Problem Approach & Steps

for Time Series Data

By Shresth Malik

Submitted by : Shresth Malik, CSE Undergrad at NIT Raipur

In order to solve the given problem. Below is the procedure involving 4 steps I will do

I. Data Preprocessing

i. Data Cleaning

- Remove duplicate & irrelevant data.
- Remove unexplainable outliers.
- Remove records with missing values of parameters (missing data) :

ii. Organize the data

If the no. of features is too large, we organise the data in different tables & establish sufficient & meaningful relationships between these tables. Although this must be avoided if possible, since it complicates the Learning part.

iii. Standardization

This step shall be done only if the range of data is too small with respect to it's magnitude.

We basically re-scale the data by using appropriate coefficients like it's mean, standard deviation, upper limit, lower limit, etc.

iv. Reduction or Augmentation

If we have too many records. Processing it all will be time consuming. So, we reduce the data appropriately by Sampling.

On the contrast if we have very less data. Then we need augmented data to make it eligible for a better Model. For that we can use GAN (Generative Adversarial Network)

v. Categorise or Sort the data according to the time stamp.

This will be helpful later to analyse the data cyclicity with week, month, year.

We can also group the data by week. And repeat the steps IV and V after integrating these records as :

a. Weekly data
b. Monthly data
c. Yearly data

II. Features

i. Feature Analysis

We analyse the features available to find their relative importance by evaluation using :

a. Sequential Forward Selection or Wrapper Method

b. Quick to compute statistics $J(X_f)$ or Filter Method
c. Embedded Method like LASSO

ii. Feature Engineering

Post Analysis, we decide whether we need to

a. Create New Features using data of old features.
b. Transform Existing Features into more suitable form.

in order to improve the performance of Machine-Learning Model
for competitve advantage OR to meet the customer's specific needs.

OR on the contrary we could Reduce the number of features if we have too many features. We can:

a. Select the most important features with highest evaluative score from Feature Analysis.
b. Combine multiple features to create one feature by either interaction or aggregation.

III. Which Model & Why ?

i. Regression (Linear or Logistic)

We could have used regression if we knew the output label. But here we have no clue which feature is the output. So discussing regression would be futile.

ii. K-Nearest Neighbors (KNN)

We could use the KNN to categorise the data points to respective labels. But the same reason stops us again. We have no labels.

Looks like we need to explore with Unsupervised Learning Methods.

iii. Clustering

Clustering seems to be a viable method for analysing data with no label (output variable).

Which type of clustering must we use ? We have

a. Centroid Based
b. Connectivity Based
c. Graph Based
d. Distribution Based
e. Density Based

We make this decision according to the type of data features we have available. For the sake of our discussion I'm going to assume we have chosen Centroid Based Clustering.

Specifically, K-Means Clustering

Modeling Begins Learning

i. We choose the number of clusters (ie K)
ii. We randomly initialise the K cluster centers.
iii. We start analysing each record and Assign each data point to it's Nearest Cluster.
By Nearest, I mean the cluster center, from which the root of sum of squares of differences in features is minumum. Identical to distance formula in N-dimensional space.
iv. When all the data points are assigned a cluster. Update the Cluster Centers to the Centroid of all the data points assigned to this Cluster.
v. Repeat Step iii & iv until there is negligible change in cluster assignment.

IV. Post Processing Analysis

Since we don't know which of the feature is the output. We can't exactly analyse the quality of the model we just created. However, we can augment new data points using GAN (Generative Adversarial Network) and check the accuracy of new data points with our predicted model.

If the results are not significant. We must create 2 agents, Generator & Detector. And imitate a competitive objective between them where their job is

a. Generator must create artificial data points.
b. Detector must detect whether a data point is real of fake.

After running a few thousand iterations. Both of the agents improve significantly in their job. And now we have a model which has learned to imitate behavior of the distribution.

This helps in improving the overall accuracy of the model.

Acknowledgement

Thank you for reading my approach towards solving the problem explained by Mr. Sandeep Yadav.

Thank you for giving me enough time, consideration & opportunity to put my knowledge into practice.

I hope to hear from you soon and discuss my further journey with your organisation as an Intern.

Yours Sincerely

Shresth Malik

CSE Undergrad NIT Raipur

Candidate for Data Science & Machine Learning Internship Role

Email : shresthmalik.official@gmail.com

Phone No : +91 88892 92088