Dimensionality Reduction Algorithms in Data Visualization

Computer Science Senior Project | Ilya Ilyankou ‘18 | Advisor: Prof. Peter Yoon | Trinity College, CT

Problem

- Multidimensional datasets are hard to understand and interpret.
- Visualization helps, but we are limited by 3-dimensional geometry.

Solution

- Apply a dimensionality reduction algorithm that decreases the number of dimensions from P to 3.
- Treat values of new dimensions as x, y, and z coordinates and plot all data points in 3D.

Principal Component Analysis

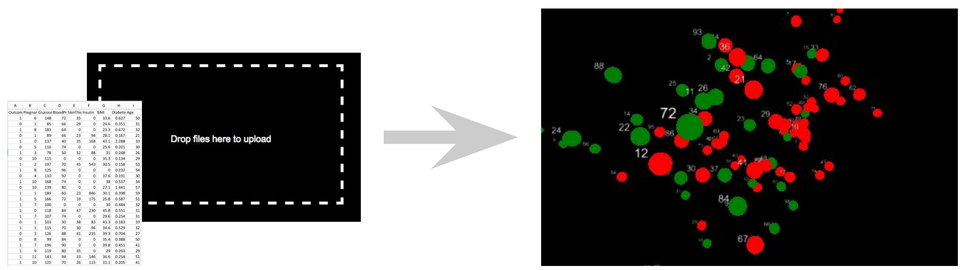
- Represents original data using new linearly uncorrelated dimensions (principal components).
- Principal components are sorted by variance, so the first few components carry most information.
- Requires computing covariance matrix and its eigenvalue decomposition.
- Computationally expensive, runs is $O(n^3)$.

A	B	C	D	E	F	G	H	I
Outcom	Pregnar	Glucose	BloodPr	SkinThic	Insulin	BMI	Diabete	Age
1	6	148	72	35	0	33.6	0.627	50
0	1	85	66	29	0	26.6	0.351	31
1	8	183	64	0	0	23.3	0.672	32
0	1	89	66	23	94	28.1	0.167	21
1	0	137	40	35	168	43.1	2.288	33
0	5	116	74	0	0	25.6	0.201	30
1	3	78	50	32	88	31	0.248	26
0	10	115	0	0	0	35.3	0.134	29
1	2	197	70	45	543	30.5	0.158	53
1	8	125	96	0	0	0	0.232	54
0	4	110	92	0	0	37.6	0.191	30
1	10	168	74	0	0	38	0.537	34
0	10	139	80	0	0	27.1	1.441	57
1	1	189	60	23	846	30.1	0.398	59
1	5	166	72	19	175	25.8	0.587	51
1	7	100	0	0	0	30	0.484	32
1	0	118	84	47	230	45.8	0.551	31
1	7	107	74	0	0	29.6	0.254	31
0	1	103	30	38	83	43.3	0.183	33
1	1	115	70	30	96	34.6	0.529	32
0	3	126	88	41	235	39.3	0.704	27
0	8	99	84	0	0	35.4	0.388	50
1	7	196	90	0	0	39.8	0.451	41
1	9	119	80	35	0	29	0.263	29
1	11	143	94	33	146	36.6	0.254	51
1	10	125	70	26	115	31.1	0.205	41

x	y	z
88	66	21
176	90	34
150	66	42
73	50	10
187	68	39
100	88	60
146	82	0
105	64	41
84	0	0
133	72	0
44	62	0
141	58	34
114	66	0
99	74	27
109	88	30
109	92	0
95	66	13
146	85	27
100	66	20
139	64	35
126	90	0
129	86	20
79	75	30

Implementation

Just drag-and-drop your dataset into a browser window to have it visualized!

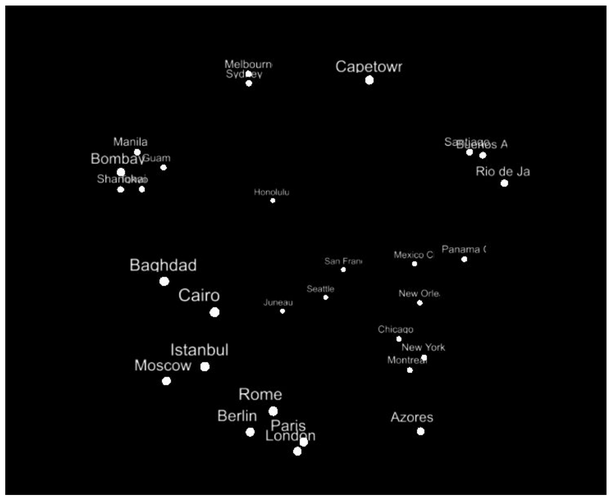


Back-end is a Python server that analyzes a dataset, performs dimensionality reduction, and returns a JSON file with x, y, and z coordinates of each data point.

Front-end uses WebGL (three.js) to create an interactive visualization of all data points in space, allowing users to zoom and rotate to explore clustered areas.

Multidimensional Scaling

- Dimensionality reduction algorithm that attempts to preserve distances among data points.
- When applied to real-world datasets of distances between cities, the resulting visualization is a real map in 2D, or a globe in 3D.
- Runs in $O(n^3)$.



Conclusion

- Visualization is a powerful technique to understand data.
- Dimensionality reduction can reveal hidden patterns and relations among data points in high-dimensional space.
- Different algorithms should be applied to different types of datasets, but there is no simple rule to determine *which* algorithm would work best.

Acknowledgements

I would like to thank **Professor Peter Yoon** for guidance and encouragement. I would also like to thank **my seminar classmates** for their feedback and suggestions.

References

[1] Farcomeni, Alessio, and Luca Greco. *Robust methods for data reduction*. CRC press, 2016.

[2] Lee, John A., and Michel Verleysen. *Nonlinear Dimensionality Reduction*. Springer Science Business Media, LLC, 2008.

[3] Khot, Tejas. "Visualizing high-dimensional data." *XRDS: Crossroads, The ACM Magazine for Students* 23, no. 2 (2016): 66-67.

[4] Grinstein, Georges, Marjan Trutschl, and Urška Cvek. "High-dimensional visualizations." *Proceedings of the Visual Data Mining Workshop, KDD*. Vol. 2. 2001.

[5] Izenman, Alan Julian. "Linear Dimensionality Reduction." *Modern Multivariate Statistical Techniques*. Springer, New York, NY, 2013. 195-236.