Mutlivariate Assignment Topic Suggestions

Andomei Smit: SMTAND051

02/03/2025

Contents

Introduction

The purpose of this document is to provide 3 topic ideas that may be used in the final Assignment for Multivariate Analysis. Below each topic will be stated with an associated research question and brief description of the data.

Topic 1: Classifying leaves based on features extracted from images of leaves

Research question: Can leaves be classified into species using features extracted from images?

Data description: The data is a set of features extracted from approximately 1584 images of plant leaves. These images consist of 16 unique images for each of the 99 unique species in the data. Three features were extracted from these images:

1. Margin Features
2. Shape Features
3. Texture Features

where each feature is made up of 64 attribute vectors. Thus in total the data has 192 $(64 + 64 + 64)$ independent variables or attribute vectors. Formally, the dimensions of the data is 1524×192. All of the variables are continuous. This link can be followed to the Kaggle page with the dataset information.

Possible Multivariate techniques: It may be attempted to apply various linear and non-linear dimension reduction techniques to the data. Thereafter various clustering algorithms could be applied to see if meaningful clusters could be extracted based on the different reduced dimension datasets (if time allows, otherwise the project could focus on an comparison of the different dimension reduction techniques).

Topic 2: Identifying key Environmental Factors in Air Pollution Across Cities

Research question: Can we identify distinct pollution profiles for different cities based on daily air quality measurements?

Data description: The data is made up of a large scale spatio-temporal dataset that measures various air pollutants and environmental factors across 52 United States Cities. These measures were taken everyday from 1 January 2019 to 11 December 2020. The variables in the data are:

1. Date the measurement was taken
2. City and State names
3. Median, Min, Max and variance of each pollutant or meteorological feature of a day (listed below)
4. Total vehicle travel distance for the sample
5. Calculated feature showing the influence of neighboring power plants
6. Some measure of domestic emissions
7. Longitude and Latitude of the cities

The meteorological features mentioned above are:

1. Temperature
2. Pressure
3. Humidity
4. Dew
5. Wind Speed
6. Wind Gust Speed

The pollutants (i.e. the dependent variables) are:

1. PM2.5
2. PM10
3. NO_2
4. O_3
5. CO
6. SO_2

Since the data has some time-series element, some time could be spent summarising the measurements into averages, such as "Average change in CO_2 over the time period" in order to move away from Longitudinal Data. In total, there are 64 dependent variables (some of which are categorical, for example 'City') and 6 independent variables (pollutants) for 52 observations (or cities, once the data is summarised to remove the time element). This link can be followed to the Kaggle page with the dataset information.

Possible Multivariate techniques: Similarly to above, it may be attempted to apply a PCA to reduce the dimensionality of the data. Thereafter various clustering algorithms could be applied to see if meaningful clusters could be extracted based on the Principal Components.It could also be attempted to apply a CCA to analyze how environmental factors influence air pollution.

Topic 3: Classifying Forest Cover Types based on Environmental Factors

Research question: Can we classify different forest cover types based on soil composition, temperature, and rainfall patterns?

Data description: The data was collected from four parks in the Roosevelt National Forest of Northern Colorado. The areas were divided into 30 meter by 30 meter blocks and for each block, different measurements were taken. Each block is treated as one observation. These measurements are a mixture of continuous and categorical variables. The continuous variables are:

1. Elevation in meters
2. Aspect in degrees azimuth
3. Slope in degrees
4. Horizontal and Vertical Distance to nearest surface water features
5. Horizontal Distance to nearest roadway
6. Horizontal Distance to nearest wildfire ignition points

The categorical variables are:

1. Hillside index at 9am, 12pm and 3pm (0 to 255 index) (although, given the number of indeces this variable has, it may still reasonably be treated as continuous)
2. Wilderness Area (4 binary columns, 0= absence or 1= presence)
3. Soil Type (40 binary columns, 0= absence or 1= presence)

The dependent variable is the cover type, i.e. the predominant type of tree in a 30 meter by 30 meter block. This is a categorical variable with 7 levels, representing 7 different types of trees (Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir and Krummholz). This link can be followed to the Kaggle page with the dataset information.

Possible Multivariate techniques: Since the independent variables are a mixture of categorical and continuous variables, the continuous variables could be summarised using PCA or some other technique. The dimensions of the categorical variables could be reduced using Multiple Correpondence Analysis (MCA). An Independent Correspondence Analysis could also be used to extract the underlying independent pattern in the soil types, since this in itself is 40 different binary variables. However, since this course mostly focusses on continuous independent variables, this topic will most likely not be chosen.