붙임3

졸업논문 최종보고서

2017 학년도 제 2 학기

제목 : 날씨에 따른 서울특별시 대중교통 이용량 예측에 관한 연구

김희진(2013310314)

2017 년 11 월 7 일

지도교수: 김 응 모 서명

계획(10)	주제(20)	개념(20)	상세(30)	보고서(20)	총점(100)
10	20	20	29	18	97

* 지도교수가 평가결과 기재

■ 요약

현대 사회에서는 다양한 이동수단 중 지하철, 버스 등의 대중교통에 대한 수요가 높은 편이다. 본 연구의 배경이 되는 서울특별시의 경우에는 출퇴근 시, 과반 수 이상이 대중교통을 이용한다. 대중교통 이용량에는 날씨, 평일-주말, 연착, 도로현황 등 여러 가지에 원인을 둔다.

본 연구에서는 여러 요인 중에서도 날씨 데이터(기온, 강수량, 미세먼지)에 초점을 두어, 날씨에 따른 대중교통 이용량의 변화 양상을 학습하여 예측하는 연구를 진행한다. 서울특별시 25개 자치구마다의 날씨 데이터와 대중교통 이용 데이터를 이용하여 Regression을 통한 데이터 학습을 진행하였으며, 학습된 모델을 통한 날씨에 따른 서울특별시 대중교통 이용량 예측에 따른 평균 오차율은 15.49%로 낮은 오차율을 가진다.

본 연구 결과는 날씨에 따른 버스와 지하철의 배차 간격 조절 등의 대중교통 배치 판단 결정에 기초자료로 사용될 것으로 기대된다.

■ 서론

가) 제안배경 및 필요성

중·고등학교 시절과 다르게 대학 생활을 시작하면서 대중교통(지하철, 버스)을 이용하는 횟수가 급격하게 증가하였다. 기본 한 달 동안 사용되는 대중교통 비용만 6만원이 넘을 정도이다. 집과 학교간의 거리가 멀다보니 학기 중에는 거의 매일 대중교통을 이용하고, 그 외에도 약속이 있을 경우 대중교통을 이용하여 약속장소로 이동한다. 물론, 이동수단으로는 대중교통 외에도 자가용 또는 택시 등을 이용할 수 있지만 다른 이동 수단에는 많은 제약이 존재하여 본인을 포함한 많은 사람들이 이동수단으로 주로 대중교통을 이용한다. 실제로 '잡코리아'에서 직장인과 대학생 873명을 대상으로 '출퇴근, 통학 시 이용하는 이동수단은 무엇인가?'라는 질문에 81.5%가 대중교통을 이용한다고 답변했다.

다른 이동수단들의 제약 조건을 설명하자면 첫째, 자가용을 이용하기 위해서는 차량 구매와 운전 면허증 취득과 같은 부가적인 요소가 많이 필요하다. 차량 구매는 주변인 또는 가족의 자가용, 쏘카(Socar)와 같은 카셰어링 또는 렌터카를 이용으로 대체 할 수 있지만, 매번 이용하기에는 가격적으로 많은 부담이 된다. 만약 차량이 준비가 되어도 많은 어려움이 존재한다. 자가용 유지비용과 출·퇴근 시 교통난, 주차 시설의 어려움, 주유비 등 많은 이유로 자가용이 존재하는 사람들도 평소 출퇴근 시 대중교통을 이용하기도 한다.

둘째, 택시를 이용하기 위해서는 가장 먼저 비용 부담이 걱정이 된다. 학생들끼리 장난스럽게 택시를 이용하면 부자라고 놀리는 것처럼 택시는 비싸다는 인식이 강한 이동수단이다. 택시비용은 거리에 비례하여 증가하고, 버스와 지하철에 비하여 많이 비싸기 때문에 개인적으로 대중교통 운영이 끝난 시간과 같은 특별한 경우에만 이용하게 된다.

셋째, 요새는 운동 겸 자전거, 전동 휠 등 퍼스널 모빌리티도 많이 이용하는 추세이다. 하지만 다른 이동수단에 비하여 직접 운행해야 하기 때문에 가까운 거리에 적합하고, 안전상의 위험도 존재한다. 또한, 비나 눈이 오거나, 기온이 너무 낮거나 높을 경우 등 기후에 많은 영향을 받기도 한다. 이러한 여러 가지 이유로 많은 사람들이 학교나 직장에 등교·하교, 출근·퇴근할 때, 약속 장소로 이동할 때 등 장소를 이동할 때 자연스럽게 대중교통을 더 선호하게 된다.

이와 같이 현대 사회에서는 다양한 이동 수단 중 지하철, 버스 등의 대중교통에 대한 수요가 높은 편이다. 이를 뒷받침하듯이 서울특별시가 2016년도에 대중교통을 이용한 시민들의 교통카드 빅 데이터를 분석한 결과, 2016년도에 총 49억 4천만여 명 즉, 하루 평균 13,491천명의 많은 시민들이 서울 지하철과 버스를 이용한 것으로 집계되었다. 이처럼 대중교통 이용객이 많은 만큼, 대중교통의 수요를 예측하고 이에 맞춰 적절하게 대중교통의 배차를 조절하는 것은 매우 중요한 이슈가 되었다.

대중교통 이용량에는 날씨, 평일-주말, 연착, 도로현환 등 여러 가지가 원인이 된다. 날씨를 제외한 다른 요인들은 현재 많은 연구가 존재 및 진행되고 있으며 대중교통 이용량의 관계가 쉽게 파악이 가능하다. 반면 날씨와의 관계를 분석한 연구의 수는 적으며, 연구 범위가 넓지 않기 때문에 본 연구에서는 여러 요인 중에서 날씨 데이터(기온, 강수량, 미세먼지)에 초점을 두어, 날씨에 따른 대중교통 이용량의 변화 양상을 학습하여 예측하는 연구를 진행하고자 한다.

2017년 1월 8일, 공공 데이터 포탈에 따르면 지난해 공공데이터 개방형 애플리케이션 프로그래밍 인터페이스(Application Programming Interface 이하 API) 중 활용신청 수가 가장 많은 것은 날씨와 대중교통 관련 API로 집계되었다. 상위 20개 API 활용신청 1만4939건 가운데 36%에 달하는 5400건이 날씨 관련 API이다. 또한 버스 위치정보 조회서비스(서울특별시, 936건)을 포함하여 대중교통 관련 7개의 API의 활용신청 수가 4292건(28.7%)으로 대중교통 API의 인기도 높았다. 그 중 버스 출발·도착, 노선, 정류소 위치 등에 관한 정보 수요가 컸다. 이처럼 날씨와 대중교통 공공데이터 활용도가 높은 것은 그만큼 국민들의 관심이 많다는 점에서 논문을 진행하기에 매우 긍정적인 효과를 준다. 따라서 두 데이터를 연관 지어 날씨에 따른 대중교통 이용량을 분석 및 활용을 한다면 대중교통을 더욱 활성화 시키는 좋은 지표가 될 것이라고 예상한다.

나) 졸업논문의 목표

평소 걸어 다니던 거리도 기온에 따라 시민들의 대중교통 이용에 관한 생각에 차이가 생긴다. 예를 들어 날씨가 너무 덥거나 너무 추울 경우 버스나 지하철을 이용하는 것이 더 편하다는 생각에 혼자 또는 동행인을 설득하여 대중교통을 이용하고자 하는 경우가 생긴다. 기온 뿐 아니라 비나 눈이 많이 내려서 교통상황이 좋지 않을 경우, 평소 자전거를 타거나 걷는 사람들은 대부분 대중교통을 이용하고, 오히려 자가용을 이용하는 사람들도 차량 세차 걱정이나 도로가 막히는 것에 대비하여 대중교통을 이용하려는 경향이 커진다. 이외에도 이번년도에 많은 이슈를 가져온 미세먼지와 초미세먼지 때문에도 대중교통을 이용하는 이용객의 수가 증가했다. 게다가 이번에 실시하게 되는 '자동요금처리시스템'으로 미세먼지가 심한 날에는 대중교통을 무료로 이용할 수 있게 하여 이용객의 수는 더욱 증가할 것으로 보인다. 이처럼 날씨의 변화에 따라 대중교통을 이용하는 사람들의 생각에 많은 영향을 끼치게 된다. 따라서 본 논문에서는 날씨에 따른 대중교통 이용량을 분석하기로 한다.

'서울특별시 열린 데이터 광장'에서 제공하는 서울특별시 월별 버스노선별, 지하철 호선별 역별 승하차 인원 정보와 서울특별시 월별 평균 기상개황(기온, 강수량, 상대습도 등) 그리고 서울특별시 월별 평균 대기오염도 정보를 이용하여 날씨와 대중교통 이용량의 관계를 분석하고자 한다.

제공되는 날씨 데이터가 자치구별로 존재하기 때문에 버스와 지하철역도 자치구별로 구분하여 진행하기로 하였다. 하지만 서울특별시에 존재하는 모든 버스 노선과 지하철역으로 분석하기에는 범위와 데이터의 크기가 너무 커져서 논문을 진행하기에 무리가 있기 때문에 사용하는 버스와 지하철의 개수를 축소시켰다. 축소시키는 기준으로는 자치구별로 대중교통 이용량이 많은 상위 4개의 지하철역과 지하철역 주변에 위치하는 버스 정류장으로 기준을 세웠다. 버스 정류장은 이름이 다양하기 때문에, 지하철 역 이름을 포함하는 버스 정류장을 역 주변에 위치하는 버스 정류장으로 가정했다.

선정된 상위 4개의 대중교통 이용량을 기준으로 자치구별 평균 한 달 대중교통 이용량을 계산하여 진행한다. 동일한 기간인 자치구별 한 달 평균 날씨는 기온과 강수량, 습도, 미세먼지 정보를 이용한다. 초기에는 정확한 예측을 위하여 날씨 정보들은 정규화를 통해 조정을 하려고 했지만 정규화를 하지 않아도 원하는 예측 결과가 나왔기 때문에 본 연구에서는 정규화를 실시하지 않았다. 위의 과정을 통해 정리한 자치구별 대중교통 이용량과 한 달 평균 날씨를 Python Regression 함수를 통해 지도 학습을 시킨다.

학습 결과 '기온, 강수량, 상대습도, 미세먼지' 정보가 주어졌을 때, 자치구 별 대중교통 이용량을 예측할 수 있다. 예측한 대중교통 이용량 정보를 이용하면, 버스와 지하철 회사에서 탄력적으로 버스나 지하철을 배치할 수 있다. 예를 들어, 평소에는

기본이 되는 일정량의 버스와 지하철을 운행한다. 그러다가 이용량이 평소와 다르게 많아지는 달이라고 예상이 되면, 추가적으로 그 달에만 버스와 지하철을 추가 배치하는 것이 가능해진다.

다) 졸업논문 전체 overview

1) 선행 연구와 관련 기술 분석
: 날씨와 대중교통 이용량을 분석한 여러 논문들이다. 본 논문에서 진행하고자 하는 기준에 충족되지 않지만, 논문을 진행하는데 많은 지표가 되어줄 논문들이다.

- ✓ "강우와 서울시 대중교통 승차인원과의 관계 분석"는 날씨 데이터 중 오직 강우 상태만을 고려하여 대중교통 이용량을 분석하였다. 또한 사용된 데이터의 기간이 201년 3월 18일부터 3월 24일로 1주일간의 자료 중 주중자료만을 사용하였다. 즉, 범위가 매우 짧다.
- ✓ "강우 상태에 따른 대중교통 이용패턴 특성연구" 또한 강우 상태만을 고려하여 대중교통 이용량을 분석하였다. 그리고 분석에 이용된 대중교통이 부산광역시 버스만을 이용하였다. 그리고 데이터 기간이 2010년 1월에서 2010년 12월로 1년의 범위를 사용하였다.
- ✓ "과거 패턴자료와 날씨계수를 이용한 버스 통행시간 예측에 관한 연구"는 분석에 사용된 대중교통의 범위가 6614번(양천공영차고지~부천생태공원)까지 운행하는 노선에서 4주간의 버스 데이터만을 사용하였다. 즉 사용된 데이터의 기간과 대중교통 범위 모두 원하는 기준에 비하여 낮았다.

또한 위의 논문들 모두 대중교통과의 관계 분석으로 끝나며 예측하는 연구는 거의 존재하지 않는다.

날씨와 대중교통간의 관계뿐 아니라 논문에서 가장 중요한 부분은 지도 학습을 시키는 부분이다. 따라서 지도 학습으로 훈련 데이터를 학습 시키고 새로운 데이터의 값을 결정하는 여러 논문들을 찾아 읽어보면서 지도 학습하는 방법을 분석한다. 또한 지도 학습 시 사용되는 여러 개의 kernel 중 'linear', 'poly', 'rbf'가 주로 사용되는 kernel이다. 각 kernel의 장·단점을 분석하고 어떤 kernel을 사용하는 것이 원하는 결과에 가까운지 판단하여 논문을 진행한다. 또한 각 kernel 뿐 아니라 사용되는 여러 지도 학습 함수들을 비교하여 가장 좋은 예측 결과를 제공하는 함수를 선택한다. 그리고 지도 학습에 주로 사용하는 Classification과 Regression 중 본 논문에 적합한 방법을 채택하여 학습을 한다.

2) 데이터 수집 및 가공

: 데이터는 '서울 열린 데이터 광장'에서 제공하는 서울특별시 월별 버스노선별, 지하철 호선별 역별 승하차 인원 정보, 서울특별시 월별 평균 기상개황 그리고 서울특별시 월별 평균 대기오염도 정보를 이용한다. 따라서 데이터를 수집하기에는 크롤링이 필요하지 않아 큰 어려움은 없다. 하지만, 제공하는 데이터를 논문에 이용하기 적절한 상태로 처리하기에 많은 시간과 노력이 걸린다. 예를 들어, 이용량이 많은 상위 4개의 역을 구하기 위해는 각 역의 총 이용량을 계산하여 비교해야한다. 또는 지하철역 근처에 존재하는 버스 정류장을 찾아내는 부분도 쉽지 않은 과정이다. 이렇게 데이터를 처리하는 과정은 주로 python으로 코딩하여 해결하였다. 하지만 코딩으로 진행하기 어려운 부분, 즉 컴퓨터가 판단하기 어려운 부분에 대해서는 직접 처리하였다.

3) 데이터 분석

: 가공한 데이터를 python 코딩을 이용하여 지도학습을 한다. 앞서 말했듯이, Classification과 Regression 중 선택한 방법과 여러 개의 kernel을 이용하여 좀 더 원하는 결과를 제공하는 kernel을 찾는다. 학습을 시킨 후에는 앞으로 날씨 변화에 따른 대중교통 이용량을 예상할 수 있다. 또한 각 날씨와 대중교통 이용량의 관계도 분석하여 이용량에 많은 영향을 주는 요인과 상관관계도 찾아낼 수 있다.

4) 논문 작성

: 위의 1)~ 3)과정을 통해 공부한 지식과 기술을 서술하고, 데이터를 처리하는 과정을 자세히 서술한다. 데이터 처리 과정에서 일어났던 여러 문제점도 분석하여 해결 방안도 작성한다. 그리고 처리된 데이터를 어떻게 분석하였는지도 서술하고, 분석 결과를 자세히 작성한다. 마지막으로 논문 결과를 이용 가능한 방향을 제시하면서 졸업 논문을 마무리하고자 한다.

■ 관련연구

1) 지도 학습

지도 학습은 label이 주어진 훈련 데이터로부터 하나의 함수를 유추하는 기계 학습 방법 중 하나이다. 훈련 데이터는 일반적으로 벡터 형태로 나타나는 입력 객체와 원하는 출력 값을 가지고 있는 예제들의 모임이다. 지도 학습은 출력 값의 형태에 따라서 Regression과 Classification로 나뉜다.

Regression은 유추된 함수 중 연속적인 값을 출력하는 것을 말한다. 즉 학습 데이터가 연속 형 변수들이며, 이를 기반으로 두 변수 사이의 모형을 구한 후에 검증 데이터의 예상 값을 측정하는 방법이다. 주로 예측하는 결과 값이 연속 값일 때 사용한다.

반대로 Classification는 주어진 입력 벡터가 어떤 종류의 값인지 표시하는 것을 말한다. 즉 학습 데이터가 Category로 분류되어 있어, 이를 기반으로 검증 데이터의 Category를 식별하는 방법이다. 주로 예측하는 결과 값이 이산 값일 때 사용된다.

이 두 지도 학습기가 하는 작업은 훈련 데이터로부터 주어진 데이터에 대해 예측하고자 하는 값을 올바르게 추측하는 것이다. 올바른 추측을 하기위해 지도 학습기가 적합한 방법을 통해 기존의 훈련 데이터로부터 알 수 없었던 상황까지도 일반화하여 처리할 수 있어야 한다. 그리고 본 논문에서는 대중교통 이용량이라는 연속 값을 예측하기 때문에, Classification 함수보다 Regression 함수를 사용하는 것이 더 적합하다.

1-1) Regression

Regression은 관찰된 연속형 변수에 대하여 두 변수 사이의 관계(모형)을 구한 뒤 적합도를 측정하는 분석 방법이다. 회귀분석은 시간에 따라 변화하는 데이터나 인과 관계의 모델링 등의 통계적 예측에 이용될 수 있다. 하지만, 대부분 가정이 맞는지 아닌지 정답과 비교하여 확인하지 않은 채로 이용되어 결과가 오용되는 경우도 있다.

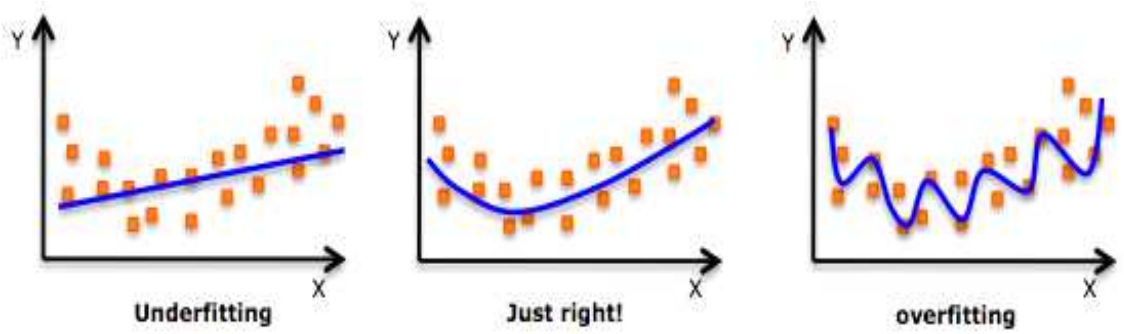
Regression은 하나의 종속변수와 독립변수 사이의 관계를 분석하는데, 만약 하나의 독립변수와의 관계를 분석할 경우, Simple-Regression이라고 한다. 반면, 여러 독립변수와의 관계를 분석할 경우, Multi-Regression이라고 한다.

Regression은 5가지의 가정을 바탕으로 이루어진다. 첫 번째, 오차 항은 모든 독립변수 값에 대하여 동일한 분산을 갖는다. 두 번째, 오차항의 평균값은 0이다. 세 번째, 수집된 데이터의 분산은 정규분포를 이루고 있다. 네 번째, 독립변수 상호간에는 상관관계가 없어야 한다. 다섯 번째, 시간에 따라 수집한 데이터들은 잡음의 영향을 받지 않아야 한다. 만약 독립변수간의 상관관계가 나타나는 경우 다중공선성

문제라고 한다.

2) 과적합

지도학습 과정에서 주의해야할 점 중 하나는 과적합이다. 과적합은 훈련 데이터 set에 너무 가깝거나 정확하게 일치하는 분석을 생성하여 새로운 데이터를 올바르게 분류하지 못하거나 앞으로의 관측을 신뢰성 있게 예측하지 못하는 것을 말한다. 즉, 과적합(Overfitting)이란 말 그대로 너무 과도하게 훈련 데이터에 대해 모델을 학습시킨 것을 말한다. 흔히 주어진 훈련 데이터에 대하여 거의 100% 정확하게 예측할 수 있으면 새로운 데이터에 대해서도 좋은 예측을 할 수 있는 좋은 모델이라고 생각한다. 하지만 실제로는 오히려 훈련 데이터에 대하여 60~70% 정확하게 예측할 수 있는 모델이 새로운 데이터에 대해서 더 좋은 예측을 할 수 있다. 아래 (그림 1)을 통해 부적합(Underfitting)과 과적합(Overfitting)의 차이점과 새로운 데이터에 대하여 예측을 정확하게 하지 못하는 것을 알 수 있다.



(그림 1) 부적합과 과적합

과적합과 반대로 부적합은 너무 적은 훈련데이터에 대해 모델을 학습 시켜서 훈련 데이터 예측률로 낮고, 새로운 데이터 예측률도 낮다. 그리고 과적합은 훈련 데이터 예측률은 매우 높으나, 새로운 데이터 예측률이 낮은 것을 알 수 있다. 따라서 지도 학습을 할 때, 과적합을 막기 위해 검증 데이터라는 것을 정의했다.

2-1) 검증 데이터

기존에 모두 훈련 데이터로 쓰던 데이터 중 일부분을 검증 데이터로 구분한다. 그리고 지도 학습 과정에서 훈련 데이터로 모델을 구축한 뒤, 검증 데이터를 새로운 데이터라고 가정하고 예측을 실시한다. 그 다음 검증 데이터의 실제 값과 예측 값을 비교하여 예측률이 가장 높은 모델을 적합한 모델이라고 학습시키는 것이다.

Python에는 train_test_split라는 함수를 이용하여 훈련 데이터와 검증 데이터를 분

류를 할 수 있다. 이 함수는 아래 (식 1)과 같은 형식으로 작동된다.

(식 1) X_train, X_test, y_train, y_test
= train_test_split(X, y, test_size=0.3, random_state=0)

X를 통해 도출하는 값을 y라고 설정한다. 예를 들어, 본 논문에서는 날씨 데이터를 가지고 대중교통 이용량을 예측하고자 한다. 따라서 X는 날씨 데이터, y는 대중교통 이용량으로 설정한다. test_size는 훈련 데이터와 검증 데이터를 나누는 비율을 의미한다. 'test_size=0.3'의 의미는 전체 데이터 중 30%를 검증 데이터로 사용한다는 의미이다. 그리고 random_state는 주어진 전체 데이터를 random하게 정렬한 후 그 중에서 일정한 기준으로 검증 데이터를 고르기 위해 사용된다. 또한 X_train과 y_train은 훈련 데이터, X_test와 y_test는 검증 데이터로 사용된다.

3) 정규화

본 논문에서 실행하고자 했던 정규화는 통계관련용어로 평준화라고 볼 수 있다. 평준화는, 다른 척도로 측정된 값을 가지는 평준화하기 전의 데이터를 공통적인 척도로 조정하는 것을 의미한다. 복잡한 경우에는, 조정 된 데이터의 전체 확률 분포를 정렬하는 것보다 정규화가 더 정교한 조정을 나타낼 수 있다. 또 다른 사용법으로는, 정규화는 통계의 이동 및 크기 조정 된 버전의 생성을 의미하며, 이렇게 정규화 된 데이터를 사용하면 특정 데이터 set의 특정 영향을 주는 부분을 찾아내어 제거하는 방식으로 비교할 수 있다.

논문에서는 날씨 데이터set의 원소인 기온, 상대습도, 강수량, 미세먼지 각각의 데이터를 정규화하려고 했다. 적용하려고 했던 정규화 (식 2)은 아래와 같다.

(식 2) $X_{normal} = (X - X_{\min}) / (X_{\max} - X_{\min})$

먼저 각 날씨 데이터 원소를 X라고 표현했다. 즉 날씨 데이터 원소의 원본 데이터 값이다. 다음 각 날씨 데이터 원소 중 최솟값을 $X_{\min}$, 최댓값을 $X_{\max}$으로 잡는다. 그 다음 원본 데이터에서 최솟값을 뺀 뒤, 최댓값과 최솟값의 차이로 나누어준 것을 X_{normal} 즉, 정규화한 날씨 데이터 원소로 설정한다.

4) SVM

SVM은 기계 학습의 분야 중 하나로 자료 분석이나 패턴 인식을 위한 지도 학습 모델이다. 주로 분류와 회귀분석을 위해 사용한다. 데이터들이 두 카테고리 중 어느 하나에 속하는지 label을 가지는 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새롭게 받는 데이터가 어느 카테고리에 속하는지 판단하는 비확률적 이진 선형 분류 모델을 만든다. 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데, SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. 즉 두 카테고리를 잘 분리할 수 있는 경계를 찾는 알고리즘이라고 이해할 수 있다. SVM은 선형 분류에서도 사용될 수 있다.

4-1) SVM 관련 코드 및 kernel

먼저 Support Vector Regression(SVR)와 관련된 python코드이다. 'sklearn.svm.SVR', 'sklearn.svm.LinearSVR', 'sklearn.svm.NuSVR'이 존재한다. SVR은 kernel을 설정할 수 있다. kernel은 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'가 존재한다. SVR은 libsvm을 기반으로 구현되며 10000개가 넘는 샘플을 두 개 이상 포함하는 데이터를 사용하기에는 부적합하다. LinearSVR는 SVR에서 kernel을 'linear'로 설정한 것과 비슷하지만, libsvm이 아닌 liblinear 기반으로 구현되므로 많은 수의 샘플을 가지는 데이터에도 사용가능하다. NuSVR는 SVR와 비슷하지만 매개 변수를 사용하여 support vector의 수를 제어 가능하다. 또한 libsvm을 기반으로 구현된다.

다음은 Support Vector Classification(SVC)와 관련된 python코드이다. 'sklearn.svm.SVC', 'sklearn.svm.LinearSVC', 'sklearn.svm.NuSVC'이 존재한다. SVR를 SVC로 바꾼 것으로 자세한 설명은 생략한다.

4-2) Kernel에 관한 간략한 설명

각 kernel에 따라 label을 구분하는 경계선이 달라진다. 'linear'일 경우, 데이터를 직선을 기준으로 구분된다. 'poly'일 경우, Polynomial의 축약형으로 데이터를 n차 곡선을 기준으로 구분된다. 'rbf'일 경우, (Gaussian) radial basis function의 축약형으로 데이터를 선을 기준으로 나누는 것이 아니라, 원형으로 나눈다. 즉, 일정 범위 안에 존재하는 데이터들은 같은 label로 구분한다. 'sigmoid'와 'precomputed'는 거의 사용하지 않는 kernel이다.

5) 상관관계

상관분석은 두 변수 간에 어떤 선형적 관계를 가지고 있는지를 분석하는 방법이다. 두 변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수간의 관계의 강도를 상관관계(Correlation)라고 한다. 이러한 상관분석에서는 상관관계의 정도를 나타내는 단위로 모상관계수 p를 이용한다.

상관관계의 정도를 파악하는 상관계수는 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다. 즉 상관계수가 높게 나온다고 두 변수 간에 인과관계가 존재하는 것은 아니다. 두 변수간의 원인과 결과의 인과관계가 있는지에 대한 것은 회귀분석을 통해 인과관계의 방향, 정도와 수학적 모델을 확인해 볼 수 있다.

5-1) 피어슨 상관계수

본 논문에서 사용된 상관계수는 피어슨 상관계수이다. 먼저 기본적으로 상관계수 p가 $0 < p \leq +1$이면 양의 상관, $-1 \leq p < 0$이면 음의 상관, $p = 0$이면 무상관이라고 정의한다.

두 변수 X, Y에 대하여, "X와 Y가 함께 변화는 정도 / X와 Y가 따로 변하는 정도"의 결과 값을 피어슨 상관계수라고 한다. 따라서 X와 Y가 완전히 동일하면 p=+1, 전혀 다르면 p=0, 반대방향으로 완전히 동일하면 p=-1을 가진다. 이것을 좀더 세분화 시켜서 해석하면, 아래 <표1>와 같다.

<표 1> 피어슨 상관계수 해석

p의 범위	해석
-1.0 < p < -0.7	강한 음적 선형관계
-0.7 < p < -0.3	뚜렷한 음적 선형관계
-0.3 < p < -0.1	약한 음적 선형관계
-0.1 < p < +0.1	거의 무시될 수 있는 선형관계
+0.1 < p < +0.3	약한 양적 선형관계
+0.3 < p < +0.7	뚜렷한 양적 선형관계
+0.7 < p < +1.0	강한 양적 선형관계

■ 제안 작품 소개

작년 2016년도에 서울특별시에서 대중교통을 이용한 시민들을 분석한 결과, 총 49억 4천만여 명으로 하루 평균 13,491천명의 많은 시민들이 서울 지하철과 버스를 이용하였다고 결과가 나왔다. 이처럼 많은 시민들이 이용하는 대중교통 이용량에는 여러 요인이 영향을 끼친다.

기본적으로 평일인지 주말인지에 따라 이용량에 많은 영향을 받으며 또한 대중교통의 연착 여부와 도로현황 그리고 날씨에 따라 많은 영향을 받게 된다. 다른 요인들에 비하여 날씨와 대중교통 이용량에 대한 관계는 쉽게 드러나지 않으며 연구가 많이 되고 있지 않다. 또한 공공데이터 개방형 API 활용 신청 수를 분석해본 결과, 날씨와 대중교통 관련 API가 가장 많은 것으로 나타난다. 즉, 이 결과는 날씨와 대중교통에 국민들의 관심이 많다는 것을 의미한다. 따라서 본 논문에서는 두 데이터를 연관 지어 날씨와 대중교통 이용량의 상관관계를 분석하고자 한다.

1) 데이터 처리

'서울특별시 열린 데이터 광장'에서 제공하는 데이터들을 사용하여 본 논문을 진행했다. 먼저 대중교통 데이터에 사용될 데이터로는 서울특별시 월별 버스노선별(이하 버스 데이터) 그리고 지하철 호선별 역별 승하차 인원 정보(이하 지하철 데이터)를 선정하였다.

그 다음 날씨 데이터에 사용될 데이터로는 서울특별시 월별 평균 기상개황(이하 기상 데이터) 그리고 서울특별시 월별 평균 대기오염도 정보(이하 대기오염도 데이터)를 선정하였다. 총 4개의 데이터들을 이용하여 날씨 데이터와 대중교통 데이터를 추출하고, 날씨와 대중교통 이용량의 관계를 분석하고자 한다.

또한 데이터 분석에는 2015년 1월부터 2017년 5월까지 총 29개월간의 데이터를 사용한다.

1-1) 날씨 데이터 처리

날씨 데이터는 기상 데이터와 대기오염도 데이터, 두 데이터를 이용하였다. 먼저 기상 데이터는 서울특별시 평균 월별 데이터로서, 다양한 기상 데이터 중 필요한 평균 기온(℃), 강수량(mm), 평균 상대습도(%)을 추출하였다. 하지만 이 기상데이터만 사용하면 월별 데이터만이 존재하기 때문에, 좀 더 세분화를 주기 위해 자치구별로 데이터를 나눌 수 있는 날씨 데이터를 하나 더 추가했다. 그래서 서울특별시 자치구별 월별 데이터로 제공되는 대기오염도 데이터를 추가적으로 이용한다. 오존농도, 초미세먼지와 같은 다양한 대기오염도 데이터에서 자치구별 미세먼지(㎍/㎥)

만을 이용하였다. 따라서 월별 자치구별 날씨 데이터를 아래 <표 2>와 같이 구성하게 되었다.

<표 2> 날씨 데이터

평균 기온(℃)	강수량(mm)	평균 상대습도(%)	미세먼지(㎍/㎥)

1-2) 대중교통 데이터 처리

대중교통 이용량으로는 버스 데이터와 지하철 데이터를 이용하였다.

1-2-1) 지하철 데이터 처리

제공된 지하철 원본 데이터에는 전국에 존재하는 지하철역이 존재하여, 검색 포탈 네이버 지도를 이용하여 지리적으로 서울특별시에 포함되지 않는 지하철 정보들은 1차적으로 제거한다. 또한 지하철역 주소를 기반으로 지하철역을 25개의 자치구로 분류한다. 이 과정에서 같은 이름의 지하철역 이지만 호선에 따라 자치구가 다르게 구분되는 지하철도 존재하는 사실도 알게 되었다. 그리고 수집된 데이터 기간 동안 지하철역 명이 변경된 지하철역들이 존재하여 통일하는 과정도 추가하였다.

지하철 이용량은 지하철 원본 데이터에 존재하는 시간별 승차인원, 하차인원을 이용하여 월 평균 승·하차인원을 계산하였다. 그 다음 월 평균 승·하차 인원을 모두 합하여 자치구별로 대중교통 이용량에 사용할 상위 k개를 선정한다. 이 과정에서 동일한 자치구에 존재하는 환승 지하철역(같은 지하철역 명이지만 호선이 다른 경우)은 한 개의 지하철역으로 합쳐서 계산했다.

이때, 자치구별로 최소 3개 이상의 지하철역이 존재하여 k의 값을 최소 3으로 선정한다. 버스 평균 승·하차 인원도 포함하여 상위 k개 역을 선정해야하므로 현 단계에서는 후보 지하철역을 포함하여 자치구별 k^*개의 역을 선정한다.

(식 3) $k^* \geq k$

k^*와 k의 관계는 (식 3)과 같으며, 본 연구에서는 자치구별 지하철역의 수를 고려하여, 아래의 (식 4)과 같이 정하였다.

(식 4) $k^* = k + 2$

1-2-2) 버스 데이터 처리

버스 원본 데이터는 약 319MB로 이대로 데이터를 처리하기에는 용량이 크기 때문에 python을 이용하여 데이터의 크기를 줄였다. 먼저 1-2-1)에서 처리한 후보 지하철역과 관련 있는 데이터만 필요하기 때문에 불필요한 버스 데이터를 제거하는 것으로 시작하였다.

수많은 버스 데이터 중 후포 지하철역 근처에 위치하는 버스 정류장만을 이용한다. 근처에 위치하는지 판별하는 기준으로는 후보 지하철역 이름을 포함하는 버스 정류장으로 선정하였다. 예를 들어 후보 지하철역에 '합정'이 존재할 경우, 버스 정류장 이름에 '합정역'이 포함된 정류장만을 사용한다. 만약 '합정역'을 포함한 정류장이 하나라도 없을 시, '합정'이 포함된 정류장으로 선정한다.

이 과정에서 29개월 전체 데이터를 비교하여 버스 정류장을 추출하는 것보다 무작위로 선정한 2015년 6월에 존재하는 버스 정류장 데이터를 기준으로 진행하는 것이 더 효율적이라고 판단했다. 따라서 선정된 달의 버스 정류장 이름만을 이용하여 필요한 버스 정류장 이름과 버스 정류장의 key가 되는 '표준 버스 정류장ID' 그리고 근처 지하철 역 이름을 엑셀로 출력시켰다. 이렇게 생성된 엑셀에 존재하는 버스 정류장들 중 직접 확인하는 과정을 통해 잘못 삽입된 데이터들은 제거하였다. 예를 들어 '동대문역'을 추출하는 과정에서 '동대문역사문화공원'이 포함된 것은 제거했다. 그 다음, '표준 버스 정류장ID'를 통해 총 5419개의 버스 정류장을 중복 제거하여 1045개의 버스 정류장으로 데이터를 정리하였다.

이제는 '표준 버스 정류장ID' 데이터를 기준으로 python을 이용하여 버스 원본 데이터에서 필요한 버스 정류장만을 찾아낸다. 이 과정에서 버스 원본 데이터에서 필요한 데이터만 추출하기 위해, '날짜, 표준 버스 정류장 ID, 근처 지하철 역 이름' 그리고 시간별로 존재하는 승·하차 인원을 더하여 '월 평균 승·하차 인원'을 계산하여 총 4가지 요소 <표 3>을 포함하여 엑셀로 출력하였다.

<표 3> 버스 데이터

날짜	표준 버스 정류장 ID	근처 지하철 역 이름	월 평균 승·하차 인원

그 다음 과정으로는 각 지하철 별로 근처에 존재하는 버스 정류장의 승·하차 인원을 계산하여 월별 지하철역별 평균 승·하차 인원 데이터를 만들었다.

1-2-3) 대중교통 데이터 처리

1-2-1)에서 선정한 자치구별 k^*개의 후보 지하철역 데이터와 1-2-2)에서 추출한 버스 데이터를 이용하여 최종적으로 상위 k개의 지하철역을 선정하였다. 먼저 버스 데이터도 29개월 총 평균 승·하차인원을 구하여 지하철역 데이터와 합쳤다. 두 데이터의 평균 승·하차인원을 기준으로 자치구별 상위 k개의 지하철역을 선정하였다. 마지막으로 날씨 데이터와 구성방식을 통일시키기 위해 먼저 월별 k개의 지하철역 대중교통 이용량의 승·하차인원을 계산했다. 그 다음 자치구별 대중교통 이용량을 측정하기 위해 앞서 계산한 승·하차인원을 이용하여 월별 자치구별 평균 대중교통 이용량(승·하차인원)을 산출했다.

본 연구에서 사용된 자치구별 대중교통 데이터 k의 값은 4를 사용하였으며, 이에 대해서는 '구현 및 결과분석'에서 자세하게 설명할 예정이다.

2) 지도 학습

1)에서 구한 날씨 데이터와 대중교통 데이터를 하나로 통합하여 하나의 학습 모델 데이터를 구축한다. 이 구축된 데이터로 지도학습을 진행하여 컴퓨터에게 학습을 시킨다. 학습 과정에서 과 적합을 막기 위해 전체 데이터를 훈련 데이터와 검증 데이터로 나누어 학습을 진행한다. Python에서 제공하는 train_test_split 함수를 사용하였으며, test_size를 0.3으로 설정하여 random하게 나누었다. 즉 전체 데이터 중 70%를 훈련 데이터, 나머지 30%를 검증 데이터로 사용하였다.

그리고 여러 가지 지도 학습 method를 이용하여 학습을 시킨 결과 Classification보다 Regression이 본 논문에 더 적합하다고 판단하였다. 자세한 이유는 '구현 및 결과분석'에서 설명하고자 한다. 따라서 본 논문에서는 Python 언어를 이용하여 'scikit-learn' 라이브러리에서 제공하는 다양한 19가지의 Regression 함수들을 사용해본 결과, 가장 적합한 GaussianProcessRegressor를 사용하여 논문을 진행하였다. 이 함수를 선정하게 된 이유도 '구현 및 결과분석'에서 자세하게 다룰 예정이다. 또한 학습 데이터의 요소로는 '기온', '강수량', '미세먼지'만을 이용하여 진행하였다. 즉, 상대습도 데이터가 학습 데이터 요소에서 빠지게 되었는데, 이 부분도 '구현 및 결과분석'에서 설명한다.

지도 학습을 한 뒤 새로운 데이터에 대하여 예측한 결과는 평균 오차율(r)로 정확도를 판단하였다. 오차율(r)은 아래의 (식 5)을 이용하여 계산한다.

(식 5) $r = (|\ \text{측 값} - \text{실제 값}|)/\text{실제 값} \times 100$

GaussianProcessRegressor를 이용한 각 오차율의 평균을 구한 결과, 약 15.49%의 낮은 평균 오차율(r)을 가진다. 자세한 구현 방법과 결과는 '구현 및 결과분석'에서 다루고자한다.

■ 구현 및 결과분석

1) 데이터 처리 과정에서의 오류

구현 과정에서 여러 번의 데이터 처리 과정 오류로 데이터를 재가공한 경험이 있다. 여러 재가공 경험 중 버스 원본 데이터를 버스 정류장 이름을 기준으로 데이터를 선정하여 1차적으로 버스 데이터를 만드는 과정이 가장 어려웠다.

주어진 버스 원본 데이터가 여러 개의 sheet로 구성 되어있는데, 일정번호 sheet를 기준(10 sheet)으로 양쪽 데이터의 데이터 타입이 다른 경우였다. 표준 버스 정류장 ID가 숫자로 구성되어있기 때문에 당연히 int type이라고 생각하고 python 코딩과 논문을 진행했다. 하지만 10 sheet부터는 int type뿐 아니라 string type도 섞여있어서 제대로 원하는 데이터를 추출하지 못하였다. 당연히 코드는 오류 없이 실행이 되었고, 엑셀로 출력이 되었기 때문에 원하는 결과 값이 출력되었다고 생각했다. 하지만 대중교통 이용량 데이터를 가공하면서 일정한 월 이후에는 버스의 대중교통 이용량이 현저히 줄어든 것을 이상하게 생각하여 분석한 결과 잘못 출력된 것이었다. 초기에 결과 엑셀을 자세히 분석하지 않아서 추후에 문제를 발견하고 데이터 가공부터 다시 한 경우였다. 다행히 이 경우는 코드 수정은 쉬운 편이었다. 모든 데이터를 string type으로 변환하여 진행한 결과 원하는 결과가 출력이 되었다.

2) 지도 학습에 사용한 method

2-1) Regression을 선택한 이유

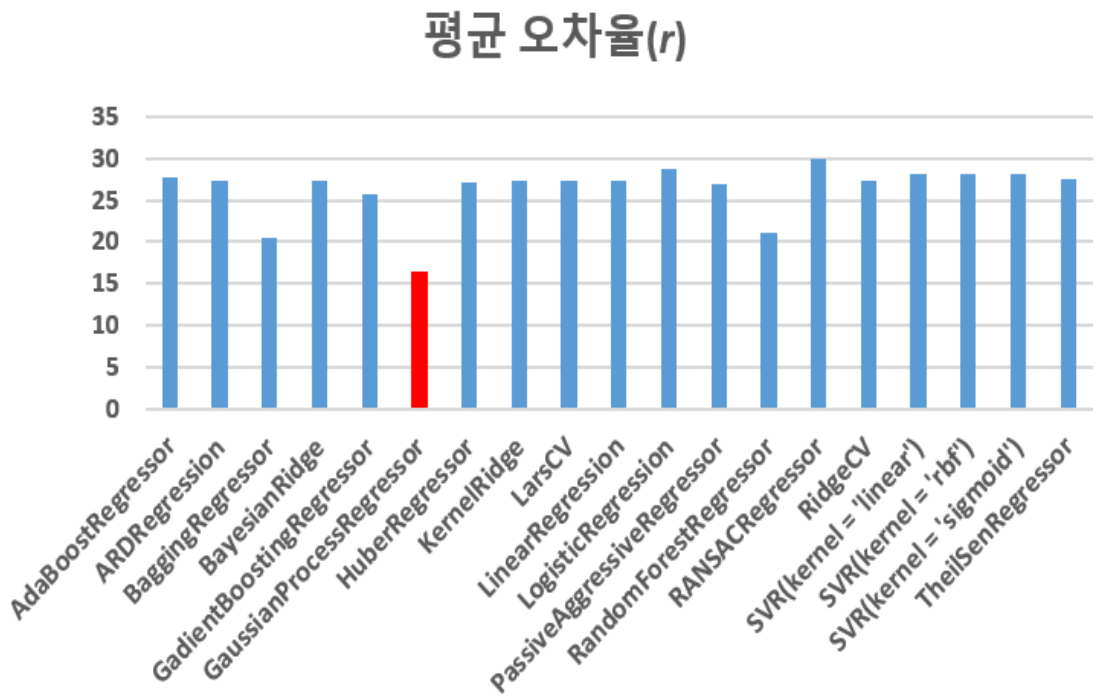
Classification의 경우, 학습 데이터가 Category로 분류되어 있어, 이를 기반으로 검증 데이터의 Category를 식별하는 방법이다. 대부분 예측하는 결과 값이 이산 값일 경우 Classification을 주로 사용한다. 예를 들어, 강아지와 고양이로 카테고리를 분류한 사진들을 학습한 뒤 새로운 데이터가 들어오면 강아지인지 고양이인지 판단하는 것이다.

반면 Regression은 주어진 학습 데이터가 연속 형 변수들이며, 이를 기반으로 두 변수 사이의 모형을 구한 후에 검증 데이터의 예상 값을 측정하는 방법이다. 따라서 대부분 예측하는 결과 값이 연속 값일 경우 Regression을 사용한다.

본 논문에서는 대중교통 이용량인 연속 값을 예측해야한다. 따라서 Classification 함수보다 Regression 함수가 더 적합하여 이를 통해 논문을 진행하였다. 만약 이용량에 따른 혼잡도(여유, 복잡)를 분류하는 것이라면 Classification을 사용하여 진행했을 것이다.

2-2) GaussianProcessRegressor를 선택한 이유

본 논문에서 진행하는 method는 'scikit-learn'에서 제공하는 19가지의 Regression method를 앞에서 처리하여 수집된 학습 데이터를 사용하여 학습 모델을 구축한다. 그리고 각 모델의 평균 오차율(r)을 계산하여 그중 가장 낮은 평균 오차율(r)을 가진 method로 선정하였다. Regression method의 수를 줄이기 위해 먼저 날씨 데이터로 선정된 '평균 기온(℃), 강수량(mm), 평균 상대습도(%), 미세먼지(㎍/㎥)'를 학습 데이터 요소로 가진 상태에서 평균 오차율(r)을 구하였다. 평균 오차율(r)을 계산한 결과는 아래 (그림 2)와 같다.



(그림 2) 각 method의 평균 오차율

(그림 2)를 보면 평균 오차율(r)이 가장 낮은 'GaussianProcessRegressor'이 본 논문에서 사용하기에 적합한 method라고 볼 수 있다. 하지만 이후 학습 데이터를 분석하기 위하여 비교 method로 평균 오차율(r)이 25% 이하인, 'RandomForest Regressor', 'BaggingRegressor'도 선정하여 논문을 진행했다.

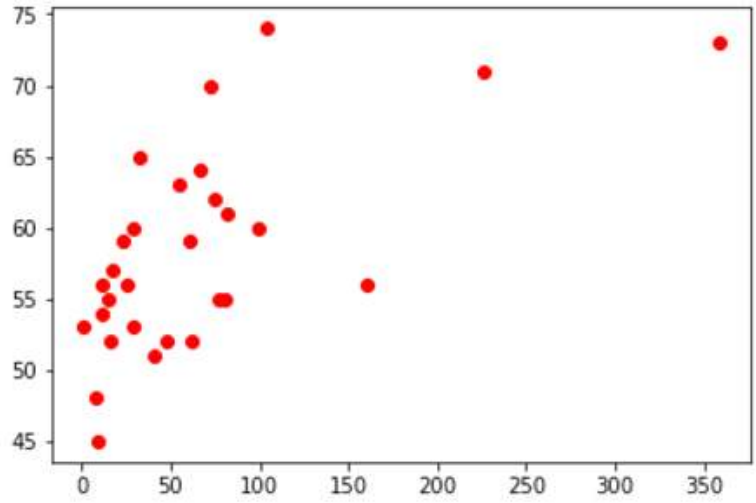
아래 (그림 3)는 세 가지 method의 평균 오차율(r)만을 비교한 그래프이다. GaussianProcessRegressor가 가장 낮은 평균 오차율(r)을 가지는 것을 확인할 수 있다.

(그림 3) Regression 함수에 따른 평균 오차율(r)

3) 학습 데이터 요소 결정

본 논문에서는 여러 가지 날씨 데이터 중 대중교통에 원인을 주는 요인으로 '평균 기온(℃), 강수량(mm), 평균 상대습도(%), 미세먼지(㎍/㎥)' 4가지의 날씨 데이터를 선정했다. 선정된 날씨 데이터 중 강수량(mm)과 상대습도(%)는 밀접한 관계가 있다고 예상을 했다. 왜냐하면 강수량이 증가하면 그만큼 습도도 증가하기 때문이다. 따라서 두 데이터의 상관관계를 분석하게 되었다.

먼저 강수량에 대한 상대습도 데이터를 그래프로 나타낸 결과 아래 (그림 4)처럼 양적 선형관계가 있음을 알 수 있다. 그리고 상관관계를 계산해 본 결과 약 +0.641로 +0.3~+0.7 사이에 존재하여 뚜렷한 양적 선형관계임을 분석 가능하다. 그 결과 학습 데이터 요소로 '강수량'과 '상대습도'를 둘 다 사용하는 것은 비효율적이라고 생각했다. 따라서 2-2)에서 선정한 총 3개의 method를 이용하여 학습 모델 구축의 간결함을 위해 강수량과 상대습도 중 하나의 요인만을 이용하도록 했다.



(그림 4) x축 : 강수량, y축 :상대습도

아래 <표 4>는 세 가지의 method의 학습 데이터 요소를 다양하게 바꾸면서 학습한 결과이다. 학습 데이터 요소에 '강수량'만을 넣어서 진행한 결과, '상대습도'만을 넣어서 진행한 결과, 두 요소를 모두 넣어서 진행한 결과로 총 3가지로 분류가 된다.

<표 4> 지도 학습 평균 오차율(r) 결과 (단위 : %)

Method Name	강수량	강수량+상대습도	상대습도
GaussianProcessRegressor	15.6163	15.6163	15.61631
RandomForestRegressor	19.5681	19.33563	19.77877
BaggingRegressor	19.46849	20.05177	19.92979

만약 'GaussianProcessRegressor'만을 이용하면 '강수량', '강수량+상대습도', '상대습도'의 평균 오차율(r)의 차이가 미미하여 다른 두 개의 method를 이용하였다. 그 결과 두 요소를 다 포함하였을 때의 평균 오차율과 '강수량'만을 포함하였을 때의 평균 오차율이 거의 비슷하거나 오히려 '강수량'만을 포함했을 때 더 정확한 것을 알 수 있다. 따라서 학습 데이터 요소로 '상대습도'를 제거하고 '강수량'만을 사용하였다.

또한 대중교통 이용량은 이전 달 이용량에 영향을 받을 것으로 예상하여 이전 달 이용량도 학습 데이터 요소로 넣어서 진행해보았다. 하지만 그 결과 오차율이 거의 0%에 다다르고, 논문의 취지인 날씨에 따른 대중교통 이용량 예측에 벗어난다고 판단하여 학습 데이터 요소로는 날씨 데이터 총 3개만을 이용하기로 결정했다.

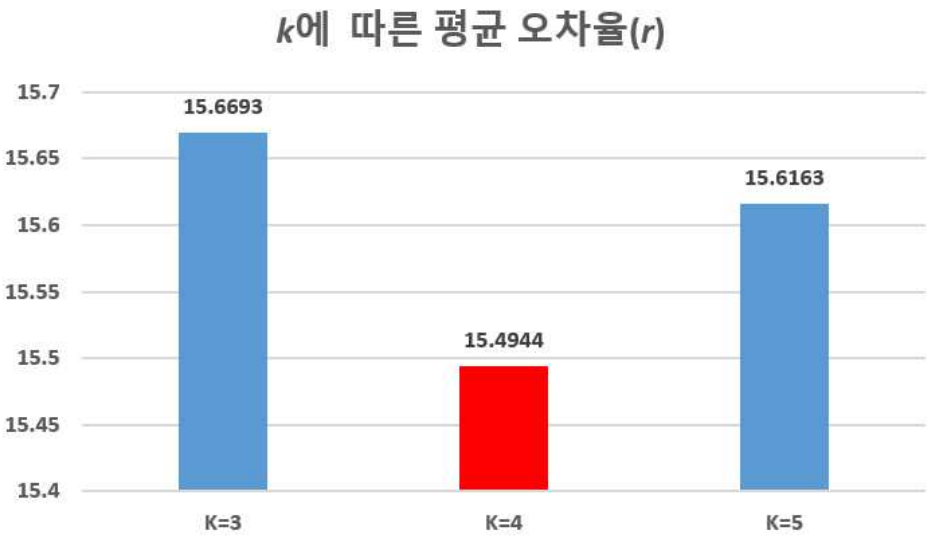
<표 5> 학습 데이터 요소 결정

평균 기온(℃)	강수량(mm)	미세먼지(㎍/㎥)

3) 지도 학습 결과

3-1) k 선정에 따른 학습 모델 평가

(그림 5)는 k값에 따라 학습 모델의 평균 오차율(r) 변화를 비교한 그래프이다. 본 논문에서 사용되는 GaussianProcessRegressor 함수와 <표 5>에서 결정된 학습 데이터를 이용하여 진행하였다.



(그림 5) k에 따른 평균 오차율(r)의 변화

결과에서 알 수 있듯이 본 논문에서는 가장 낮은 평균 오차율(r) 15.4944%을 가지는 k=4일 때를 사용한다. k=3일 때는 15.6693%로 k=4일 때보다 평균 오차율(r)이 높았다. 이유는 k=4가 되면서 학습 데이터로 더 많은 정보들이 제공되기 때문에 예측 율이 높아진다. 하지만 k=5일 때는 15.6163%로 k=4일 때보다 평균 오차율(r)이 높았다. 이유는 데이터의 과적합이 발생하여 오히려 예측 율이 떨어지게 된 것

이다.

따라서 본 논문에서는 k=4를 사용하여 논문을 진행했다. 즉 대중교통 이용량 데이터에 사용되는 자치구별 지하철역을 4개로 설정한 것이다. 만약 3개밖에 없는 자치구는 그대로 3개만을 사용하였다.

3-2) 논문 결과

<표 6>은 GaussianProcessRegressor을 사용하여 '평균 기온, 강수량, 미세먼지, 대중교통 이용량'을 학습 데이터로 가지는 학습 모델을 구축한 뒤, 검증 데이터를 이용하여 모델을 검증한 결과이다.

<표 6>에 사용된 용어를 간략하게 설명하면, '1_test'는 검증 데이터의 값으로 실제 대중교통 이용량이다(이하 실제 값). '2_pred'는 지도 학습 결과로 예측한 값이다(이하 예측 값). '3_sub'는 (실제 값 -예측 값)의 절대 값이다(이하 오차). '4_rate'는 (오차/실제 값X100)으로 오차율을 뜻한다.

<표 6> 지도 학습 결과

	1_test	2_pred	3_sub	4_rate
1	1297894	1741135	443240	25.45697
2	1742700	1708277	34424	1.975326
3	1017756	1349691	331935	24.59341
4	922260	948990.1	26730	2.816679
5	1671879	1663782	8098	0.484365
		. . .		
214	931660	931660	1	0.000107
215	999405	867250.3	132155	13.22337
216	1698054	1532882	165173	9.727194
217	1164737	1601316	436578	27.26372
218	781559	794623.6	13064	1.64405

본 논문에서 검증 데이터로 사용된 데이터는 총 218개로 15.49%의 낮은 평균 오차율(r)을 가진다.

4) 활용 방안 및 향후 연구 계획

특정한 달의 '기온', '강수량', '미세먼지' 데이터가 주어졌을 때, 월 대중교통 이용량을 예측할 수 있다. 이렇게 예측한 이용량 정보를 이용하면, 버스와 지하철 회사에서 탄력적으로 버스나 지하철을 배치할 수 있다. 예를 들어, 이용량이 증가하는 달이라고 예측 되면, 그 달에만 버스와 지하철을 추가적으로 배치하는 것이다. 따라서 본 논문 결과는 날씨에 따른 버스와 배차 간격 조절 등의 대중교통 배치 판단 결정에 기초자료로 사용될 것으로 기대된다.

하지만, 더 효율적인 추가 배치를 위해서는 본 논문을 더 보완해야한다. 왜냐하면 배치를 어느 시간대에 추가 배치할지 결정하는 것과 버스와 지하철의 추가 배치 비율을 측정하기 힘들기 때문이다. 따라서 시간대 별 이용량을 추가하고 대중교통으로 통합한 데이터가 아닌 버스와 지하철 각각의 이용량을 데이터에 추가한 뒤 학습시키면, 더 효율적인 결과 데이터를 얻게 된다.

또한 일 별 대중교통 이용량을 측정하고 싶을 경우에는 월 별 데이터 대신에 일 별 기상정보와 대중교통 이용량 정보로 바꾸어 학습 시키면 된다. 하지만 시간대 별 이용량 또는 일 별 기상 정보와 대중교통 이용량을 데이터에 추가시키면 데이터의 용량이 매우 커지므로 학습하는 시간이 많이 소요된다는 단점이 있다.

■ 결론 및 소감

처음 논문을 시작할 때, 지도 학습, 딥러닝에 대하여 강의를 들은 적이 있고, 데이터 또한 직접 크롤링을 해서 얻어내는 것이 아닌 제공되는 데이터 set을 이용하기 때문에 논문이 쉽게 진행될 것이라고 예상했다. 하지만, 이론적으로 배운 지도학습과 실제로 구현되는 지도 학습은 예상과 많이 달랐다.

가장 예상을 벗어난 점은 요소들의 상관관계를 분석하는 과정이었다. 강수량과 상대습도의 상관관계를 분석한 그래프를 보았을 때, 교재에서 보았던 선명한 양적 선형관계 그래프와 달라 두 데이터간의 상관관계가 미미하다고 생각했다. 하지만 분명 두 데이터간의 상관관계가 존재한다고 생각하여 나의 방법이 잘못 되었다고 생각했다. 따라서 원하는 결과를 얻어내기 위해 많은 검색과 노력을 해보았지만 실패하여 연구실 석사님께 도움을 요청하기까지 이르렀다. 그 결과 강의나 교재에서 배우는 데이터들의 양이 논문에서 진행하는 양보다 훨씬 많고, 대부분의 결과는 이상적인 결과를 그려낸 것이라는 것을 깨닫게 되었다.

또한, 논문에 이용하고자 하는 데이터 프레임과 제공받은 데이터 set의 프레임이 달라 데이터 처리 과정에서 많은 시간이 소요되었다. 특히 버스 원본 데이터는 크기도 매우 커서 한번 Python으로 실행하여 처리하는데 최소 6시간이 걸렸다. 처리 과정에서 데이터의 형식이 각 엑셀 sheet마다 달라서 python코딩이 제대로 읽지 못하는 경우도 생겼다. 그 결과 논문 진행 도중에도 오류를 발견하여 데이터를 계속 수정하는 과정을 겪게 되었다.

처음에는 새로운 코드를 짜는 것이 가장 어렵다고 생각하면서 논문을 진행했다. 하지만, 이론상으로 맞았다고 생각하는 코드의 결과가 잘 못 나왔을 때 오류 부분을 찾아서 수정하는 것이 더 어렵고 복잡한 과정이라는 것을 느끼게 되었다. 올바른 코드라고 생각하게 되는 나만의 잘못된 생각을 버리고 오류부분을 찾는 것은 생각보다 어려운 과정이었다. 좀 더 나의 코드를 객관적으로 보는 훈련이 부족해서 어려웠던 것 같다. 따라서 이 논문 뿐 아니라 앞으로 다른 프로젝트를 할 때, 코드를 짜는 경우 좀 더 많은 논리과정과 테스트 과정을 거쳐서 진행하고 객관적인 시선을 가지도록 노력할 것이다.

■ 참고문헌

[1] 공공데이터포털. (2017). http://www.data.go.kr/main.do (2017-8-20 방문)

[2] 서울특별시청. (2017). "교통카드 데이터로 본 '16년 서울 대중교통 이용현황" http://spp.seoul.go.kr/main/news/news_report.jsp#list/1 (2017-8-31 방문)

[3] 박상우. (2016). "직장인, 대학생 출퇴근·통학 이동수단1위 '대중교통'" 잡코리아 홈페이지, http://www.jobkorea.co.kr/GoodJob/News (2017-8-31 방문)

[4] 이광섭, 엄진기, 유소영, 민재홍, 양근율. (2014). "강우와 서울특별시 대중교통 승차인원과의 관계 분석" 한국철도학회 학술발표대회논문집, 252-257

[5] 박근영, 이시복. (2012). "강우 상태에 따른 대중교통 이용패턴 특성연구" 대한토목학회논문집 D, 32(1D), 23-31

[6] 박명근, 고강섭, 곽재덕, 이한열, 박성수 (2010). "과거 패턴자료와 날씨계수를 이용한 버스 통행시간 예측에 관한 연구" 대한교통학회 학술대회지, 63, 253-257

[7] 최상기, 이종호, 오승훈. (2013). "기상조건이 대중교통수요에 미치는 영향에 관한 연구" 대한토목학회논문집, 33(6), 2447-2453

[8] 서울특별시 열린 데이터 광장. (2017). http://data.seoul.go.kr/ (2017-7-10 방문)

[9] scikit-learn. (2017). http://scikit-learn.org/stable/ (2017-9-3 방문)