

시계열 데이터

Contradictory, My Dear Watson 소개

TABLE OF CONTENTS

목차 소개

01 | 시계열 데이터란?

02 | 시계열 데이터를 다루는 방식

03 | 시계열 데이터 문제 소개

04 | 7~8주차 문제 소개

시계열 데이터란?

기상청 날씨 예측



PC 사용 기록



전제 & 결론



“

시계열 데이터란?

”

시계열 데이터 = 시간이나 순서에 따라 일련의 연관성을 가진 데이터



시계열 데이터의 특징

시간에 관해 순서가 매겨짐
연속한 값들이 서로 상관관계를 가짐

시계열 데이터의 불규칙성

추세, 계절성, 주기 등 다양한 양상을 보이는 데이터기 때문에, 명확한 규칙을 찾아 바로 예측하기 어려움

RNN과의 관계

RNN은 순차 데이터 처리에 효과적이기 때문에, 시계열 데이터의 시간 의존성에 대해서도 효과적인 모델링이 가능

관점	문제 양상	예시
시간 정보를 가지는 변수/컬럼 존재 여부	과거 / 현재 등 특정 시점을 기준으로 분리해서 이용	전제 - 결론 제시 후 참/거짓 예
학습/테스트 데이터의 시계열에 따른 분할 여부	과거의 일련의 데이터를 학습용으로 제시 테스트 데이터는 미래 또는 특정 시점에서의 예측	날씨 예측
컬럼별로 시계열의 목적 변수 존재 여부	사용자별 과거 이용 시간 등 시계열 목적 변수/컬럼의 존재 이를 이용해 미래의 이용 시간을 예측	전기 사용료 예측

출처 : 데이터가 뛰어노는 AI 놀이터, 캐글, 상위 랭킹 진입을 위한 필살기(도서)

03

시계열 데이터 문제 소개

SMARCLE

Bike Sharing Demand

Forecast use of a city bikeshare system

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Overview

Start

May 29, 2014

Close

May 30, 2015

Description

[Get started on this competition through Kaggle Scripts](#)

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

Bike Sharing Demand

- 2년간 시간별 자전거 대여 데이터를 토대로 예측하는 문제
- 매달 1~19일에 대한 데이터가 학습 셋으로 제공되며, 20일부터 말일까지의 대여된 자전거 수를 예측해야 함
- 날씨와 계절의 경우 라벨 인코딩된 상태로 제공됨

<https://www.kaggle.com/competitions/bike-sharing-demand>

Predict Future Sales

Final project for "How to win a data science competition" Coursera course

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Overview

Start

6 years ago

Close

8 months to go

Description

This challenge serves as final project for the "How to win a data science competition" Coursera course.

In this competition you will work with a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - **1C Company**.

We are asking you to predict total sales for every product and store in the next month. By solving this competition you will be able to apply and enhance your data science skills.

Predict Future Sales

- 2013~2015년도 상품 판매 데이터를 보고 2015년 11월 각 상점의 상품에 대한 월간 판매 데이터를 예측하는 문제
- 판매량을 예측하는 문제이기 때문에 회귀 문제를 겸함
- 시간에 따라 판매되는 상품 종류가 달라질 수 있다는 점에 유의
- 초기 학습 데이터가 여러 개의 csv 파일로 나누어 제공

<https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales>



KAGGLE · GETTING STARTED CODE COMPETITION · ONGOING

Contradictory, My Dear Watson

Detecting contradiction and entailment in multilingual text using TPUs

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Overview

∞ This competition runs indefinitely with a rolling leaderboard. [Learn more.](#)

Description

"...when you have eliminated the impossible, whatever remains, however improbable, must be the truth"

-Sir Arthur Conan Doyle

Our brains process the meaning of a sentence like this rather quickly.

We're able to surmise:

Contradictory, My Dear Watson

- 전제 - 결론 데이터가 주어지고, 이 두 데이터 간 상관관계를 파악하는 문제
- label에서 0은 참, 2는 거짓, 1은 두 데이터 간 상관관계를 알 수 없음을 의미
- 총 15개의 언어로 작성된 문장 데이터이기 때문에, NLP 문제도 겸하고 있음 = NLP와 RNN 양쪽을 전부 사용해보자!

Q&A

발표를 들어주셔서 감사합니다.

지금 답변드리지 못한 부분은 이번 주 내로 슬랙을 통해 답변드리겠습니다!