About Me

- Name : Görkem Temizel
- University : Senior Student, Bachelor of Computer Engineering, Kocaeli University
- Project: Dimensionality Reduction for Studying Diffuse Circumgalactic Medium
- Contact : gorkemtemizell@gmail.com
- Time Zone : GMT+03:00
- Github : https://github.com/MrFinchh
- Technical Experiences

Abstract

This project aims to apply dimensionality reduction techniques for the dataset simulated quasar absorption spectra. We will begin with simple dimensionality reduction techniques such as univariate feature selection methods, wrapper, and embedded-based feature selection methods. Then we're going to focus on data transformation methods also known as matrix factorizations. In this section, we can focus on linear and nonlinear methods. Linear methods contain PCA, fast ICA, LDA, etc. Nonlinear methods contain Nonlinear PCA, Kernel PCA, Auto Encoders, t-sne, etc. When the program nears the end we're going to look for correlation-based feature selection methods and at the end of the project we can compare all methods and their results. With those results, we can select one of them to propose.

Why this project ?

CGM is simply the component of galaxies that will help us understand the evolution of galaxies more closely. The CGM is a source for a galaxy's star-forming fuel, a galactic feedback and recycling site, and perhaps the primary regulator of galactic gas supply. [1] All redshifts and information extracted from the electromagnetic spectrum show that CGM gas is the key to galaxy evolution. The CGM response involves the regulation of gas flow in and out of galaxy evolution problems. Although these problems can be solved in other ways, their high-z responses and solutions require understanding the CGM and the flows that fed it in all cosmic ages. [2]

Because of all these, understanding CGM will lead to a better understanding of galaxy evolution, so the more meaningful and fast we interpret the data obtained from CGM-related simulations, the more agile and accurate decisions we can make. With dimensionality reduction, we can get results about CGM faster and with a preserved success rate.

Why me ?

I believe that I am the most suitable candidate for this project, as I have been working on dimension reduction for a long time. As a result of these studies, I believe that I have grasped this subject well in terms of both theoretical and implementation. This is where my difference from other people emerges. While others are content with applying the most familiar methods, PCA, correlation-based methods, I in turn aim to evaluate all the options available and maintain success while doing them. I can say, relying on the success I have shown with the Chi-square feature selection method in the previous assessment task, I believe that I will get good results in this task and I am ready to work for it.

Recent Researches on CGM

Studies on stellar explosions show that Starbursts appear to act as an important baryon reservoir for C IV absorbers for CGM. The high detection rate of such highly ionized material is an indication that the CGM can affect radii as large as 200 kpc. These results were mentioned in the study as the first direct and observational evidence of gaseous states of local stellar explosions, and it was mentioned that they could be useful in answering questions about galaxy evolution.[3]

In a study covering the H I Lyman boundary of 14 CGM systems, it is mentioned that the new distant ultraviolet spectra from the COS-Halos questionnaire of 13 quasars were analyzed. This study applies the Monte-Carlo Markov chain approach to estimate the metallicity of CGM gas. In this study, it was mentioned that the transmission of ionizing radiation alone with CGM gas is 70 ± 7%, there are unexpected increases in NH1 gas metallicity at 99.9% confidence interval and this corresponds to early enrichment.[4]

Finally, in two studies, various studies were carried out on reducing the dimensions of CGM data with NMF (non negative matrix factorization) method.[5, 6]

Technical Details

In this section, I will talk about the algorithms and tools that I will use in the important stages of the project.

Dimensionality Reduction Algorithms

1. **Feature Selection**
    a. Filter Based Methods
    b. Wrapper Based Methods
    c. Embedded Based Methods
2. **Data Transformation (Matrix Factorization, Data Projection)**
    a. Linear Transformations
    b. Nonlinear Transformations
3. **Correlation Based Feature Selection**

Pros and Cons

Filter Based Methods

- Pros
  - Easy to apply and understand.
  - After selection we can check which features selected. So we can interpret which feature is important.
  - It is up to us how many features we can select or eliminate.
  - Not expensive for computational cost.
- Cons
  - Not oriented to models.

Wrapper Based Methods

- Pros
  - Model oriented.
  - We can get selected feature names and rankings(in RFE).
  - It can work with cross-validation techniques to compare and accept metrics for comparing.
- Cons
  - Takes huge amount of time.
  - Very expensive for computational cost.

Embedded Based Methods

- Pros
  - Can get feature names and feature importances. This can help also interpreting features.
- Cons
  - Not the cheapest also the most expensive for computational cost.

Linear Transformations

- Pros
  - Most of them work very fast.
  - Helps to remove collinearity.
  - Not expensive for computational cost.
- Cons
  - We can’t get feature names because actually it’s create new features.
  - Most of them need scaling.
  - Can only detect linear relations.

Nonlinear Transformations

- Pros
  - Can detect nonlinear relations.
- Cons
  - We can’t get feature names because actually it’s create new features.

Correlation Based Feature Selection

- Pros
  - Easy to understand.
  - Works fast, low computational cost.

- Cons
    - Only works for linearly correlated data.
    - Can be misleading.

Deliverables

Deliverables include data analysis, data visualizations, feature engineering, feature selection and data transformations, model selection, hyperparameter tuning, final model development and well documented jupyter notebooks and python files.

Data Analysis
- Skewness, Kurtosis, Z-Score Analysis
- Correlation analysis
- Nan values, empty values, infinity values, outlier analysis

Data Visualization
- ROC Curve
- Confusion Matrix
- Distribution of Data
- Errors, Scores in model selection, evaluation and final model development
- Correlation Heatmap

Feature Engineering
- Creating new features
- Testing new feature’s importance

Feature Selection and Data Transformations
- Filter, Wrapper, Embedded based methods
- Linear, Nonlinear data transformations
- Correlation based methods

Model Selection, Hyper Parameter Tuning, Final Model Development
- Include many models to compare errors and scores.
- Tune 2-3 models and seek the best model based on metrics.
- Test final model and show the final result.

Documentation
- All studies will be documented for easy follow-up and understanding.

Timeline

I can work 20 hours a week. I plan to use these 20 hours a week, working 4 hours 5 days a week.

Before June 7 - Community Bounding

At this stage, I will be in contact with the community and gather information about the route the works will follow. I will try to collect more professional information about feature selection in the company where I am currently working and in the places where I do internship, and try to maximize the work I will prepare for GSoC.

June 7 - June 14

- If the data is not a blackbox, that is, if the names of the properties are available, search for these properties and become familiar with the data in terms of meaning, to recognize the data.
- To work on the analysis of the data. It is among my works to note the predictions we will gain from this statistical analysis.
- To search for empty data and to perform correlation analysis.
- Asking questions about the data and trying to develop approaches to create new features.

June 14 - June 21

- Realization of visualization of statistical analysis.
- Detection of outlier data and, if necessary, correction.
- Correction or removal of NaN data.

June 21 - June 28

- To reveal the features I plan to create.
- Creating new properties and examining the correlations of these properties.
- To test new features in simple models, to examine their importance priorities and to observe the contribution of these features to the model.

June 28 - July 5

- Testing the success of the models by working with the default versions of many models on data without dimensioning.
- Again, to observe the success of many models by applying the default versions of filter based feature selection methods.
- Visualizing the transactions made and making it easier to compare them in terms of metrics.

July 5 - July 12

- Application of Wrapper based feature selection methods, interpretation of results.
- Application of embedded based feature selection methods, interpretation of results.
- Visualization of the results according to the feature selection processes.

July 12 - July 19

- Reviewing the work I have done and arranging the missing places and fixing them.

- To make sure that the studies are traceable, to update the works in terms of documentation.
- Communicating with the mentors and reflecting their evaluations on the studies.

July 19 - July 26

- Performing Hyperparameter tuning processes.
- Comparing the results according to metrics and deciding on the actual model.
- Visualizing the results and presenting which parameter space is decided through visualizations.

July 26 - August 2

- Apply linear data transformations as known as feature extraction or matrix factorization, feature selection methods like PCA, fast ICA, LDA, OLS etc.
- Creating a traceable dimension reduction table by comparing the obtained results according to metrics.
- Visualization of the results of feature selection results.

August 2 - August 9

- Apply nonlinear data transformations methods like Auto Encoders, t-sne etc.
- Examining the results obtained in dimension reduction and metrics.
- Visualization of the feature selection results.

August 9 - August 16

- Apply correlation based feature selection methods like VIF, FCBFS and linear correlation.
- Presenting all feature selection processes with various metrics and values in a table so that they can be followed easily.
- Visualization of the feature selection results.
- Since this week is the last week, I will talk to the mentors and ask if there is any point in the project that they would like me to correct or change and I will make changes in this direction.

August 16 - August 23

This week, I will try to briefly explain the summary of the work, its position at the beginning and the end, the actions taken. This week, I will make improvements and convey the work to make my work understandable.

Technical Experiences

June 2019 - July 2019

- Starting to work with OpenCV, numpy and relevant libraries for image preprocessing.
- Understanding of the basics of distributed systems and working with it.

November 2019 - May 2020

- Worked as part time DBA in Kuveyt Turk Participation Bank Research and Development center.

May 2020 - August 2020

- Completed many Data Science courses and made projects to improve myself.
- I joined the Deep Learning Turkey community after finishing Google Machine Learning Crash course successfully.

August 2020 - Present

- I joined our school's artificial intelligence and simulation systems research and development laboratory in August 2020 and I am still working there.
- In the school laboratory, we are working on 3 different subjects under the title of Radiomics. Here, we carry out studies in order to get more successful results by reducing them with feature selection processes from data with more than 110 features. Our article writing process continues.

September 2020 - Present

- Our school laboratory undertook the consultancy duties of a company within Kocaeli Technopark and at the beginning of this consultancy, I, a graduate student and our assistant professor supported this company in the field of data science.
- The goal of the project was to produce an artificial intelligence study on the locations of the customers and the density of the locations for taxi drivers. Since the graduate student graduates after a while, we continue to work with my assistant professor on the project.

November 2020 - February 2021

- I worked at Tiko as a machine learning technical leader intern. Throughout my role, I supported my teammates on machine learning and tried to answer their questions. At the end of my internship, the pipeline I realized for Tiko was appreciated by the authorities and I started working as a part time machine learning practitioner at Tiko.

February 2021 - Present

- At Tiko, I am part of a team of 2 people responsible for all data science processes. In this team, we organize and update the whole pipeline from end to end. In a short time, we not only solve the old mistakes and provide an improvement of up to 6% in the project, we continue to offer solutions.
- In terms of feature selection, we have accelerated our processes without any loss of success as a result of approximately 50% size reduction in the data with more than 110 columns at Tiko. We are still continuing our work in this area.
- I started doing an internship at Arçelik to complete my school internship. Here, I had the opportunity to work with AWS-Deep Racer in Reinforcement Learning, which I have not worked before.

- I also conducted a review on Quality Intelligence projects at Arçelik. While doing this review, I learned about many issues in the field of NLP.
- I conducted an NLP-based study with the Customer of Voice project in Arçelik within a period of 2 weeks.

References

[1]. Tumlinson, J., Peeples, M. S., & Werk, J. K. (2017). The circumgalactic medium. Annual Review of Astronomy and Astrophysics, 55, 389-432.

[2]. Mas-Ribas, L., Riemer-Sørensen, S., Hennawi, J. F., Miralda-Escudé, J., O'Meara, J. M., Pérez-Ràfols, I., ... & Webb, J. K. (2018). Origin of Metals around Galaxies. I. Catalogs of Metal-line Absorption Doublets from High-resolution Quasar Spectra. The Astrophysical Journal, 862(1), 50.

[3]. Borthakur, S., Heckman, T., Strickland, D., Wild, V., & Schiminovich, D. (2013). The impact of starbursts on the circumgalactic medium. The Astrophysical Journal, 768(1), 18.

[4]. Prochaska, J. X., Werk, J. K., Worseck, G., Tripp, T. M., Tumlinson, J., Burchett, J. N., ... & Tejos, N. (2017). The COS-halos survey: metallicities in the low-redshift circumgalactic medium. The Astrophysical Journal, 837(2), 169.

[5]. Zhu, G., & Ménard, B. (2013). Calcium H & K induced by galaxy halos. The Astrophysical Journal, 773(1), 16.

[6]. Murga, M., Zhu, G., Ménard, B., & Lan, T. W. (2015). Calcium H&K and sodium D absorption induced by the interstellar and circumgalactic media of the Milky Way. Monthly Notices of the Royal Astronomical Society, 452(1), 511-519.