Data Exploration and Networks

Topics covered so far

1. Data Exploration and Visualization
    a. Hypothesis Testing
    b. PCA
    c. t-SNE
2. Graphs & Networks
    a. Adjacency matrix
    b. Connected components
    c. Centrality measures

Discussion Questions

1. What are multiple testing issues and why do they occur?
2. Why do we need to do dimensionality reduction?
3. How is PCA different from t-SNE algorithm?

Multiple testing issue and their corrections

Multiple testing problem

- This problem arises when multiple hypothesis are tested simultaneously
- The number of false positives increases as you test more number of hypotheses

Following are the correction methods that can be used to deal with this problem:

- **Bonferroni correction**
  - It states that the corrected significance level for all the test combined is **α/m.** where **m** is the total number of hypothesis tests performed
  - Reject null hypothesis H_0 when **p-value ≤ α/m or m*p-value ≤ α**
- **Holm-Bonferroni correction**
  - Sort p-values in increasing order: p(1) ≤ · · · ≤ p(m), The corrected significance level for the ith test is **α/(m-i+ 1)**
  - Reject null hypothesis H_0 **p(i) ≤ α/(m-i + 1) or (m – i + 1)*p(i) ≤ α**
- **Benjamini-Hochberg correction:**
  - Sort p-values in increasing order: p(1) ≤ · · · ≤ p(m), The corrected significance level for the ith test is **α*i/(m)**
  - Reject null hypothesis H_0 **p(i) ≤ α*i/(m) or m*p(i)/i ≤ α**

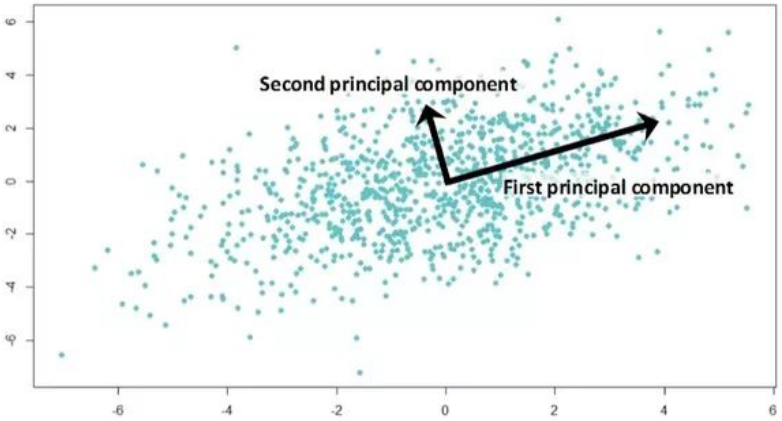
Need for dimensionality reduction

- Dimensionality reduction is the process to reduce the number of dimensions in the feature space.
- In machine learning, we tend to add many features to get more accurate results. However, after a certain point, the performance and robustness of the model start decreasing and computational complexity starts increasing as we increase the number of features. This is called the curse of dimensionality where the sample density decreases exponentially with the increase of dimensionality.
    - We use dimensionality reduction to transform the data into low dimensions while keeping most of the information intact.
    - It also helps us to visualize the high dimensional data in 2D & 3D.
- There are the following techniques we can use for dimensionality reduction:
    - PCA
    - t-SNE

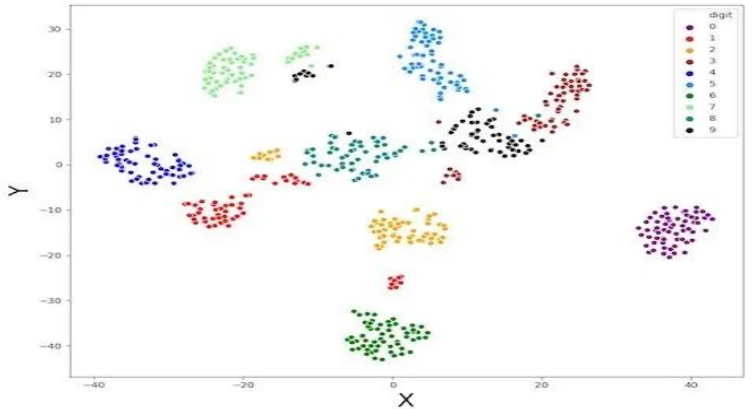
PCA and t-SNE

Principal component analysis (PCA) is a dimensionality reduction technique used for the identification of a smaller number of uncorrelated variables known as principal components from a larger dataset. The technique is widely used to emphasize variation and capture strong patterns in a dataset.

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets.



PCA



t-SNE

Image Source

Image Source

Difference between PCA and t-SNE

PCA	t-SNE
It tries to capture the linear structure in the data	It tries to capture the non-linear structure in the data
It focuses to preserve the global structure of the data	It focuses to preserve the local structure (i.e., clusters) of the data
There are no hyperparameters involved in PCA	There are some hyperparameters like perplexity, no. of dimensions, etc. in t-SNE
PCA works by separating points as far as possible based on the highest variance	t-SNE works by grouping points as close as possible based on the characteristics of the point
It might easily get affected by outliers	It can handle outliers as well

Case Study - PCA & t-SNE

Discussion Questions

1. Why do we study Graphs & Networks?

2. What is an adjacency matrix and How do we interpret it?

Why do we study Graphs and Networks

A Graph is basically the study of relationships. It has certain nodes (vertices) and links (edges) that create these relationships.

It can be used to model and create many types of relations and processes in physical, social, biological, and information systems, and has a wide range of applications:

- Community networks (through social media)
- Google maps
- DNA/RNA sequencing
- Search engine rankings

Example: This friendship network shows us a bunch of friends, the networks they belong to, and the social cliques they are part of.

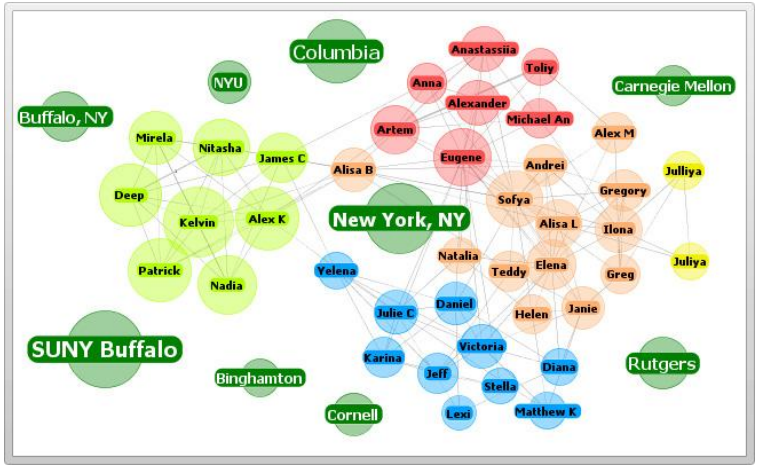
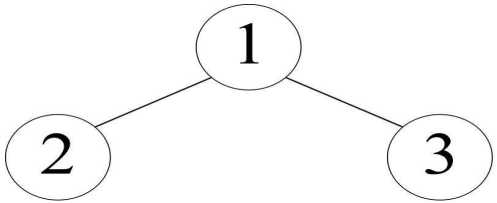


Image Source

Adjacency matrix

We can represent the graph as an adjacency matrix, where the row and column indices represent the nodes, and the entries in the matrix represent the absence or presence of an edge between the nodes.

Example: For a graph 2-1-3, the adjacency matrix will be - $\begin{pmatrix} 0,1,1 \\ 1,0,0 \\ 1,0,0 \end{pmatrix}$



An adjacency matrix is the best representation of a graph into a mathematical form that tells us whether there are any edges between all sets of nodes. The diagonal of this matrix will always be zero if there are no self-loops in the network.

Case Study - Caviar

Appendix

Degree and its calculation

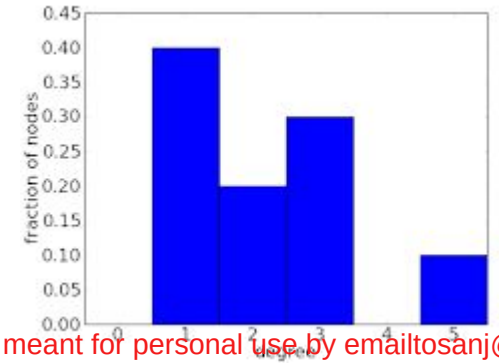
The degree of a node refers to the number of edges that are connected to it. In a directed graph, you can calculate the in-degree and out-degree which means incoming and outgoing connections of a node.

In simple words, it is a popularity measure. The higher the degree, more central the node is.

We can calculate the average degree of a network by using the formula **2*(m/n)**, where **m** is the number of edges and **n** is the number of nodes.

Degree distribution: It is a probability that the random chosen node has k number of connections.

In this graph, you can observe how the degree is varying with the fraction of nodes.



Centrality measures

Centrality measures capture the importance of a node's position in a network. There are the following types of centrality measures:

1. **Degree centrality:** It is a measure of the popularity of a node in a network. It does not capture the quality vs quantity.
2. **Propagated degree (eigenvector) centrality:** It measures the importance of a node in a graph with respect to the importance of its neighbors. If a node is connected to highly important nodes, it will have a higher score as compared to a node that is connected to less important nodes.
3. **Closeness centrality:** It tracks how close a node is to another by measuring the distance between them. In other words, it measures the node efficiency in terms of connection to other nodes.
4. **Betweenness centrality:** It measures the importance of a node in a network based on how many times it occurs in the shortest path between all pairs of nodes in a graph. It measures the extent to which a node lies on paths between other nodes.

Real life example of a network and its centrality measures

In a **social network:**

- High degree centrality - most popular person who can quickly connect with the wider network
- High eigenvector centrality - most popular person who has a good social network with another popular person
- High closeness centrality - a person who can influence the whole network most quickly
- High betweenness centrality - a person who influences the flow around the network, i.e., removal of that person can break the network

Great Learning

Happy Learning !