석사학위논문

# 딥러닝 알고리즘을 이용한 제주도 관광객 수 예측

# Forecasting the number of tourists in Jeju Island using Deep learning Algorithm

최 민 정

한양대학교 대학원

2017년 2월

석사학위논문

# 딥러닝 알고리즘을 이용한 제주도 관광객수 예측

# Forecasting the number of tourists in Jeju Island using Deep learning Algorithm

지도교수 차 경 준

이 논문을 이학 석사학위논문으로 제출합니다.

2017 년 2 월

한양대학교 대학원

응용통계학과

최 민 정

이 논문을 최민정의 석사학위 논문으로 인준함

2017 년 2 월

심 사 위 원 장 : 최 정 순 (인)

심 사 위 원 : 박 영 선 (인)

심 사 위 원 : 차 경 준 (인)

한양대학교 대학원

# 목 차

# 표 목차

# 그림 목차

# 국문요지

## 딥러닝 알고리즘을 이용한 제주도 관광객 수 예측

한양대학교 대학원

응용통계학과

최민정

지역 관광객 수의 예측은 해당 지역의 생산유발, 고용유발 등 경제적 효과에 큰 영향을 미친다. 특히, 관광산업이 전체 산업의 70%를 차지하는 제주도의 관광객 수를 정확하게 예측하는 것은 제주도 관광산업의 발전에 좋은 기초자료로 사용될 수 있는 중요한 연구라고 할 수 있다. 하지만 관광수요의 예측에 주로 사용되어 온 시계열 모형을 이용할 시 여러 정상성 등 여러 가정을 만족해야 한다는 한계점이 있다.

이에 본 논문에서는 최근 이미지 인식, 음성 인식, 자연어 처리 등 다양한 분야에서 우수한 성능을 보여주고 있는 딥러닝(Deep learning)을 사용하여 제주도를 방문하는 관광객 수를 여행 목적별로 예측한다. 또한 최근 많은 분야에서 연구되고 있는 인터넷 검색어 자료를 제주도 관광객 수의 예측변수로 사용하며 국내 최대 검색 점유율을 가지는 네이버의 검색어의 검색량을 사용한다. 딥러닝 알고리즘 중 심층신경망(Deep Neural Network; DNN)을 이용하고 심층신경망의 여러 단점들을 해결해주기 위해 Dropout정규화와 활성화 함수로는 ReLu함수를 심층신경망에 적용한다. 또한 딥러닝 알고리즘의 예측력을 확인하기 위하여 기존에 사용되던 계절형 ARIMA 모형과 비교한다. 분석에

사용한 자료는 2007년 1월부터 2016년 7월까지의 월별 자료이며 분석 결과 인터넷 검색어 변수를 이용한 딥러닝 모형이 계절형 ARIMA모형과 좋거나 비슷하다는 것을 알 수 있었다.

# 1. 서론

## 1.1 연구배경 및 목적

한국문화관광연구원의 '관광산업의 경제효과 분석'에 따르면 국내 관광산업은 생산 및 소득유발, 고용 및 취업유발 등 경제적 효과에 큰 영향을 미친다. 이에 국내의 각 지역 관광지에서는 관광객을 많이 유치하고자 관광지 홍보, 관광 상품 개발 등 관광정책수립에 힘을 쓰고 있다. 따라서 이러한 관광정책수립, 기업전략수립 등 여러 분야의 기초자료로 쓰일 수 있는 관광수요의 예측은 중요한 연구주제라 할 수 있다 (송다영, 2015). 그 중에서도 특히 관광산업이 전체 산업의 70%를 차지하는 제주도의 관광객 수는 매년 꾸준히 증가하고 있다. 고태호 등 (2011)은 제주도 방문 관광객의 소비지출이 제주지역 경제에 미치는 생산 유발효과, 부가가치 유발효과, 취업유발 효과는 모두 꾸준히 증가하는 추세라고 하였다. 따라서 본 논문에서는 제주도의 관광객 수를 여행목적별로 예측하고자 한다.

관광수요 예측방법은 크게 정성적 예측법(qualitative technique)과 정량적 예측법(quantitative technique)으로 구분할 수 있다. 정성적 예측법은 전문가의 주관적인 견해를 이용하는 방법으로 객관적이지 못하다는 비판을 받지만, 대규모의 큰 변화가 있어 과거의 시계열자료만으로는 예측할 수 없는 경우와 장기 예측을 하는 경우에 적합하다. 예측 방법으로는 델파이 예측법(Delphi thechniqures), 전문가 판단모형(judgment-aided models), 시나리오 설정법(senario writing methods) 등이 있다. 정량적 예측법은 계량적인 자료를 이용

하여 예측하는 방법으로 시계열 모형과 인과모형, 공간 상호작용모형으로 구분한다 (조광익, 1999).

최근 인터넷 검색어를 활용한 연구가 다양한 분야에서 시행되고 있다. 인터넷 검색어 자료는 실시간으로 자료가 축적되기 때문에 신속, 정확한 수요예측에 많이 쓰인다 (최재혁, 신창섭, 2015).

따라서 본 논문에서는 사람들이 여행을 준비하는 단계에서 인터넷 검색을 통해 여행지역정보 수집을 한다는 사실에 기반 해 인터넷 검색정보를 제주도 관광객 수를 예측하는 변수로 사용하고자 한다.

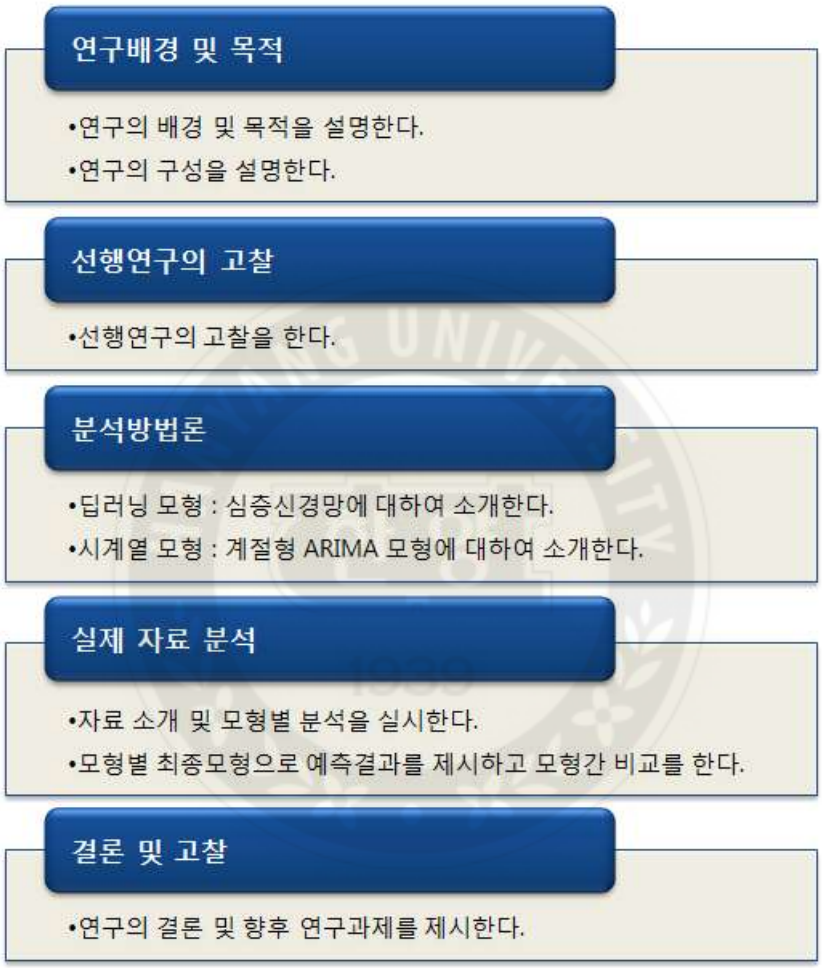
기존의 관광수요 예측에 관한 연구는 계절형 ARIMA 모형을 많이 사용해 왔다. 하지만 시계열 ARIMA 모형을 이용할 시 여러 가정들을 충족해야 된다는 한계점이 있다. 최근 기존의 인공신경망의 한계점들을 극복해 다양한 분야에서 우수한 결과를 보여주고 있는 딥러닝(Deep learning)방법이 대두되고 있다.

따라서 최종적으로 본 논문에서는 인터넷 검색어을 이용해 비모수적 방법인 딥러닝 알고리즘에 기반한 제주도 관광객 수 예측을 제안 하고 기존의 모수적 분석방법인 계절형 ARIMA 모형과 비교하고자 한다.

## 1.2 연구의 구성

본 논문은 총 5장으로 구성되어 있다. 1장에서는 연구의 배경 및 목적에 대하여 서술하고, 2장에서는 지역 관광객 수 예측에 관한 선행연구의 고찰을 한다. 3장에서는 딥러닝과 시계열 모형에 대한 이론적 배경을 살펴보고 4장에서는 실제 자료분석 결과를 서술하였다. 마지막 5장에서는 연구의 결론과 고찰을 서술한다.

본 논문의 흐름도는 다음 [그림 1.1]과 같다.




[그림 1.1] 본 논문의 흐름도

# 2. 선행연구의 고찰

관광수요의 예측은 관광산업의 발전에 큰 영향을 끼치는 만큼 다양한 방법을 이용하여 이루어져 왔다. 관광수요를 예측하기 위해 일반적으로는 계량적 기법을 활용한다 (김성태, 2014). 계량적 기법의 대표적인 방법으로는 시계열 모형과 인과모형이 있다. 김영우와 손은호 (2006)는 시계열 모형인 계절 ARIMA 모형을 이용해 경주 방문객의 수요예측에 관한 연구를 하였다. 이충기와 송학준 (2007)은 단계적 회귀모델, Winters 지수평활모델, ARIMA모델, ARIMA Invertion모델을 사용해 국내 1990년부터 2006년 까지 외래관광객 입국동향을 비교하였다. 그 결과 ARIMA Invention 모델이 가장 최적인 것으로 분석되었다.

비모수적 방법으로 관광수요를 예측한 방법으로는 Law와 Au (1999)가 인공신경망 모형으로 홍콩으로 여행하는 일본인 관광수요를 예측한 연구가 있으며, Plamer 등 (2006)도 인공신경망 모형으로 관광수요를 예측하였다. 하지만 아직까지 딥러닝 알고리즘으로 관광수요를 예측한 연구사례는 드물다. Kuremoto 등 (2013)은 딥러닝 알고리즘 중 Deep belief network(DBN)를 이용해 시계열 예측을 하고 다층퍼셉트론(Multi layer perceptron; MLP)과 비교하였다. 그 결과 DBN이 더 우수함을 증명하였다.

한편, 인터넷 검색어를 활용한 선행연구도 활발히 진행되고 있다. Anvik과 Gjelstad (2010)은 노르웨이 정부의 실업률 자료와 구글 트렌드의 실업 관련 인터넷 검색어 자료를 활용하여 실업률을 조기에 예측하는 모형을 연구하였다. 또한 Mccallum와 Bury (2013)는 구글인사이트를 이용해 환경에 대한 대중들의 관심도 변화에 관한 연구를 진행하였다. 국내에서 권치명 등 (2015)이

국내 실업률 예측에 국내 포털사이트인 네이버에서 얻은 검색어 자료를 시계열 ARIMA모형에 포함시켜 인터넷 검색어 자료를 포함하지 않은 ARIMA 모형보다 더 나은 예측력을 나타내는 것을 보였다. 또한 최재혁 등 (2015)은 네이버에서 얻은 검색어 자료를 활용해 휴양림 이용객 현황과 인터넷 검색어의 상관관계 분석을 실시하여 인터넷 검색어와 빅데이터 간에 밀접한 상관관계가 있음을 나타내었다.

# 3. 분석방법론

3장에서는 본 논문에서 사용할 분석 방법론에 대하여 알아본다. 3.1장에서 딥러닝 모형에 대하여 기술하고 3.2장에서는 시계열 모형에 대해 기술하였다.

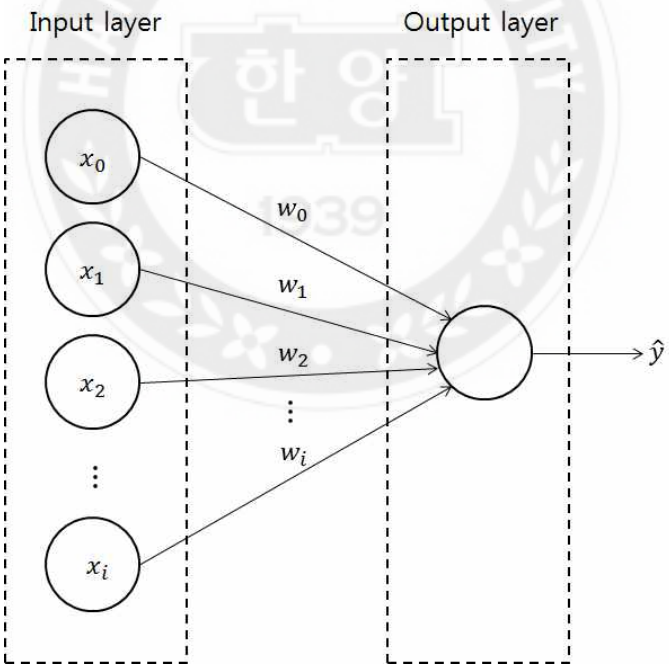
## 3.1 딥러닝(Deep Learning)

딥러닝(Deep Learning)은 기존의 인공신경망 구조에서 은닉층이 많아질 때 사용하는 알고리즘을 말한다. 딥러닝에는 다양한 알고리즘이 있지만, 대부분의 경우 대표적인 몇 가지 구조들에서 파생된 것이다. 대표적인 알고리즘으로는 심층신경망(Deep Neural Network; DNN), 제한된 볼츠만 머신(Rescricted Boltzmann Machine; RBM), 합성곱신경망(Covolutional Neural Network; CNN), 순환신경망(Recurrent Neural Network; RNN), 심층신뢰신경망(Deep Belief Network; DBN), 심층 Q-네트워크(Deep Q-Networks) 등이 있다. 또한 Srivastava (2014)는 Dropout 정규화 알고리즘을 연구하면서 과적합(Overfitting) 문제를 해결할 수 있는 방법을 제안했다.

### 3.1.1 심층신경망(Deep Neural Network)

심층신경망(Deep Neural Network; DNN)은 입력층(input layer)과 출력층(output layer) 사이에 여러개의 은닉층(hidden layer)들로 이루어진 인공신경망이다. 인공신경망은 인간의 뇌구조인 신경망과 비슷한 구조를 가지는 통계학적 학습 알고리즘이다. 심층신경망과 인공신경망의 가장 큰 차이점은 층의 개수(신경망의 깊이)이다. 보통 2개 이상의 은닉층을 갖는 경우에 심층신경망이라고 한다.

인공신경망 중 가장 간단한 구조로는 퍼셉트론(perceptron)이 있다.




[그림 2.1] 퍼셉트론의 구조도

[그림 3.1]의 퍼셉트론은 입력층과 출력층으로 구성되어 있다. 입력층에서는 데이터를 벡터 형식으로 입력노드(input node)에 입력한다. 가중치 $w_i$는 입력된 데이터 $x_i$와 출력층의 출력값 $\hat{y}$을 최적으로 일치하게 만들어주는 링크로서, 훈련데이터를 이용해 모형을 적응시킨다 (Tan 등, 2007). 식 (3.1.1)은 위의 [그림 3.1]을 수학적인 식으로 표현한 것이다.

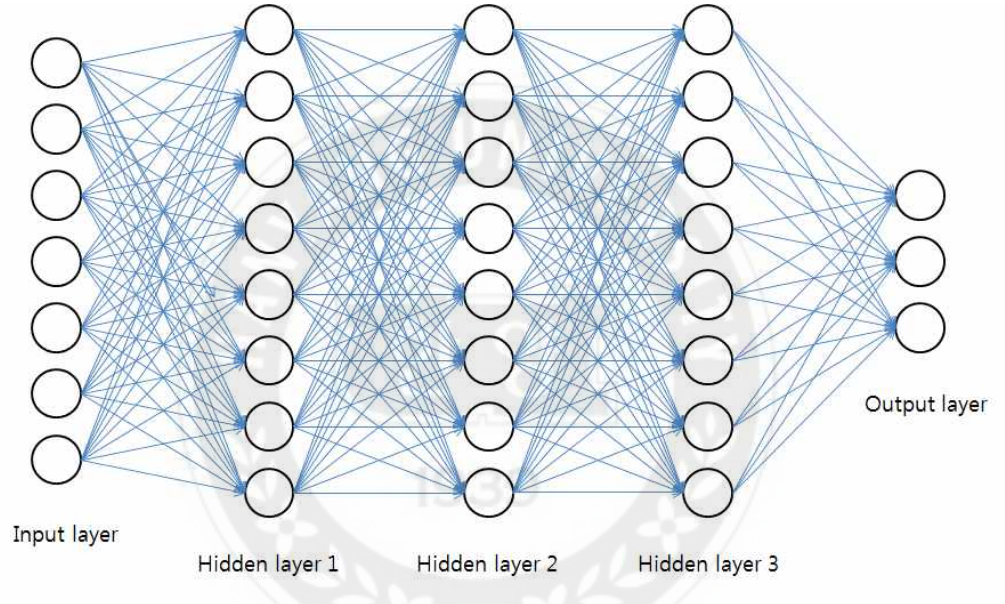
$$\hat{y} = f(\sum_{i}^{n} w_i x_i + w_0 x_0), \tag{3.1.1}$$

여기서 $n$은 입력벡터의 크기이고 $x_0$는 바이어스 입력값, $w_0$는 바이어스 기울기, $f$는 활성화 함수(activation function)이다. 하지만 Minsky와 Papert (1969)는 단층 퍼셉트론은 선형 분리만 가능한 알고리즘으로 XOR같은 비선형 문제는 해결 할 수 없다는 것을 지적했다. 이러한 문제를 해결하기 위해 Werbos (1974)는 역전파(Backpropagation) 알고리즘과 입력층과 출력층 사이에 은닉층을 추가하는 방법을 제안하였다.

역전파 알고리즘은 인공신경망을 학습시키는 방법이다. 역전파 알고리즘은 전파단계와 가중치 수정단계로 구성된다. 전파 단계는 정해진 가중치를 이용해 훈련 데이터로부터 출력값을 출력하고, 목적값과 출력값의 차이인 오차를 계산하여 각 층에 전달하는 단계이고 가중치 수정단계에서는 전파된 오차를 이용하여 가중치를 수정한다. 즉,

$$\delta = \frac{1}{2}(\hat{y} - y)^2, \tag{3.1.2}$$
$$w(t+1) = w(t) - \alpha \frac{\partial \delta}{\partial w(t)}.$$

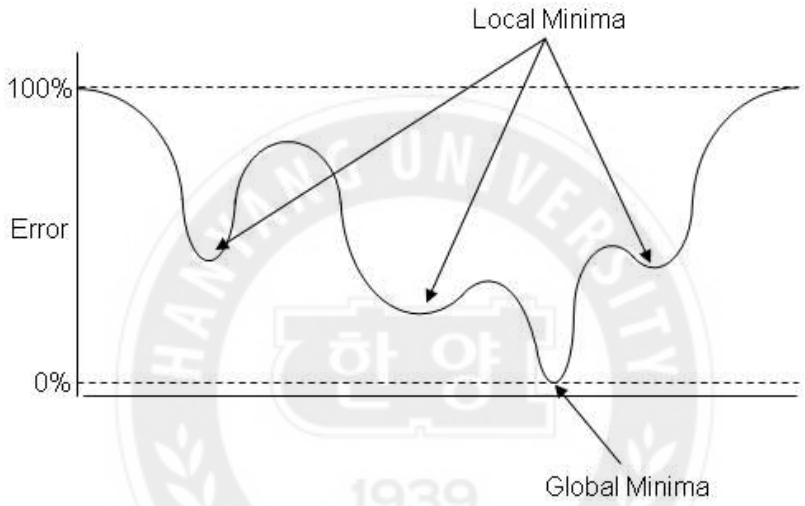
식 (3.1.2)에서 $y$는 실제 목적값이고 $\hat{y}$는 출력값으로 이 오차값 $\delta$을 측정한다. 그 후 $\delta$가 최소화 되도록 출력층으로부터 입력층 방향으로 되돌아가면서 가중치를 수정한다. $w(t)$는 $t$단계에서의 신경망 가중치 값을 말하고 $w(t+1)$은 갱신 후의 가중치 값이다. 이와 같은 방법으로 가중치를 갱신하여 가중치가 거의 일정하거나, 오차함수가 최소화 될 때까지 반복을 한다 (이재성, 2016).




[그림 3.2] 심층신경망의 구조도

현대에는 데이터가 많아지고 복잡해지면서 인공신경망의 층을 깊게 쌓아야 한다. 하지만 [그림 3.2]와 같은 은닉층의 수가 3개 이상인 심층신경망은 층이 깊어지면서 여러 가지 문제점이 발생한다. 첫 번째 문제점은 오차가 [그림 3.3]과 같이 지역 최소값(Local minima)에 빠져 전역 최소값(Global minima)에 도달하지 못해 학습 시간이 오래 걸린다는 것이다(http://mnemstudio.org). 둘째는, 과적합 문제가 있다. 과적합은 모형이 훈련데이터에 너무 가깝게 맞추

어져서 시험데이터의 예측력은 떨어지는 것을 말한다. 셋째는, 초기값을 어떻게 설정하느냐에 따라 수렴이 되지 않고 진동 또는 발산하는 문제가 있다. 넷째, 기울기 변화량이 매우 작아 신경망을 효과적으로 학습시키지 못하고 오차가 최소화 되지 못한 채 수렴해 버려 가중치가 갱신되지 않는 'Vanishing Gradient Problem'이 발생한다 (이재성, 2016).

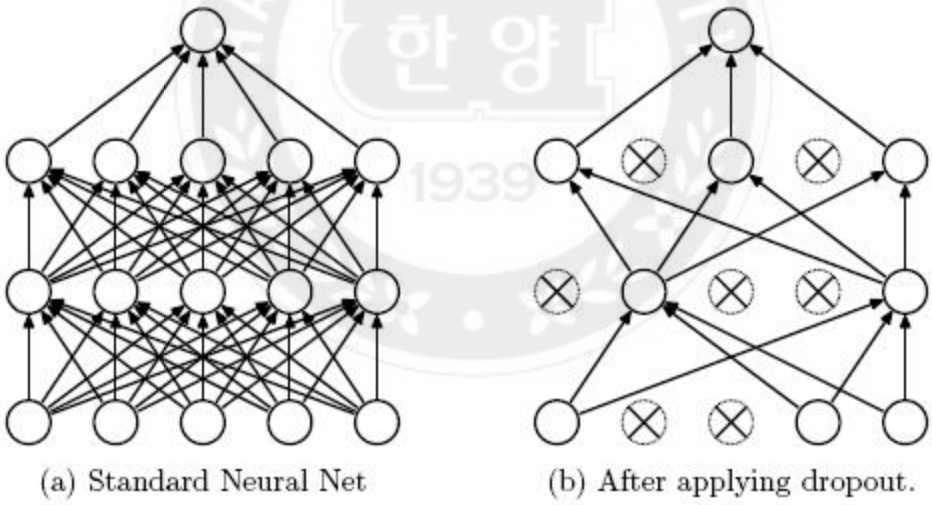



[그림 3.3] 전역 최소값과 국소 최소값

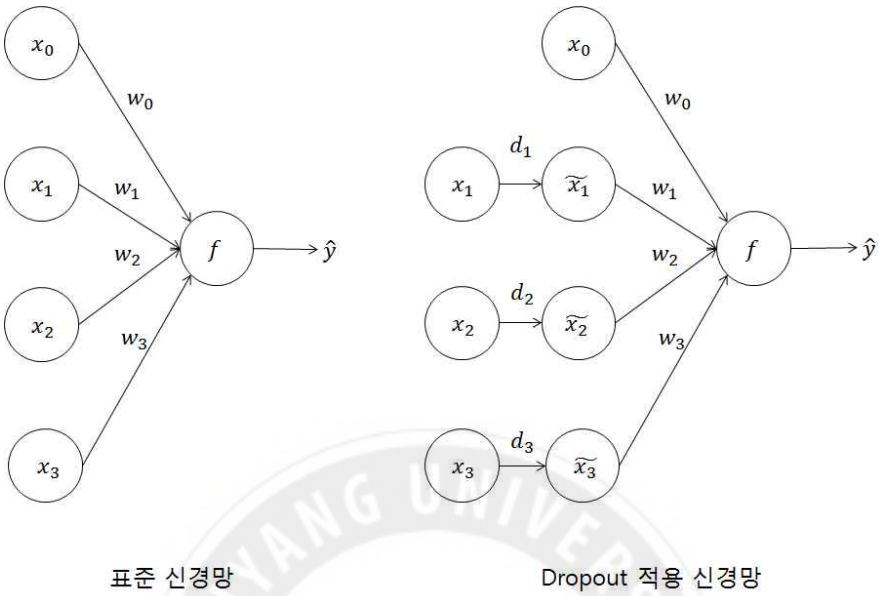
이러한 심층신경망의 문제점을 해결하기 여러 방법들이 연구되고 있다. Hinton 등 (2006)은 제한된 볼츠만 머신을 이용해 데이터를 사전학습(pre-training)시키고 심층신뢰신경망를 학습시키는 방법을 고안해 내기도 했지만 최근 Dahl 등 (2013)은 기존의 심층신경망에 학습과정에서 일부 노드 사이의 연결을 무작위로 끊음으로써 과적합 문제를 해결할 수 있는 Dropout 정규화를 제안하였다. 또한 활성함수로 Rectified Linear Function을 사용하여 학습시간을 줄이고 'Vanishing Gradient Problem'도 해결할 수 있게 되었다 (Nair 등, 2010).

### 3.1.2 Dropout 정규화

Droptout 정규화는 과적합의 문제를 해결하기 위한 기법중 하나이다. [그림 3.4]의 (a)와 같이 2개의 은닉층을 가지는 신경망이 있을 때, Dropout 정규화를 적용하면 [그림 3.4]의 (b)와 같이 일부의 노드를 무작위로 선택하여 사용한다. 즉, 신경망을 학습할 때 은닉층의 모든 노드를 사용하는 것이 아니라 무작위로 선택된 노드만을 사용하는 것이다. 생략된 노드는 학습에 영향을 끼치지 않고, 일반적으로 50~80%정도의 노드를 사용한다 (Srivastava 등, 2014).




[그림 3.4] (a) 2개의 은닉층을 가지는 신경망과
(b) 신경망에 Dropout 적용 후




[그림 3.5] 표준 신경망과 Droupout 적용 신경망

[그림 3.5]와 같이 신경망에 Dropout 정규화를 적용한다는 것은 표준신경망의 노드에 베르누이분포를 따르는 변수를 무작위로 곱해주는 것으로 생각할 수 있다. 여기서 베르누이 랜덤변수는 노드가 존재할 확률이 $p$, 평균이 $p$, 분산이 $p(1-p)$인 변수이다. 입력노드에 이 랜덤변수를 곱해주면 결과적으로 랜덤변수의 값에 따라 노드가 줄어들게 된다. 식 (3.1.3)은 Dropout 정규화의 적용을 도식화 한 것이다.

$$\begin{aligned} d_i &\sim Bernoulli(p), \\ \widetilde{x_i} &= d_i \times x_i, \\ \hat{y} &= f(w_i \widetilde{x_i} + w_0 x_0), \end{aligned} \tag{3.1.3}$$

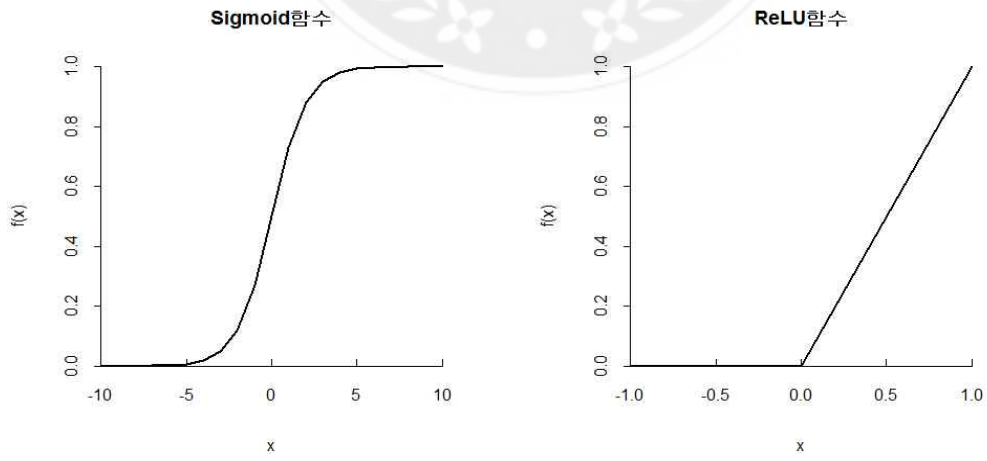
여기서 $d_i$는 베르누이분포를 따르는 변수이고, $x_i$는 입력값, $\widetilde{x_i}$는 Droupout 적용 후 노드, $w_i$는 가중치, $f$는 활성함수, $\hat{y}$는 출력값을 말한다.

### 3.1.3 Rectified Linear Unit(ReLU) 활성함수

기존에 자주 사용되는 활성함수로 [그림 3.6]의 Sigmoid함수가 있다. 하지만 Sigmoid함수는 극한으로 갈수록 기울기가 작아져서 가중치가 갱신되지 않는 'Vanishing Gradient Problem'이 발생한다. 즉, 기울기하강을 여러층으로 해나갈 때 마다 오차가 소멸되는 것이다. 이러한 문제점을 해결하기 위해 Nair 등(2010)은 sigmoid함수 대신 [그림 3.6]의 Rectified Linear Unit(ReLU) 활성함수를 제안했다.

$$f(x)=\begin{cases}0\,, & x<0\\ x\,, & x\geq 0\end{cases}. \tag{3.1.4}$$

ReLU 활성함수는 식 (3.1.4)처럼 0보다 작을 때는 0을 사용하고, 0보다 크거나 같은 값에 대해서는 해당 값을 그대로 사용한다. ReLU 활성함수는 기울기가 0또는 1로 학습이 되기 때문에 오차가 100%로 전파된다.




[그림 3.6] Sigmoid함수와 ReLU함수

## 3.2 계절형 ARIMA 모형

대부분의 시계열 자료들은 시간의 흐름에 따라 추세를 가지고 증가하거나, 분산이 커지는 등 정상이 아닌 경우가 많다. 이러한 자료를 비정상시계열이라고 하는데, 비정상시계열을 분석하기 위해서는 먼저 정상시계열로 변환해 주어야 한다. 정상시계열로 변환하기 위해서는 자료가 나타내는 특성에 따라 로그변환이나 차분 및 계절차분 등을 실시한다.

일정한 계절적인 주기를 가지고 변하는 시계열 모형의 분석은 삼각함수, 지시함수를 이용한 회귀모형 또는 Winters의 계절형 지수평활법도 있지만 이러한 방법들은 계절형 시계열 자료의 성분들이 서로 독립일 경우에 사용이 가능하다. 하지만 우리가 일반적으로 분석해야 하는 시계열 자료는 그 성분들이 확률적이거나 다른 성분들과 상관이 있는 경우로, 주로 자기회귀(Autoregressive)와 이동평균(Moving Average)과 차분과정이 복합된 ARIMA 모형을 이용해 분석한다 (조신섭, 손영숙, 2009).

먼저 시계열자료 $Y_t$가 차수 $p$를 가지는 자기회귀(Autoregressive; AR)모형 AR($p$)를 따른다면 그 형태는 식 (3.2.1)과 같다.

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \\ &= (1 + \phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p) Y_t + \varepsilon_t, \end{aligned} \qquad (3.2.1)$$

여기서 $Y_t$는 시계열자료이고 $\phi_1, \phi_2, \cdots, \phi_p$는 자기회귀모형의 모수, $\varepsilon_t$는 평균이 0이고, 분산이 $\sigma_\varepsilon^2$인 백색잡음(white noise)을 따르는 오차항이다. 또한 0보다 큰 정수 $i$에 대해 $Y_{t-i}$는 $B^i Y_t$라고 표현 한다. 식 (3.2.1)에 보이는 것과 같이 자기회귀모형은 종속변수인 $Y_t$가 $Y_t$와 시차가 각각 $1, 2, \cdots, p$인 변수들로

이루어져 있다.

만약 $Y_t$ 가 차수 $q$를 가지는 이동평균(Moving Average; MA)모형 MA($q$)를 따른다면 그 형태는 식 (3.2.2)와 같다.

$$\begin{aligned} Y_t &= \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \\ &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\varepsilon_t, \end{aligned} \tag{3.2.2}$$

여기서 $\theta_1, \theta_2, \cdots, \theta_q$는 이동평균모형의 모수이다. 이동평균모형은 과거 $q$개의 오차항들과 $t$시점에서의 오차항으로 이루어진 모형이다.

일반적인 시계열자료는 자기회귀모형 또는 이동평균모형만으로는 추정하기가 힘들다. 따라서 식 (3.2.3)과 같이 자기회귀모형과 이동평균모형의 특성을 동시에 포함하는 자기회귀이동평균(autoregressive moving average; ARMA) 모형으로 추정할 수 있다.

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} \\ &\quad + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}. \end{aligned} \tag{3.2.3}$$

이러한 ARMA 모형이 비정상시계열 일 때, ARMA 모형에 차분을 한 모형을 ARIMA 모형이라고 한다. 식 (3.2.4) 차분연산자 $\nabla$에 대한 식이다.

$$\begin{aligned} \nabla Y_t &= (1 - B) Y_t = Y_t - Y_{t-1}, \\ \nabla^2 Y_t &= (1 - B)^2 Y_t = (1 - 2B + B^2) Y_t \\ &= Y_t - 2 Y_{t-1} + Y_{t-2}. \end{aligned} \tag{3.2.4}$$

즉, 차분은 원래의 자료에 전 시점의 자료를 빼준 것을 의미한다. 이러한 차분은 정상시계열이 될 때 까지 한다.

시계열자료의 특성은 현재의 상태가 과거 및 미래의 상태와 밀접한 관계를 가지고 있다는 것이다. 즉, 시간의 흐름에 따라 시계열자료들이 서로 독립이 아니며 이러한 경우 시계열자료는 자기상관관계를 가진다고 한다. 여기서 ACF는 시차에 따른 상관정도를 나타내기 위해 식 (3.2.5)의 자기공분산함수(autocovariance function)나 식 (3.2.6)의 자기상관계수(autocorrelation function)를 의미한다.

$$\gamma_k = Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)], \tag{3.2.5}$$

$$\rho_k = Corr(Y_t, Y_{t+k}) = \frac{Cov(Y_t, Y_{t+k})}{\sqrt{Var(Y_t)\,Var(Y_{t+k})}}, \tag{3.2.6}$$

여기서 $Y_t$는 t시점에서의 시계열 자료이고 $k$는 시차, $Cov(Y_t, Y_{t+k})$는 $Y_t$와 $Y_t$사이의 공분산함수를 나타낸다.

$$\begin{aligned} \gamma_0 &= Cov(Y_t, Y_t) = E[(Y_t - \mu)(Y_t - \mu)] \\ &= E[(Y_t - \mu)^2] = Var(Y_t). \end{aligned} \tag{3.2.7}$$

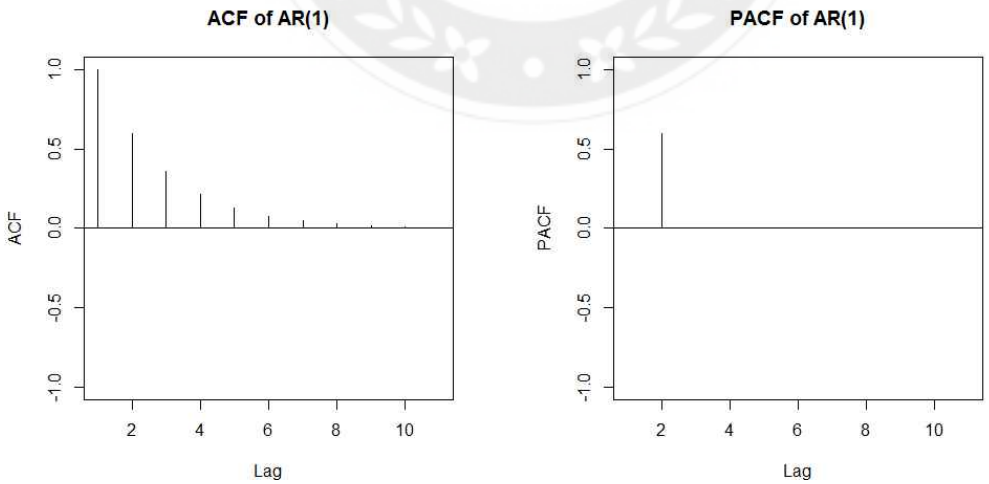
특히 위의 식 (3.2.7)에서와 같이 시차가 0일 때 자기공분산함수는 $Y_t$에서의 분산과 같으므로 다음 식 (3.2.8)과 같은 관계가 있음을 알 수 있다.

$$\rho_k = \frac{\gamma_k}{\gamma_0}. \tag{3.2.8}$$

PACF는 부분자기상관함수(partial autocovariance function)로 ACF와 같이 $Y_t$와 $Y_{t+k}$의 관련성을 판단할 때 유용하게 사용되는 통계량이다. PACF는 $Y_t, Y_{t+1}, \cdots, Y_{t+k-1}, Y_{t+k}$가 관측되었을 때 $Y_{t+1}, \cdots, Y_{t+k-1}$의 효과를 배제하고 $k$시차만큼 떨어진 $Y_t$와 $Y_{t+k}$만의 순수한 상관관계를 나타낸다. $Z$를 시간이라고 하고 $Z$의 효과를 배제한 후의 $X$와 $Y$사이의 부분상관계수는 다음 식 (3.2.9)와 같다.

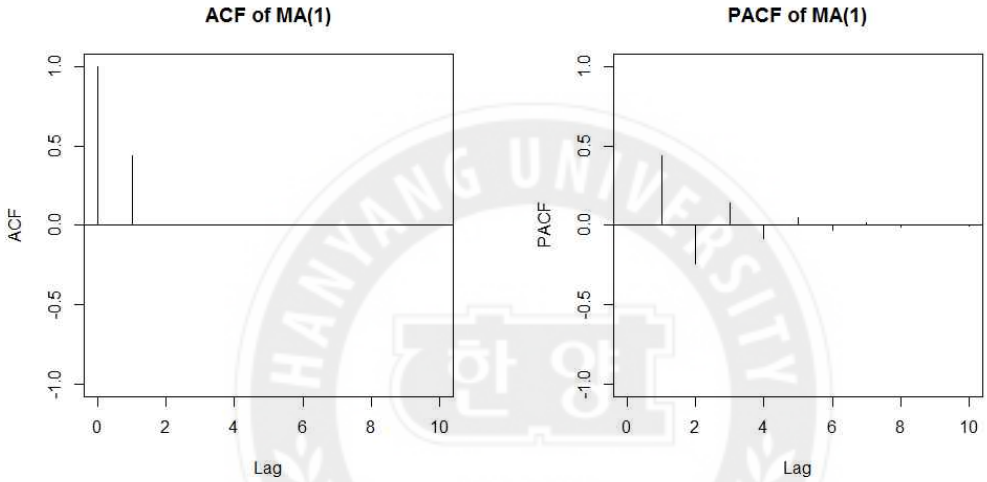
$$\rho_{XY,Z} = \frac{E\{[X-E(X|Z)][Y-E(Y|Z)]\}}{\sqrt{[X-E(X|Z)]^2[Y-E(Y|Z)]^2}}. \qquad (3.2.9)$$

AR($p$) 모형일 때, ACF는 시차가 커질수록 지수적으로 감소하고 PACF는 시차 $p$이후에는 0에 가까운 값들을 가진다. [그림 3.7]은 AR(1) 모형의 ACF와 PACF의 이론적인 형태이다. ACF는 지수적으로 감소를 하고 PACF는 시차 1까지만 값을 갖고 그 이후에는 0인 것을 볼 수 있다.




[그림 3.7] AR(1) 모형의 ACF와 PACF의 이론적인 형태

또한 MA($q$) 모형일 때는 AR($q$) 모형과는 반대로 ACF가 $q$시점 이후에 0이고 PACF는 시차가 커질수록 지수적으로 감소한다. 예를 들어, MA(1) 모형의 ACF와 PACF는 [그림 3.8]과 같다. ACF는 지수적으로 감소하고 있고 PACF는 시차 1 이후로 값이 존재하지 않는 것을 확인 할 수 있다.




[그림 3.8] MA(1) 모형의 ACF와 PACF의 이론적인 형태

따라서, ACF와 PACF는 모형의 차수를 식별하는데 사용되는 유용한 도구이다 (조신섭, 손영숙, 2009).

### 3.2.1 승법계절모형

ARIMA 모형에 계절성을 함께 고려한 모형을 계절형 ARIMA 모형 이라고 한다. 계절형 ARIMA 모형은 Box와 Jenkins (1970)에 의해 제안된 모형으로 계절적 주기를 갖는 시계열을 모형화하기 위하여 가장 널리 사용되고 있다. 시계열이 순수하게 계절주기 $s$를 갖는 경우 식 (3.2.10)과 같은 모형을 고려해 볼 수 있다.

$$\begin{aligned} \Phi(B^s)(1-B^s)^D Y_t &= \delta + \Theta(B^s)\varepsilon_t, \\ \Phi(B^s) &= (1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}), \\ \Theta(B^s) &= (1-\Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}), \end{aligned} \tag{3.2.10}$$

여기서 $\Phi_1, \Phi_2, \cdots, \Phi_P$와 $\Theta_1, \Theta_2, \cdots, \Theta_Q$는 모수이고 $s$는 일정한 계절주기, $\delta$는 모형의 평균, $D$는 계절 차분의 수이다.

승법계절 ARIMA 모형이란 시계열자료 $\varepsilon_t$가 백색잡음이고 식 (3.2.11)과 같이 형태를 갖는 모형을 말한다.

$$\phi(B)\Phi(B^s)(1-B^s)^D(1-B)^d Y_t = \delta + \theta(B)\Theta(B^s)\varepsilon_t, \tag{3.2.11}$$

여기서 $d$는 일반 차분의 수이다. 이 모형을 간단하게 승법계절모형이라고 하고 $Y_t \sim \text{ARIMA}(p,d,q)(P,D,Q)_s$라고 나타낸다.

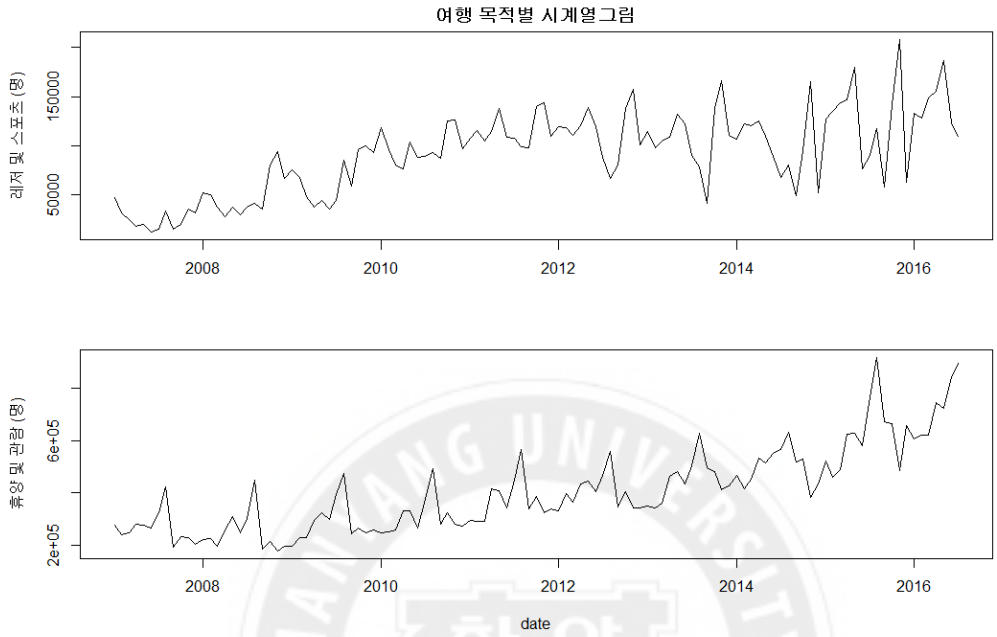
# 4.실제 자료 분석

4장에서는 실제 자료에 대한 설명과 실제 자료 분석을 수행한 결과를 비교하였다. 4.1장에서는 자료 및 검색어 변수 선정에 대해 설명하고 4.2장에서는 모형의 결과 및 두 모형 간 비교를 한다.

## 4.1 자료 설명 및 변수 선정

제주특별자치도관광협회는 일별, 월별 입도관광객 수 자료를 제공하고 있다. 월별 자료는 여행 형태별, 여행 목적별로 분류해 제공하고 있다.

본 논문에서 분석 및 예측에 사용된 자료는 2007년 1월부터 2016년 7월까지 여행 목적별 제주도 입도관광객 수 중 레저 및 스포츠, 휴양 및 관람의 월 단위 자료를 이용하였다. 외국인 관광객의 경우에는 국적별로 구분해 제공하고 있지만, 본 논문에서 예측변수로 사용되는 인터넷 검색어 변수는 국내 포털 사이트의 자료이기 때문에  외국인 관광객의 수와는 관련이 없어 외국인 관광객의 수는 제외하도록 한다. [그림 4.1]은 2007년 1월부터 2016년 7월까지 여행 목적별 시계열 그림이다.




[그림 4.1] 여행 목적별 시계열 그림 (2007년 1월~2016년 7월)

### 4.1.1 검색어 선정

제주도 여행과 관련된 인터넷 검색어를 선정하기 위해서 권치명 등 (2015)이 사용한 방법과 같이 SNS와 블로그 자료에서 '제주도 여행'과 '제주도 레저'의 연관어를 찾아냈으며, 추가적으로 네이버에서 제공하는 '제주도 여행'과 '제주도 레저'의 연관검색어 중 검색량이 많은 상위의 검색어들로 선정하였다. 국내 여러 포털 사이트 중 네이버가 가장 점유율이 높은 검색엔진으로 2015년 국내 포털 3사의 검색 점유율 기준으로 네이버의 검색 점유율은 87.2%이다(윤영철 등, 2015). 따라서 본 분석에서는 네이버 데이터 랩(Data Lab)에서 제공하는 검색량을 변수로 사용하였다. 네이버 데이터 랩에서 제공하는 검색량은 일별 실제 검색량이 아닌, 검색하는 기간 중 최대 검색량을 100으로 환산해 나머지를 그에 맞춰 환산한 상대적인 지수(Index)이다. 이 중 검색량이 0에 가깝거나 제주도 관광객 수와 상관성이 낮은 검색어는 제외하였다. 최종 선정된 검색어는 다음 [표 4.1]와 같다.

[표 4.1] 여행 목적 별 인터넷 검색어

| 여행목적 | 검색어 |
|---|---|
| 휴양 및 관람 (n=6) | 제주도 항공권 |
| | 제주도 지도 |
| | 제주도 호텔 |
| | 제주도 바다 |
| | 제주도 가볼만한 곳 |
| | 제주도 렌트 |
| 레저 및 스포츠 (n=3) | 제주도 항공권 |
| | 제주도 잠수함 |
| | 제주도 낚시 |

인터넷 검색어 자료는 주 단위로 나타나 있다. 따라서 월 단위로 변경하기 위해 Anvik 와 Gjelstad (2010)이 사용한 가중치 부여 방식을 활용 하였다. 예를 들어 [표 4.2]는 검색어 '낚시'에 대한 네이버 검색 지수의 일부이다. 2015년 6월 29일부터 7월 5일까지의 주 중에 6월의 일수는 2일이고, 6월의 월 수는 30일 이기 때문에 2를 30으로 나누어 그 주에 해당하는 70이라는 검색 지수에 이 값을 곱한다. 그 후, 월별로 나온 값들을 다 더해준 값([표 4.2]의 Transformed value)을 월별 검색량 지수로 사용한다.

[표 4.2] 가중치 부여 방식에 대한 예(2015년 6월～2015년 7월)

| Month | Period | Days | $a$ | $b$ | $a \times b$ | Transformed value |
|---|---|---|---|---|---|---|
| | | | Weighted rate | Keyword value | | |
| 2015년 6월 (30) | 06/01～06/07 | 7 | 7/30 | 52 | 12.13 | 12.13+13.77 +14.93 +15.63+4.67 =61.13 |
| | 06/08～06/14 | 7 | 7/30 | 59 | 13.77 | |
| | 06/15～06/21 | 7 | 7/30 | 64 | 14.93 | |
| | 06/22～06/28 | 7 | 7/30 | 67 | 15.63 | |
| | 06/29～07/05 | 2 | 2/30 | 70 | 4.67 | |
| 2015년 7월 (31) | 06/29～07/05 | 5 | 5/31 | 70 | 11.29 | 11.29+16.71 +14.90+16.26 +14.23 =73.39 |
| | 07/06～07/12 | 7 | 7/31 | 74 | 16.71 | |
| | 07/13～07/19 | 7 | 7/31 | 66 | 14.90 | |
| | 07/20～07/26 | 7 | 7/31 | 72 | 16.26 | |
| | 07/27～08/02 | 5 | 5/31 | 63 | 14.23 | |
| 2015년 8월 | 07/27～08/02 | 2 | 2/30 | 63 | 4.20 | ⋮ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |

## 4.2 모형의 결과

본 절에서는 딥러닝 모형과 계절형 ARIMA 모형으로 제주도 입도객수의 예측을 하고 모형의 결과를 비교하였다. 본 논문에 쓰인 자료는 2007년 1월부터 2016년 7월까지 월별 자료이다. 이중 2007년 1월부터 2014년 12월 까지 자료를 훈련데이터로 사용하고 2015넌 1월부터 2016년 7월까지 데이터를 시험데이터로 사용했다.

딥러닝 모형의 은닉층 개수와 계절형 ARIMA 모형의 차수에 따라 어느 모형이 더 좋은지 결과가 달라질 수 있기 때문에 각 모형 내의 최적 모형을 찾고 최종적으로 모형 간 비교를 하였다.

본 연구의 분석 프로그램으로는 R 3.1.3을 이용하였으며 'h2o(http://www.h2o.ai)', 'tseries(https://CRAN.R-project.org/package=tseries)', 'forecast(http://github.com/robjhyndman/forecast)' 패키지를 이용하여 분석을 진행하였다.

### 4.2.1 딥러닝 모형의 결과

딥러닝 모형에서의 최적모형을 찾기 위해서 여러 조합의 은닉층의 수, 노드의 수, Dropout의 비율을 변화시키며 시행하였다. 또한 입력층에도 Dropout을 적용할 시 은닉층에만 Dropout을 적용한 것 보다 더 예측력이 높다고 알려져 있다 (Hinton 등, 2012). 본 연구에서도 은닉층에 Dropout을 적용한 것은 물론 입력층에도 Dropout을 적용한 것과 적용하지 않은 것을 비교해 보았다. 모형 비교는 식 (4.2.1)의 RMSE(Root mean square error)로 하였다.

$$RMSE = \sqrt{MSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}}, \qquad (4.2.1)$$

여기서 $n$은 데이터의 개수, $y$는 실제값이고 $\hat{y}$는 모형의 훈련을 통해 나온 추정값이다.

여행 목적별 딥러닝 모형의 결과는 [표 4.3]과 [표 4.4]에 나타내었으며, [표 4.3]의 레저 및 스포츠에서는 각 50개씩 노드를 갖는 3개의 은닉층으로 이루어진 심층신경망에 입력층 Dropout 비율을 0, 은닉층 Dropout 비율을 0.5씩으로 정하고, 활성함수를 사용해 훈련 반복시행횟수(epochs)를 200번 진행하였을 때 모형의 RMSE가 29,796.5로 가장 낮아 최종모형으로 선택하였다. 본래 입력층에 Dropout을 적용한 것이 더 예측력이 높다고 알려져 있지만 이 자료의 경우 변수가 3개로 많지 않아 입력층에 Dropout을 적용하는 것이 의미가 없다는 것으로 나타난다.

[표 4.4]의 휴양 및 관람에서는 각 100개씩 노드를 갖는 3개의 은닉층으로 이루어진 심층신경망에 입력층 Dropout 비율을 0.2, 은닉층의 Dropout 비율을

0.5씩으로 정하고 반복시행횟수를 200번 진행한 것이 모형의 RMSE가 50,528.72로 가장 낮게 나와 최종 딥러닝 모형으로 선정하였다. [표 4.5]는 최종 딥러닝 모형 시험데이터를 예측해 RMSE와 MAE를 나타낸 표이다. MAE(Mean absolute error)는 예측력을 측정하는 또 다른 측도로서 다음 식 (4.2.2)과 같다.

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y_i} - y_i|}{n}. \tag{4.2.2}$$

[표 4.3] 딥러닝 모형의 RMSE : 레저 및 스포츠

| 레저 및 스포츠 | | | | |
|---|---|---|---|---|
| hidden layer<br>Dropout_input<br>Dropout_hidden | 반복시행횟수 | | | |
| | 50 | 100 | 150 | 200 |
| 50-50-50<br>0.2<br>0.5-0.5-0.5 | 34,978.03 | 33,368.95 | 37,295.62 | 34,616.37 |
| 100-100-100<br>0.2<br>0.5-0.5-0.5 | 35,221.18 | 41,336.89 | 34,099.66 | 34,186.72 |
| 50-50-50<br>0<br>0.5-0.5-0.5 | 32,767.78 | 31,384.58 | 36,400.36 | 29,503.69 |
| 100-100-100<br>0<br>0.5-0.5-0.5 | 35,320.88 | 35,088.11 | 32,525.29 | 31,580.25 |

[표 4.4] 딥러닝 모형의 RMSE : 휴양 및 관람

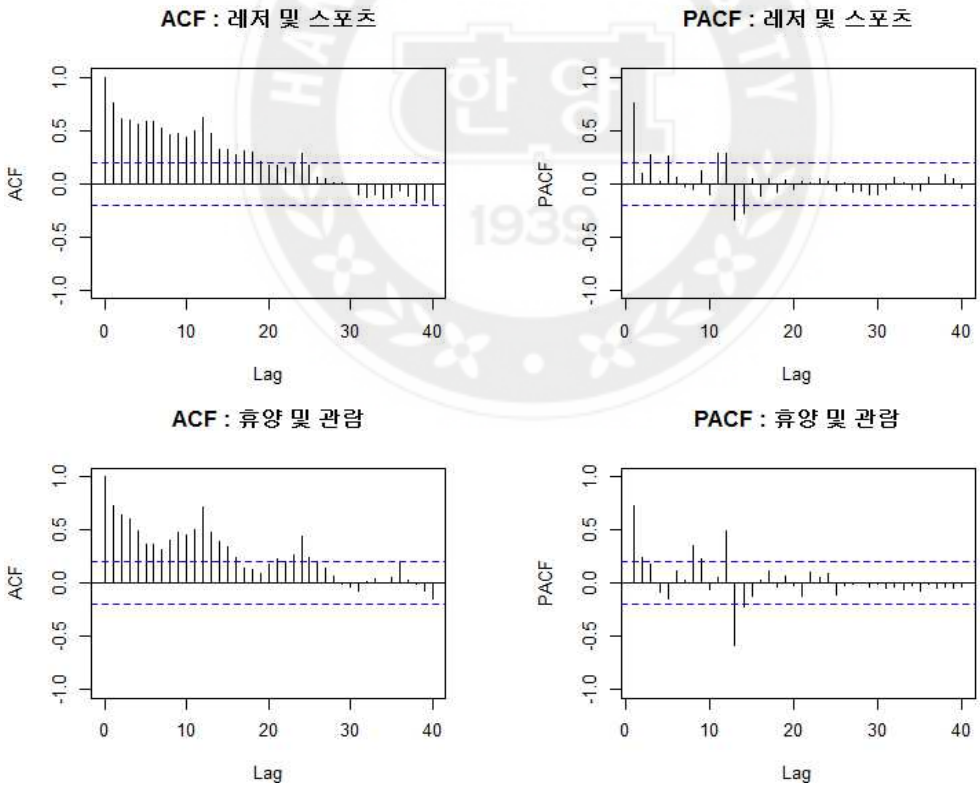
| 휴양 및 관람 | | | | |
|---|---|---|---|---|
| hidden layer<br>Dropout_input<br>Dropout_hidden | 반복시행횟수 | | | |
| | 50 | 100 | 150 | 200 |
| 50-50-50<br>0.2<br>0.5-0.5-0.5 | 64,715.10 | 54,079.74 | 52,829.94 | 54,627.35 |
| 100-100-100<br>0.2<br>0.5-0.5-0.5 | 59,676.67 | 53,604.78 | 53,363.84 | 50,411.36 |
| 50-50-50<br>0<br>0.5-0.5-0.5 | 52,935.42 | 55,700.42 | 52,335.70 | 54,118.23 |
| 100-100-100<br>0<br>0.5-0.5-0.5 | 51,968.69 | 56,667.59 | 50,847.84 | 52,429.84 |

[표 4.5] 최종 딥러닝 모형의 시험데이터에 대한 예측력

| | 레저 및 스포츠 | 휴양 및 관람 |
|---|---|---|
| RMSE | 51,290.79 | 159,932.7 |
| MAE | 44,014.02 | 126,505.2 |

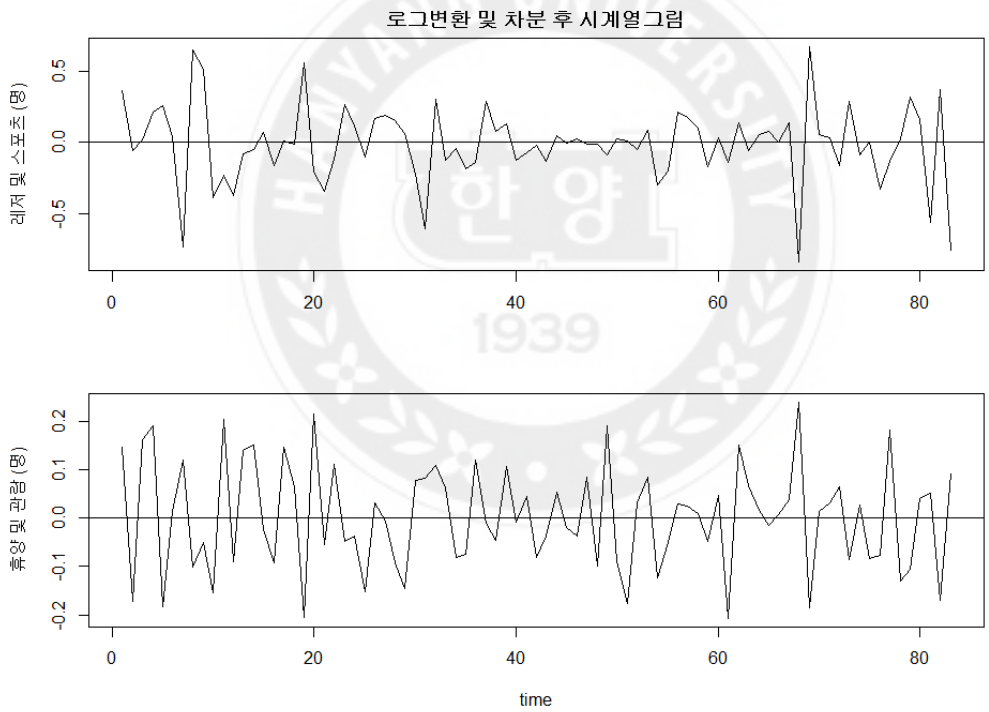
### 4.2.2 계절형 ARIMA 모형의 결과

4.1의 자료설명 및 변수 설정의 [그림 4.1]에서 볼 수 있듯이 두 여행 목적 모두 시간이 지날수록 증가하는 추세가 있고, 분산이 커지는 경향이 있으며 계절성이 보이는 것을 확인 할 수 있다. [그림 4.2]는 두 여행 목적별 자료 ACF와 PACF를 나타낸 것이다. 두 여행 목적 모두 ACF가 천천히 감소하고 Time lag가 12, 24, 36인 곳에서 값이 치솟는 것을 볼 수 가 있다. 따라서 두 비정상 시계열 자료의 정상화를 위해 훈련 데이터에 로그변환과 1차분 및 계절차분을 실시하였다.

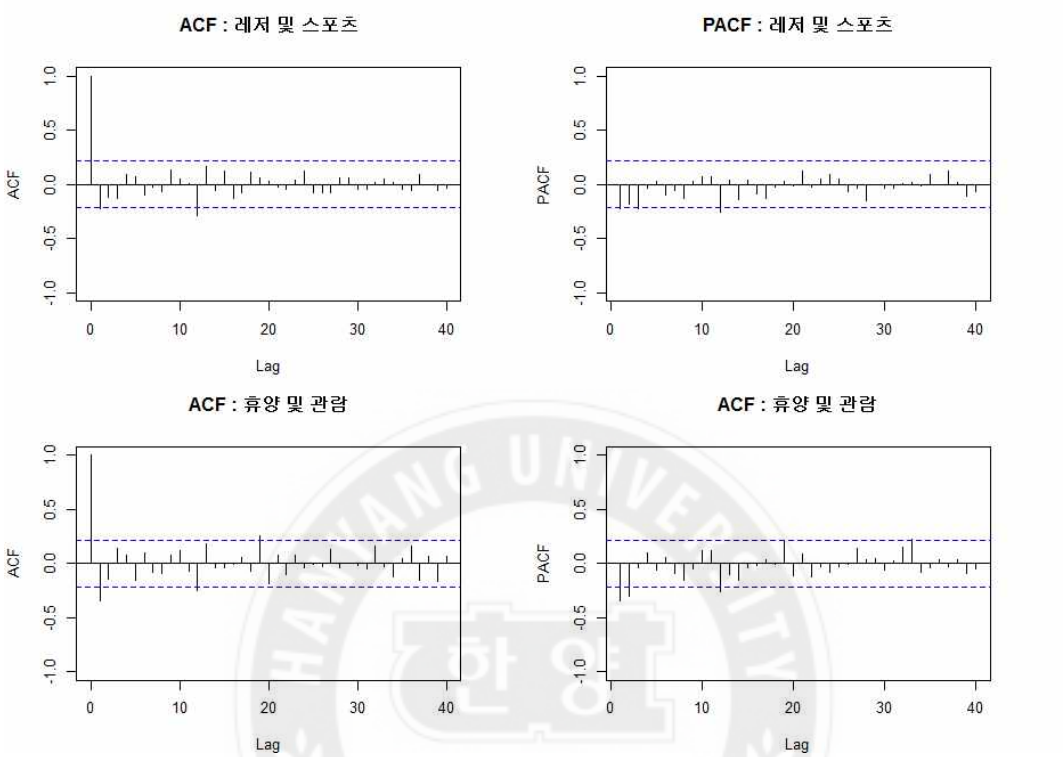



[그림 4.2] 여행 목절 별 ACF 및 PACF

로그변환 및 계절차분, 비계절 차분을 진행한 후 그린 시계열 그림과 ACF, PACF은 [그림 4.3]과 [그림 4.4]이다. [그림 4.3]을 보면 차분 후의 시계열 그림이 추세 없이 랜덤하게 분포 되어 있는 것을 볼 수 있다. 또한 [그림 4.4]의 ACF, PACF를 통해서 두 자료 모두 정상시계열 형태로 나오는 것을 볼 수 있다. 또한 정상성을 검정하는 Augmented Dickey-Fuller(ADF) 검정을 했을 때 유의수준 0.05에서 두 자료 모두 p-value가 0.01 이하로 나와 정상시계열이라는 검정결과를 얻었다.




[그림 4.3] 여행 목적 별 로그변환 및 차분 후 시계열그림




[그림 4.4] 로그변환 및 차분 후 ACF, PACF

따라서 원 시계열 자료에 ARIMA모형을 적합하는데, 비계절변동과 계절변동을 함께 갖고 있으므로 승법계절모형을 적용한다. 차수의 식별을 위해 [그림 4.4]의 ACF, PACF를 참고하면 '레저 및 스포츠'는 $ARIMA(0,1,1)(1,1,0)_{12}$, '휴양 및 관람'은 $ARIMA(0,1,1)(0,1,1)_{12}$이 적당해 보인다. 따라서 위 차수의 주변 차수들로 모델링을 진행하여 [표 4.7]와 [표 4.8]의 AIC(Akaike information criterion)와 BIC(Bayesian information criterion) 및 RMSE 비교를 통해 최종 시계열 모형을 정하였다.

AIC와 BIC는 같은 모형 내에서 모형이 얼마나 적합한지 판별하는 모형 적합도로서 다음 식 (4.2.3)과 같이 정의된다.

$$\begin{aligned} \text{AIC} &= 2k - 2\ln(L), \\ \text{BIC} &= -2\ln L + k\ln(n), \end{aligned} \tag{4.2.3}$$

여기서 $L$은 모형의 최대우도함수이고, $k$는 추정된 모수의 개수, $n$은 관측된 데이터의 수이다. AIC와 BIC는 값이 더 낮을수록 모형이 더 적합함을 의미한다.

[표 4.6] 레저 및 스포츠

| 레저 및 스포츠 | AIC | BIC | RMSE |
|---|---|---|---|
| $ARIMA(1,1,1)(1,1,0)_{12}$ | 3.36 | 13.04 | 31,966.61 |
| $ARIMA(0,1,1)(0,1,1)_{12}$ | 8.66 | 15.61 | 31,942.74 |
| $ARIMA(0,1,1)(1,1,0)_{12}$ | 6.13 | 13.39 | 36,083.78 |
| $ARIMA(1,1,0)(0,1,0)_{12}$ | 18.43 | 23.26 | 34,035.05 |

[표 4.7] 휴양 및 관람

| 휴양 및 관람 | AIC | BIC | RMSE |
|---|---|---|---|
| $ARIMA(0,1,1)(0,1,1)_{12}$ | -151.42 | -144.16 | 18,817.04 |
| $ARIMA(0,1,1)(0,1,0)_{12}$ | -144.16 | -139.32 | 20,058.99 |
| $ARIMA(1,1,0)(0,1,0)_{12}$ | -138.7 | -133.86 | 19,877.61 |
| $ARIMA(1,1,1)(0,1,1)_{12}$ | -149.44 | -139.76 | 21,736.65 |

[표 4.6]과 [표 4.7]의 결과 중 AIC, BIC와 RMSE가 가장 낮은 모형을 최종 모형으로 선택하였다. 레저 및 스포츠는 $\text{ARIMA}(1,1,1)(1,1,0)_{12}$ 모형이

AIC, BIC, RMSE값이 가장 낮아 최종모형으로 선택하였고 휴양 및 관람도 AIC, BIC, RMSE이 가장 낮은 ARIMA$(0,1,1)(0,1,1)_{12}$으로 선택하였다.

시계열분석에 관한 시험데이터의 RMSE와 MAE는 [표 4.8]에 나타내었다. 레저 및 스포츠는 RMSE가 64,996.1, MAE가 59,290.5이고 휴양 및 관람에 대해서는 RMSE가 147,444.4, MAE가 129,457.2이다.

[표 4.8] 시계열 시험데이터

| | 레저 및 스포츠 | 휴양 및 관람 |
|---|---|---|
| RMSE | 64,996.1 | 147,444.4 |
| MAE | 59,290.5 | 129,457.2 |

### 4.2.3 모형의 예측결과 비교

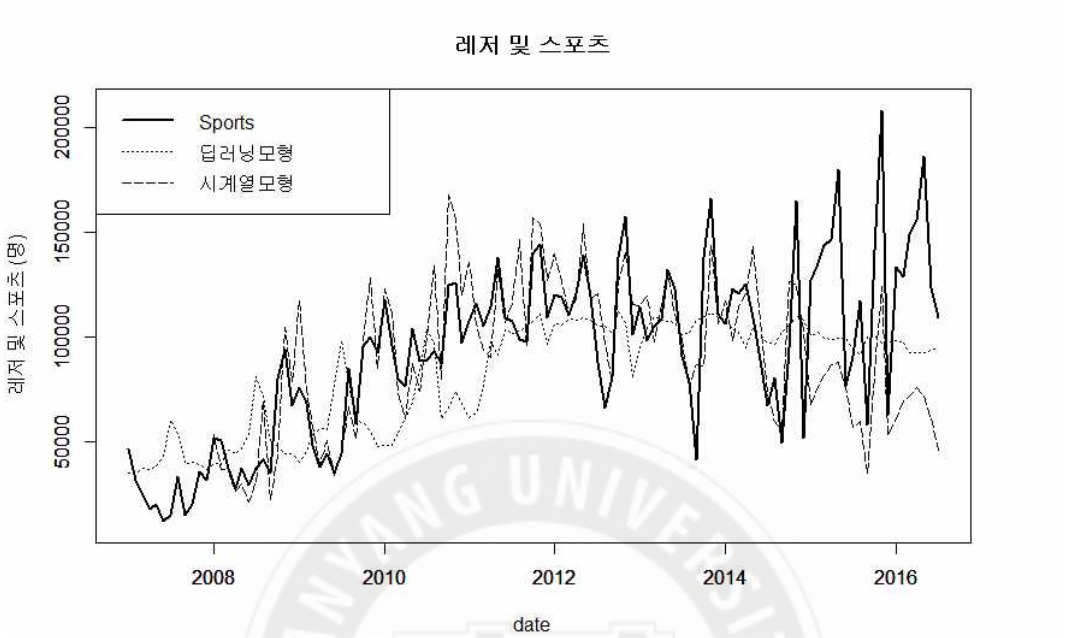
다음의 [표 4.9]는 [표 4.5]와 [표4.9]의 결과를 이용하여 각 모형의 시험데이터에 대한 예측 결과를 비교한 것이다.

[표 4.9] 최종 모형의 시험데이터에 대한 예측력 비교

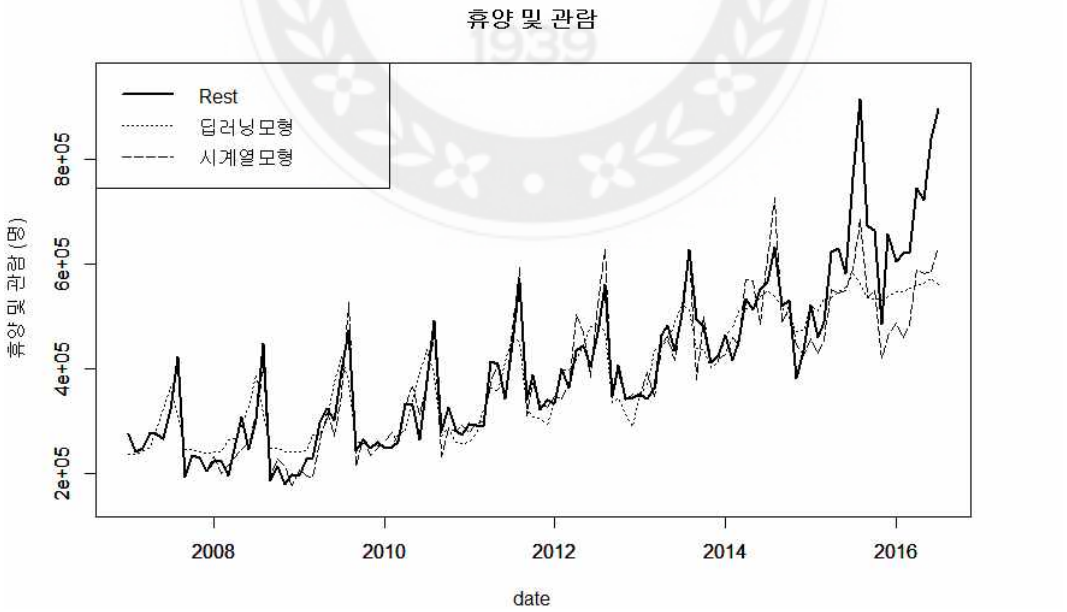
| | 레저 및 스포츠 | | 휴양 및 관람 | |
|---|---|---|---|---|
| | 딥러닝 모형 | 시계열 모형 | 딥러닝 모형 | 시계열 모형 |
| RMSE | 51,290.79 | 64,996.1 | 159,932.7 | 147,444.4 |
| MAE | 44,014.02 | 59,290.5 | 126,505.2 | 129,457.2 |

[표 4.9]에서 볼 수 있듯이 레저 및 스포츠에서 딥러닝 모형의 RMSE와 MAE가 시계열 모형의 RMSE와 MAE보다 더 낮아 딥러닝 모형이 예측을 더 잘하고 있다고 판단할 수 있다. 휴양 및 관람에서는 딥러닝 모형의 RMSE가 시계열 모형의 MAE보다 조금 더 높지만 MAE는 126,505.2로 시계열모형보다 조금 더 낮은 값을 보였다.

[그림 4.5]와 [그림 4.6]은 총 구간에 대한 예측 그림이다. [그림 4.5]는 레저 및 스포츠에 대한 딥러닝 모형과 시계열 모형 그리고 원 자료의 시계열 그림이다. 그림을 보면 추세는 시계열 모형과 잘 맞지만 예측오차 측면에서는 딥러닝 모형이 더 잘 맞는 것을 알 수 있다. [그림 4.6]의 휴양 및 관람에서도 마찬가지로 추세는 시계열 모형과 더 잘 맞아 보이지만 예측오차+ 측면에서는 딥러닝 모형과 비슷하다는 것을 알 수 있다.




[그림 4.5] 레저 및 스포츠의 딥러닝 모형과 시계열 모형 예측 그림




[그림 4.6] 휴양 및 관람의 딥러닝 모형과 시계열 모형 예측 그림

# 5. 결론 및 고찰

기존까지 시계열모형으로 관광수요 예측을 한 것과는 다르게 본 논문에서는 최근 음성인식, 이미지인식, 자연어처리 등의 다양한 분야에서 우수한 성능을 보이고 있는 딥러닝 모형으로 인터넷 검색어 변수를 이용해 여행 목적별 제주 관광객 수의 예측을 하고 기존의 시계열 모형과 비교를 하였다. 분석에 사용된 자료는 2007년 1월부터 2016년 7월까지의 월별 자료이다. 2007년 1월부터 2014년 12월까지의 자료는 모형 훈련데이터로 사용하고 2015년 1월부터 2016년 7월까지의 자료를 시험데이터로 사용하여 예측력을 비교하였다.

분석결과 훈련데이터를 이용한 모형훈련에서는 RMSE와 MAE 측도로 비교를 하였을 때 시계열모형이 실제의 자료와 더 낮은 차이를 보였지만 시험데이터를 사용한 두 목적별 예측에서는 모두 인터넷 검색어 변수를 사용한 딥러닝 모형의 예측력이 비슷하거나 좋았다. 모형의 적합에서 훈련데이터에 모형이 잘 맞는 것도 중요하지만 본 논문의 목적은 예측이기 때문에 인터넷 검색어 변수를 사용한 딥러닝 모형의 활용 가능도가 있다고 본다.

본 논문에서 사용한 자료는 총 9년 7개월간의 월별 자료를 사용하여 데이터의 수가 많지 않았다는 점에서 다른 분야에서 이용된 딥러닝의 성능보다 좋은 결과를 보이지 못한 것으로 보이며, 추후에 더 많은 기간의 자료를 가지고 분석을 시행한다면 더 좋은 예측력을 보일 것으로 사료된다.

# 참고 문헌


1. Anvik, C., Gjelstad, K. (2010). “Just Google it : Forecasting norwegian unemployment figures with web queries”, Center for Research in Economics and Management, Working paper N, 11.

2. Box, G.E.P., Jenkins, G. M. (1970). “Time series analysis: rorecasting and control”, Holden-day, San Francisco.

3. Dahl, G. E., Sainath, T. N. and Hinton, G. E. (2013). "Improving deep neural networks for LVCSR using rectified linear units and dropout," IEEE International Confedence Acoustics on Speech and Signal Processing(ICASSP), pp.8609-8613.

4. Hinton, G. E., Osindero, S., Tech, Y. W. (2006). “A fast learning algorithm for deep belief nets”, Neural computation, 18(7), pp.1527-1554.

5. Hinton, G. E., Srivastava N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2012). “Improving neural networks by preventing co-adaptation of feature detectors”, arXiv preprint arXiv:1207.0580.

6. Kuremoto, T., Kimura, S., Kobayashi, K., Obayashi, M. (2014), “Time series forecasting using a deep belief network with restricted Boltzmann machines”, Neurocomputing, 137, pp.47-56.

7. Law, R., & Au, N. (1999). “A neural network model to forecast Japanese demand for travel to Hong Kong”, Tourism Management, 20(1), pp.89-97

8. Mccallum, M.L, Bury, G.W. (2013). “Google search patterns suggest declining interest in the evironment”, Biodiversity and Conservation, 22(6-7), pp.1355-1367.

9. Minsky M.L., Papert, S. A. (1969). “Perceptrons”, Cambridge, MA: MIT Press.

10. Nair, V., Hinton, G. E. (2010). “Rectified linear units improve restricted boltzmann machines”, In Proceedings of the 27th International on Machine Learning (ICML-10), pp.809-814.

11. Palmer, A., Montano, J. J., Sese, A. (2006). “Designing an artificial neural network for frecasting tourism time series”, Tourism Management, 27, pp.781-790.

12. Srivastava N., Hinton, G. E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). “Dropout: A simple way to prevent neural networks from overfitting”, The Journal of Machine Learning Research, 15(1), pp.1929-1958.

13. Tan, P.N., Steinbach, M., Kumar, V. (2007). “Introduction to Data Mining”, Pearson Education, Addison Wesely.

14. Werbos, P.J. (1974). “Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences”, PhD thesis, Harvard University.

15. 고태호, 강정미, 임정현 (2011). “제주지역 관광산업의 경제적 효과 분석”, 제주발전연구원.

16. 권치명, 황성원, 정재운 (2015). “실업률 예측을 위한 인터넷 검색 정보의 활용”, 한국시뮬레이션학회 논문지, 24(2), pp.31-39.

17. 김성태 (2014). “관광수요예측에 대한 실증연구-패널 데이터 분석기법을 중심으로”, 관광레저연구, 26(1), pp.115-129.

18. 김영우, 손은호 (2006). “계절 ARIMA Model을 이용한 경주방문객의 수요예측에 관한 연구”, 호텔경영학연구, 15(1), pp.309-326.

19. 송다영 (2015). “계절형 ARIMA모형을 이용한 여행목적 및 형태별 제주관광객수 예측에 관한 연구”, 제주대학교 대학원 전산통계학과.

20. 윤영철, 윤석민, 방상인, 배진아, 변종석, 심미선, 양승찬, 이민규, 이재진, 정준희, 김유정 (2015). “여론집중도조사보고서”, 제 2기 여론집중도조사위원회.

21. 이재성 (2016). “심층 신경망의 발전 과정과 이해”, 정보와 통신, 33(11), pp.40-48.

22. 이충기, 송학준 (2007). “최적 시계열 수요예측 모델선정에 관한 연구”, 관광학연구, 31(6), pp.289-311.

23. 조광익 (1999). “관광수요 예측 및 경제적 파급효과 분석-강원 역사문화촌을 중심으로”, 한국관광연구원.

24. 조신섭, 손영숙 (2009). “SAS/ETS를 이용한 시계열분석(3rd edition)”, 율곡출판사.

25. 최재혁, 신창섭 (2015). “빅데이터를 이용한 휴양림 이용객현황과 인터넷 검색어의 상관관계 분석”, 한국산림휴양학회지, 19(4), pp.13-23.

# ABSTRACT

## Forecasting the number of tourists in Jeju Island using Deep learning Algorithm

Choi, Min Jung
Dept. of Applied Statistics
The Graduate School
Hanyang University

Forecasting the number of local tourists has a great impact on economic effects such as production and employment inducement in the region. Especially, the tourism industry in Jeju Island accounts for 70% of the entire industry. Thus forecasting the number of tourists in Jeju Island is an important research that can be used as a good basis for the development of tourism industry in Jeju Island. However, there are limitations to satisfy various assumptions such as stationary when using the time series model which is mainly used for forecasting tourism demand.

In this thesis, deep learning algorithm, which shows superior performance in various fields such as image recognition, voice recognition, and natural language processing, is used to forecast the number of tourists in Jeju Island by travel purpose. In addition, internet search data which is being

studied in many fields recently is used as a predictor of the number of tourists in Jeju Island. For this purpose, we adopt the Naver which is the largest search site in korea. Deep Neural Network(DNN) among the deep leargning algorithms is used in this thesis. In order to solve the disadvantages of DNN, the ReLu function and dropout normalization is applied to DNN. Also, to confirm the predictive power of the deep learning algorithm, it is compared with seasonal ARIMA model. The data used for the analysis is from January 2007 to July 2016. The analysis shows that the deep learning model using internet search data is similar to or better than the seasonal ARIMA model.

# 연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2016년12월27일

학위명 : 석사

학과 : 응용통계학과

지도교수 : 차경준

성명 : 최민정

**한 양 대 학 교 대 학 원 장 귀 하**

# Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

DECEMBER 27, 2016

| | |
|---|---|
| Degree : | Master |
| Department : | DEPARTMENT OF APPLIED STATISTICS |
| Thesis Supervisor : | Cha Kyung Joon |
| Name : | Choi min jung (Signature) |