Peer Review Report

1. The statistic explanation of the US crime data (YAN, Ningyu; YAN, Bokai; LAI, Yanming)
2. Summary: The report presents a statistical analysis of the crime rate in the US, aiming to identify the primary factors contributing to the average crime rate using crime data. The dataset used in the analysis includes 85 American cities' crime information from 1969 to 1992, which underwent preprocessing techniques to simplify it. The study then utilized principal component analysis (PCA), Isometric Mapping (Isomap), and uniform manifold approximation and projection (UMAP) techniques to reduce the number of variables to 5 and 11. Visualization results from the three methods revealed that the 85 cities were well separated into several groups. The report concludes that the US crime rate is influenced by several factors, such as population structure, police numbers, mayoral term, and society welfare. Finally, the linear regression model was used to establish a relationship between the "principal components" and the average crime rate.
3. Strengths: The report provided a clear and concise summary of the statistical analysis conducted on US crime rate using PCA, Isomap, and UMAP. It presented the results in a visually appealing manner with the aid of figures and tables. Instead of using either linear or non-linear methods for dimensionality reduction, the report compared and discussed the advantages and limitations of PCA, Isomap, and UMAP methods in data reduction. Finally, the report included a linear regression model to predict the average crime rate.
4. Weaknesses: The report did not discuss the limitations and potential biases of the methodology used or the results obtained. And it did not provide a discussion of the implications of the findings for the broader field of criminology or policy-making. In addition, during data preprocessing, data with missing values was simply deleted. Instead of deleting them, perhaps imputing them such as replacing them with mean would leave more available data for further analysis?
5. Evaluation on Clarity and quality of writing:3. The report was well and clearly written in general. It is mostly easy to understand and the logic is smooth to follow. The figures look pretty and the color choices are appalling. The equations are clearly written as well. The sentences are with little grammatical mistakes. However, most of the figure legends and descriptions are either lacking or not clear. For example, I am confused with what message the authors wanted to confer through figure 3 and 4. They were not clearly explained in the main text and they have exactly the same legends which is quite misleading.
6. Evaluation on Technical Quality (3): Overall, the experiments well thought out and convincing, and the results are technically sound. However, the report did not provide a clear explanation of the criteria used to select the number of components for each data reduction method. The source codes are clear and able to be executed. The report did not discuss the limitations and potential biases of the methodology used or the results obtained. In addition, the authors only stated that they reduced the number of variables from 21 to 5 and 11 using PCA, Isomap, and UMAP, respectively. The selection of the number of components is crucial in data reduction methods, and the lack of explanation may affect the reliability of the results. Finally, while the report cited relevant papers and compared the presented work to them, there is room for additional discussion and critique of the literature. The authors could provide a more detailed comparison between their approach and existing literature, highlighting the strengths and limitations of each method.
7. Overall rating: 3
8. Confidence on your assessment: 3
9. Visualization and Dimensionality Reduction Techniques for US Crime Data (Yingxue XU, Jiaxin ZHUANG, Fengtao ZHOU)
10. Summary: The report discusses the analysis of crime data for 59 major cities in the United States from 1969 to 1992. The analysis involved using innovative multivariate analysis techniques, including data preprocessing, dimensionality reduction, and data visualization methods such as Principal Component Analysis (PCA), Sparse PCA, Multidimensional Scaling (MDS), Isometric Mapping (ISOMAP), and Locally Linear Embedding (LLE). The results showed that the crime number is influenced by various socio-economic, demographic, and environmental factors, and PCA had good performance in terms of data reduction and visualization.
11. Strengths: Firstly, the report presents practical significance by providing valuable and actionable information that can inform evidence-based strategies for crime prevention and intervention, ultimately contributing to creating safer and more secure communities. Secondly, it takes a meticulous approach by plotting the three-component results, enabling better understanding of the data, and identifying differences in key parameters. Thirdly, the report presents its analysis and results in a clear and concise manner using scientific language and clear figures.
12. Weaknesses: Firstly, the different types of crimes were aggregated into a total number of crimes, assuming that the underlying parameters have similar effects on all crime types. This assumption may not be accurate, and the analysis could potentially miss important differences in crime patterns across different types of crimes. Secondly, while the report covers various socio-economic, demographic, and environmental factors that could potentially influence crime patterns, it does not extensively discuss potential confounding variables that could also impact the results. This could limit the comprehensiveness of the analysis and the accuracy of the conclusions drawn. It did not go in detail about the interpretation of the results, and it did not discuss whether the goal of the project is achieved.
13. Evaluation on Clarity and quality of writing: 4. The report is generally clearly written and well-organized, with a straightforward style and concise language. The use of figures and examples is appropriate and helpful in illustrating key points and presenting the results of the analysis. However, there are also a few minor typos, such as "manifold learning" being spelled as "maniford learning". In terms of style and grammar, certain sentences could be simplified and clarified such as "multifaceted outcome" in the introduction and "maximize the effectiveness of our results" in data preprocessing.
14. Evaluation on Technical Quality: 3. Most of the results appear to be technically sound. However, I do not think that the analysis is quite well-supported by the theoretical analysis and experimental results. Firstly, in the results, the authors stated that “SPCA had a similar distribution of different classes with the PCA method. This indicates that both methods perform well in data”. The evidence is not good enough to state that both methods perform we. Secondly, the report evaluated different methods by looking at the graph and whether the dots are mixed together, in particular the orange dots, green dots, and red dots. However, this criterion does not distinguish the performance of each method in my opinion. Perhaps a more unbiased criterion could be incorporated. Thirdly, the results for using the LLE method seem questionable, and the report did not give a reason nor hypothesis for such a strange result. Finally, perhaps due to the limitation of a poster, the authors did not elaborate on the strengths and weaknesses and did not compare with relevant papers. The source code is clear and easy to follow.
15. Overall rating: 3.
16. Confidence on your assessment: 3.
17. Explore and Play with SNPs Data for Fun (CUI Yiran)
18. Summary: The study applied various dimensionality reduction techniques, including PCA, MDS, kernel-PCA, and random projections, to a dataset of 129 SNPs data. The results of PCA and MDS indicated a linear relationship between the SNPs and people's migration history. However, the performance of kernel-PCA with the radial basis function kernel decreased as the variance of the kernel function increased due to the increased similarity of the data. Random projections were found to be useful in reducing computation costs and providing a universal embedding control, as demonstrated by the Johnson and Lindenstrauss lemma.
19. Strengths: Firstly, the report presented the results of implementing different data reduction techniques on SNPs data in a clear and concise manner, making it easy for readers to understand the findings. Secondly, the author did in-depth exploration of different techniques providing insights into their comparisons. Thirdly, the report included several visual aids, such as figures and tables, to illustrate the results and make them more accessible to readers. Lastly, the author explained technical concepts, including dimensionality reduction methods and the Johnson and Lindenstrauss lemma, in a clear and concise manner, making them more accessible to readers who may not be familiar with them.
20. Weaknesses: The report did not provide a detailed discussion on the interpretation of the results and how they can be applied to real-world scenarios. Further, it did not compare its findings with those of other studies that have explored the performance of data reduction techniques on SNPs data, which can limit its context and relevance. Lastly, the writing of this report needs more polishing. Certain sentences are redundant and hard to understand. And typos can be found in the key results of the report resulting in confusions.
21. Evaluation on Clarity and quality of writing: 2. Overall, the report is clearly written and well-organized. The author provided clear explanations of the different data reduction techniques and their findings. The use of examples and figures is also effective in illustrating the results. However, there are a few areas where the report could be improved. Firstly, the report could benefit from a more detailed discussion on the interpretation of the results and their real-world implications. The author could expand on how these techniques could be applied in various fields, such as medicine, genetics, or other industries. Secondly, the writing of the report could be benefitted from more polishing. Certain sentences could be simplified and there are several typos. For example, in line 102, "due to curiosity" should be "out of curiosity." In line 104, there is a missing space between "10" and "-3." In line 119, "it is benefited" should be "it benefits." The “Result” in the last sentence of abstract should be in lower case. And I assume that the author meant “migration” when “mitigation” was written in sentences describing the key results.
22. Evaluation on Technical Quality: 4. The codes provided are clear and easy to follow. The results appear technically sound and there are no obvious flaws in the reasoning. The claims made by the authors are mostly supported and the experiments are well-designed. However, the author’s statement “the SNPs inherit a linear relationship with people’s migration history” is a bit confusing. On one hand, the claim itself is hard to comprehend. On the other hand, the claim is supported with little explanation and evidence. More justifications on this claim would be needed. One suggestion on comparing the dimensionality reduction methods’ performances is to use an additional unbiased parameter, such as the adjusted rand index, instead of by looking at the resulted graph. The report did little to compare its findings with those of other studies that have explored the performance of data reduction techniques on SNPs data. Therefore, another suggestion is to discuss and compare with relevant papers in the field.
23. Overall rating: 3.
24. Confidence on your assessment: 3.
25. Statistical Analysis on Authors and Word Trend of NIPS Papers from 1987 to 2017 (DUNDA, Gerry Windiarto Mohamad)
26. Summary: The report analyzed NIPS papers from 1987 to 2017 to identify trends in the usage of keywords, research themes, and publication patterns in the field of machine learning. The study found that the word usage of machine learning has undergone significant changes over time, and there is a weak positive correlation between community size and publication count. The report also revealed non-trivial clusters through document clustering using MDS and word mover’s distance metric. Overall, the report provides valuable insights into the evolution of the machine learning field and highlights the importance of considering community size and using advanced analysis methods when studying publication patterns in scientific communities.
27. Strengths: The study used advanced statistical techniques to analyze a large dataset of NIPS papers, which provides valuable insights into the evolution of the machine learning field. It presented the findings in a clear and concise manner, making it easy for readers to understand the research themes, publication patterns, and clustering results. The report also provided valuable insights into the importance of considering community size and using advanced analysis methods when studying publication patterns in scientific communities. The findings can be useful for researchers and policymakers in the field of machine learning.
28. Weaknesses: While the report discussed the non-trivial and fairly spreading clusters found through document clustering, it did not provide much detail on the interpretation of these clusters or their potential implications. In addition, the author did not provide a detailed discussion of the limitations of the study, such as potential biases in the data or limitations of the analysis methods used. Finally, the report did not provide much discussion on how they relate and compareto other research in the field.
29. Evaluation on Clarity and quality of writing: 4. Overall, the report is clearly written and well-organized, with a good use of examples and figures to illustrate the findings. Although, figure 4 is missing a figure title and a detailed description under each figure would be better. The report is free of major style and grammar issues, and the references and formatting appear to be correct. However, there are a few minor typos and errors: In the conclusion, "themes over time" should be changed to "themes over the years". The sentence "using the Louvain and Leiden algorithms demonstrated a weak positive correlation" should be changed to "demonstrated that there is a weak positive correlation". In the “Document Clustering” part, the figure number is lacking.
30. Evaluation on Technical Quality: 4. The source codes are clear and easy to follow. In addition, the report provides sufficient details on the methods used, making it possible for other researchers to reproduce the study and validate the findings. The results appear to be technically sound, with appropriate statistical methods used for the analysis. The reasoning appears to be sound, with claims supported by theoretical analysis and experimental results. The experiments are well thought out and convincing, with multiple analysis methods used to provide a comprehensive understanding of the trends in the field of machine learning. In the discussion part, the evaluation appears to be appropriate, citing relevant papers and comparing them to the presented work. I do have one small question out of curiosity with your choice of MDS for clustering. Why did you choose to use MDS instead of other methods?
31. Overall rating: 4.
32. Confidence on your assessment: 3.
33. Math5473 Mini-Project: Explore Geographical Variation Of Human Gene With SNPs Data (Zhanmiao Huang, Wencan Xia, Yuanhui Luo)
34. Summary: The report explored the relationship between genetic variation and geographic variation using single-nucleotide polymorphisms (SNPs) data. It described and utilized different dimension reduction methods, including Principal Component Analysis (PCA), Multidimensional Scaling (MDS), Random Forest (RF), and Extra Trees (ET), and presented the results obtained from applying these methods. The report evaluated the importance of different SNPs for predicting regions with random forest and extra trees and presents a case study focusing on the genetic variability of populations in China and its neighboring areas.
35. Strengths: The report presented the methods used and the results obtained in a clear and concise manner, making it easy for readers to understand the approach taken and the findings. In addition, the report used multiple dimension reduction methods and statistical learning methods to analyze the SNPs data, providing a comprehensive analysis of the relationship between genetic variation and geographic variation. Furthermore, it presented a detailed case study focusing on the genetic variability of populations in China and its neighboring areas, which adds to the understanding of the relationship between genetics and geographic location.
36. Weaknesses: The report did not provide a discussion of the limitations of the study, such as potential biases in the data or limitations of the methods used. In addition, it did not provide much discussion on the broader context of the findings, such as how they relate to other research in the field or how they may impact the broader field of genetics and genomics.
37. Evaluation on Clarity and quality of writing: 4. Overall, the report is clearly written, well-organized, and uses appropriate examples and figures to illustrate the findings. The report is free of major style and grammar issues, and the references and formatting appear to be correct. However, there are a few minor typos and errors that could be corrected to improve the clarity of the paper: In the first sentence of the introduction, "is a substitution" should be changed to "refers to a substitution". In the first sentence of the methodology section, "To deal with the high dimensionality and extract crucial information from SNPs data" should be changed to "To deal with the high dimensionality of SNPs data and extract crucial information". In the conclusion, "dimensionality reduction methods with SNPs data" should be changed to "dimensionality reduction methods for SNPs data".
38. Evaluation on Technical Quality: 4. The results appear to be technically sound based on the information presented in the report, and the claims are mostly supported by both theoretical analysis and experimental results. The experiments are well-designed. One part missing is the assessment of the strengths and weaknesses of their approach. The report effectively cites relevant literature. Although more comparisons could be made. The source codes are clear and easy to follow.
39. Overall rating: 4.
40. Confidence on your assessment: 3.