Machine Learning Crash Course

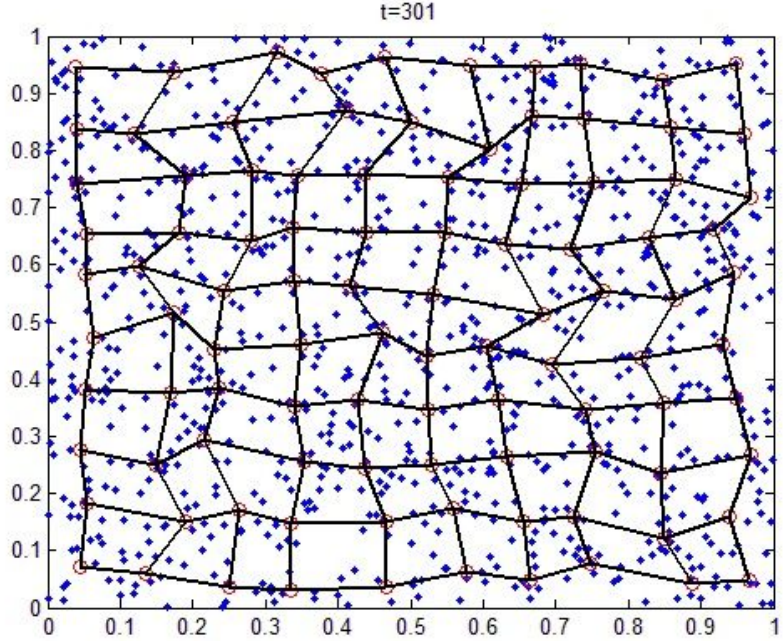
WEEK 8

"Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, we will have multiplied the intelligence, the human biological machine intelligence of our civilization a billion-fold."
~ Ray Kurzweil

"By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it." ~ Eliezer Yudkowsky

Topics

- Unsupervised Learning
  - Clustering
    - Introduction
    - K-Means
    - Optimization Objective
    - Random Initialization
    - Choosing Cluster Amount
- Dimensionality Reduction
  - Reasons for Usage
    - Data Compression
    - Visualization
  - Principal Component Analysis (PCA)
    - Summary



1.
Unsupervised Learning

1a. Clustering

Introduction

- Learning algorithms without needed input labels "y" in the usual (x,y) pair
- Good for finding structures within data and for visualization
- We consider finding a set of clusters to group together similar data

K-Means

- Choose the number "k" clusters you want
  - k will also be the number of centroids to consider
  - Choose random points for the centroids to represent

- Iterate through all your data: *O(n)*
  - Assign each data point to the most similar centroid

- Iterate for k centroids: *O(k)*
  - Calculate the average for all points in the "bag" *O(n)*
  - Reassign the centroids to the corresponding averages
  - **REPEAT UNTIL CONVERGENCE "*OR*" TERMINATION**

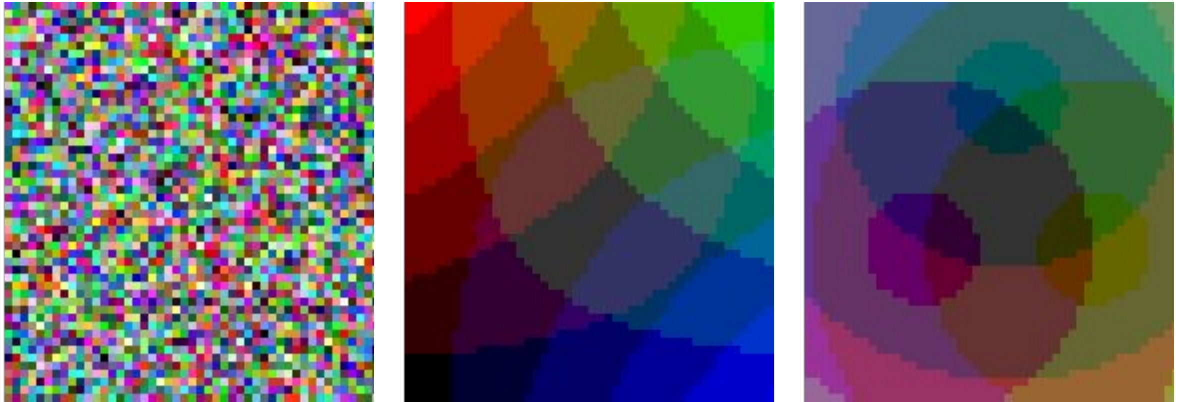
Optimization Objective

- Min(Cost Function) - Min(Distortion Function)
- Given “m” samples
- Cost equals the average of the squared length of the difference between some “x” data point and some “u_c” centroid

Random Initialization

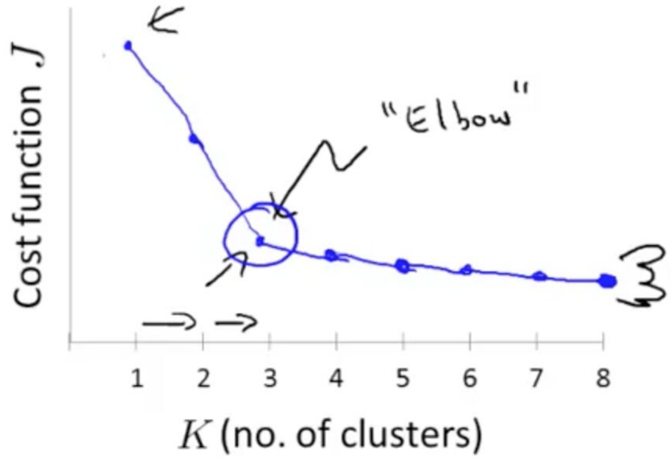
- The algorithm will sometimes come up with different solutions depending on this stochastic initialization

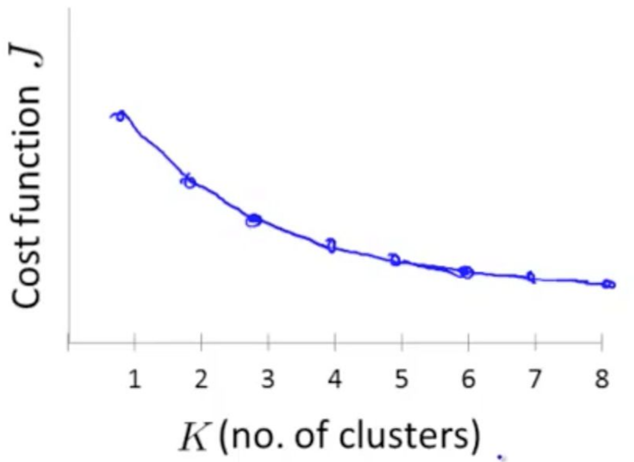
- The algorithm can get stuck at a “bad” local optima



Amount of Clusters

- k should always be less than the number of samples
- **Elbow Method**



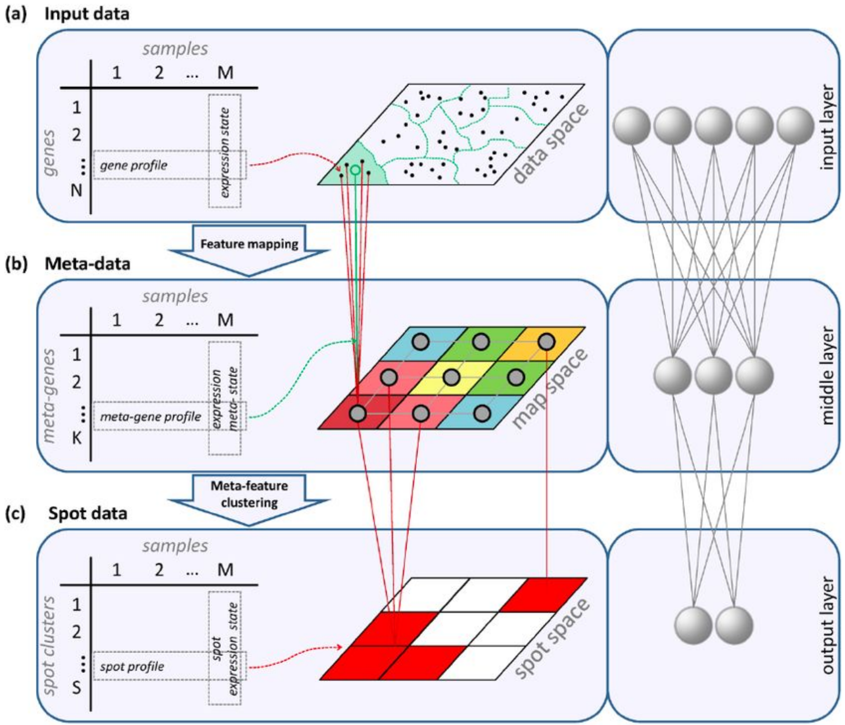


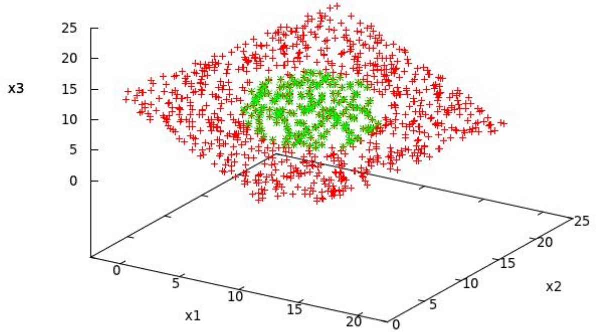
2.

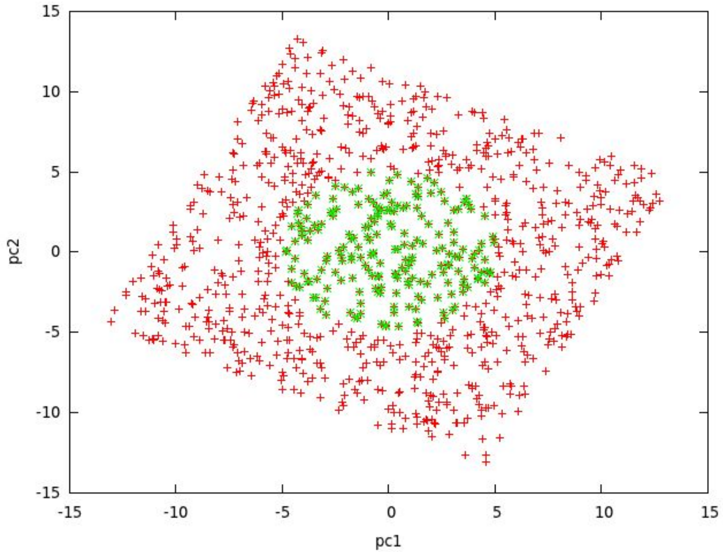
Dimensionality Reduction

Reasons for Usage

- Data Compression **(EXAMPLE)**
- Data Visualization **(EXAMPLE)**







SUMMARY (DEMO)

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

e=nx1 vector of ones
$X=A-(1/n)ee^TA$
$Y=X^TX$
<u>Depending on the Dataset Used:</u>
$\Sigma=(1/n)Y$
$\Sigma=(1/(n-1))Y$

- Preprocessing Step:
  - Feature Scaling / Mean Normalization
- Compute the vectors to project the data on
  - Use Singular Value Decomposition
  - Use the obtained "U" matrix to obtain the kx1(dim) vector z
  - **Equivalently:**
    - **Calculate the Covariance Matrix**
    - **Calculate the eigenvectors of Σ**
    - **Use $U=[u_1,u_2,u_3,\ldots,u_n] \epsilon R^{nxn}$ to get $z=reduce[u_1,u_2,\ldots,u_k]^Tx=[u^T_1;u^T_2;\ldots;u^T_k]x$**
- Choosing k principal components
  - Initialize small to minimize the averaged square projection error

Supervised vs Unsupervised

Let's review some concepts

Unsupervised Learning

Learning without labels!

K-Means

Clustering w/ K clusters & centroids and using a similarity metric to "bag" similar data to the centroids together.

Initializations

Random initializations are usually better, despite (the programmer) having to run it more to find a more optimal solution.

Dimensionality Reduction

Mapping N dimensional data to a lower dimension as specified by the programmer. Highly recommended!

PCA

Calculate the z "kx1 (dim)" vector using the transformed(reduced(covariance_matrix))x

Amount of Clusters

Run the program many times to find the elbow in the cost function plot.

TIME FOR CODE!

K-Means & Kohonen SOMs

CODE DEMO 1

https://github.com/ECE-Engineer/MachineLearning-BigData-Project/blob/master/365/src/main/java/GUI.java

CODE DEMO 2

http://cs.oswego.edu/~kzeller/Portfolio/coursework/csc466/AI.html

Thanks!

Any questions?

You can find me at:

cs.oswego.edu/~kzeller

https://github.com/ECE-Engineer

(Course Material) →
https://github.com/ECE-Engineer/Machine-Learning-Foundations

Credits

- https://www.ritchieng.com/
- https://nghiaho.com/?page_id=1030
- https://www.researchgate.net/figure/Two-step-data-compression-using-SOM-machine-learning-Firstly-the-input-data-are_fig1_262232523
- https://www.mathworks.com/matlabcentral/fileexchange/46481-self-organizing-map-kohonen-neural-network