STAT 432 Final Project

Detecting Volcanoes on Venus via Classification (Where are the Volcanoes?!!)

Team : Mamamia!
Di Ye (diye2), Hao Wang (haow2),Hannah Pu (chpu2)

11/6/2018

Contents

Dataset Description:

The data was downloaded from Kaggle, which is originally from NASA's Magellan spacecraft database. 9734 images were captured by the spacecraft and converted to pixel (110x110, from 0 to 255), where every image is one row of 12100 columns (all the 110 rows of 110 columns). Images can contain more than one volcanoes or maybe none. The 9000+ images are separated to four datasets (file names : *train_images*, *train_labels*, *test_images*, and *test_labels*):

Image dataset (*train_images* and *test_images*)

Train_images : 7000 images as train data with 12100 variables;
Test_images : 2734 images as test data with 12100 variables; All the variables correspond to the pixel image, 110 pixel * 110 pixel = 12100.

Label dataset (*tain_labels* and *test_labels*)

Both *train_labels* and *test_label* datasets include the following labels:
1. *Volcano?* : if in the image there are volcanoes (Main target), 1 (yes) or 0 (no)
(If Volcano? = 0, the following three categories would be "nan")
2. *Type* : 1= definitely a volcano,2 =probably, 3= possibly, 4= only a pit is visible
3. *Radius* : is the radius of the volcano in the center of the image, in pixels
4. *Number Volcanoes* : The number of volcanoes in the image
For this project, we will focus mainly on predicting whether each image has a volcanoe or not. In addition, if the classification prediction goes well, we will also construct model to predict the number of volcanoes in the images.

Data Import:

Data downloaded from Kaggle were csv files, there are four data files in total. The four data files were imported into R, dimensions of the four data files are as follows:
train_images : 7000 observations and 12100 variables
train_labels : 7000 observations and 4 variables
test_images : 2734 observations and 12100 variables

test_labels : 2734 observations and 4 variables
Initial observations of the four data files are printed as following (due to large number of variables, for the *train_images* and *test_images* files, only first 18 variables are printed):

1. *train_images*

```
head(train.x[,1:18])
```

```
##     V1  V2  V3  V4 V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
## 1:  95 101  99 103 95  86  96  89  70 104 115  96  89 102 109 108 102 104
## 2:  91  92  91  89 92  93  96 101 107 104  92  81  76  83  88  93  91  92
## 3:  87  70  72  74 84  78  93 104 106 106  94  79  96  88  86  89  98  94
## 4:   0   0   0   0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 5: 114 118 124 119 95 118 105 116 123 112 110 113 119 112 105 121 117 125
## 6:  79  95  90  82 73  74  77  75  82  87  84  81  83  69  72  84  93  84
```

2. *train_labels*

```
head(train.y)
```

```
##    Volcano? Type Radius Number Volcanoes
## 1:        1    3  17.46               1
## 2:        0  NaN    NaN             NaN
## 3:        0  NaN    NaN             NaN
## 4:        0  NaN    NaN             NaN
## 5:        0  NaN    NaN             NaN
## 6:        0  NaN    NaN             NaN
```

3. *test_images*

```
head(test.x[,1:18])
```

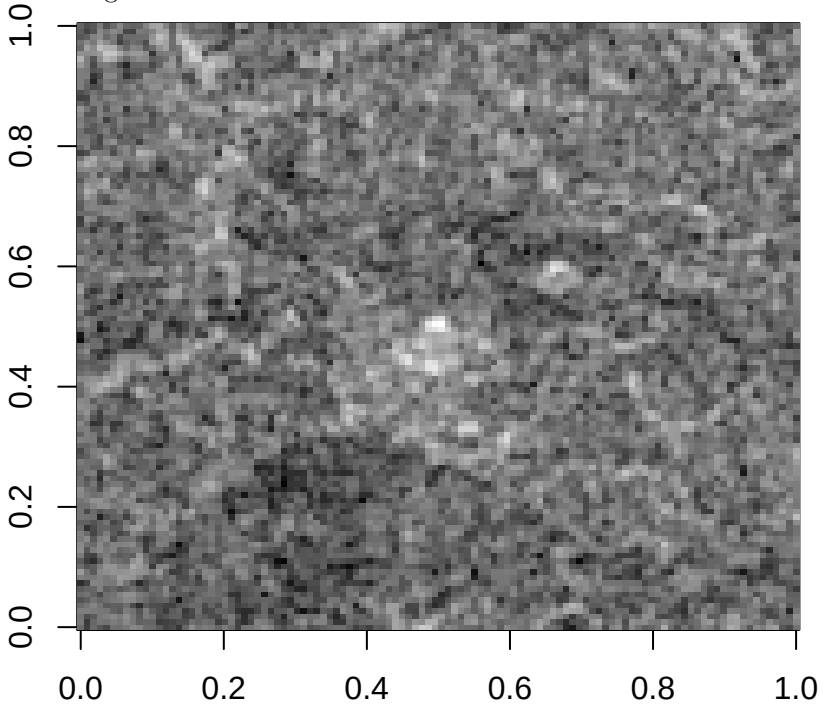
```
##     V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
## 1: 107 116 108 101 107 109 108 110 100 109 118 115 111 121 114  94  98 100
## 2:  93  95  98 100  90 100 108  98  90 103 107  93  90  91 104 113 104 102
## 3: 108 108  92 116 116 140 126 104 112 103 107 107 100 116 107 118 117 121
## 4: 165 164 156 159 151 120 103 114 114 116  95  87  81  65  70  69  69  71
## 5: 105 106  84 115 121 103  94 108 103  91  95 102  90  96 105  92  86  99
## 6: 127 135 128 125 124 136 128 136 146 136 115 104 105 131 126 127 132 121
```

4. *test_labels*

```
head(test.y)
```

```
##    Volcano? Type Radius Number Volcanoes
## 1:        0  NaN    NaN             NaN
## 2:        0  NaN    NaN             NaN
## 3:        1    1  17.00               1
## 4:        0  NaN    NaN             NaN
## 5:        1    3  15.13               1
## 6:        0  NaN    NaN             NaN
```

The image of first observation



Dataset Background:

Finding Volcanoes On Venus.
Kaggle. https://www.kaggle.com/amantheroot/finding-volcanoes-on-venus/data

Methods:

In our project, we are interested in detecting the volcanoes on Venus by analyzing and classifying the images.
- Through our project, we are planning to apply k means clustering method to classify the images of the volcanoes into different categories.
- We will also utilize the linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA) in our project.
Our ultimate goal is to find the best method and build the best model that performs the best classification and has the minimum classification error to classify the images and match up with our label.

Interpretation of our model and prediction

This dataset is originally from NASA's Magellan spacecraft database. For this project, we are using Image datasets (*train_images* and *test_images*) as our features and Label data datasets (*train_labels* and *test_labels*) as our labels. Each label contains 3 variables as mentioned in the Data Description: *Volcano?*, *Type*, *Radius*, *Number Volcanoes*. The variable *Volcano?* is 1 meaning that there is a volcano in the image. Other 3 variables further describe the volcano in the image. However, the variable *Volcano?* is 0 meaning that there is no volcano in the image. Other 3 variables would be not available(NA). First of all, We aim in constructing mainly classification model to predict whether there exist a valcano on each image. If the classification model works well, we will continue to further doing analysis based on the images which we identify as containing at least one volcano to predict the number of volcanoes in the images.
Our target model response are "Is there Volcanoe or not" (this will be done using classfication model) and

"Number of Volcanoes" (this will be done using regression model). Prediction error for classification model will be calculated using classification error, whereas the prediction error for regression model will be calculated using root mean square error (RMSE).

Challenges:

- We are dealing with large datasets (roughly 400 MB in total).
- We will have data visualization by converting the pixel observations into images.
- We will learn volcano knowledge to help us facilitate the process of classifying the volcanoes on venus.
- If we have more time, we want to further identity the number of volcanoes in each image rather than simply detecting if volcano exists in an image.