Air Pollution Prediction based on multicollinearity

Shanika Iroshi Nanayakkara
Deakin University

03/06/2022

Overview

Problem Definition

Air pollution predictors

Challenges

Challenges

Proposed Model

Step One -Preprocessing

Step Two - Model training

Step Three - Model Prediction and Evaluation

Evaluation Results

Conclusion

Problem Definition

Air pollution predictors

Defn

Predict air pollution composition in atmosphere in the future.

- Use predictors such as Temperature,Humidity,Sensor data.
- Response variables are carbon monoxide,benzene,notrous oxide

Linear Regression vs Rigid Regrssion

Linear Regression

- Linear regression presents relationship as a straight line.
- Show correlation between two variables (one predictor for response variable/variables).
- Response should be continuous and independent variable(s) (predictor variables) can be continuous or discrete.

Rigid Linear Regression

- Use to implement multicollinearity of predictor variables which are highly correlated each other objects in the whole dataset.
- Add penalty values to reduce the loss or error of linear regression cause by bias and/or variance of the variables.

Challenges

Group Outlying Aspects Mining

- Focus on differences between groups.
- Multiple points.

Outlying Aspects Mining

- Concentrates on differences between objects.
- One point.

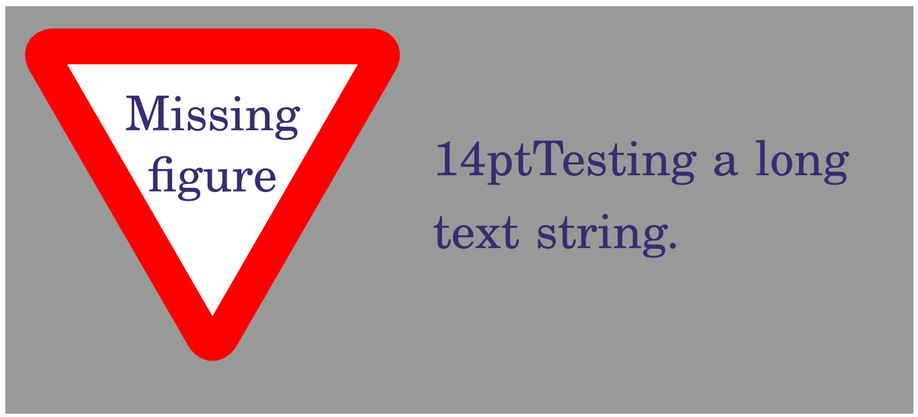


Figure 1: Group Outlying Aspects Target

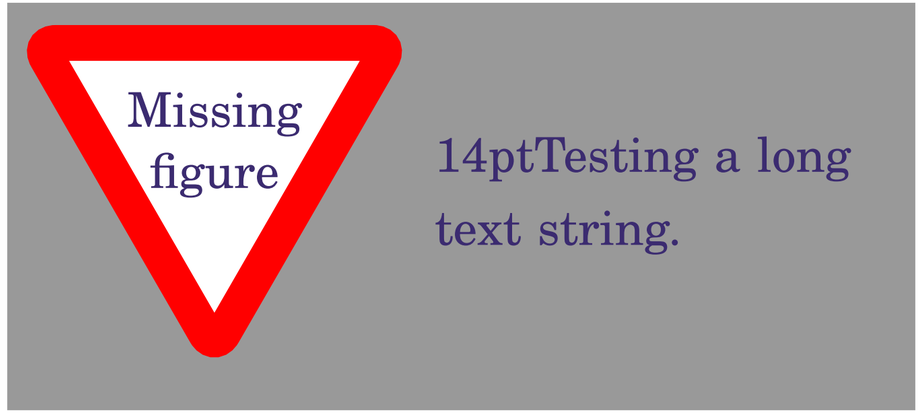


Figure 2: Outlying Aspects Target

Challenges

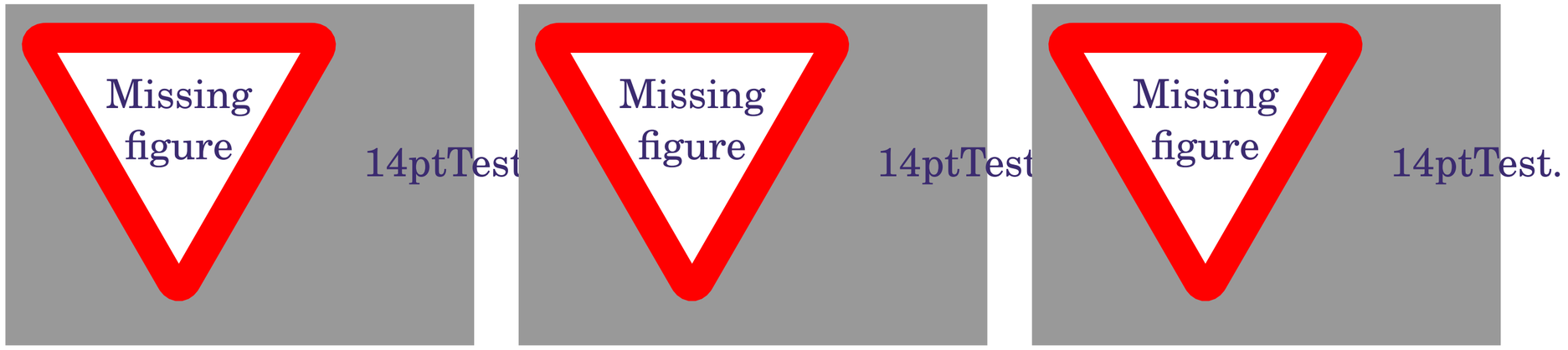
- Reasons for multicollinearity in predictors.
  - Inaccurate use of different types of variables.
  - Poor selection of questions or null hypothesis.
  - Variable repetition.
  - A dependent variable selection.
  - High correlation.
  - Use of dummy variables.

Proposed Model

Framework of Proposed Model:

Figure 3: Framework of Proposed Model

Step One -Preprocessing



(a) f_1 (b) f_2 (c) f_3

Figure 4: Histogram of G_q on three features

Step Two - Model training

- Calculate Earth Mover Distance
  - Represent one feature among different groups
  - Purpose: calculate the minimum mean distance

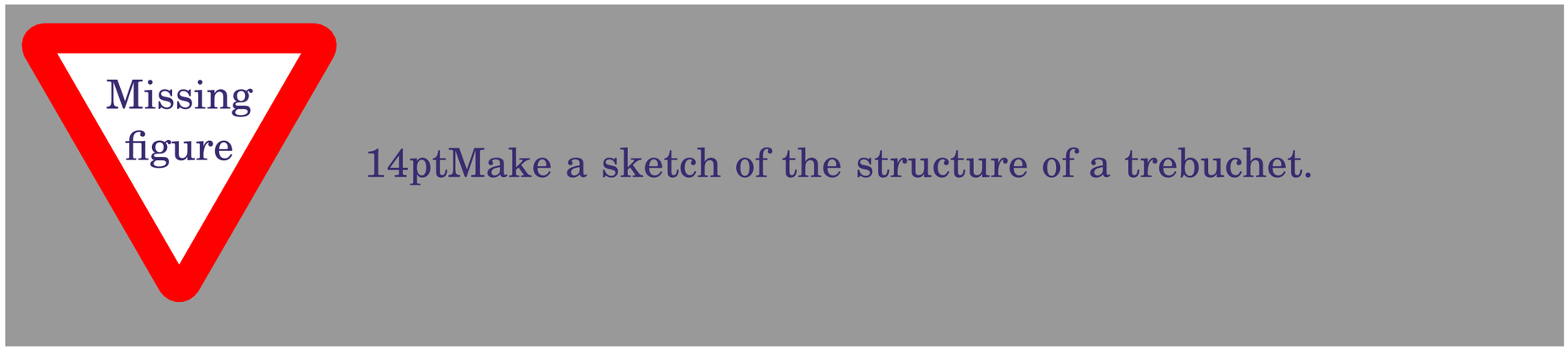


Figure 5: EMD of one feature

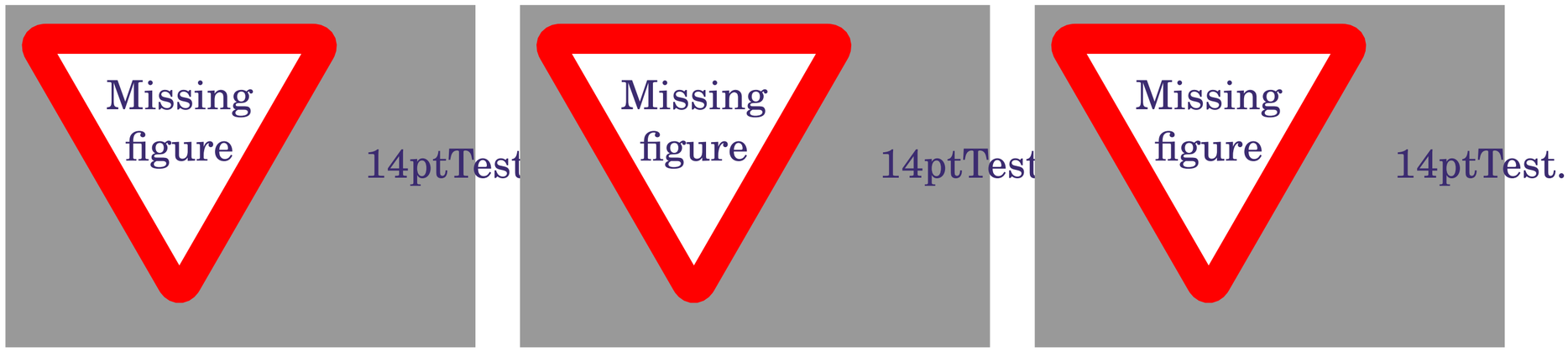
Step Three - Model Prediction and Evaluation

- Identify group outlying aspects mining based on the value of outlying degree.
- The greater the outlying degree is, the more likely it is group outlying aspect.

Illustration

Evaluation Results

Evaluation



(a) f_1 (b) f_2 (c) f_3

Figure 6: Histogram of G_q on three features

Conclusion

Conclusion

- Problem Definition: we propose prediction model to predict air pollution components which are specified as carbon monoxide, benzene and nitrous oxide. Algorithm Consequently, we use rigid linear regression instead general linear regression method.
- Strategies :We uses multiple predictors such as temperature, absolute and relative humidity and five sensor data. We identify that these predictors are correlated among each other. Therefore, we reveals the need of handling multicollinearity. We clearly show the performance efficiency of proposed models using determi- nation of coefficient, bias and variance scores for each models.
- Recommendations : Evaluate how well these prediction models behave on adding more noisy data. How to affect prediction accuracy by noisy data. Increase the amount of predictors to predict air pollution

Questions?

Contact Information

Associate Professor Gang Li
School of Information Technology
Deakin University, Australia

GANGLI@TULIP.ORG.AU

TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING