On the relation between Big Data and Machine Learning

Leandro Ariza-Jiménez, MSc.

Advisors:

Olga Lucía Quintero, Ph.D.

Javier Correa Álvarez, Ph.D

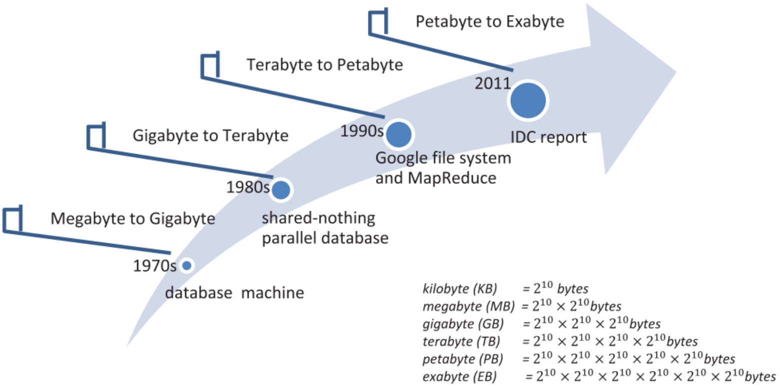
Nicolas Pinel Peláez, Ph.D.

Mathematical Modeling Research Group
School of Mathematical Sciences
Universidad EAFIT

May 30, 2016

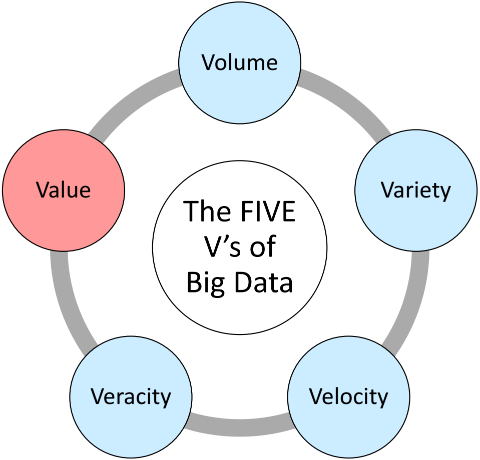
What is Big Data?

Buzzword vs. Data-explosion trend [12]



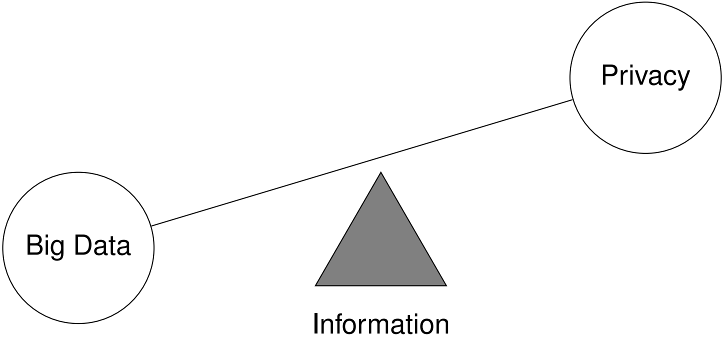
Source: [6]

A definition for Big Data



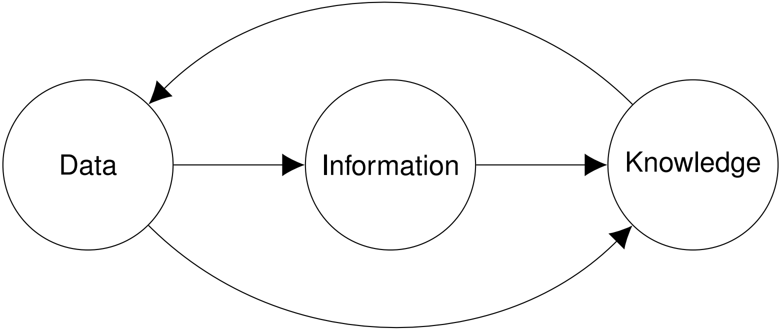
Source: Adapted from [6, 4, 11]

Big Data vs. Privacy



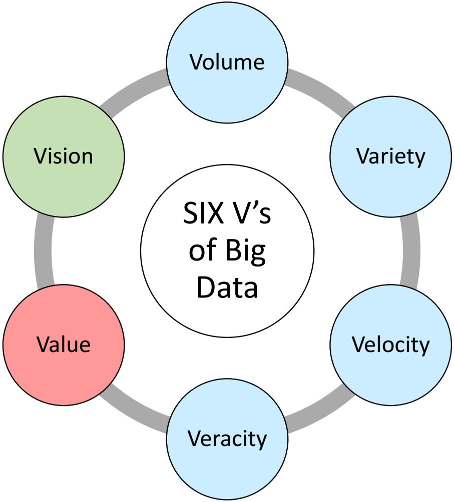
Source: Based on [11]

Artificial Intelligence



"Knowledge from data through information"

Our vision



Volume
Variety
Velocity
Veracity
Value
Vision
SIX V's
of Big
Data

What is Machine Learning?

- Machine Learning (ML) refers to a branch of the Artificial Intelligence field.
- ML concerns to the study and construction of algorithms with the ability to learn from the existing data:

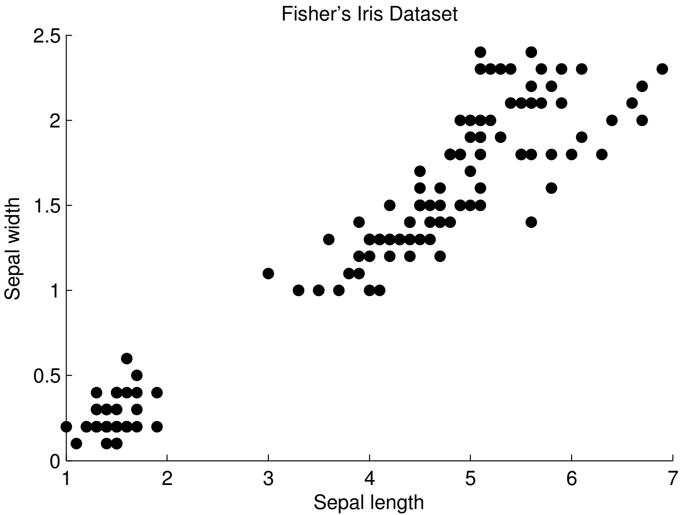
  > "A machine learns to perform a task $\mathcal{T}$ if its performance as measured by $\mathcal{P}$ increases with the experience $\mathcal{E}$ [1].

- Paradigms of learning:
  - Supervised
  - Unsupervised
  - Semisupervised

Big Data and ML relation

- ML can be used to provide us with intelligent analysis of Big Data.
- ML can contribute to every attribute of Big Data.
- ML has had to adapt to Big Data challenge:
  - Becoming scalable from single-machine implementation to cluster-of- machines implementations.
  - Being able to do parallel-processing of large volumes of data.

Clustering

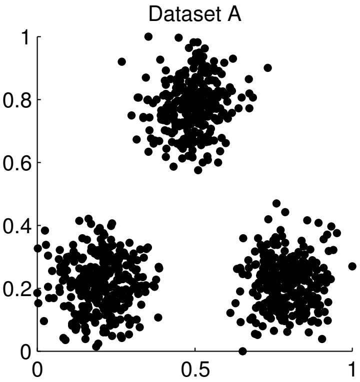


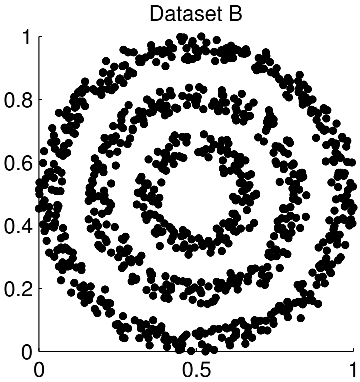
Fisher's Iris Dataset
Sepal width
Sepal length
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7

Clustering algorithms

- Clustering algorithms to describe and discuss:
  - K-means clustering
  - Mountain clustering
  - Subtractive clustering
  - Fuzzy C-means (FCM) clustering
  - Spectral clustering
- Assumptions:
  - Number of clusters is known a priori.
  - Euclidian distance is the similarity measure.

Artificial datasets





K-means clustering I

- Input: n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$; number of clusters K.
- Output: Cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$; membership matrix U.
- Steps:
  1. Select randomly K data points from the dataset as cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$.
  2. Determine the entries u_{ij} of the membership matrix U:

$$u_{ij} = \begin{cases} 1, & \|\mathbf{x}_j - \mathbf{c}_i\|^2 \leqslant \|\mathbf{x}_j - \mathbf{c}_l\|^2, \ l \neq i \\ 0, & \text{otherwise} \end{cases}$$

  3. Update the center of each cluster:

$$\mathbf{c}_i = \frac{1}{N_i} \sum_{j=1}^{n} u_{ij} \mathbf{x}_j$$

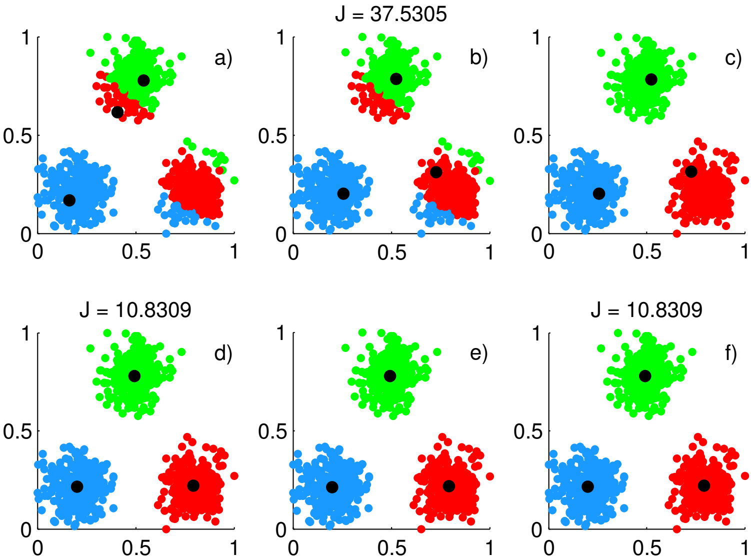
K-means clustering II

4. Compute the cost function J:

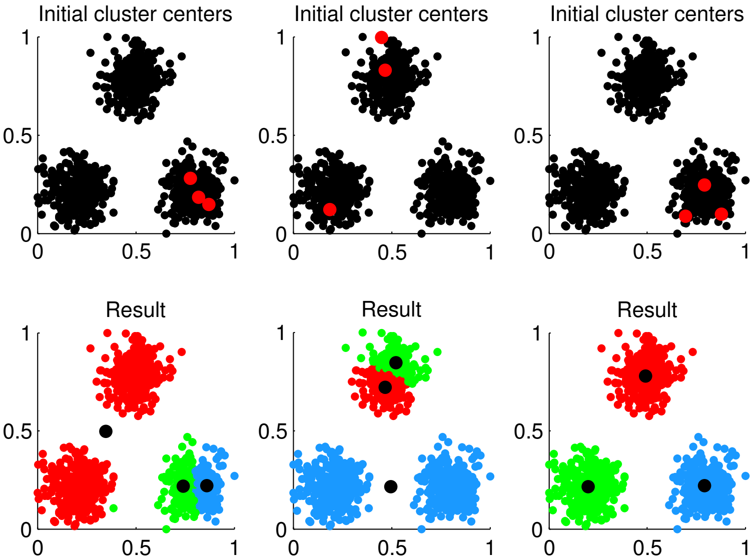
$$J = \sum_{i=1}^{K} \sum_{j=1}^{n} u_{ij} \left\| \mathbf{x}_j - \mathbf{c}_i \right\|^2$$

5. Repeat steps 2 to 4 until cost function J converges.

K-means clustering: Step-by-step



K-means clustering: Sensitivity



Mountain clustering I

- Input: n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$; constants σ and β; number of clusters K.
- Output: Cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$
- Steps:
  1. Form grid on the data space. Let V be the set of all of the points or nodes where the grid lines intersect each other.
  2. Set $i = 1$. Compute the value of the mountain function m_i at each point $\mathbf{v} \in V$ as follows:

     $$m_i(\mathbf{v}) = \sum_{j=1}^{n} \exp\left(-\frac{\left\|\mathbf{v} - \mathbf{x}_j\right\|^2}{2\sigma^2}\right)$$

     where σ is an application specific constant.
  3. Determine the point $\mathbf{v}$ at which the function m_i reaches the highest value and designate this point as the cluster center $\mathbf{c}_i$.

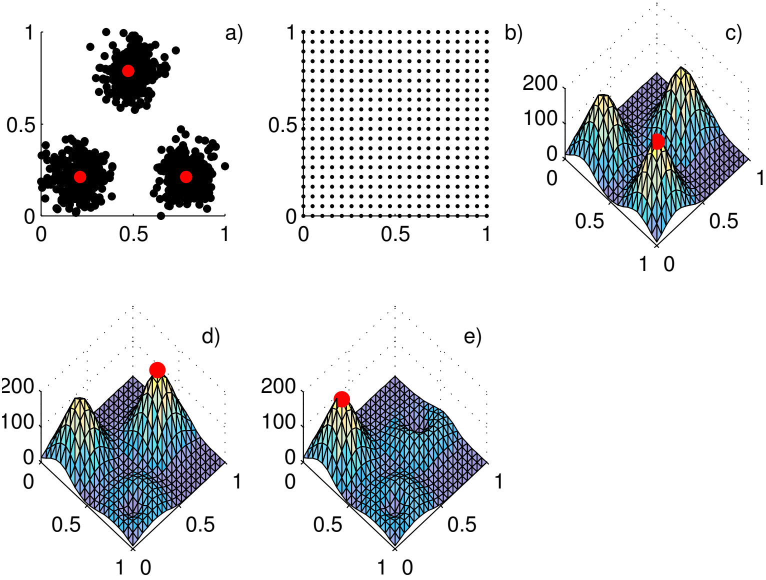
Mountain clustering II

4. Compute the value of the new mountain function m_{i+1} at each point $\mathbf{v} \in V$ as follows:

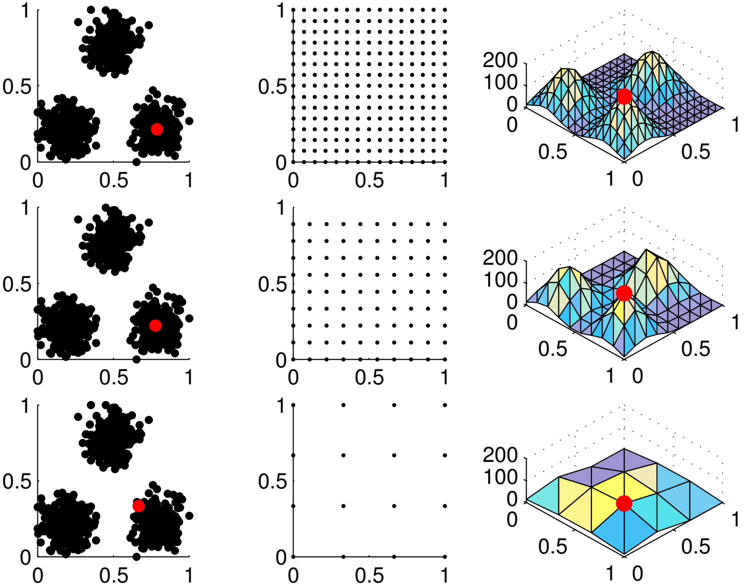
$$m_{i+1}(\mathbf{v}) = m_i(\mathbf{v}) - m_i(\mathbf{c}_i) \exp\left(-\frac{\left\|\mathbf{v} - \mathbf{c}_i\right\|^2}{2\beta^2}\right)$$

   where β is an application specific constant.
5. Set $i = i + 1$.
6. Repeat steps 3 to 5 while $i \leqslant K$.

Mountain clustering: Step-by-step ($\alpha=\beta=0.1$)



Mountain clustering: Grid fineness ($\alpha=\beta=0.1$)



Subtractive clustering I

- Input: n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$; positive radii r_a and r_b; number of clusters K.
- Output: Cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$
- Steps:

1. Set $i = 1$. Calculate a density measure D_i at each data point $\mathbf{x}_j$ as follows:

$$D_i\left(\mathbf{x}_j\right) = \sum_{l=1}^{n} \exp\left(-\frac{\left\|\mathbf{x}_j - \mathbf{x}_l\right\|^2}{\left(r_a/2\right)^2}\right)$$

   where r_a is a positive constant.
2. Find the data point $\mathbf{x}_j$ with the hightest density measure D_i and designate it as the cluster center $\mathbf{c}_i$.

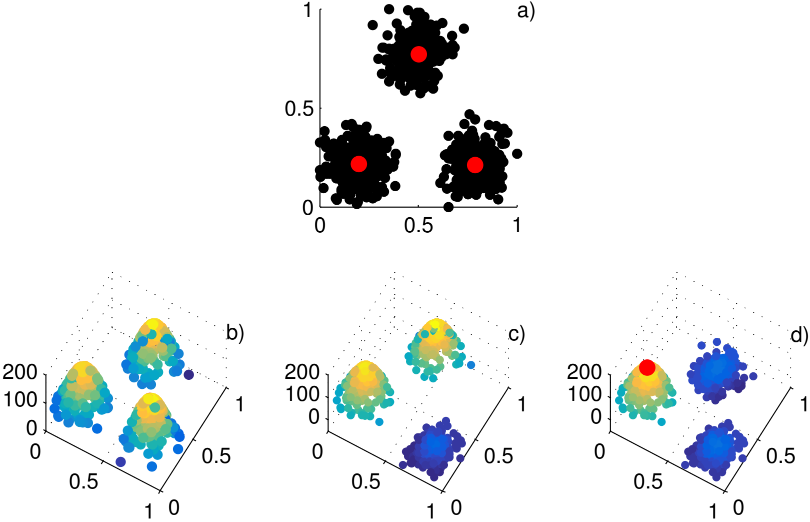
Subtractive clustering II

3. Calculate a new density measure D_{i+1} at each data point $\mathbf{x}_j$ as follows:

$$D_{i+1}\left(\mathbf{x}_j\right) = D_i\left(\mathbf{x}_j\right) - D_i\left(\mathbf{c}_i\right)\exp\left(-\frac{\left\|\mathbf{x}_j - \mathbf{c}_i\right\|^2}{\left(r_b/2\right)^2}\right)$$

   where r_b is a positive constant.
4. Set $i = i + 1$.
5. Repeat steps 2 to 4 while $i \leqslant K$.

Subtractive clustering: Step-by-step ($r_a = 0.3, r_b = 1.5 r_a$)



FCM clustering I

- Input: n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$; number of clusters c; fuzzification parameter m.
- Output: Cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$; membership matrix U.
- Steps:
  1. Initialize randomly the fuzzy membership matrix U.
  2. Calculate the cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_c$:

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} (u_{ij})^m \mathbf{x}_j}{\sum_{j=1}^{n} (u_{ij})^m}$$

  3. Determine the entries u_{ij} of a new the fuzzy membership matrix U:

$$u_{ij} = \left[\sum_{l=1}^{c} \left(\frac{\|\mathbf{c}_i - \mathbf{x}_j\|}{\|\mathbf{c}_l - \mathbf{x}_j\|} \right)^{2/(m-1)} \right]^{-1}$$

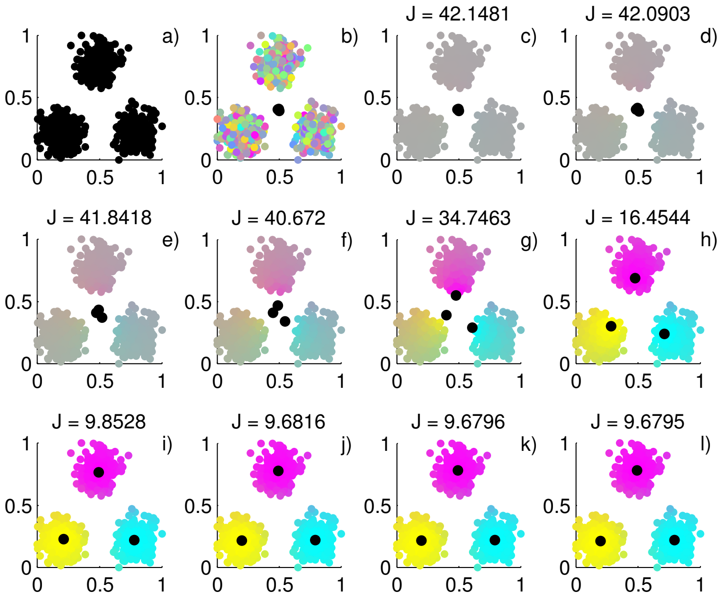
FCM clustering II

4. Compute the cost function J:

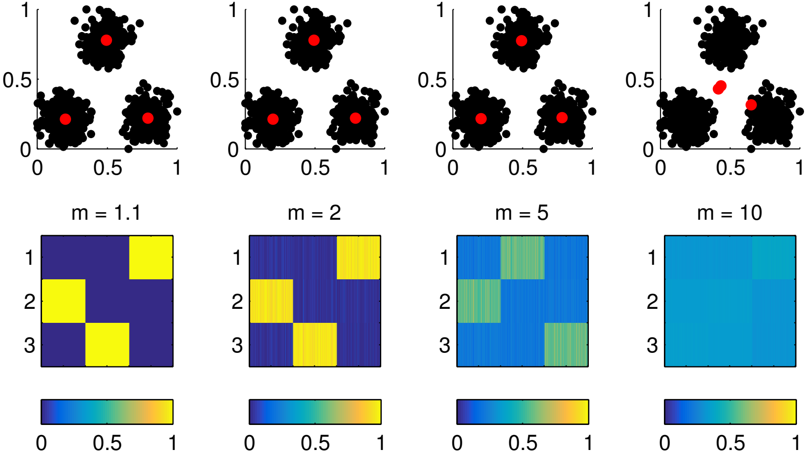
$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m \|\mathbf{x}_j - \mathbf{c}_i\|^2$$

5. Repeat steps 2 to 4 until cost function J converges.

FCM clustering: Step-by-step ($m = 2$)



FCM clustering: Effect of *m*

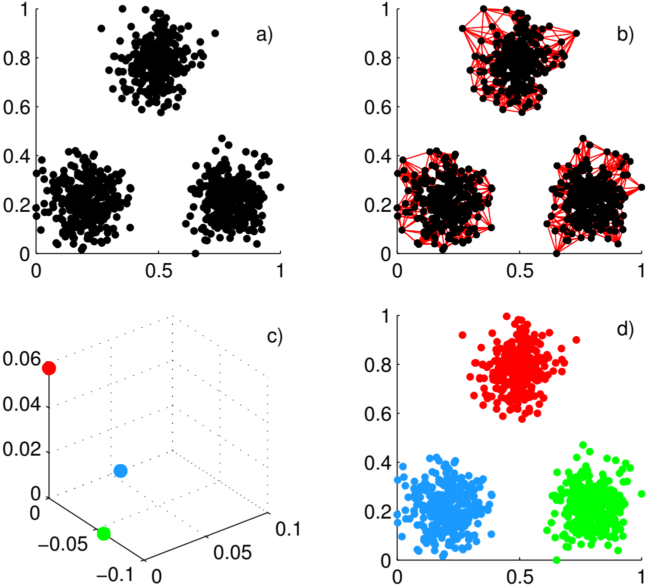


Spectral clustering I

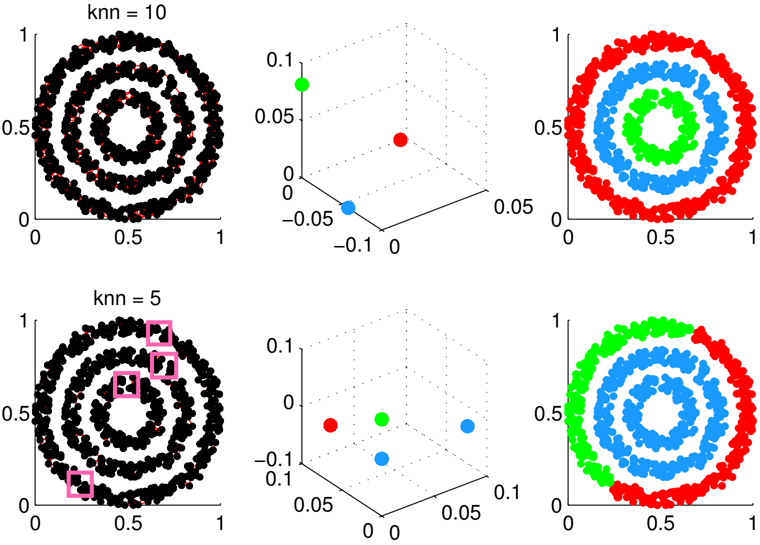
- Input: n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$; number of k-nearest neighbors; number of clusters K.
- Output: K clusters assignments
- Steps:

1. Create an adjacency matrix $W = [w_{ij}]$, $i, j = 1 \ldots n$, for the dataset based on the *k-nearest neighbors approach*.
2. Construct the degree matrix $D = \mathrm{diag}\,(d_1, \ldots, d_n)$, where $d_i = \sum_{i=1}^{n} w_{ij}$.
3. Compute the Laplacian matrix $L = D - W$.
4. Find $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K$ the first K eigenvectors of L with the smallest eigenvalues, and form the matrix $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K]$.
5. Consider the rows of U as a new set of data poins $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, and cluster them into K clusters using the K-means algorithm.
6. Assign the original data point $\mathbf{x}_i$ to the cluster k if and only if the data point $\mathbf{y}_i$ was assigned to cluster k.

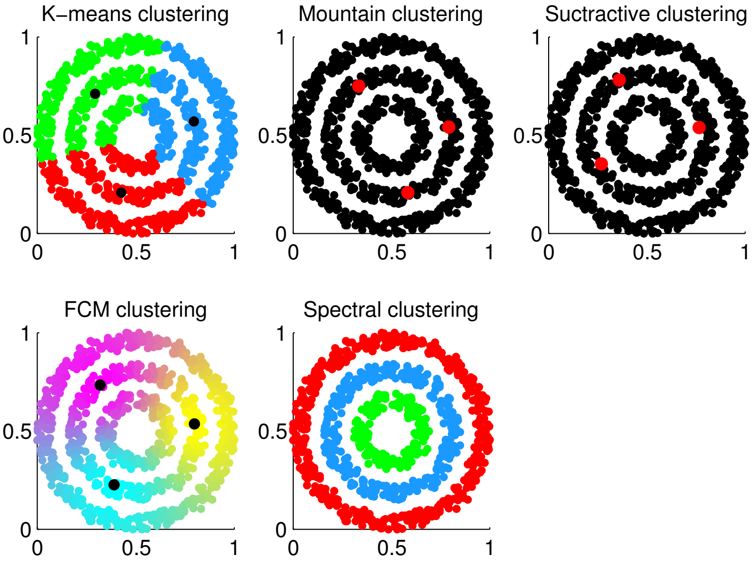
Spectral clustering: Step-by-step ($knn = 10$)



Spectral clustering: *knn* effect



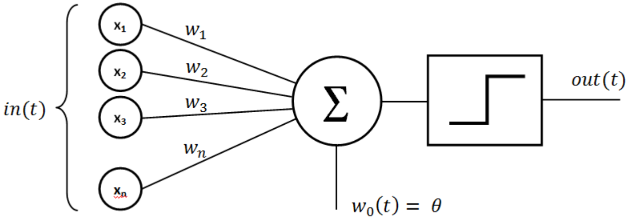
Clustering performance on Dataset B



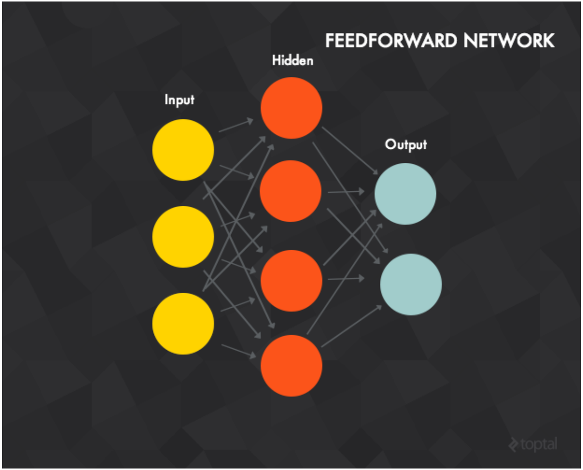
Deep Learning

- Emerging area of ML.
- Typically uses Artificial Neural Networks (ANN).
- It is about learning...
  - with deep architectures,
  - and overcoming problems of these arquitectures.

Deep Learning: Perceptron

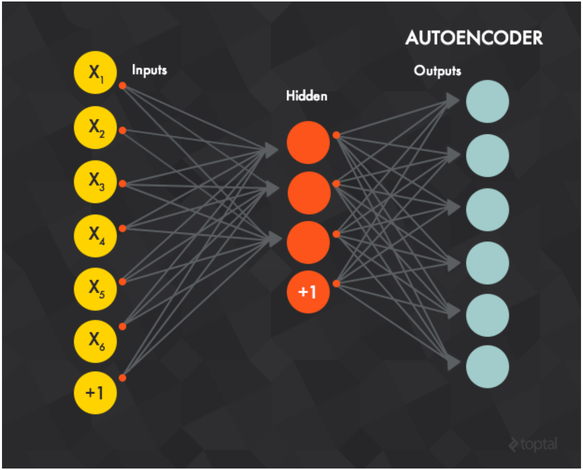


Deep Learning: Multilayer Perceptron



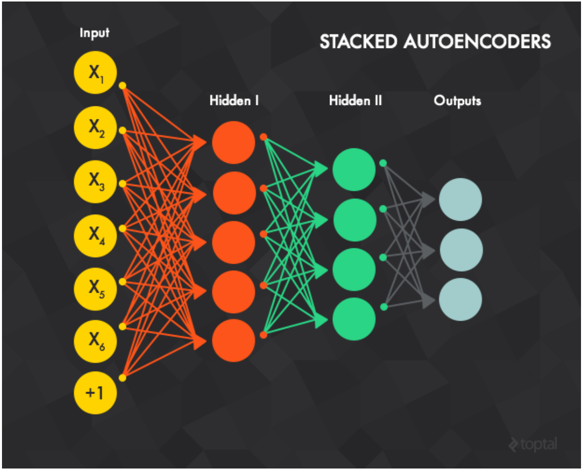
Source: https://assets.toptal.io/uploads/blog/image/333/toptal-blog-image-1395721488746.png

Deep Learning: Autoencoder



Source: https://assets.toptal.io/uploads/blog/image/340/toptal-blog-image-1395769623098.png

Deep Learning: Stacked Autoencoders



Source: https://assets.toptal.io/uploads/blog/image/335/toptal-blog-image-1395721542588.png

Future work

- To implement Local Spectral Clustering (LSC) [**?**].
- To apply LSC on finding local community structures in large networks.
- To attend Cornell's Program for Research Experience:
  - 2016 topic is Deep Learning

Bibliography I

[1] Juan Carlos Cardona-Gómez, Leandro Fabio Ariza-Jimérnez, and Juan Carlos Gallego-Gómez.
Cell Biology - New Insights, chapter A Proposal for a Machine Learning Classifier for Viral Infection in Living Cells Based on Mitochondrial Distribution.
InTech, 2016.

[2] Stephen L. Chiu.
Fuzzy Model Identification Based on Cluster Estimation.
Journal of Intelligent and Fuzzy Systems, 2(3):267–278, 1994.

[3] Geoff Dougherty.
Pattern Recognition and Classification.
Springer New York, 2013.

Bibliography II

[4] Chris Eaton, Dirk Deroos, Tom Deutsch, George Lapis, and Paul Zikopoulos.
Understanding Big Data.
McGraw-Hill, 2012.

[5] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl.
Cluster Analysis.
Wiley, 5th edition edition, 2011.

[6] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li.
Toward Scalable Systems for Big Data Analytics: A Technology Tutorial.
IEEE Access, 2:652–687, 2014.

[7] Jyh-Shing Roger Jang, Chuen-TsaiSun, and Eiji Mizutani.
Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence.
Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1997.

Bibliography III

[8] Bryan F. J. Manly.
Multivariate Statistical Methods: A Primer.
Chapman & Hall/CRC Press, 3rd edition, 2005.

[9] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss.
On Spectral Clustering: Analysis and an Algorithm.
In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors,
Advances in Neural Information Processing Systems 14,
pages 849–856, 2001.

[10] Daniel E. O'Leary.
Artificial Intelligence and Big Data.
IEEE Intelligent Systems, 28(2):96–99, 2013.

[11] Daniel E. O'Leary.
Big Data and Privacy: Emerging Issues.
IEEE Intelligent Systems, 30(6):92–96, 2015.

Bibliography IV

[12] Gil Press.
A Very Short History of Big Data.
http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#574303b355da, May 2013.
Accesed: 24/04/2016.

[13] Ulrike von Luxburg.
A Tutorial on Spectral Clustering.
Statistics and Computing, 17(4):395–416, 2007.

[14] Jeremy Watt.
An Intro to Spectral Clustering.
https://ece.uwaterloo.ca/~nnikvand/Coderep/spectralclustering-1.1/spectral_clustering_draft_33.pdf, March 2013.
Accesed: 14/03/2016.

Bibliography V

[15] Rui Xu and Donald C. Wunsch.
Clustering Algorithms in Biomedical Research: A Review.
IEEE Reviews in Biomedical Engineering, 3:120–154, 2010.

[16] R. R. Yager and D. P. Filev.
Approximate Clustering Via the Mountain Method.
IEEE Transactions on Systems, Man, and Cybernetics, 24(8):1279–1284, Aug 1994.