여름방학 인공지능 프로젝트 : 학기말 보고서 작성 & 프로젝트 진행

< 인공지능 학기말 보고서 제출 내역 >		
logistic Regression 폴더 내	LogisticRegression.ipynb	로지스틱 회귀를 이용하여 작성한 주피터 노트북 파일입니다
	daily_traffic.csv & precipitation.csv	데이터를 가져오기 위해 사용한 CSV 파일입니다
해당 보고서	인공지능 학기말 보고서로, 프로젝트를 진행한 내역과 주제 선정 학습 내역 등을 작성한 파일입니다.	

* 위 제출 내역에 문제가 있다면, https://github.com/jaehyeok3017/2022_SunrinAI_VacationProject 에서도 확인 가능합니다.
* 개념 소개 시 발췌한 자료들의 출처는 모두 보고서 마지막 페이지에 작성되어 있습니다.

01. 프로젝트 선정

1학기 인공지능 프로젝트로, K-최근접 이웃 회귀와 다항 회귀를 이용한 <외부 요인과 교통량 관계에 대한 분석>을 진행하였습니다.

여기서 유류 가격이라는 요인을 분석하기 위해, 강수량 데이터를 제외하여 그래프와 예측 모델을 만들었었는데,
강수량이라는 요인이 교통량이라는 요인에 영향을 미칠 것이라고 생각 했었고, 여러 뉴스나 생활 속 경험을 통해 이미 검증된 사실이었기에, 강수량을 제외하고 진행하였습니다.

허나, 의문이 들었던 점은 진짜 강수량이 교통량에 영향을 주는가? 라는 질문에는 전혀 답변을 할 수 없었습니다.
그리하여, 이번 여름방학 프로젝트에서는 강수가 있을 때와 없을 때를 구분하여 교통량을 기준으로 강수량을 예측할 때, 성능이 제대로 나오는 지 판단하여 위의 질문에 답해보려고 합니다.

02. 프로젝트 진행

주피터 노트북 환경을 이용하여 진행하였으며, 코드 실행 결과는 모두 클래스룸에 업로드한 [LogisticRegression.ipynb] 파일을 확인하시거나, 아래 작성된 보고서 결과를 참고하시면 됩니다.

[1] 데이터 수집
사용한 데이터는 아래와 같습니다. (1학기 프로젝트의 데이터)
- 권역 별 이용 차량 : http://data.ex.co.kr/portal/traffic/trafficRegion#
- 기상 자료 : https://data.kma.go.kr/climate/RankState/selectRankStatisticsDivisionList.do

	지점번호	지점명	일시	강수량(mm)
0	108	서울	2020-01-01	0.1
1	108	서울	2020-01-02	NaN
2	108	서울	2020-01-03	NaN
3	108	서울	2020-01-04	NaN
4	108	서울	2020-01-05	NaN
...	...	...	...	...
894	108	서울	2022-06-13	18.0
895	108	서울	2022-06-14	0.0
896	108	서울	2022-06-15	18.5
897	108	서울	2022-06-16	0.7
898	108	서울	2022-06-17	NaN

먼저, 기상 자료를 살펴보면 강수량을 기준으로 0 혹은 Nan으로 표시된 부분은 강수량이 없는 부분이고, 0.1 이상부터는 해당 날짜의 강수량이 있는 것을 확인할 수 있습니다.

따라서, 분류를 할 때
강수량이 있다면 → 1
강수량이 없다면 → 0으로 하여 분류하였습니다.

위를 토대로 봤을 때, 학습에 사용할 분류 방법으로 "로지스틱 회귀"를 사용하면
좋을 것 같다고 판단하였습니다. (0과 1로 단순 분류되어 있기 때문)

[2] 로지스틱 회귀 (Logistic Regression)란 무엇인가?

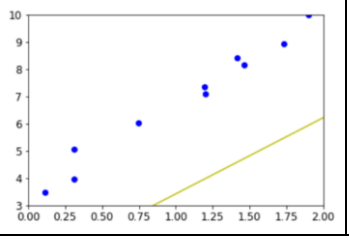
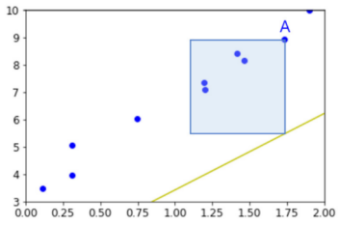
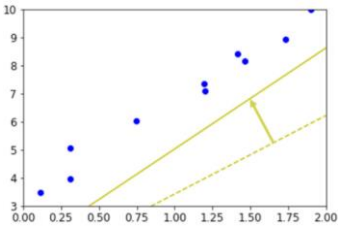
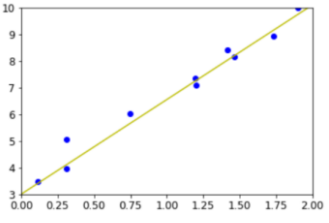
위에서 로지스틱 회귀를 이용하면 좋을 것 같다고 판단하였는데, 그렇다면 로지스틱 회귀라는 방법은 무엇인지 조사해봤습니다.

지도 학습	- 독립 변수, 종속 변수가 **모두 존재하는 데이터**를 제공하여 학습시키는 방법, 정답 특성을 **레이블(label)**라고 함	
	회귀	판단하려는 종속 변수의 값이 **숫자 형태**인 경우의 지도 학습 방법으로 결과는 특정한 **값**을 도출함 Ex. 독립 변수로는 공부 시간, 종속 변수로 시험 점수를 주어 공부 시간 별 시험 점수 데이터 생성
	분류	판단하려는 종속 변수의 값이 **문자 형태**인 경우의 지도 학습 방법으로 결과는 특정한 **그룹**을 도출함 Ex. 독립 변수로는 공부 시간, 종속 변수로는 합격 여부를 주어 시간에 따른 합격 여부 데이터 생성

해당 방법은 지도학습 중, 분류에 해당합니다.

학기 중에 선형 회귀 분석을 배웠었는데 일반 선형 회귀 같은 경우, 종속 변수의 데이터가 범주형이 아닌 것을 알 수 있습니다.

(아래는 클래스룸에 업로드 된 PPT 수업 자료 중 일부를 발췌했습니다.)

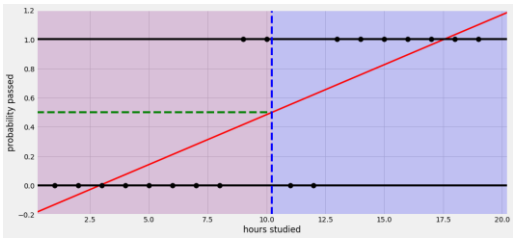
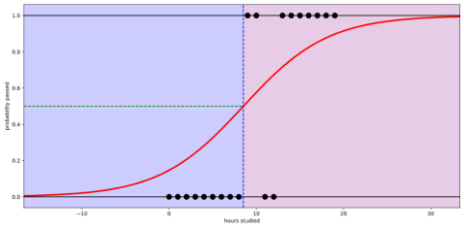
회귀 알고리즘 : 선형 회귀 알고리즘의 절차			
① 일차함수로 표현되는 추세선 정하기 (무작위 기울기)	② 평균 제곱 오차를 통해 비용 함수 구하기	③ 경사 하강법으로 비용함수가 최소일 때 추세선 기울기 구하기	④ 최종 추세선(최소 비용)을 통해 데이터 값을 예측하기
			

허나, 로지스틱 회귀(Logistic Regression) 같은 경우에는, 사건의 발생 가능성을 예측하는 데 사용되는 통계 기법입니다.

여기서 확률은 0과 1의 사잇값이 나오게 되어 일반 회귀 분석과는 달리 종속변수의 값이 제한적이라는 부분이 다르게 됩니다.

그리하여 로지스틱 회귀 분석은 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류 기법으로 볼 수 있게 되는 것입니다.

그렇다면, 로지스틱 회귀는 왜 사용하게 되는 것일까요?

[그래프 1] 선형 회귀를 이용했을 때의 그래프	[그래프 2] 로지스틱 회귀를 이용했을 때의 그래프
	

위의 그래프는 어떤 학생이 공부하는 시간에 따라 시험에 합격할 확률을 나타낸 그래프입니다.

여름방학 인공지능 프로젝트 : 학기말 보고서 작성 & 프로젝트 진행

만약, 왼쪽처럼 선형회귀를 이용하게 된다면 공부한 시간이 적으면 시험에 통과를 하지 못하고, 공부한 시간이 많으면 시험에 통과한다고 설명할 수 있습니다.

다만, 잘 보면 공부를 계속하면 계속할수록, 안 하면 안 할수록 각각 양의 방향, 음의 방향 무한대로 뻗어가는 형태를 띄게 됩니다. 또한, 공부를 2시간도 안 하게 되면 시험에 통과할 확률이 0도 안 된다고 하여 논리에 맞지 않는 것을 볼 수 있습니다.

반면, 로지스틱 회귀를 오른쪽처럼 이용하게 되면, 확률이 0과 1 사이의 값으로 그려지게 되며 논리에 적합함을 알 수 있습니다
이제 로지스틱 회귀를 더 자세하게 알아보도록 하겠습니다. 로지스틱 회귀는 아래와 같이 세 종류가 있습니다.

이진 로지스틱 회귀	범주형 응답에 대해 가능한 결과는 두 가지 뿐으로, 학생이 합격하거나 불합격한다는 예시가 있습니다.
다항 로지스틱 회귀	응답 변수에 순서가 없는 3개 이상의 변수가 포함될 수 있습니다. 레스토랑에서 특정 음식을 선호하는 지 예측하는 예시가 있습니다.
순서 로지스틱 회귀	다항 회귀와 마찬가지로 3개 이상의 변수가 있을 수 있고, 측정에 순서가 포함됩니다. 예를 들자면, 1에서 5까지의 척도로 호텔을 평가하는 경우가 있습니다.

위와 같은 로지스틱 회귀의 종류가 있는데, 프로젝트에 사용할 회귀 방법은, 강수량이 있는가를 중심으로 다룰 것이기 때문에
이진 로지스틱 회귀 방법을 이용하여 모델을 생성하도록 하겠습니다.

이러한 로지스틱 회귀 방법에서는 데이터가 특정 범주에 속할 확률을 예측하기 위해 아래와 같은 단계를 거치게 됩니다.
1) 모든 속성(feature)들의 계수(coefficient)와 절편(intercept)을 0으로 초기화한다.
2) 각 속성들의 값(value)에 계수(coefficient)를 곱해서 log-odds를 구한다.
3) log-odds를 sigmoid 함수에 넣어서 [0,1] 범위의 확률을 구한다.

최종적으로 로지스틱 회귀 분석 방법을 사용한다면 얻게 되는 장점과 단점을 소개하고 실제 프로젝트로 들어가도록 하겠습니다.

장점	- 매우 효율적이고 엄청난 양의 계산 리소스를 필요로 하지 않으며, 쉽게 해석할 수 있고 입력 기능 확장의 필요가 없습니다. - 쉽게 구현되고 학습되기 쉬우므로 다른 복잡한 알고리즘의 성능을 측정하는데 도움이 되는 기준이 됩니다.
단점	- 비선형의 문제를 해결하는 데 사용할 수 없으며, 더 좋고 복잡한 예측을 생성할 수 있는 다른 알고리즘이 많습니다. - 데이터 표시에 크게 의존해, 범주형 결과를 예측하는 데만 사용할 수 있습니다.

여름방학 인공지능 프로젝트 : 학기말 보고서 작성 & 프로젝트 진행

[3] 프로젝트 진행

3-1) 데이터 전처리

다른 부분들은 모두 1학기 프로젝트 부분과 다름이 없고, 아래 코드 부분만 다르게 됩니다.

precipitation.loc[(precipitation['강수량(mm)'] >= 0.1), '강수 여부'] = '1' precipitation.loc[(precipitation['강수량(mm)'] < 0.1), '강수 여부'] = '0'	강수가 0.1 이상이면 비가 내린 것이므로 1로 판단, 아니라면 0으로 판단

3-2) 로지스틱 회귀 분석

from sklearn.model_selection import train_test_split x = result['전국'] y = result['강수 여부']	데이터를 train과 test로 나누기 위해서 sklearn의 train_test_split를 import 해줍니다.
x = x.to_numpy() x = x.reshape(-1, 1) y = y.to_numpy() y = y.reshape(-1, 1) x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=42)	데이터를 나눠주기 위해서 numpy 형태로 바꿔준 다음, 형태를 맞춰주고 train_test_split을 이용하여 데이터를 훈련과 테스트 데이터로 나눠줍니다

from sklearn.linear_model import LogisticRegression
 	
# 로지스틱 회귀 모델 학습 	
model = LogisticRegression(penalty = 'l2')
model.fit(x_train, y_train)
	LogisticRegression을 사용하기 위해 import,
	LogisticRegression() 함수와 fit() 함수를 이용하여
	모델을 학습시켜 생성합니다

from sklearn.metrics import accuracy_score
 	
# 로지스틱 모델 학습 성능 비교 	
y_pred = model.predict(x_test)
 	
# 정확도 측정 	
accuracy_score(y_pred, y_test)
	정확도를 측정하기 위해 import를 해주고,
	정확도를 측정합니다.
	
	[실행결과]
	0.7428571428571429 → 약 74%의 정확도를 보임

03. 프로젝트 결론 & 느낀점

해당 프로젝트는 1학기 때 프로젝트로 선정하였던 <유류 가격과 교통량 관계에 대한 분석>에서 강수량 데이터를 제외한 것이 타당한 것인가?를 주제로 진행하였던 여름방학 프로젝트였습니다.

결론적으로, 강수량 데이터에 따라 교통량을 예측하는 것이 약 75%의 성능을 보임에 따라 연관성이 있음이 증명되었고, 강수량 데이터가 교통량에 영향을 미치면 유류가격이라는 요인으로 분석하는 데 오차를 보일 수 있으므로 제외하는 것이 타당하다고 결론을 지을 수 있습니다.

해당 프로젝트를 진행하면서 느낀 점은 1학기 때 단순히 진행하면서 "에이, 당연히 강수량 데이터는 영향을 많이 미칠 거니깐 제외 해 주어야지"라고 생각하며 안일하게 넘긴 부분에 있어 발표 당시 선생님께서 "왜 강수량 데이터를 제외했나요?"라고 질문 하셨을 때 잘 답변하지 못한 기억이 있는데, 왜 제외해야 하는 지 정확한 이유를 분석한 것 같아 1학기 프로젝트의 의문점을 풀어줌과 동시에 신뢰성을 더 높여줄 수 있는 프로젝트였던 것 같습니다.

따라서, 이러한 분석을 진행했을 때 단편적으로 끝내는 것이 아니라 왜?라는 의문점을 갖고 계속 진입하여 궁금증을 풀어가면서 신뢰성을 높이는 프로젝트를 앞으로 진행 할 필요성이 있다고 판단하였습니다.

또한, 수업시간에 진행하는 지도학습 중 회귀와 분류 이외에도 다양한 회귀와 분류 기법이 있음을 알게 되었고, 구글링이나 인터넷 서칭을 해보면서 다양한 지식들을 늘려 나가는 것이 얻어가는 것이 많다고 느낀 프로젝트였습니다. 앞으로 수업 시간을 통해 지식을 습득하면 궁금증에 대해서는 해소하면서 지식을 습득 해야 겠다는 계기가 된 프로젝트였던 것 같습니다.

자료 출처
- 인공지능과 미래사회 클래스룸 PPT 자료 중 그래프들 및 개념들

- 로지스틱 회귀 개념 정리
https://losskatsu.github.io/statistics/logistic-regression/#%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1-%ED%9A%8C%EA%B7%80%EB%B6%84%EC%84%9D-%EA%B0%9C%EB%85%90-%EC%A0%95%EB%A6%AC

- 로지스틱 회귀(Logistic Regression) 쉽게 이해하기 → [그래프 1], [그래프 2] 사진 및 개념 본문 발췌
https://hleecaster.com/ml-logistic-regression-concept/
- 로지스틱 회귀란 무엇입니까? → 로지스틱 회귀의 종류 부분 발췌
https://www.tibco.com/ko/reference-center/what-is-logistic-regression

- 로지스틱 회귀
https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/