HEART: Statistics and Data Science with Networks

Joshua Agterberg

Johns Hopkins University

Outline

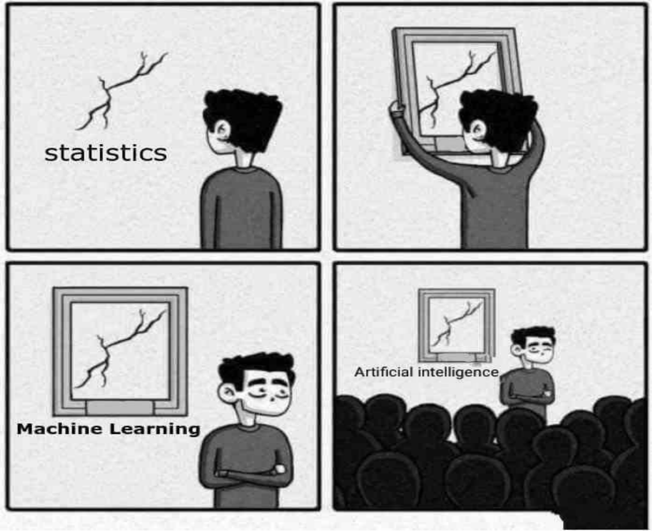

Figure: Source: https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3

- Broadly speaking, there are two areas of machine learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
    - Regression (continuous response)
    - Classification (categorical response variable)
- Unsupervised learning:
    - No specific response variable
    - Dimensionality Reduction
    - Clustering

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning

Crash Course on Data Science and Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning

Crash Course on Data Science and Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction

Crash Course on Data Science and Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

- Machine Learning can be closer to engineering or closer to statistics

Crash Course on Data Science and Machine Learning

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems
- Even still, we do not understand everything about linear regression!

Crash Course on Data Science and Machine Learning

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems
- Even still, we do not understand everything about linear regression!
- I am happy to discuss this more with anyone

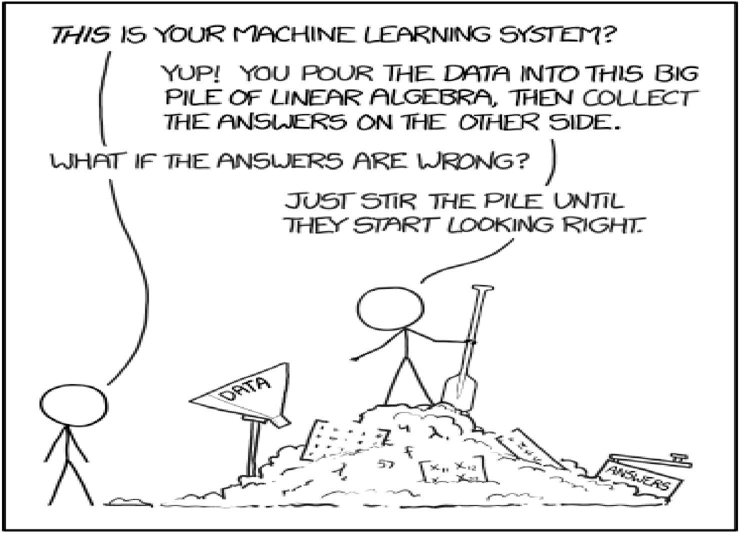

Figure: Source: https://xkcd.com/1838/

Clustering

- Clustering assumes data come from a mixture and seeks to estimate the clusters

Clustering

- Clustering assumes data come from a mixture and seeks to estimate the clusters
- Examples of Clustering Algorithms:
  - K-Means (uses only means)
  - Expectation Maximization Algorithm (Mixtures of Gaussians)
  - K-Medoids
  - and more!

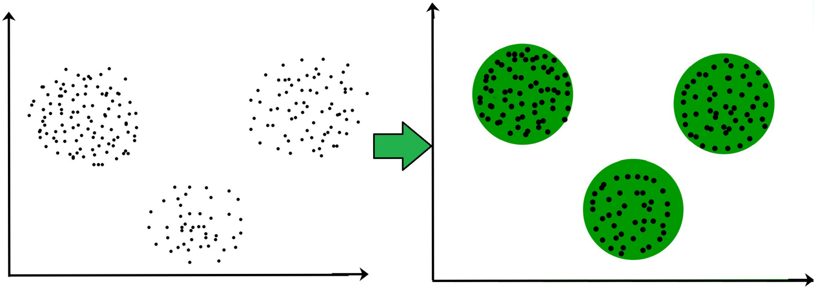

Figure: Source:
https://www.geeksforgeeks.org/clustering-in-machine-learning/

Dimensionality Reduction

- Goal: start with some complex, high-dimensional data and want to obtain representations of each data point in a smaller dimension

Dimensionality Reduction

- Goal: start with some complex, high-dimensional data and want to obtain representations of each data point in a smaller dimension
- Examples:
  - Manifold Learning
  - Principal Components Analysis
  - Spectral Embeddings

Dimensionality Reduction

- Goal: start with some complex, high-dimensional data and want to obtain representations of each data point in a smaller dimension
- Examples:
  - Manifold Learning
  - Principal Components Analysis
  - Spectral Embeddings
- Why would we want to do this?

Dimensionality Reduction for Graphs

- Start with $n \times n$ Adjacency matrix

Dimensionality Reduction for Graphs

- Start with $n \times n$ Adjacency matrix
- Obtain a *graph embedding* of dimension $n \times d$
- Idea is d is small (e.g. for SBM it is the rank of the SBM)

How to do dimensionality reduction for graphs?

1. Let S be a similarity matrix (e.g. adjacency matrix or Laplacian matrix)

Dimensionality Reduction for Graphs

How to do dimensionality reduction for graphs?

1. Let S be a similarity matrix (e.g. adjacency matrix or Laplacian matrix)
2. Compute biggest d eigenvalues and corresponding eigenvectors of S, call them $\hat{u}_1, \ldots, \hat{u}_d$ and $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$

Dimensionality Reduction for Graphs

How to do dimensionality reduction for graphs?

1. Let S be a similarity matrix (e.g. adjacency matrix or Laplacian matrix)
2. Compute biggest d eigenvalues and corresponding eigenvectors of S, call them $\hat{u}_1, \ldots, \hat{u}_d$ and $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$
3. Set either $\hat{U}$ or $\hat{X}$ as the graph embedding where $\hat{U}$ is the $n \times d$ matrix whose columns are $\hat{u}_i$ or $\hat{X}$ as the $n \times d$ matrix

$$\hat{X} := \hat{U}\hat{\Lambda}^{1/2}$$

Dimensionality Reduction for Graphs

Practical considerations:

- This requires knowing d in advance – but we know how to choose d now
- For large graphs, computing eigenvectors and eigenvalues can be computationally intensive, so may want to use `irlba` or randomized SVD algorithms that reduce computation time
- For this class, the SVD and eigendecomposition are essentially the same (SVD works on rectangular matrices, but eigendecompositions only work on square matrices)
- Can be modified to obtain a general procedure for general data by computing a similarity matrix S between data points (e.g. using a Gaussian kernel or other method)

Spectral Clustering and Community Detection

Now we get to spectral clustering:

- Starting with a graph, obtain $n \times d$ graph embedding matrix $\hat{X}$ or $\hat{U}$ whose rows are vertex representations
- Cluster the rows of this matrix using K-means or other clustering method

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables
- Lots of theory exists in fixed-dimension, high-dimension, and more

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables
- Lots of theory exists in fixed-dimension, high-dimension, and more
- Highly intuitive explanation in terms of covariances and singular value decompositions

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables
- Lots of theory exists in fixed-dimension, high-dimension, and more
- Highly intuitive explanation in terms of covariances and singular value decompositions
- First component of PCA maximizes the variance along that direction

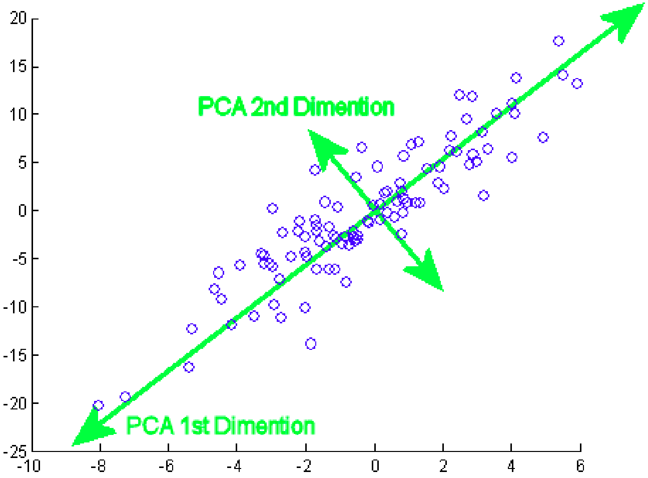

Figure: Source: https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$
- Manifold Learning seeks to uncover this manifold structure

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$
- Manifold Learning seeks to uncover this manifold structure
- Lots of algorithms exist (see Wiki on nonlinear dimensionality reduction)

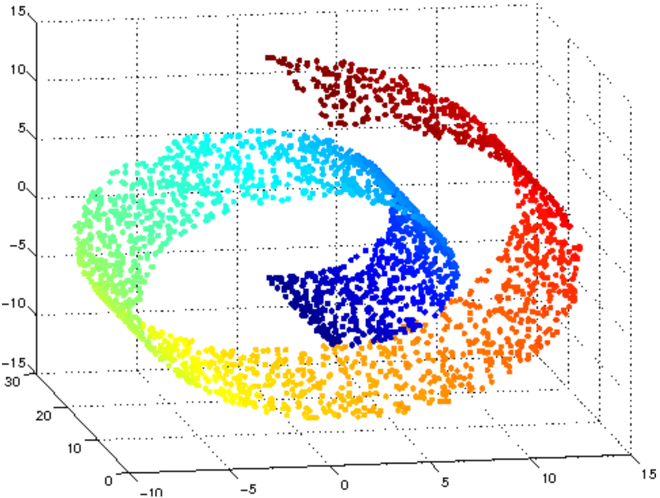

Figure: Source: https://www.semanticscholar.org/paper/Algorithms-for-manifold-learning-Cayton/100dcf6aa83ac559c83518c8a41676b1a3a55fc0/figure/0