Machine Learning I: Fractal 2

Executive M.Tech. in AI for Working Professionals
Semester 1, 2021

Instructor

Rajendra Nagar

Assistant Professor

204, Dept. of EE, IIT Jodhpur

Education: B.Tech. (IIT Jodhpur) and Ph.D. (IIT Gandhinagar)

Research Interests: Computer Vision & Graphics and 3D Shape Analysis

Email: `rn@iitj.ac.in`

Homepage: http://home.iitj.ac.in/~rn/

Content

Clustering

k-means clustering, Spectral Clustering.

Parameter Estimation

Maximum Likelihood and Bayesian Parameter Estimation, Gaussian Mixture Modeling, EM-algorithm.

Feature Selection and Dimensionality Reduction

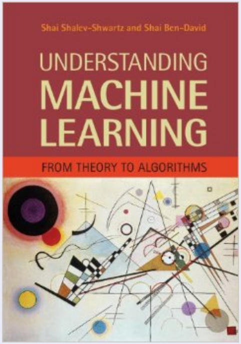
Principal Component Analysis, Linear Discriminant Analysis, Independent Component Analysis, SFFS, SBFS,Distance-based methods, Linear Discriminant Functions

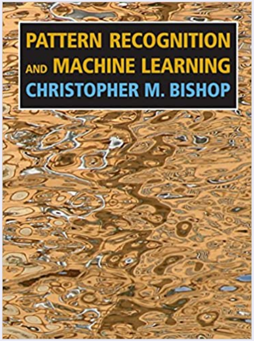
Evaluation

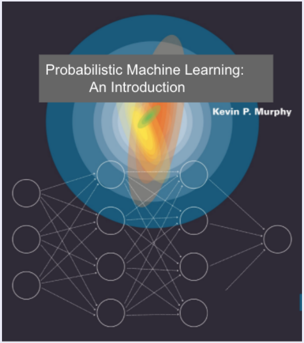
Minor 2	20%
Quizzes	5%
Programming Assignment	5%

Reading Material

Books







Similar Course

Machine Learning, CMU

Conference Papers

ICML, NeurIPS, CVPR, ICCV etc.

Course Material and Q&As

Google Classroom

- Assignments and quizzes will be posted here.

Course Website

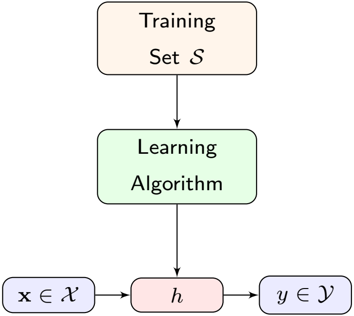
- https://sites.google.com/iitj.ac.in/ml1f2-21/
- Recordings, slides, notes, and reference material will be uploaded after every class.
- Chapter-wise reference to every lecture and further readings.
- Text and Reference books.

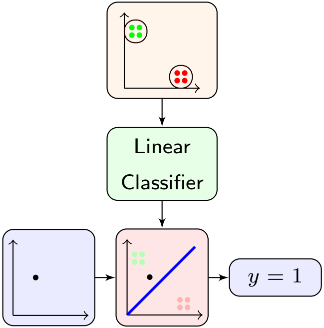
Q&As or Doubt Sessions

- Post on the Google Classroom.
- Email to me: `rn@iitj.ac.in`.
- A doubt session can also be arranged.
- Contact ours are already scheduled.

Supervised Learning

Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ be a training set. Here, $\mathbf{x}_i$ is the i^{th} training input, e.g. an image, and y_i is the corresponding label, e.g. "cat". Let $\mathcal{X}$ be the set of all inputs and let $\mathcal{Y}$ be the set of all possible output labels and let $h : \mathcal{X} \to \mathcal{Y}$ be a predictor. Then, our goal is to find h such that $h(\mathbf{x}_i)$ is equal to the true label of the input $\mathbf{x}_i$.

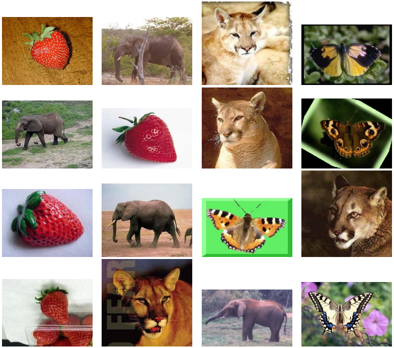


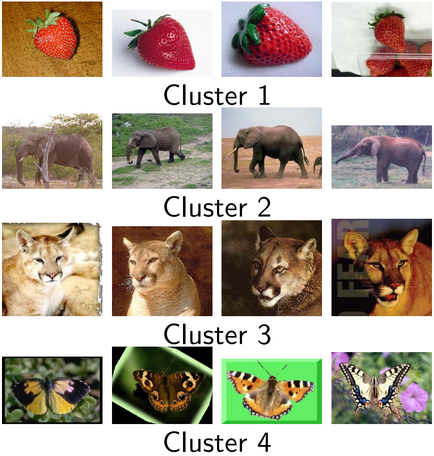


Unsupervised Learning

The dataset does not contain any labeled points. The task is to learn meaningful information without any labels.

Clustering



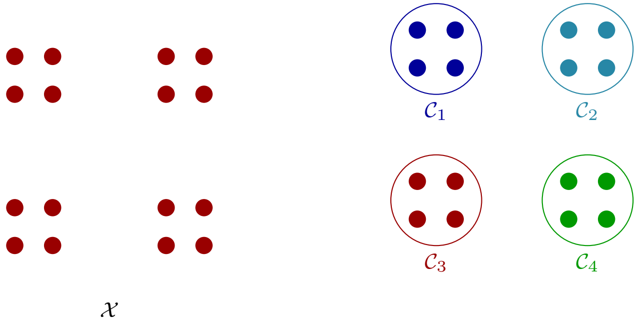


Clustering

- Identifying meaningful groups among the data points without using any supervision.
- For example, computational biologists use similarities in gene expressions to cluster genes.
- Retailers cluster customers, on the basis of their customer profiles, for the purpose of targeted marketing.
- Astronomers cluster stars on the basis of their spatial proximity.
- Clustering is the task of partitioning a set into groups of points such that similar points end up in the same group and dissimilar points are separated into different groups.

Clustering Problem

- Given a set of data points, $\mathcal{X}$, and a distance function over it. That is, a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ that gives distance between two points.
- Our goal is to partition the input dataset set $\mathcal{X}$ into k subsets/cluster/groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ such that $\cup_{i=1}^{k} \mathcal{C}_i = \mathcal{X}$, and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \forall i \neq j$.

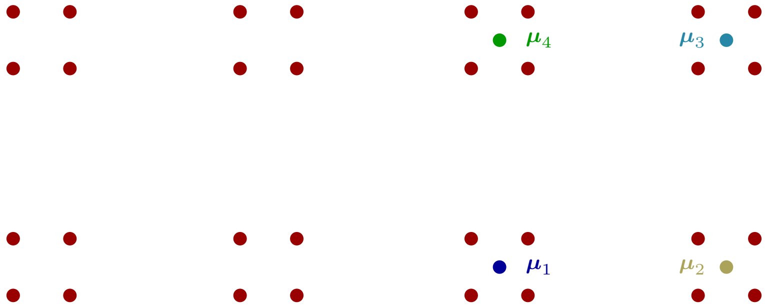


Clustering Algorithms

- Centroid models: k-Means
- Graph-based models: Spectral Clustering
- Distribution models: Gaussian Mixture Models
- Density models: DBSCAN
- Connectivity Based: Hierarchical Clustering
- Neural models: Self-organizing map

k-Means: Problem Formulation[1,2]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of data-points, where $\mathbf{x}_i \in \mathbb{R}^m$. We want to partition $\mathcal{X}$ into groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar points. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centers), where $\boldsymbol{\mu}_i \in \mathbb{R}^m$.

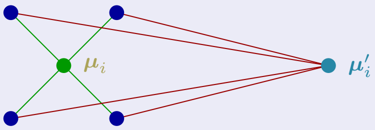


[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means Clustering

- How do we choose the cluster centers $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$?
- Let us assume that the groups/clusters $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ are given to us.
- Now, consider the i-th cluster $\mathcal{C}_i$.
- How do you find the group representative $\boldsymbol{\mu}_i$ for this group?
- How about the one who has best friendship with every member of the group?
- The best $\boldsymbol{\mu}_i$ should have as minimum as possible distance from all points of the cluster $\mathcal{C}_i$.



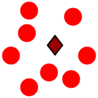
$$\begin{aligned} \text{Sum of distances} &= \sum_{\mathbf{x} \in \mathcal{C}_i} d(\mathbf{x}, \boldsymbol{\mu}_i) \\ \boldsymbol{\mu}_i^\star &= \arg\min_{\boldsymbol{\mu}_i} \sum_{\mathbf{x} \in \mathcal{C}_i} d(\mathbf{x}, \boldsymbol{\mu}_i) \\ &= \frac{\sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x}}{|\mathcal{C}_i|}. \end{aligned}$$

- Now lets consider all the clusters together, then the best group representatives can be found as

$$(\boldsymbol{\mu}_1^\star, \ldots, \boldsymbol{\mu}_k^\star) = \arg\min_{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k} \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} d(\mathbf{x}, \boldsymbol{\mu}_j).$$

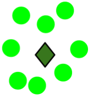
Given the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$, it is easy to find the respective optimal centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$.

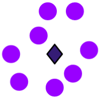
$$\boldsymbol{\mu}_1 = \frac{\sum_{\mathbf{x} \in \mathcal{C}_1} \mathbf{x}}{|\mathcal{C}_1|}$$



$$\boldsymbol{\mu}_2 = \frac{\sum_{\mathbf{x} \in \mathcal{C}_2} \mathbf{x}}{|\mathcal{C}_2|}$$

$$\boldsymbol{\mu}_3 = \frac{\sum_{\mathbf{x} \in \mathcal{C}_3} \mathbf{x}}{|\mathcal{C}_3|}$$

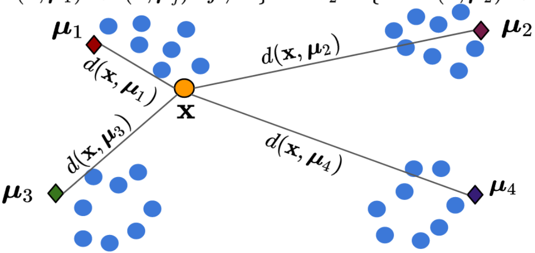




$$\boldsymbol{\mu}_4 = \frac{\sum_{\mathbf{x} \in \mathcal{C}_4} \mathbf{x}}{|\mathcal{C}_4|}$$

Given the optimal centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, it is easy to find the clusters $\mathcal{C}_1, \ldots \mathcal{C}_k$.

$\mathcal{C}_1 = \{\forall \mathbf{x} : d(\mathbf{x}, \boldsymbol{\mu}_1) < d(\mathbf{x}, \boldsymbol{\mu}_j) \ \forall j \neq 1\}$ $\quad$ $\mathcal{C}_2 = \{\forall \mathbf{x} : d(\mathbf{x}, \boldsymbol{\mu}_2) < d(\mathbf{x}, \boldsymbol{\mu}_j) \ \forall j \neq 2\}$



$\mathcal{C}_3 = \{\forall \mathbf{x} : d(\mathbf{x}, \boldsymbol{\mu}_3) < d(\mathbf{x}, \boldsymbol{\mu}_j) \ \forall j \neq 3\}$ $\quad$ $\mathcal{C}_4 = \{\forall \mathbf{x} : d(\mathbf{x}, \boldsymbol{\mu}_4) < d(\mathbf{x}, \boldsymbol{\mu}_j) \ \forall j \neq 4\}$

Algorithm 1 k-Means Algorithm

Input: $\mathcal{X} \subset \mathbb{R}^m$, Number of clusters k

Initialize: Randomly choose initial centroids $\boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_k^{(0)}$

while not converged **do**

  for $i \in [k]$ **do**

    $\mathcal{C}_i^{(t+1)} \leftarrow \left\{ \forall \mathbf{x} \in \mathcal{X} : d(\mathbf{x} - \boldsymbol{\mu}_i^{(t)}) < d(\mathbf{x}, \boldsymbol{\mu}_j^{(t)}) \; \forall j \in [k] \backslash \{i\} \right\}$

    $\boldsymbol{\mu}_i^{(t+1)} \leftarrow \frac{1}{|\mathcal{C}_i^{(t+1)}|} \sum_{\mathbf{x} \in \mathcal{C}_i^{(t+1)}} \mathbf{x}$

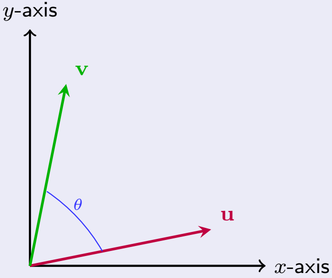
    $t \leftarrow t + 1$

  end for

end while

Linear Algebra Basics

Let $\mathbf{u} = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix}^\top$ and $\mathbf{v} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^\top$ be two vectors in $\mathbb{R}^n$.



$$
\begin{aligned}
\textbf{2-Norm: } \|\mathbf{v}\|_2 &= \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \\
\textbf{Inner-product: } \langle \mathbf{u}, \mathbf{v} \rangle &= \mathbf{u}^\top \mathbf{v} = \mathbf{u}^\top \mathbf{v} = \langle \mathbf{v}, \mathbf{u} \rangle \\
&= u_1 v_1 + u_2 v_2 + \cdots + u_n v_n \\
&= \|\mathbf{u}\|_2 \times \|\mathbf{v}\|_2 \cos\theta \\
\mathbf{v}^\top \mathbf{v} &= v_1^2 + v_2^2 + \cdots + v_n^2 \\
&= \|\mathbf{v}\|_2^2 \\
\textbf{Distance: } \|\mathbf{u} - \mathbf{v}\|_2 &= \sqrt{(\mathbf{u} - \mathbf{v})^\top (\mathbf{u} - \mathbf{v})} \\
&= \sqrt{\mathbf{u}^\top \mathbf{u} - 2\mathbf{u}^\top \mathbf{v} + \mathbf{v}^\top \mathbf{v}}.
\end{aligned}
$$

Orthogonal Vectors: Two unit norm vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ will be orthogonal to each other, if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ or $\mathbf{u}^\top \mathbf{v} = 0$.

Gradient: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Then, its gradient $\nabla f \in \mathbb{R}^n$ is defined as

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial v_1} & \frac{\partial f}{\partial v_2} & \cdots & \frac{\partial f}{\partial v_n} \end{bmatrix}^\top .$$

$$\nabla_{\boldsymbol{\mu}_i} f = \mathbf{0} \Rightarrow \nabla_{\boldsymbol{\mu}_i} \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} \|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2 = \mathbf{0}, \ \forall i \in [k]$$

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}_i} \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} \|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2 &= \sum_{\mathbf{x} \in \mathcal{C}_i} \nabla_{\boldsymbol{\mu}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2 \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \nabla_{\boldsymbol{\mu}_i} \left((\mathbf{x} - \boldsymbol{\mu}_i)^\top (\mathbf{x} - \boldsymbol{\mu}_i) \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \nabla_{\boldsymbol{\mu}_i} \left((\mathbf{x}^\top - \boldsymbol{\mu}_i^\top)(\mathbf{x} - \boldsymbol{\mu}_i) \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \nabla_{\boldsymbol{\mu}_i} \left(\mathbf{x}^\top \mathbf{x} - \boldsymbol{\mu}_i^\top \mathbf{x} - \mathbf{x}^\top \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \nabla_{\boldsymbol{\mu}_i} \left(\mathbf{x}^\top \mathbf{x} - 2\boldsymbol{\mu}_i^\top \mathbf{x} + \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i \right)
\end{aligned}$$

$$
\begin{aligned}
\nabla_{\boldsymbol{\mu}_i} \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} \|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2 &= \sum_{\mathbf{x} \in \mathcal{C}_i} \nabla_{\boldsymbol{\mu}_i} \left(\mathbf{x}^\top \mathbf{x} - 2\boldsymbol{\mu}_i^\top \mathbf{x} + \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} (-2\mathbf{x} + 2\boldsymbol{\mu}_i) \\
\Rightarrow \sum_{\mathbf{x} \in \mathcal{C}_i} (-2\mathbf{x} + 2\boldsymbol{\mu}_i) &= \mathbf{0} \\
\Rightarrow \sum_{\mathbf{x} \in \mathcal{C}_i} \boldsymbol{\mu}_i &= \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x} \\
\Rightarrow \boldsymbol{\mu}_i \sum_{\mathbf{x} \in \mathcal{C}_i} 1 &= \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x} \\
\Rightarrow \boldsymbol{\mu}_i |\mathcal{C}_i| &= \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x} \\
\Rightarrow \boldsymbol{\mu}_i &= \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x}
\end{aligned}
$$