Data processing

UNIT I

OUTLIER ANALYSIS | STATISTICAL MODEL

Third quarter.

Data 3.A

Teacher: Didier Gamboa.

Team:

- Oswaldo Armin Chan Gonzales.
- Hector Arturo Hernandez Escalante.
- Guillermo Antonio Montes de Oca Rueda.
- Samuel Isaac Venegas Santamaría.
- Alexa Victoria Canche Anaya.
- Lisette Ruíz Peña.

Introduction

- Hey everyone! today we're pleased to be here with you with a tutorial about outliers' detection with statistical method.
- To begin with, what is first an outlier? An outlier is a data that deviates significantly from the rest of the other data, as if it were generated by a different mechanism. We may refer to data objects that are not outliers as "normal" or expected data. Similarly, we may refer to outliers as "abnormal" data.
- In Figure 12.1, most objects follow a roughly Gaussian distribution. However, the objects in region R are significantly different. It is unlikely that they follow the same distribution as the other objects in the data set. Thus, the objects in R are outliers in the data set.

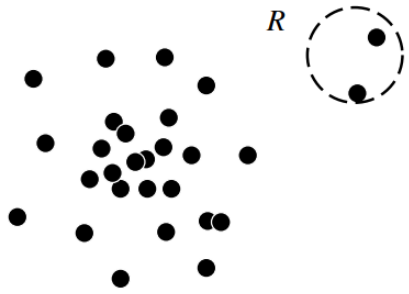


Figure 12.1 The objects in region *R* are outliers.

- Now that we have understood what an outlier is, let's explain the different types that exists, in order to be better at identifying them. We have three different types: global, contextual and collective.
- Global outliers are the easiest ones, since they are globally as their name says, the deviates from the rest of the hole data set. A perfect example would be again the figure 12.1 that we already shown, that from the whole data (the points) the two points in R are the global outliers.
- On the other hand, contextual outliers, depends on the context (who would've guessed, right?), as all the outliers their deviate significantly but here in a specific context. For example: temperature is 8° C, is that an outlier? I don't know, if we'd day that is in Yucatan, Mexico it would definitely be, since we are always melting here.
- Lastly, collective outliers, in the simplest words, would be like a global outlier, but instead they act like a subgroup of the hole data set, with its own mechanism, but they individually may not be considered as outliers.
- An example would be like the Figure 12.2 where the black points together are outliers, but each one of them compared to the rest of the data may not be an outlier.

Collective outliers:

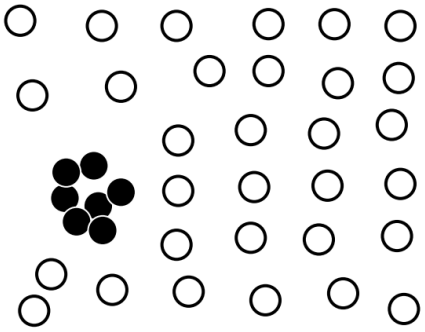


Figure 12.2 The black objects form a collective outlier. *(Jiawei Han & et.al, 2012)*

Statistical Methods

- We are now experts in the outlier department, but to get our diplomas we may need some tools to actually be able to identify them, don't you think? And today we are going to concentrate in the statistical methods.
- And what are they? The general idea behind statistical methods for outlier detection is to learn a generative model fitting the given data set, and then identify those objects in low-probability regions of the model as outliers.
- An example would be taking Figure 12.1 (again), the data follows a Gaussian distribution, for the points in the R region we could try to fit the Gaussian distribution, but the since is very low is unlikely that the those data were generated by the Gaussian model, thus they are outliers.
- However, there are many ways to learn generative models. In general, statistical methods for outlier detection can be divided into two major categories: parametric methods and nonparametric methods, according to how the models are specified and learned.
- So, that we can have our PhD in outliers we need to understand those two categories.
- A parametric method assumes that the normal data objects are generated by a parametric distribution with parameter Θ. The probability density function of the parametric distribution f (x, Θ) gives the probability that object x is generated by the distribution. The smaller this value, the more likely x is an outlier.
- A nonparametric method does not assume an a priori statistical model. Instead, a nonparametric method tries to determine the model from the input data. Note that most nonparametric methods do not assume that the model is completely parameter-free. (Such an assumption would make learning the model from data almost mission impossible.) Instead, nonparametric methods often take the position that the number and nature of the parameters are flexible and not fixed in advance. Examples of nonparametric methods include histogram and kernel density estimation.

Algorithm

- The method that today will be approaching would **parametric**, specifically for **univariate outliers** based on normal distribution, but what it is univariate data? Data involving only one attribute or variable are called univariate data. For simplicity, we often choose to assume that data are generated from a normal distribution. We can then learn the parameters of the normal distribution from the input data and identify the points with low probability as outliers. Let's start with univariate data. We will try to detect outliers by assuming the data follow a normal distribution.
- In this case we will practice with the simplest form to detect outliers, which is using **maximum likelihood**. Using the mean and the standard deviation we can find outliers in a set of values.
- We'll assume that the average of the data follows a normal distribution, which is determined by two parameters: the mean, μ, and the standard deviation, σ. We can use the maximum likelihood method to estimate the parameters μ and σ. That is, we maximize the log-likelihood function

$$lnL(\mu, \sigma^2) = \sum_{i=1}^{n} lnf[x_i|(\mu, \sigma^2)] = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(2\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

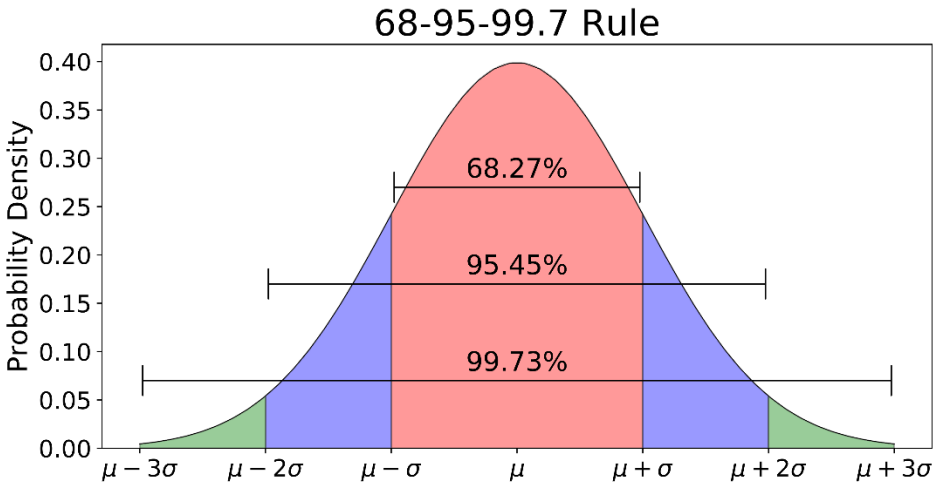
Maximum likelihood methods to estimate the parameter μ and σ

- Taking derivatives with respect to μ and σ and solving the resulting system of first-order conditions leads to the following maximum likelihood estimates:

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- After that what we'll do is that μ ± 3σ region contains 99.7% data under the assumption of normal distribution. So, if it's more than that will take it as an outlier.



- A standard deviation from the mean on the right plus a standard deviation from the mean on the left should contain 68% of your data. Similarly, if instead of one mean deviation, you use 2 standard deviations, you should include 95% of your data, and if you use 3 standard deviations, you should contain 99.7% of the total values. This means that any value that you grab out of those 3 standards deviations, must be within the 0.3% of remaining data, which can be considered as outliers.

Applications (General)

- I n this case, what we are going to is load a data set from 'sklearn', the most popular one is the 'iris'. We are going to take two variables: one that we already know has an outlier and the other doesn't.

```
iris = datasets.load_iris()
```

```
def flowers(value):
    flower = iris.target_names[value]
    return flower

flowers_function = lambda x: iris.target_names[x]
```

```
iris_data = pd.DataFrame(iris.data, columns = iris.feature_names)
iris_data["species"] = iris.target
iris_data.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

- Now, we choose the sepal width and the sepal length variables and their values

```
sepalwidth = iris_data['sepal width (cm)'].values
display('sepal width: {0}'.format(sepalwidth))
sepallength = iris_data['sepal length (cm)'].values
display('sepal length: {0}'.format(sepallength))
```

```
'sepal width: [3.5 3.  3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 3.7 3.4 3.  3.  4.  4.4 3.9 3.5\n 3.8 3.8 3.4 3.7 3.6 3.3 3.4 3.  3.4 3.
5 3.4 3.2 3.1 3.4 4.1 4.2 3.1 3.2\n 3.5 3.6 3.  3.4 3.5 2.3 3.2 3.5 3.8 3.  3.8 3.2 3.7 3.3 3.2 3.2 3.1 2.3\n 2.8 2.8 3.3 2.4
2.9 2.7 2.  3.  2.2 2.9 2.9 3.1 3.  2.7 2.2 2.5 3.2 2.8\n 2.5 2.8 2.9 3.  2.8 3.  2.9 2.6 2.4 2.4 2.7 2.7 3.  3.4 3.1 2.3 3.
2.5\n 2.6 3.  2.6 2.3 2.7 3.  2.9 2.9 2.5 2.8 3.3 2.7 3.  2.9 3.  3.  2.5 2.9\n 2.5 3.6 3.2 2.7 3.  2.5 2.8 3.2 3.  3.8 2.6 2.2
3.2 2.8 2.8 2.7 3.3 3.2\n 2.8 3.  2.8 3.  2.8 3.8 2.8 2.8 2.6 3.  3.4 3.1 3.  3.1 3.1 3.1 2.7 3.2\n 3.3 3.  2.5 3.  3.4 3. ]'

'sepal length: [5.1 4.9 4.7 4.6 5.  5.4 4.6 5.  4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1\n 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.  5.
5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.\n 5.5 4.9 4.4 5.1 5.  4.5 4.4 5.  5.1 4.8 5.1 4.6 5.3 5.  7.  6.4 6.9 5.5\n 6.5 5.7 6.3 4.9
6.6 5.2 5.  5.9 6.  6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1\n 6.3 6.1 6.4 6.6 6.8 6.7 6.  5.7 5.5 5.5 5.8 6.  5.4 6.  6.7 6.3 5.6
5.5\n 5.5 6.1 5.8 5.  5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3\n 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.
6.9 5.6 7.7 6.3 6.7 7.2\n 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.  6.9 6.7 6.9 5.8 6.8\n 6.7 6.7 6.3 6.5 6.2 5.9]'
```

- We assume that the average temperature follows a normal distribution, which is determined by two parameters: the mean, μ, and the standard deviation, σ. Where n is the total number of samples. Taking derivatives with respect to μ and σ and solving the resulting system of first-order conditions leads to the maximum likelihood: mean and standard deviation.

```
mean_sw = sepalwidth.mean()
var_sw = np.var(sepalwidth)
std_sw = sepalwidth.std()

print('Sepal width')
print('Mean: {0}'. format(mean_sw))
print('Variance: {0}'. format(var_sw))
print('Standard deviation: {0}'. format(std_sw))
```

```
Sepal width
Mean: 3.0573333333333337
Variance: 0.1887128888888889
Standard deviation: 0.4344109677354946
```

```
mean_sl = sepallength.mean()
var_sl = np.var(sepallength)
std_sl = sepallength.std()

print('Sepal length')
print('Mean: {0}'. format(mean_sl))
print('Variance: {0}'. format(var_sl))
print('Standard deviation: {0}'. format(std_sl))
```

```
Sepal length
Mean: 5.843333333333334
Variance: 0.6811222222222223
Standard deviation: 0.8253012917851409
```

- The mean and the standard deviation are now calculated, the next step is to seek for the existence of outliers. What we are going to do now it's apply the rule of '68–95–99.7' [more info here], that says to follow a normal distribution the 99.7% of the data supposed to remains within the rule: $\mu \pm 3\sigma$.
- So we have:

```
def outliers(data, data_mean, data_std):
    maximum_right = data_mean + 3*(data_std)
    maximum_left = data_mean - 3*(data_std)
    outliers_list = [i for i in data if (i > maximum_right) or (i < maximum_left)]
    return outliers_list
```

- Now we try the function outliers for both of our variables.

```
1 outliers(sepalwidth, mean_sw, std_sw)
```

```
[4.4]
```

```
1 outliers(sepallength, mean_sl, std_sl)
```

```
[]
```

– Finally, we see that the only variable with an outlier is sepal width, the advantage of this function is that we can see how many and what they are. So, with that we conclude today's video. We hope that you've learned a lot and that you are now experts in outliers and statistical methods. See you next time.

Tasks:

1. Outlier Investigation----------Done by Samuel
2. Basic Classification Investigation----------Done by Samuel
3. Statistical Method Investigation----------Done by Héctor
4. Algorithm investigation and paraphrasing----------Done by Alexa
5. Algorithm investigation----------Done by Oswaldo
6. Application----------Done by Alexa
7. Application----------Done by Lisette
8. Re-write all the information to change it to a script----------Done by Alexa
9. Re-do the example----------Done by Alexa
10. Make a presentation for the video----------Done by Lisette and Héctor
11. Voices in the video----------Done by Oswaldo and Guillermo
12. Edition of the video----------Done by Samuel

Bibliography

Jiawei Han, Micheline Kamber, Jian Pei. (2012) Data Mining. Concepts and Techniques. Third edition. Morgan Kaufmann Publishers of Elsevier. (Chapter 12)

Schaum's Outline of Business Statistics. McGraw Hill Professional. 2003. p. 359,