CSIC 5011: Project 1 Review

bcaiaa

April 2023

1 The first report by ZHONG Guangzheng

This report proposed a distance function for the ordinary regression dimensional reduction problem, in which everyday information in the classes is considered. The resulting representation in low-dimensional space is clear and easy to separate. This paper introduces the distance function in 3-dimensional reduction methods: MDS, Isomap, and t-SNE, and performs experiments on the Hand-Written Digits dataset. That is because the proposed distance function can be applied in the dimensional methods which use Euclidean distance.

From the comparison of MDS and Isomap, the different classes can be easily separated by the proposed distance, and the order of classes almost can also be persevered. But when comparing the proposed distance function with Euclidean distance in t-SNE methods, the improvement of the proposed distance function is not satisfactory enough, and the ordinal information is ignored.

The evaluation on Clarity and Quality of Writing is 5. This report is quite neat and written since this is a mini-project. I do not have any suggestions for this. Maybe add more details on the distance function of Isomap and t-SNE. **The Evaluation on Technical Quality is 4**. Claims are well-supported by theoretical analysis or experimental results. It's possible for others to replicate these results. The author assessed both the strengths and weaknesses of his approach. But no relevant papers were cited, discussed, or compared to the presented work.**The overall rating is 4**. **The confidence in my assessment is 2**.

2 The second report by CUI Yiran

This report explores the applcation of PCA, MDS, kernel-PCA, and random projection with both PCA and MDS by SNPs data. PCA and MDS show good indicators of human mitigation history but this implies the linearity in the data. For the kernel-PCA, the radial basis function kernel is selected with various variances(σ). It is clear that the performance of kernel-PCA becomes worse with σ increase. In the case of σ lower than N_F (which is the inverse of feature size), results are good and similar to direct PCA, but latter data points start to concentrate on one point. For random projections combined with PCA and MDS, it is clear that random projections work well and it

provides similar results as that of direct PCA/MDS. And this algorithm is significantly faster than the case without implementing random projections.

The evaluation on Clarity and Quality of Writing is 4. This report is clearly written. However, the description of dataset is wrong since the SNPs from Quanhua MU and Yoonhee Nam consists of 1043 subjects. There is typo in line 98(It is noticeable that implementing PCA on MDS results **in** only a coordinate rotation). **Evaluation on Technical Quality is 5**. Claims are well-supported by theoretical analysis and experimental results. But no relevant paper is cited, discussed, or compared to the presented work. **Over rating is 4**. **Confidence on my assessment is 2**.

3 The third report by MA-Tang-Ruan-Huang

This report uses some dimensional reduction methods: PCA, MDS, t-SNE, ISOMAP, LLE, UMAP, robust PCA, and random projection to analyze the SNPs dataset. From the visualization results, t-SNE, PCA, and MDS can separate the whole dataset into 7 regions well, and others show relatively bad results. The comparison based on Adjusted Rand Index (ARI) shows that random projection, robust PCA, and t-SNE rank top with high ARI scores. What's more, the reduced dimension was set as 5957 for random projection considering the sample number and epsilon. Finally, we showed that the chosen method t-SNE can indeed be used for ancestry prediction in an additional 1000 Genomes Project dataset with the ARI score is 0.709.

The evaluation on Clarity and Quality of Writing is 5. This group submitted a poster, but it's clearly written and well-explained. **Evaluation on Technical Quality is 5**. Claims are well-supported by theoretical analysis and experimental results. But no relevant paper is cited, discussed, or compared to the presented work. **Over rating is 4**. **Confidence on my assessment is 2**.

4 The fourth report by Lin-Liu

This report uses PCA, JS&MES estimator, Lasso, and Tree-based model to find which features or factors affect the number of crime events most and how these features affect the crime. However, PCA is not so appropriate for this problem and data. James-Stein & MSE estimator show that economic level, unemployment rate, and the population of different ages and black affected crime during 1970-1992. For Lasso, the percentage of black people and capital spent on education and welfare are selected as core factors. For the tree-based method, the number of sworn and civil police officers employed by the city and the percentage of black people is important.

The evaluation on Clarity and Quality of Writing is 4. This report has some typos, like in the description of Figure 1. The combinations of variables in PCA are not so clear. **Evaluation on Technical Quality is 4**. No relevant paper is cited, discussed, or compared to the presented work. **Over rating is 4**. **Confidence on my assessment is 1**.

5 The fifth report by Yan-Yan-Lai

The goal is to find the main contributions to the average crime rate using the crime data. The original dataset includes 85 American cities' crime information from 1969 to 1992 with many missing values and useless data. Firstly, the dataset is simplified using pre-processing techniques, like removing the data of different cities' names and years, the ratio of total crime number and total population instead of the detailed classification, etc. Then, principal component analysis (PCA), Isometric Mapping (Isomap) and uniform manifold approximation and projection (UMAP) methods are used to reduce the number of variables from 21 to 5 and 11. Visualization results show that the 85 cities are well separated into several groups by the three methods. Finally, the linear regression model is applied to find the relationship between 'principle components' and the average crime rate. In short, we find that the US crime rate is affected by many factors, such as the population structure, police number, mayoral term, society welfare, etc.

The evaluation on Clarity and Quality of Writing is 4. Figure s-1 and Figure s-2 are not pointed out in this report. **Evaluation on Technical Quality is 4**. No relevant paper is cited, discussed, or compared to the presented work. **Over rating is 4**. **Confidence on my assessment is 1**.