November 20 2024

Data Visualization:

- Can paint a picture and illuminate, but could also be misleading.
- Florida has not become safer due to stand your ground laws!
- Nicolas Cage and people drowning in pools???
- Histogram, density plots, box plots and a range of others can be useful when plotting distributions

Alternatives to dimensionality reduction other than PCA

- Done by projecting the points from high dimensions to low dimensions

A lot of the time we want to create clusters

- Distances in original data may not be meaningful
- We want some sort of embedding that preserves clustering (Linear projection -PCA - is only one time of embedding)
- Embedding - Display of low(er) level data.

t-distributed Stochastic Neighborhood Embedding (t-SNE)

- Define distances between a point x to a point Y using a Gausian function centered at X
- Differences in small and large distances tend to get really squished.
- Tends to emphasize intermediate distances - clusters!

Clear t-SNE supremacy shown!

Profesor backtracked on aforementioned supremacy

- Preserving distance vs Preserving structure

Admissions case study!

- How would we encode features?
- What features would we want to optimize?
- Discount rate
- Alumni Donations

`Machine Learning and Data Mining

- Supervised learning
- - Makes use of examples where we know the underlying truth
- Unsupervised Learning

- - Learn underlying structure or features without labeled training data
- - - Clustering based on features only
- - - Dimensionality reduction, compressing data (removing feature correlation)
- - - Structured prediction : model latent variables
- - Eg, biology: clustering

Clustering

- learn about the structure of our data
- Cluster new data (prediction)
- Goal : minimize within cluster similarity
- Clustering genes with similar expression patterns