PYNALYTICS APPLICATION: GUIDE FOR REGRESSION, NAIVE BAYES AND K-MEANS

Table of Contents

Regression, Naive Bayes and K-means

Regression, naive Bayes and k-means are the processes that fall under advanced data analytics which were used to implement the showing of the visualizations and numerical results from structured data to the application. Below are information about each process and how they work and what they generate.

A. Regression

Regression is one of the most widely used statistical techniques in data analytics. It involves the identification of the relationship between one or more independent variables and a dependent variable (Ostertagová, 2012). Typically, regression analysis is used for modeling the relationship between variables, predicting a target variable and testing hypotheses (Hyon, 2017). In simple linear regression, each observation consists of only two variables of interest known as the independent and dependent variable, while in multiple linear regression, each observation may consist of multiple variables where there may be several independent variables to be paired with only one dependent variable. Multiple regression models describe how a single response variable depends linearly on a number of predictor variables (Hyon, 2017). Polynomial regression is a special technique for instances having only one independent variable. It is still considered to be a linear model, but the curve being fitted in polynomial regression is quadratic instead of linear (Agarwal, 2018).

Main points:

1. An independent variable and the dependent variable is used to obtain the regression equation. The main objective is to see if there is a relationship as well as the correlation between the two variables.

2. For each of the processes and concept in regression analysis, scatter plots will be created on the graph using the values of the dependent variable and the chosen independent variable. For simple linear regression, a least squares regression line can be generated and seen on the graph. The least squares regression line is the best-fitting line formed in between the points based on a position where it is closest to all the points. Additionally, in order for a line to be considered as the regression line, the value of the distance between each of the points and the line formed between the points must be at its minimum.

3. By looking at the least squares regression line generated using simple linear regression, it is shown whether there is a definite relationship between the variables. When the line is leaning upward right, there is a positive linear relationship where the independent variable goes up and dependent variable also goes up. When the line is leaning downward right then there is a negative linear relationship where the independent variable goes up and dependent variable goes down. When there is a horizontal least regression line, there is no correlation. This means that the independent variable is not important to the dependent variable where at any point, a person can choose any independent variable she or he wants but the dependent variable will always be around the same value. The horizontal least squares regression line only shows that the independent variable values are not important, but it may show a definite trend. There is a good-fit line, thus, the independent variable cannot possibly have any relationship with the dependent variable (Stapel, 2018) The slope of the least squares regression line represents the effect of the independent variable on the dependent variable.

4. Polynomial regression is used for cases where a simple linear regression model is the best choice (Ostertagová, 2012). It can be used in handling complex distribution of data considering that not all linear models are used to fit non-linear data. A line generated on simple linear regression may not be able to capture the patterns in the data which can be an example of under-fitting. Under-fitting is based on high bias. A bias is the error generated by the least squares regression model's assumptions about fitting the data. On the creation of the framework, the relationship between the independent variable and the dependent variable is visualized as 2nd degree polynomial in the independent variable. A quadratic equation was generated under this process (y = ax^2 + bx + c) where a, b and c are constants and the graph will be dependent to the constant a. The graph will be composed of a parabola. The larger constant a becomes, the wider the parabola is. If a is positive or greater than zero then a parabola opens upward and if a is negative or less than zero, the parabola opens downward (Dawkins, 2018).

5. A summarized table for simple and multiple linear regression can be generated. It contains numerical results of the adjusted r-squared, Pearson's R, and the p-value. The summarized table for polynomial regression includes numerical results of Pearson's R and R-squared.

- R-squared or the coefficient determination is the percentage of variance that the least squares regression model explains. It is a statistical measure of how well the regression line approximates the true or real data points. It ranges from 0% to 100%. The adjusted r-squared serves as a main basis of prediction. The adjusted r-squared was based on the number of observations and the degrees of freedom of the residuals. The adjusted r-squared is the r-squared statistic, it was adjusted for the number of the coefficients in the least squares regression model. The values of it is used to have a comparison of the models having a different number of coefficients.
- Pearson's R - There is a "const" word on the OLS statistical model results meaning a constant term. The researchers of Pynalytics microframework added a constant term because without it, the model will be forced to go through the origin. Also, after applying the constant term, the y-intercept will be shown on the table. The Pearson's r is an index that ranges from -1 to 1. This research has used the Pearson's correlation degree of correlation (Sedgwick, 2012) on explaining the numerical results, specifically on the ways of looking in correlation of the data. The figure below explains the degree.

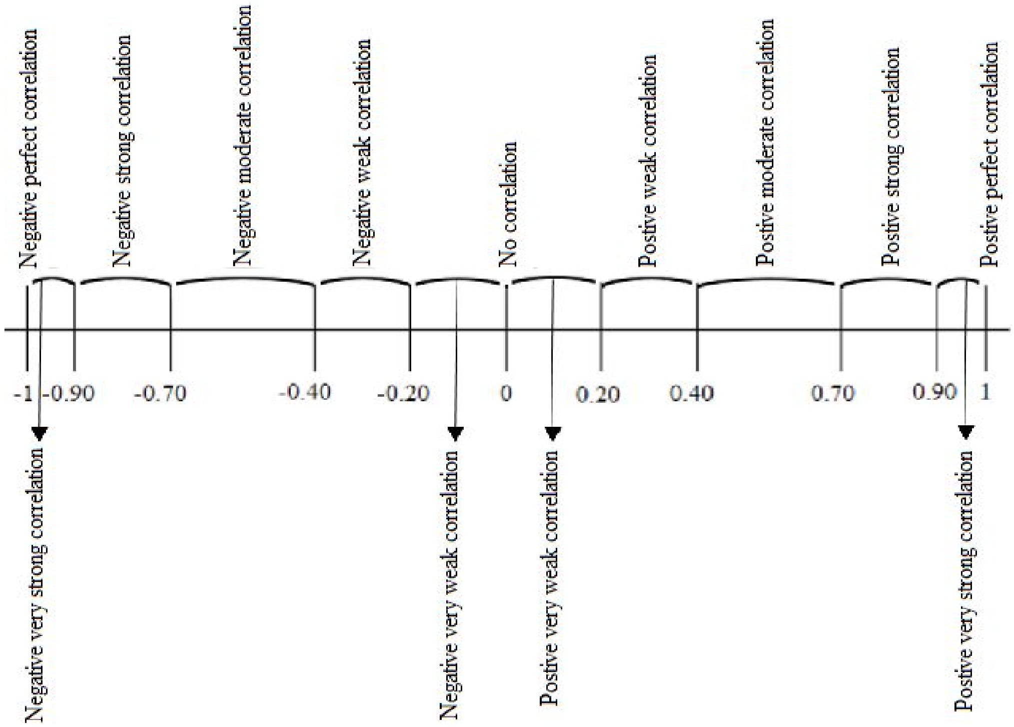
Negative perfect correlation
Negative strong correlation
Negative moderate correlation
Negative weak correlation
No correlation
Postive weak correlation
Postive moderate correlation
Postive strong correlation
Postive perfect correlation
-1
-0.90
-0.70
-0.40
-0.20
0
0.20
0.40
0.70
0.90
1
Negative very strong correlation
Negative very weak correlation
Postive very weak correlation
Postive very strong correlation

Figure 1: Pearson's correlation degree of correlation

Based on the figure above, a perfect correlation is having either a positive one or negative one value. A very strong correlation lies between the Pearson’s r of positive and negative 0.99 to 0.90. A strong correlation lies between the Pearson’s r of positive and negative 0.70 to 0.89.A moderate correlation lies between the Pearson’s r of positive and negative 0.40 to 0.69.A weak correlation lies between the Pearson’s r of positive and negative 0.20 to 0.39.A very weak correlation lies between the Pearson’s r of positive and negative 0.01 to 0.19. (Garambas, 2015) There is no relationship when the Pearson’s r is zero or simply no correlation. As its value gets closer to plus or minus one, the relationship is strong. When the value between two variables is one or negative one, it indicates either a positive or negative perfect linear relationship. The meaning of a positive Pearson’s r means that as the variable 1 increases or decreases, the variable 2 also increases or decreases. If there is a positive correlation, the variable moves in the same direction. Meanwhile, negative correlation means that as the variable 1 increases, the variable two decreases and vice versa, which indicate that the variables move in opposite directions when it exists.

- The t scores and p-values which are used for hypothesis test. The t scores are the measure of how “Statistically significant” the coefficient is. P >|t| is the p-value. If the p-value for a variable is less than the significance level of 0.05, the sample data provide evidence to have a rejection of the null hypothesis. A low p-value means it is less than 0.05 and it will be considered that the relationship of the two that are tested is “Statistically significant”. This explains that you can reject the null hypothesis. Meanwhile, a high p-value which means it is greater than 0.05, then the relationship of the two that are tested will be predicted as “Statistically Insignificant”.

B. Naive-Bayes

Naive Bayes classification is the application of the naive Bayes classifier. The naive Bayes classifier is a machine learning model that is used to predict a class based on specific features. It is based on Bayes theorem which states that an event can happen given that another event has occurred. However, it is called naive because the assumption when using the naive Bayes classifier is that predictors and features are independent and do not affect each other (Gandhi, 2018).

Main points:

Naive Bayes classification was used to create a predictive model with the purpose of predicting the degree of the number of data. It uses the principles of Bayes' theorem which states that an event can happen given that another event has occurred.

However, it is called naive because the assumption when using the naive Bayes classifier is that predictors and features are independent and do not affect each other (Gandhi, 2018)

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots x_n \mid y)}{P(x_1, \dots, x_n)}$$

Figure 2: Formula of naive Bayes

Below are the steps that were followed along with the scikit-learn library functions used:

1. **Scaling**

   The features were first scaled using MinMaxScaler to normalize the data.

2. **Binning**

   If the given value or the class/target features were continuous/numerical data, it is necessary to transform and assign their values into n number of classes/bins to apply naive Bayes classification. There are many strategies that can be used for binning and three of which are provided by the KBinsDiscretizer function (KBinsDiscretizer, n.d.) and 2 of them were used for this research.

   - Uniform - used when bins have identical widths/ranges
   - K-means - used when the values in each bin have the same nearest center of a 1d k-means cluster

   When the data used for the manipulation are unbalanced. This makes most of the data certain to be grouped into a single bin. In order to make up with this, it was necessary to have the number of bins as low as possible. Opting to go with 2 bins and 2 classes (low and high) is not ideal because the maximum and the minimum number of the data can be significantly varied which makes it critical to have inaccurate predictions. This makes 3 the optimal number of classes so as to have a compromise between the imbalanced and significantly varied data.

3. Feature Selection

Feature selection was done using the RFECV (recursive feature selection with cross validation) which uses supervised estimators to provide coef_ or feature_importance_ values as a basis for selecting and using k folds cross validation to determine the best number of features. The estimators used was Logistic Regression.

4. Modelling and Cross-Validation

The data from the structured csv file can serve as features and target features for modelling. Since the structured csv file contains continuous data, Gaussian naive Bayes was used. In Gaussian naive Bayes, there is an assumption of normal distribution. This makes it best used for data sets where all the features are continuous (Albon, 2017). Thus, a Gaussian naive Bayes model was created using the GaussianNB function. (Pedregosa et al.,2011)

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Figure3 : Formula of Gaussian naive Bayes

K-fold cross-validation was used for training the model in order to avoid overfitting (“3.1. Cross-validation: evaluating estimator performance, n.d”). For the number of folds, 5 or 10 was used since both yield test error rates that do not suffer from excessive high bias and from very high variance (James et al., 2013).

5. Presentation and Evaluation

A confusion matrix plot was created to see the number of points that were predicted successfully and unsuccessfully by the model. At the same time, the performance of the model was computed using the balanced_accuracy_score function. The balance accuracy score computes the average recall of each class which is an essential metric for evaluating imbalanced data sets. (Buitinck et al., 2013). Since cross-validation was applied, the accuracy and average recall in the confusion matrix were not the same.

$$\texttt{balanced-accuracy}(y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \sum_i 1(\hat{y}_i = y_i)\hat{w}_i$$

Figure 4: Formula of balance-accuracy

C. K-means

K-means clustering is a method used for clustering data. Clustering means grouping data together in a manner that data in one cluster are more similar to each other than to data in other clusters. In k-means clustering, the data set is grouped based on a predefined k number of clusters. Each cluster is represented by a centroid which is the data point found in the middle of a cluster (Paul, 2018).

Main Points:

K-means clustering was used to cluster the data and generate centroids. This enabled the summarization of the data for the purpose of studying their patterns. The data came from preselected combinations of the features. The data was normalized before it was clustered into a selected k number of clusters. Silhouette analysis was applied in feature selection and also in measuring the performance of the clustering.

- **Silhouette Analysis**

Silhouette analysis is used to measure the performance of clustering through generating a silhouette coefficient. The silhouette coefficient is a measure of the distance between the clusters. It has a range of -1 to 1. A silhouette coefficient close to 1 indicates that the clusters are far from each other. Meanwhile, a silhouette coefficient close to 0 indicates that the clusters are overlapping or that the separation between the clusters is becoming less distinct. A negative value implies that data might be in the wrong cluster (“Selecting the number of clusters with silhouette analysis on KMeans clustering”, n.d.).

Found on the next page are the steps that were followed for performing K-means clustering along with the functions used from the scikit-learn Python library:

1. Data normalization

The data was normalized using the MinMaxScaler. The purpose of normalization is to transform the data into an ordinary scale without causing distortion (Jaitley, 2018).

2. Selection of k

The value chosen as the k number of clusters was 3 to indicate each cluster as either “low”, “medium”, or “high”.

3. Feature selection using silhouette analysis

In machine learning algorithms, the word “feature” refers to a component being used. Feature selection was used to select these other data. Silhouette analysis was applied to select the best data to be used in clustering. The silhouette coefficient was generated through using the silhouette_score function.

4. K-means clustering

The features were clustered according to the chosen value of k. The data points nearest each other were clustered together. The centroid of each cluster is shown through a centroid chart. Since the centroids represent each cluster, a pattern regarding each data was interpreted from them. In addition, scatter plots for the features used and the silhouette coefficient from clustering all the features are generated to show the performance of the clustering. The clustering was accomplished through the use of the KMeans function while the silhouette coefficient was generated again through the use of the silhouette_score function.

References

Agarwal, A. (2018, October 8). Polynomial Regression [Web log post]. Retrieved April 12, 2019, from https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

3.1. Cross-validation: Evaluating estimator performance ... (n.d.). Retrieved April 1, 2019, from https://scikit-learn.org/stable/modules/cross_validation.html

Albon, C. (2017, December 20). Gaussian naive Bayes Classifier. Retrieved from https://chrisalbon.com/machine_learning/naive_bayes/gaussian_naive_bayes_classifier/

Bronshtein, A. (2017, May 08). Simple and Multiple Linear Regression in Python. Retrieved March 20, 2019, from https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... & Layton, R. (2013). API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238.

Dawkins, P. (2018, June 2). Parabolas. Retrieved February 27, 2019, from http://tutorial.math.lamar.edu/Classes/Alg/Parabolas.aspx

Gandhi, R. (2018, May 5). naive Bayes Classifier. Retrieved from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Garambas, Z. (2015). Basic Probability and Statistics. Baguio City.

Hyon, J. (2017). Regression analysis. Regression Analysis - Basics. Retrieved April 2, 2019, from http://www.uta.fi/sis/mtt/mttta6/reg17lecturen.pdf?fbclid=IwAR1mdg5JoaCaYlNRFQmhv_BbwwZg4u9SR8h9SVfkLrU1F9QDV4GcpyrK-58

Jaitley, U. (2018, October 7). Why Data Normalization is necessary for Machine Learning models. Retrieved from https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (p. 184). Retrieved from https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

KBinsDiscretizer. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.KBinsDiscretizer.html#sklearn.preprocessing.KBinsDiscretizer

Nerdy. (2018, May 12). What Is R Squared And Negative R Squared. Retrieved April 1, 2019, from http://www.fairlynerdy.com/what-is-r-squared/

Ostertagová, E. (2012). Modelling using polynomial regression. Procedia Engineering, 48, 500-506.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Prettenhofer, P. (2014, February 07). Ordinary Least Squares in Python. Retrieved April 1, 2019, from https://blog.datarobot.com/ordinary-least-squares-in-python

Sedgwick, P. (2012). Pearson's correlation coefficient. Bmj, 345, e4483.

Selecting the number of clusters with silhouette analysis on KMeans clustering. (n.d.) Retrieved April 13, 2019, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Stapel, E. (2018). Scatterplots: Correlation, Outliers, and Model Types. Retrieved April 1, 2019, from https://www.purplemath.com/modules/scattreg2.htm

Zhao, Y., Wong, Z. S., & Tsui, K. L. (2018). A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. Journal of Healthcare Engineering, 2018, 1-11. doi:10.1155/2018/6275435