숭실대학교 Soongsil University

지도학습을 이용한 타이타닉호 생존자 예측하기

숭실대학교 베어드교양대학
서유화 교수
yhsuh@ssu.ac.kr

목차

- 서포트벡터 머신 동작원리 이해
- 서포트벡터 머신을 이용한 지도학습 모델 구현
  - 타이타닉호 생존자 예측
- 학습 피쳐(Feature) 및 파라메터 최적화

학습 목표

- 머신러닝에서 지도학습 분류 모델의 원리를 이해할 수 있다.
- 서포트벡터머신을 이용한 간단한 학습모델을 파이썬으로 구현할 수 있다.
- 머신러닝에서 예측결과에 영향을 주는 요인들에 대해 설명할 수 있다.

지난시간 배운 내용

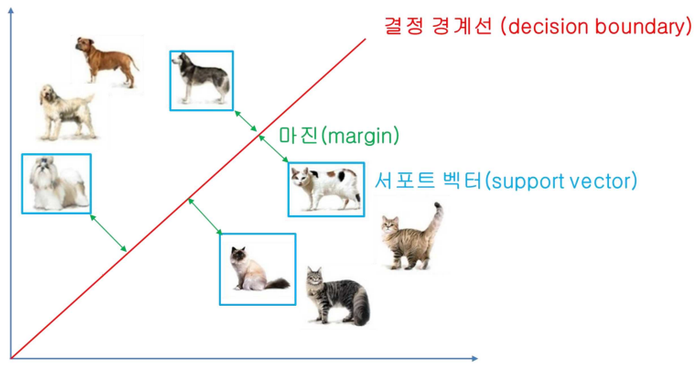
주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 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 (구글티쳐블머신) 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	

숭실대학교 Soongsil University

지도학습 알고리즘 (SVM)

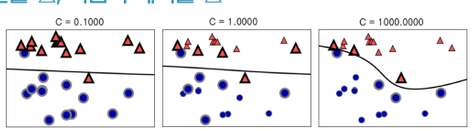
서포트 벡터 머신(Support Vector Machine)

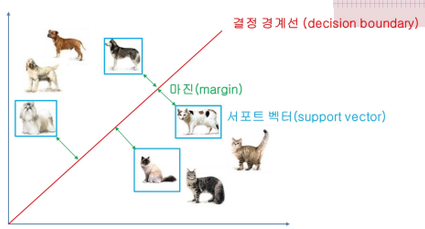
- 서로 다른 분류 값을 결정하는 경계선(결정 경계선)을 결정하는 알고리즘
- 마진을 최대로 하는 결정 경계를 찾는 것



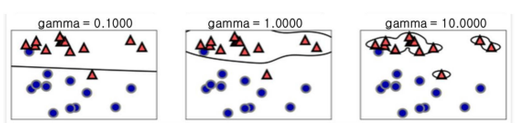
SVM 파라메터

- **비용(Cost) :** 마진조절 변수
  - 비용이 작을수록 마진이 넓어짐
  - 비용이 클수록(마진이 작음) 더 구부러진 곡선 형태의 결정경계
  - 마진을 ▼, 학습시 에러율 ▼
  - 마진을 △, 학습시 에러율 △

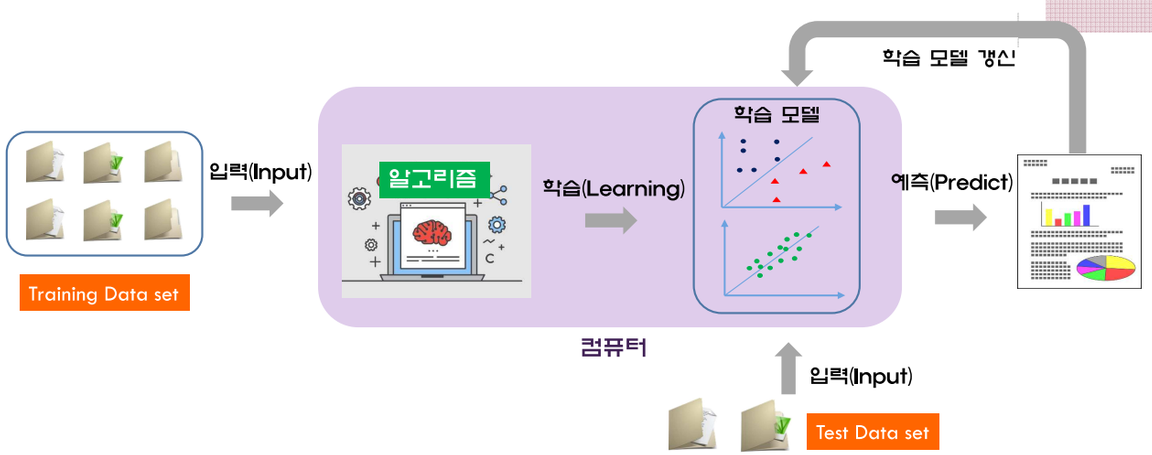




- **감마(Gamma) :** 학습 데이터 포인트들이 결정 경계에 영향을 끼치는 범위를 조절해주는 변수
  - 감마 값이 크면 많은 데이터포인트들이 가까이 있는 것으로 고려되어 결정 경계가 작아지고 구부러짐

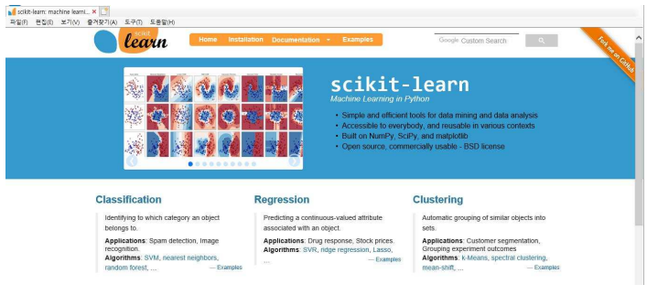


지도학습 과정



Scikit-learn 라이브러리

- Scikit-learn(사이킷 런)은 대표적인 머신러닝 라이브러리
  - 분류, 회귀, 클러스터링 등 다양한 알고리즘 제공
- Scikit-learn 홈페이지 : https://scikit-learn.org/
- 아나콘다에 포함돼 있음



지도학습(SVM)을 이용한 머신러닝 기본 과정

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 기본 과정

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)
```


학습 피쳐(Feature) 및 파라메터 최적화

- 데이터를 증가시켜 수행할 경우 정확도는 어떤 차이가 있을까?
- 학습 피쳐(특징데이터, 변수)를 변경시켜 수행할 경우 정확도는 어떤 차이가 있을까?
- SVM 분류기에서 C(비용)과 gamma 값을 변경하여 머신러닝을 수행해보고 어떤 값이 가장 좋은 결과를 갖는지 알아보자
- 예측 정확도가 가장 높은 학습 모델과 데이터 수, 피쳐 및 파라메터 값은 무엇인가?

정리

- 머신러닝은 데이터의 양이 많을 수록 높은 정확도를 보인다
- 학습데이터에서 높은 정확도를 보이는 것이 항상 좋은 것은 아니다.
  - Why? : 학습데이터에 대한 높은 정확도는 미지의 데이터에 대해서는 낮은 성능(과적합)을 발생시킬 수 있다.
- 학습을 하는 변수(Feature: 특징데이터)에 따라 예측 결과가 달라질 수 있다.
- 학습모델은 다양한 실험 파라메터(비용, 감마 등)와 다양한 특징데이터의 무수한 실험을 통해 완성된다.

다음시간에 배울 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	