탐색
     - 온라인 소스코드에서는 상관계수만을 참조하여 피처를 선택했음
     - 열이름.value_counts() : 해당열에 대해 존재하는 그룹별로 개수를 세줌
     - 열이름.value_counts().plot(kind='bar') : value_counts()의 결과를 막대 그래프로 표현
     - sns.contplot(x='열이름', hue='카테고리이름', data=데이터가 저장된 변수명)
       : x의 데이터 개수를 hue에 설정된 카테고리 별로 나눠 그 개수의 그래프를 생성
     - 데이터가 저장된 변수명.groupby (['열이름1', '열이름2'])
       : 데이터의 열이름1을 기준으로 먼저 그룹핑을 하고 그 안에서 열이름2를 기준으로 다시 그룹핑
     - 데이터가 저장된 변수명.mean()
       : 각 행열별로 데이터의 평균값을 구함

학습데이터 탐색 결과

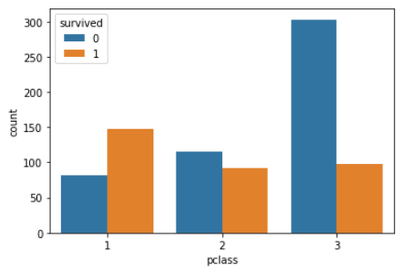
- 전체 학습데이터(836명)에서 사망자는 60%(499명), 생존자는 40%(337명)
- 그룹별(좌석등급별) 분포
  - 1등급 좌석 생존자가 가장 많고 3등급 클래스 사망자가 가장 많음

```
0    499
1    337
Name: survived, dtype: int64
```

<AxesSubplot:>



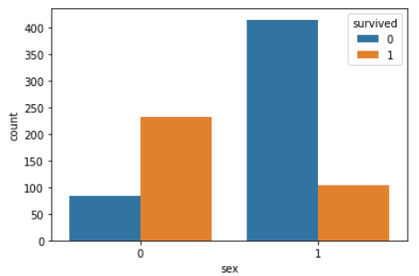

```
3    401
1    228
2    207
Name: pclass, dtype: int64
```

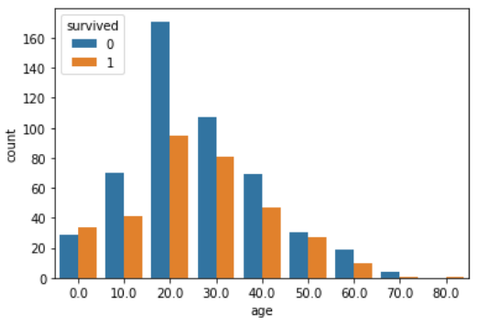


학습데이터 탐색 결과

- 그룹별(성별, 나이별) 분포
  - 여성의 생존율이 월등히 높음
  - 20,30 대 탑승객이 가장 많고 가장 많이 사망

```
1    519
0    317
Name: sex, dtype: int64
```





학습데이터 탐색 결과

- 그룹별(좌석등급, 성별) 분포
  - 1등급 탑승객은 평균연령이 가장 높음
  - 1등급 좌석 여성이 가장 생존율이 높고, 2,3등급 클래스 남성이 가장 생존율이 낮음
  - 모든 등급 좌석에서 여성의 생존율이 남성보다 매우 높음

pclass	sex	age	sibsp	parch	fare	survived
1	0	33.063063	0.531532	0.513514	107.873913	0.954955
	1	37.692308	0.401709	0.264957	72.561397	0.350427
2	0	24.047619	0.511905	0.595238	23.503324	0.869048
	1	27.317073	0.365854	0.227642	20.068631	0.154472
3	0	18.852459	0.688525	0.852459	14.925170	0.442623
	1	21.397849	0.494624	0.308244	12.265201	0.157706

학습데이터 탐색 결과

- 그룹별(좌석등급별, 나이별)분포
  - 영,유아 어린이 생존율이 모든 등급에서 높음
  - 20,30대의 생존율이 높으나 탑승객의 수도 가장 많음
  - 1등급 80대 생존율이 가장 높고, 2,3등급에서는 70대 생존율이 가장 낮음

pclass	age	sex	sibsp	parch	fare	survived
1	0.0	0.500000	0.500000	2.000000	116.704150	0.500000
	10.0	0.277778	0.777778	0.722222	114.755789	0.777778
	20.0	0.461538	0.538462	0.487179	94.440813	0.743590
	30.0	0.433962	0.301887	0.207547	100.533726	0.716981
	40.0	0.666667	0.425926	0.222222	69.234804	0.574074
	50.0	0.512821	0.538462	0.333333	83.738992	0.615385
	60.0	0.600000	0.450000	0.800000	99.517085	0.400000
	70.0	0.500000	0.500000	0.000000	64.177100	0.500000
	80.0	1.000000	0.000000	0.000000	30.000000	1.000000
2	0.0	0.571429	0.928571	1.357143	28.439886	1.000000
	10.0	0.500000	0.250000	0.375000	23.711283	0.500000
	20.0	0.565217	0.536232	0.260870	21.872342	0.420290
	30.0	0.615385	0.288462	0.269231	17.886058	0.423077
	40.0	0.592593	0.407407	0.518519	24.777778	0.407407
	50.0	0.733333	0.200000	0.200000	16.191667	0.200000
	60.0	0.800000	0.600000	0.200000	22.770000	0.200000
	70.0	1.000000	0.000000	0.000000	10.500000	0.000000
3	0.0	0.531915	2.063830	1.319149	22.085368	0.404255
	10.0	0.666667	0.753623	0.434783	14.142090	0.217391
	20.0	0.753165	0.202532	0.126582	9.884597	0.234177
	30.0	0.698795	0.337349	0.445783	12.594227	0.253012
	40.0	0.657143	0.371429	1.171429	15.307263	0.142857
	50.0	1.000000	0.000000	0.000000	7.618067	0.000000
	60.0	0.750000	0.000000	0.000000	11.956250	0.250000
	70.0	1.000000	0.000000	0.000000	7.762500	0.000000

타이타닉생존자 예측 자세히 들여다 보기

결론 : 학습데이터 피처 탐색결과 생존율에 가장 많은 영향을 주는 피쳐는 pclass와 sex이다.
(age는 상관계수가 낮은편이고 pclass와 sex보다는 영향력이 명확하지 않아 제외)

pclass와 sex 변수만을 선택해서 학습 모델을 만들어 보자.

- 지도학습을 이용한 타이타닉호 생존자 예측하기(온라인)과 동일

5. 학습용데이터셋 학습데이터와 정답(레이블) 나누기
6. 테스트용데이터셋 테스트데이터와 정답(레이블) 나누기
7. SVM모델 학습
8. 테스트데이터로 예측하기

타이타닉생존자 예측 자세히 들여다 보기

- 지도학습을 이용한 타이타닉호 생존자 예측하기(온라인)에서 추가된 코드

9. 테스트 데이터의 예측 결과 그래프로 나타내보기 (4번에서 추가된 코드 이용)
   - 좌석등급(pclass)에서 등급별 생존자 수 그래프로 확인
   - 성별(sex)에 따른 생존자 수 그래프로 확인
   - 연령별 생존자 수 그래프로 확인
   - 좌석등급, 성별에 따른 생존율 확인

예측결과

- 3등급 좌석 클래스의 사망가능성이 매우 높다
- 여성은 생존가능성은 남성보다 월등히 높고, 남성의 사망가능성은 여성보다 월등히 높다
  - 테스트 데이터 예측 결과 : 여성 모두 생존, 남성 모두 사망
- 20,30대 탑승객의 사망가능성이 가장 높다.
- 모든 좌석 등급에서 남성의 사망가능성은 매우 높다.
  - 테스트 데이터 예측 결과 : 여성 모두 생존, 남성 모두 사망

생각할 문제

- 학습데이터의 탐색 결과와 비교해 테스트 데이터의 예측 결과는 어떠한가?
- 만약 침몰하는 배에서 탑승객의 수보다 적은 한정된 수의 구명조끼만을 가지고 있다고 가정하자.
  - 생존가능성을 높이기 위해 위와 유사한 방식의(정답과 동일하게 예측한 정확도를 기반으로 선별하여 피쳐들을 적용) 머신러닝 모델을 적용하여 예측된 결과에 따라 구명조끼를 먼저 분배한다면 어떠할까?
  - 유사한 사례) 재범예측, 신용예측, 채용심사

생각할 문제

- 간과한 요소
  - 타이타닉호에서 높은 등급클래스의 좌석에서 높은 생존율을 보인 것은 귀족이나 고소득의 탑승자가 먼저 구명보트에 태워졌을 가능성이 높다.
  - 여성이 높은 생존율을 보인 것은 당시 시대 통념상 젊은 남성들의 희생으로 어린이와 여성이 먼저 구명보트에 태워졌을 가능성이 높다.
- 이러한 방식의 인공지능의 학습모델을 사용하는 경우 데이터 내 존재하는 차별(여성과 어린이, 귀족, 고소득자가 먼저 구명보트에 태워짐)에 의해 남성차별적이거나 빈부차별적인 결과를 예측할 것이다.
- 향후 그러한 불행에서 누구를 먼저 구조할지를 결정하는 경우 차별은 더 강화될 것이다.

결론

- 차별을 포함하고 있는 데이터를 학습한 인공지능의 예측결과를 의사결정에 사용하는 경우 유사한 결과를 예측함으로써 기존의 차별을 더 강화시킬 수 있다.