Conceptual and Empirical Exploration of t-SNE

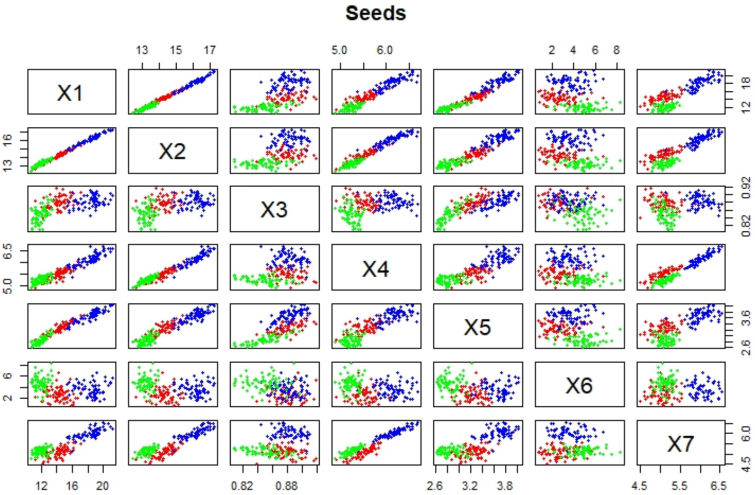
Lucy Chen, Allyson Ling, Shuting Zang

Advisor: Xiaodong Li

Seeds Dataset

- From UC Irvine Machine Learning Repository
- Geometric features of different wheat seeds
- 210 observations, 7 attributes, 3 classes
- Link for dataset: https://archive.ics.uci.edu/ml/datasets/seeds

Seeds Dataset



Seeds Dataset

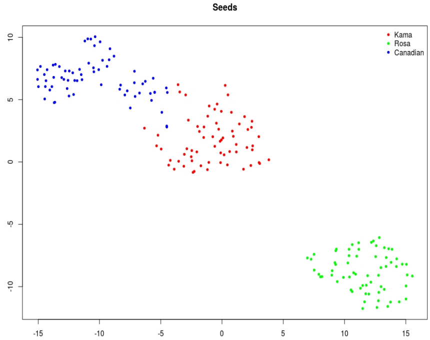


Figure: t-SNE

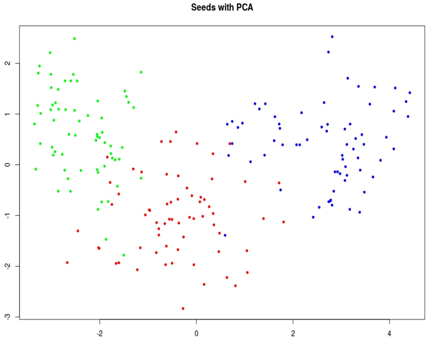


Figure: PCA

What is t-SNE?

- t-distributed stochastic neighbor embedding
- t-SNE is a non-linear dimension reduction technique
- Steps:

$$X \to D \to P \to Q \to Y$$

  - X $=$ Data matrix ($\leq$ 40 attributes)
  - D $=$ Pairwise distance matrix
  - P $=$ Probability matrix in high dimension
  - Q $=$ Probability matrix in low dimension
  - Y $=$ Data matrix in low dimension

X $\rightarrow$ Pairwise Distance Matrix $\rightarrow$ P

$$Perp(P_i) = e^{H(P_i)} \qquad H(P_i) = -\sum_j p_{j|i} log(p_{j|i})$$

$$p_{j|i} = \frac{-exp(||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} -exp(||x_i - x_k||^2/2\sigma_i^2)}$$

Perplexity

$$Perp(P_i) = e^{H(P_i)}$$

- Measure of effective number of neighbors
- Function of Shannon entropy
    - Measure of uncertainty of an outcome from a set of possible events
    - Number of bits needed to store information from dataset

Pairwise Distance Matrix → P

$Perp(P_i) = e^{H(P_i)}$, $H(P_i) = -\sum_j p_{j|i} log(p_{j|i})$, $D = \{\frac{p_1}{\sum_i p_i}, \ldots, \frac{p_k}{\sum_i p_i}\}$
$p_j = e^{-\beta_j d_{ij}}$

$$
\begin{aligned}
H(D) &= -\sum_j \frac{p_j}{\sum_i p_i} log \frac{p_j}{\sum_i p_i} \\
&= -\sum_j \frac{p_j}{\sum_i p_i} (log\, p_j - log \sum_i p_i) \\
&= -\sum_j \frac{p_j}{\sum_i p_i} log\, p_j + log(\sum_i p_i) \frac{p_j}{\sum_i p_i} \\
&= log \sum_i p_i - \frac{1}{\sum_i p_i} \sum_j p_j logp_j \\
&= log \sum_i p_i - \frac{1}{\sum_i p_i} \sum_j \beta_j d_j e^{-\beta_j d_j}
\end{aligned}
$$

P $\rightarrow$ Q $\rightarrow$ Y

- First set $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$ and diagonal entries to 0
- Then use gradient descent algorithm

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}$$

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||y_k - y_i||^2)^{-1}}$$

Real Data Implementation

Wisconsin Breast Cancer Dataset

- From UC Irvine Machine Learning Repository
- Breast Cancer Diagnosis: 357 benign, 212 malignant
- Ten real-valued features are computed for each cell nucleus, and attributes include the mean, standard error, and "worst" or largest (mean of the three largest values) of these features

Wisconsin Breast Cancer Dataset

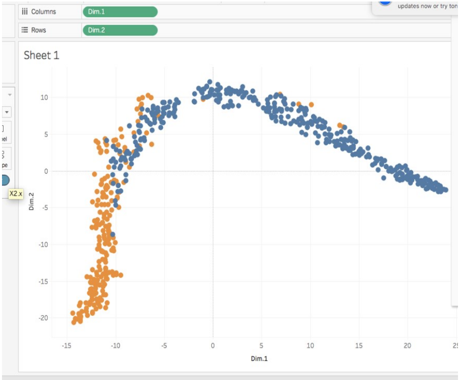


Figure: Original

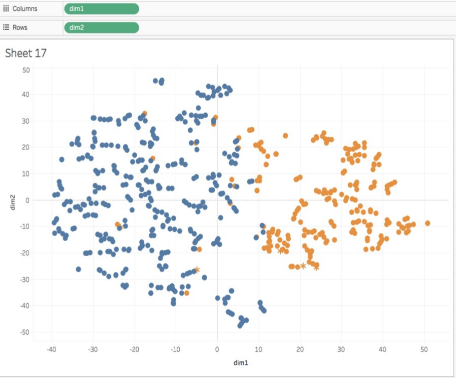


Figure: Scaled

Wisconsin Breast Cancer Dataset

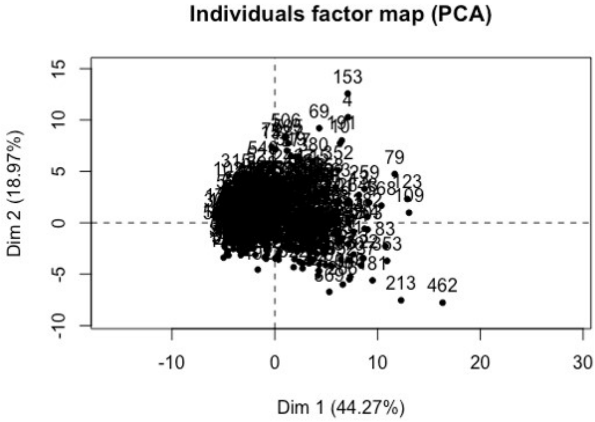
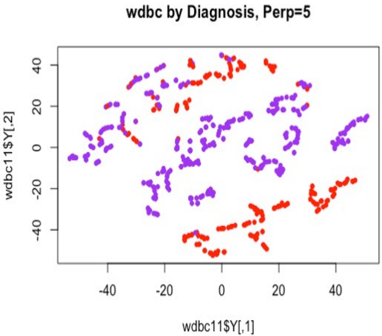
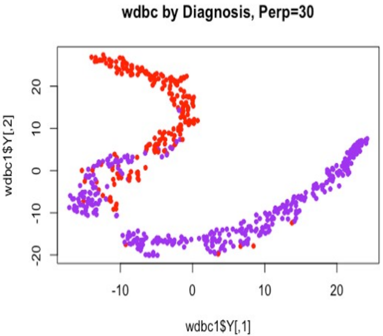


Figure: 2D PCA

Wisconsin Breast Cancer Dataset

Finding the proper perplexity

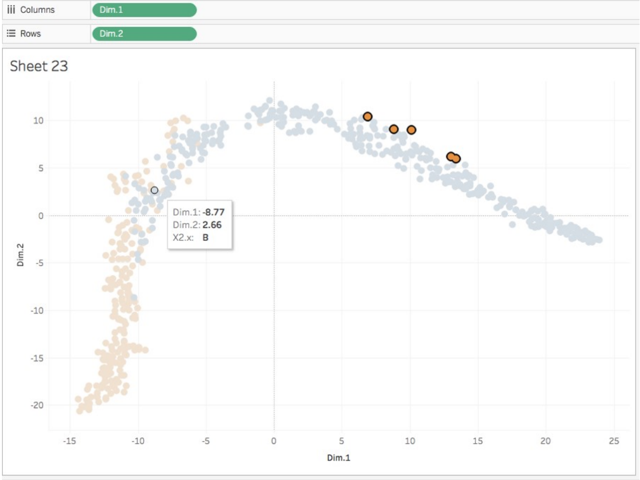




For this dataset, higher perplexity looks better for clustering

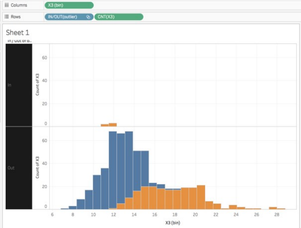
Wisconsin Breast Cancer Dataset

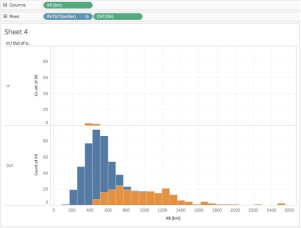
Find the anomalies, and trace them back to the original dataset

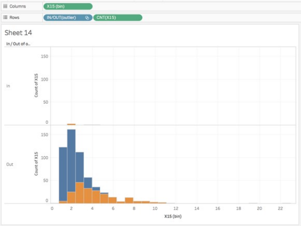


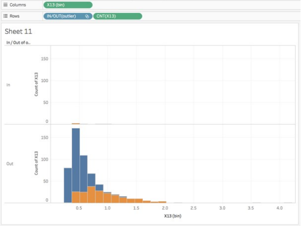
Blue for benign, orange for malignant

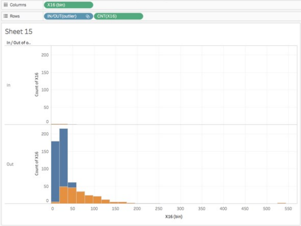
Wisconsin Breast Cancer Dataset



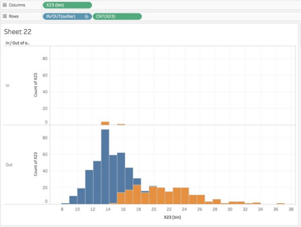


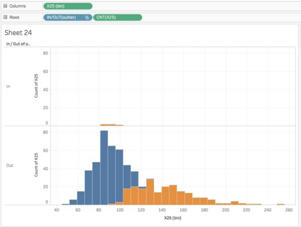


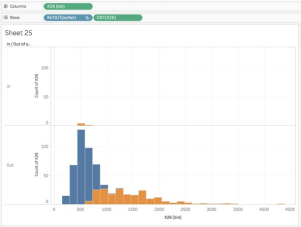




Wisconsin Breast Cancer Dataset







By the visualized clusters, we would like the doctor to reconsider the diagnosis, since the patient may be misdiagnosed.

Mean/SD/Worse of radius
Mean/SD/Worse of perimeter $\rightarrow$ Features about size of cancer cell
Mean/SD/Worse of area

Vertebral Column Dataset

- From UC Irvine Machine Learning Repository
- Features distinguishing spinal disorders
- 310 observations, 6 attributes, 3 classes
- Link for dataset:
  https://archive.ics.uci.edu/ml/datasets/Vertebral+Column

Vertebral Column Dataset

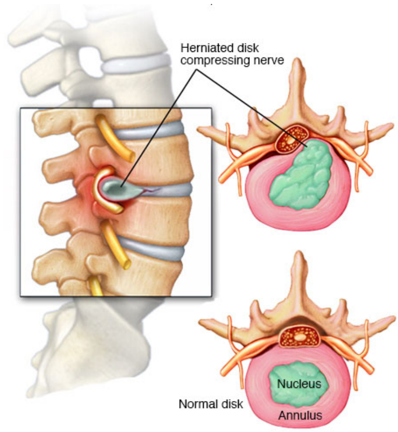


Figure: Disk Hernia and Normal

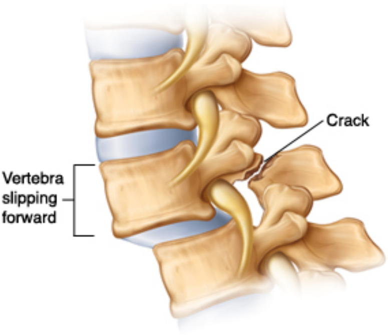


Figure: Spondylolisthesis

Vertebral Column Dataset

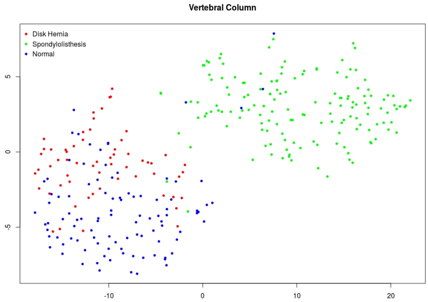


Figure: t-SNE

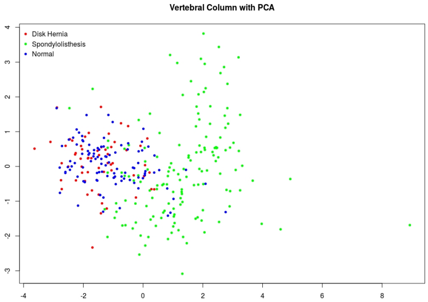


Figure: PCA

Vertebral Column Dataset

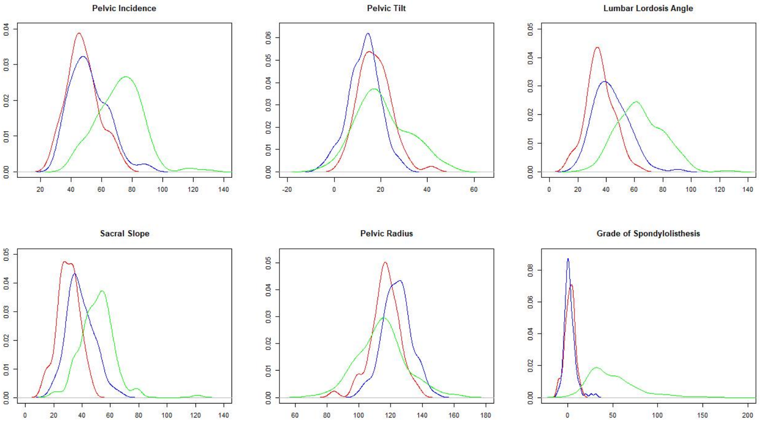
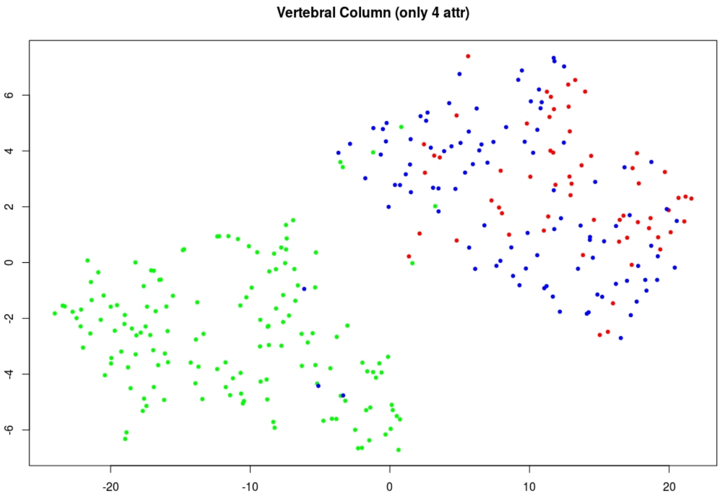


Figure: Density plots of each attribute

Vertebral Column Dataset



Vertebral Column Dataset

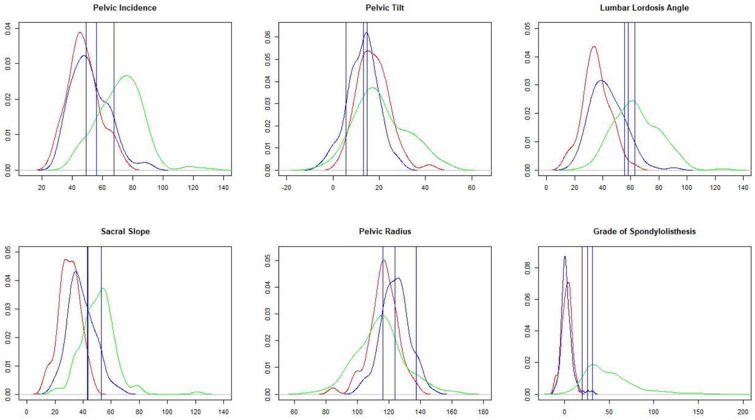


Figure: Density plots of each attribute with anomalies

References

- Maaten, Laurens van der. Hinton, Geoffrey. *Visualizing Data using t-SNE*. https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
- Shannon, Claude E. *A Mathematical Theory of Communication*. http://math.harvard.edu/ ctm/home/text/others/shannon/entropy/entropy.pdf

Datasets:

- Seeds: https://archive.ics.uci.edu/ml/datasets/seeds
- Breast Cancer: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- Vertebral Column: https://archive.ics.uci.edu/ml/datasets/Vertebral+Column