Data Mining

Lab sheet N°4: Dimensionality Reduction

ENSIA 2023-2024

Part 2: Exercises on the Chapter Data - Part 2

1. Comment on the use of a box plot to explore a data set with four attributes: age, weight, height, and income.

A great deal of information can be obtained by looking at (1) the box plots for each attribute, and (2) the box plots for a particular attribute across various categories of a second attribute. For example, if we compare the box plots of age for different categories of ages, we would see that weight increases with age.

2. Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed.

If the line representing the median of the data is in the middle of the box, then the data is symmetrically distributed, at least in terms of 75% of the data between the first and third quartiles. For the remaining data, the length of the whiskers and outliers is also an indication, although, since these features do not involve as many points, they may be misleading.

3. How might you address the problem that a histogram depends on the number and location of the bins?

The best approach is to estimate what the actual distribution function of the data looks like using kernel density estimation. This branch of data analysis is well-developed and is more appropriate if the widely available, but simplistic approach of a histogram is not sufficient.

4. Discuss the advantages and disadvantages of using sampling to reduce the number of data objects that need to be displayed. Would simple random sampling (without replacement) be a good approach to sampling? Why or why not?

Simple random sampling is not the best approach since it will eliminate most of the points in sparse regions. It is better to undersample the regions where data objects are too dense while keeping most or all of the data objects from sparse regions.

5. Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on techniques such as PCA and SVD.

The dimensionality of PCA or SVD can be viewed as a projection of the data onto a reduced set of dimensions. In aggregation, groups of dimensions are combined. In some cases, such as when days are aggregated into months or the sales of a product are aggregated by store location, the aggregation can be viewed as a change of scale. In contrast, the dimensionality reduction provided by PCA and SVD does not have such an interpretation.

6. Describe how you would create visualizations to display information that describes the following types of systems.

Computer networks. Be sure to include both the static aspects of the network, such as connectivity, and the dynamic aspects, such as traffic.

The connectivity of the network would best be represented as a **graph**, with the nodes being routers, gateways, or other communications devices and the links representing the connections. The bandwidth of the connection could be represented by the width of the links. Color could be used to show the percent usage of the links and nodes.

The distribution of specific plant and animal species around the world for a specific moment in time.

The simplest approach is to display each species on a separate **map** of the world and to shade the regions of the world where the species occurs. If several species are to be shown at once, then **icons** for each species can be placed on a map of the world.

The use of computer resources, such as processor time, main memory, and disk, for a set of benchmark database programs.

The resource usage of each program could be displayed as **a bar plot** of the three quantities. Since the three quantities would have different scales, proper scaling of the resources would be necessary. For example, resource usage could be displayed as a percentage of the total.

Alternatively, we could use **three bar plots**, one for each type of resource usage. On each of these plots, there would be a bar whose height represents the usage of the corresponding program. This approach would not require scaling.

Another option would be to display a **line plot** of each program's resource usage. For each program, a line would be constructed by (1) considering processor time, main memory, and disk as different x locations, (2) letting the percentage resource usage of a particular program for the three quantities be the y values associated with the x values, and then (3) drawing a line to connect these three points. An ordering of the three quantities needs to be specified, but is arbitrary. For this approach, the resource usage could be displayed on the same plot.

The change in occupation of workers in a particular country over the last thirty years. Assume that you have yearly information about each person, which also includes gender and level of education.

For each gender, the occupation breakdown could be displayed as an **array of pie charts**, where each row of pie charts indicates a particular level of education and each column indicates a particular year. The time gap between each column could be 5 or ten years.

Alternatively, we could order the occupations and then, for each gender, compute the cumulative percent employment for each occupation. If this quantity is plotted for each gender, then the area between two successive lines shows the percentage of employment for this occupation. If a color is associated with each occupation, then the area between each set of lines can also be colored with the color associated with each occupation. A similar way to show the same information would be to use a sequence of **stacked bar graphs**.

Another option could be the use of **animated bar charts** to display the evolution over years:

https://www.visualcapitalist.com/cp/animated-most-valuable-brands-from-2000-2022/