International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Applied SPSS for Data Forecasting of Flowers Species Name

Aung Cho[1], Aung Si Thu[2], Aye Mon Win[3]

[1,2]University of Computer Studies, Maubin, Myanmar
[3]University of Computer Studies, Hinthata, Myanmar

How to cite this paper: Aung Cho | Aung Si Thu | Aye Mon Win "Applied SPSS for Data Forecasting of Flowers Species Name" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1496-1498, **https://doi.org/10.31142/ijtsrd26665**

IJTSRD26665

ABSTRACT

SPSS is powerful to analyze data clustering and forecasting. This paper intends to support people who are interesting the species of flowers the benefits of data forecasting with applied SPSS. It showed the species value forecasting based on sepal length and sepal width. As SPSS's background algorithms, it showed the KNN algorithm for data clustering and data forecasting. It includes one sample data was downloaded from Google and was analyzed and viewed. It used IBM SPSS statistics version 23 and PYTHON version 3.7.

KEYWORDS: *SPSS is powerful to analyze data clustering and forecasting*

1. INTRODUCTION

Nowadays, flowers businesses are competing with others not to lose their market places in local and external regions. To avoid the loss of market places they should use data science technology. This paper used SPSS integrated with Python software. it showed the KNN algorithm for data clustering and data forecasting that includes three tables, one graph and one data analytical view.

1.1 SPSS

SPSS, standing for Statistical Package for the Social Sciences, is a powerful, user friendly software package for the manipulation and statistical analysis of data. The package is particularly useful for students and researchers in psychology, sociology, psychiatry, and other behavioral sciences, containing as it does an extensive range of both univariate and multivariate procedures much used in these disciplines.

The text is not intended in any way to be an introduction to statistics and, indeed, we assume that most readers will have attended at least one statistics course and will be relatively familiar with concepts such as *linear regression*, *correlation*, *significance tests,* and simple *analysis of variance*. Our hope is that researchers and students with such a background will find this book a relatively self-contained means of using SPSS to analyze their data correctly.[2]

1.2 KNN Algorithm [1]

Nearest Neighbor Analysis is a method for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity. Cases that are near each other are said to be "neighbors." When a new case (holdout) is presented, its distance from each of the cases in the model is computed. The classifications of the most similar cases – the nearest neighbors – are tallied and the new case is placed into the category that contains the greatest number of nearest neighbors. You can specify the number of nearest neighbors to examine; this value is called *k*. The pictures show how a new case would be classified using two different values of *k*. When *k* = 5, the new case is placed in category *1* because a majority of the nearest neighbors belong to category *1*. However, when *k* = 9, the new case is placed in category *0* because a majority of the nearest neighbors belong to category *0*. Nearest neighbor analysis can also be used to compute values for a continuous target. In this situation, the average or median target value of the nearest neighbors is used to obtain the predicted value for the new case.

1.3. Scatter grams

SPSS has plotted the relationship between two variables:

1. revision: How much revision an individual did for an exam (scored from 0 to 30 hours) by
2. Score: Exam Score (marked out of 30). on to a chart called a scatter gram.

These two variables are not identical, but you would expect a strong positive relationship between them. When you have a 'positive relationship', the low values on one variable tend to go with low values on the other variable and high values on one variable tend to go with high values on the other. Here, we would expect that people doing only a small amount of revision for an exam should tend to score low on the exam, and people doing a lot of revision should tend to score high. When you have a 'negative relationship', the low values on one variable tend to go with high values on the other variable and high values on one variable tend to go with low values on the other.[4]

2. Algorithm

KNN Algorithms [1]

Notation

The following notation is used throughout this chapter unless otherwise stated:

Y	Optional $1\times N$ vector of responses with element y_n, where $n=1,\ldots,N$ indexes the cases.
X^0	$P^0\times N$ matrix of features with element x^0_{pn}, where $p=1,\ldots,P^0$ indexes the features and $n=1,\ldots,N$ indexes the cases.
X	$P\times N$ matrix of encoded features with element x_{pn}, where $p=1,\ldots,P$ indexes the features and $n=1,\ldots,N$ indexes the cases.
P	Dimensionality of the feature space: the number of continuous features plus the number of categories across all categorical features.
N	Total number of cases.
$N_j, j=1,2,\cdots,J$	The number of cases with $Y=j$, where Y is a response variable with J categories
$\hat{N}_j$	The number of cases which belong to class j and are correctly classified as j.
$\hat{N}^*_j$	The total number of cases which are classified as j.

Preprocessing

Features are coded to account for differences in measurement scale.

Continuous

Continuous features are optionally coded using adjusted normalization:

$$x_{pn} = \frac{2\left(x^0_{pn} - \min\left(x^0_p\right)\right)}{\max\left(x^0_p\right) - \min\left(x^0_p\right)} - 1$$

where x_{pn} is the normalized value of input feature p for case n, x^0_p is the original value of the feature for case n, $\min\left(x^0_p\right)$ is the minimum value of the feature for all training cases, and $\max\left(x^0_p\right)$ is the maximum value for all training cases.

Categorical

Categorical features are always temporarily recoded using one-of-c coding. If a feature has c categories, then it is is stored as c vectors, with the first category denoted (1,0,...,0), the next category (0,1,0,...,0), ..., and the final category (0,0,...,0,1).

Training

Training a nearest neighbor model involves computing the distances between cases based upon their values in the feature set. The nearest neighbors to a given case have the smallest distances from that case.

Euclidean Distance. The distance between two cases is the square root of the sum, over all dimensions, of the weighted squared differences between the values for the cases.

$$Euclidean_{ih} = \sqrt{\sum_{p=1}^{P} w_{(p)}\left(x_{(p)i} - x_{(p)h}\right)^2}$$

3. Testing

A. Table-1: Given Data

Sepal Length	Sepal Width	Species
5.3	3.7	setosa
5.1	3.8	setosa
7.2	3	virginica
5.4	3.4	setosa
5.1	3.3	setosa
5.4	3.9	setosa
7.4	2.8	virginica
6.1	2.8	versicolor
7.3	2.9	virginica
6	2.7	versicolor
5.8	2.8	virginica
6.3	2.3	versicolor
5.1	2.5	versicolor
6.3	2.5	versicolor
5.5	2.4	versicolor

B. Unknown Point (Holdout data)

Sepal_Length=	5.2
Sepal_Width=	3.1
Species	?

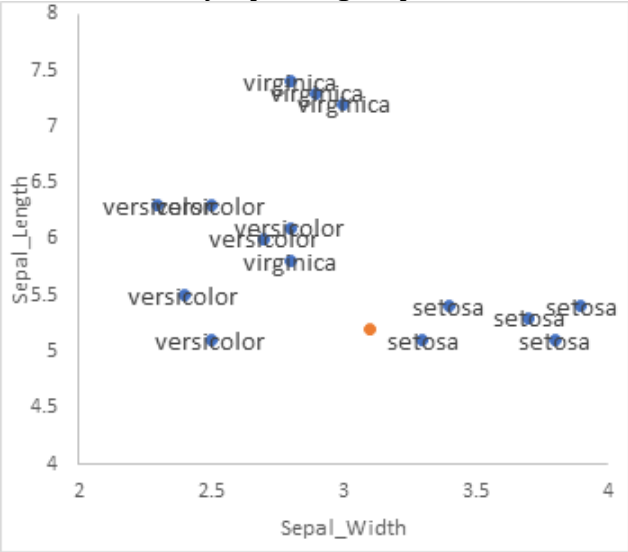
C. Table-2: Calculate distances and nearest(Training)

Rank(K)	Euclidean Distance	Species
3	0.60828	setosa
6	0.70711	setosa
13	2.00250	virginica
2	0.36056	setosa
1	0.22361	setosa
8	0.82462	setosa
15	2.22036	virginica
10	0.94868	versicolor
14	2.10950	virginica
9	0.89443	versicolor
5	0.67082	virginica
12	1.36015	versicolor
4	0.60828	versicolor
11	1.25300	versicolor
7	0.76158	versicolor

D. Table-3: Ascending Order by Rank(K)

K	Species
1	setosa
2	setosa
3	setosa
4	versicolor
5	virginica
6	setosa
7	versicolor
8	setosa
9	versicolor
10	versicolor
11	versicolor
12	versicolor
13	virginica
14	virginica
15	virginica

E. Graph-1: Given species and unknown species distribution by Sepal_Length*Speal_Width



F. Analytical views

At first, trained the data of table-1 with unknown data (holdout data) by using KNN algorithm and got table-2 that includes the distances between given data and unknown point.

In second, sorted the distances with ascending order by nearest distances or Rank(K) to get table-3 that is clear to look at result. The nearest species group is setosa-group for unknown data.

As final, look at table-3 and graph-1. You can decide the species value for unknown data that is setosa.

4. Conclusion

SPSS data analysis tools are valuable in social science, flowers business and marketing fields. It is very good for presentation report by graphical design. It showed the flowers species value forecasting based on sepal width and sepal length by using KNN algorithm that can cluster and forecast unknown data based on known data and then they can get their goal with right flower species and can avoid the waste of time to know flowers species and the loss of market places in local and global regions by using SPSS software.

References

[1] IBM SPSS Statistics 24 Algorithms pdf book [book style]

[2] A handbook of statistical analyses using SPSS / Sabine, Landau, Brian S. Everitt, ISBN 1-58488-369-3 [book style]

[3] SPSS For Dummies®, 2nd Edition, ISBN: 978-0-470-48764-8 [book style]

[4] SPSS for Social Scientists Robert L. Miller, Ciaran Acton, Deirdre A. Fullerton and John Maltby, ISBN 0–333–92286–7 [book style]

<Profile>

Aung Cho received the B.A.(Eco) degree from Yangon University in 1987 and M.I.Sc.(Information Science) degree from University of Computer Studies, Yangon in 2001. After got Master degree, I served as a teacher at the software, information science and application departments of the computer universities. I am now with University of Computer Studies, Maubin.

Aung Si Thu received the B.Sc.(Hons)(Chemistry) degree from Magwe University in 2003 and M.I.Sc.(Information Science) degree from University of Computer Studies, Yangon in 2009. After got Master degree, I served as a teacher at the software, information science and hardware departments of the computer universities. I am now with University of Computer Studies, Maubin, Myanmar.