Machine Learning I: Fractal 2

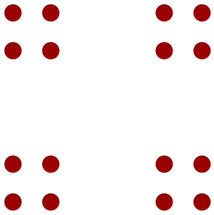
Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur

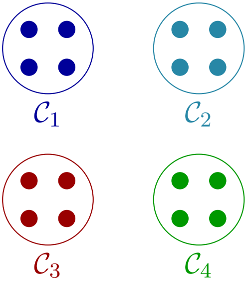
These slides are prepared from the following book:
Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

Clustering

Input: A set of elements, $\mathcal{X}$, and a distance function to measure similarity.
Objective: A partition of the input domain $\mathcal{X}$ into groups $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ of similar elements such that $\cup_{i=1}^{k} \mathcal{C}_i = \mathcal{X}$, and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \forall i \neq j$.



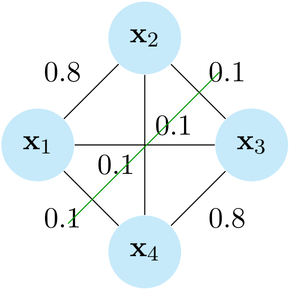
$\mathcal{X}$



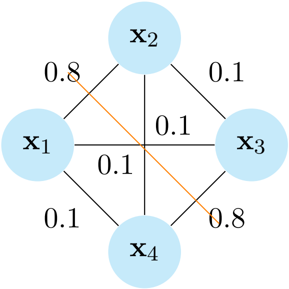
$$\mathcal{X} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_4$$

Spectral Clustering

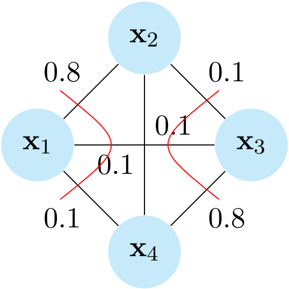
$$\text{RatioCut}(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \sum_{i=1}^{k} \frac{1}{|\mathcal{C}_i|} \sum_{r \in \mathcal{C}_i} \sum_{s \notin \mathcal{C}_i} \mathbf{W}_{r,s}.$$



$\text{RatioCut}(\mathcal{C}_1, \mathcal{C}_2) = 0.2$



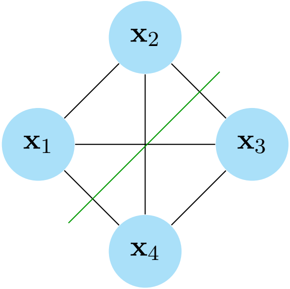
$\text{RatioCut}(\mathcal{C}_1, \mathcal{C}_2) = 0.9$



$\text{RatioCut}(\mathcal{C}_1, \mathcal{C}_2) = 1.0$

$$\min_{\mathcal{C}_1, \ldots, \mathcal{C}_k} \text{RatioCut}(\mathcal{C}_1, \ldots, \mathcal{C}_k)$$

Consider the Graph Laplacian matrix $\mathbf{L}$ of the graph constructed on $\mathcal{X}$.



Cluster Assignment Matrix

Let $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be the clustering and $\mathbf{H} \in \mathbb{R}^{n \times k}$ be a matrix such that

$$\mathbf{H}_{i,j} = \frac{1}{\sqrt{|\mathcal{C}_j|}} \mathbb{1}_{[i \in \mathcal{C}_j]}.$$

$$\mathbf{H} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \in \mathbb{R}^{n \times k}. \text{ Here, } k = 2.$$

Claim

The columns of the matrix $\mathbf{H}$ are orthonormal to each other and

$$\text{RatioCut}(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \text{trace}(\mathbf{H}^\top \mathbf{L} \mathbf{H}).$$

Problem Formulation

$$\min_{\mathcal{C}_1,\ldots,\mathcal{C}_k} \text{RatioCut}(\mathcal{C}_1,\ldots,\mathcal{C}_k) \Leftrightarrow \min_{\mathbf{H}\in\mathbb{R}^{n\times k},\mathbf{H}^\top\mathbf{H}=\mathbf{I}} \text{trace}(\mathbf{H}^\top\mathbf{L}\mathbf{H}).$$

Rayleigh quotient

$$\mathbf{v}^\star = \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^n,\mathbf{v}^\top\mathbf{v}=1} \mathbf{v}^\top\mathbf{L}\mathbf{v}$$

$$\begin{aligned} f(\mathbf{v}) &= \mathbf{v}^\top\mathbf{L}\mathbf{v} + \lambda(1-\mathbf{v}^\top\mathbf{v}) \\ \nabla f &= 2\mathbf{L}\mathbf{v} - 2\lambda\mathbf{v} \\ \mathbf{L}\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{v}^\top\mathbf{L}\mathbf{v} &= \lambda. \end{aligned}$$

Therefore, we have to minimize λ such that $\mathbf{L}\mathbf{v} = \lambda_1\mathbf{v}$. Hence, $\mathbf{v}^\star =$ eigenvector of the matrix $\mathbf{L}$ corresponding to the smallest eigenvalue $= \mathbf{u}_1$.

Rayleigh quotient

$$\mathbf{v}^\star = \underset{\mathbf{v}^\top\mathbf{u}_1=0,\mathbf{v}^\top\mathbf{v}=1}{\arg\min} \quad \mathbf{v}^\top\mathbf{L}\mathbf{v}$$

Solution

$$\begin{aligned} f(\mathbf{v}) &= \mathbf{v}^\top\mathbf{L}\mathbf{v} + \lambda(1-\mathbf{v}^\top\mathbf{v}) \\ \nabla f &= 2\mathbf{L}\mathbf{v} - 2\lambda\mathbf{v} \\ \mathbf{L}\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{v}^\top\mathbf{L}\mathbf{v} &= \lambda \end{aligned}$$

Therefore, we have to minimize λ such that $\mathbf{L}\mathbf{v} = \lambda\mathbf{v}$ and $\mathbf{v}^\top\mathbf{u}_1 = 0$. $\mathbf{v}^\star =$ eigenvector of the matrix $\mathbf{L}$ corresponding to the second smallest eigenvalue $= \mathbf{u}_2$.

Rayleigh quotient

$$\mathbf{v}^\star = \underset{\mathbf{v}^\top\mathbf{u}_i=0,\forall i<k,\mathbf{v}^\top\mathbf{v}=1}{\arg\min} \quad \mathbf{v}^\top\mathbf{L}\mathbf{v}$$

Solution

$$\begin{aligned} f(\mathbf{v}) &= \mathbf{v}^\top\mathbf{L}\mathbf{v} + \lambda(1-\mathbf{v}^\top\mathbf{v}) \\ \nabla f &= 2\mathbf{L}\mathbf{v} - 2\lambda\mathbf{v} \\ \mathbf{L}\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{v}^\top\mathbf{L}\mathbf{v} &= \lambda \end{aligned}$$

We have to minimize λ such that $\mathbf{L}\mathbf{v} = \lambda\mathbf{v}$ and $\mathbf{v}^\top\mathbf{u}_i = 0, \forall i < k$. $\mathbf{v}^\star =$ eigenvector of the matrix $\mathbf{L}$ corresponding to the k^{th} smallest eigenvalue $= \mathbf{u}_k$.

Rayleigh quotient

$$\underset{\substack{\mathbf{v}_1,\ldots,\mathbf{v}_k \\ \mathbf{v}_i^\top \mathbf{v}_j=\delta_{ij}}}{\arg\min} \sum_{i=1}^{k} \mathbf{v}_i^\top \mathbf{L}\mathbf{v}_i$$

Here, $\delta_{ij} = 1$, if $i = j$ and $\delta_{ij} = 0$, if $i \neq j$.

Solution

$$\begin{aligned}
f(\mathbf{v}_1, \ldots, \mathbf{v}_k) &= \sum_{i=1}^{k} \mathbf{v}_i^\top \mathbf{L}\mathbf{v}_i + \sum_{i=1}^{k} \lambda_i (1 - \mathbf{v}_i^\top \mathbf{v}_i) \\
\nabla_{\mathbf{v}_i} f &= 2\mathbf{L}\mathbf{v}_i - 2\lambda \mathbf{v}_i \\
\mathbf{L}\mathbf{v}_i &= \lambda_i \mathbf{v}_i \\
\mathbf{v}_i^\top \mathbf{L}\mathbf{v}_i &= \lambda_i
\end{aligned}$$

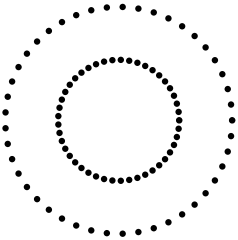
Therefore, we have to minimize $\sum_{i=1}^{k} \lambda_i$ such that $\mathbf{L}\mathbf{v}_i = \lambda \mathbf{v}_i$ and $\mathbf{v}_i^\top \mathbf{v}_j = 0$ if $i \neq j$. Hence, $\mathbf{v}_i^\star =$ eigenvector of the matrix $\mathbf{L}$ corresponding to the i^{th} smallest eigenvalue $= \mathbf{u}_i$.

Problem

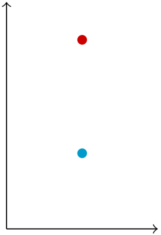
$$\mathbf{H}^{\star} = \underset{\mathbf{H}\in\mathbb{R}^{n\times k}, \mathbf{H}^{\top}\mathbf{H}=\mathbf{I}}{\arg\min} \quad \mathrm{trace}(\mathbf{H}^{\top}\mathbf{L}\mathbf{H}).$$

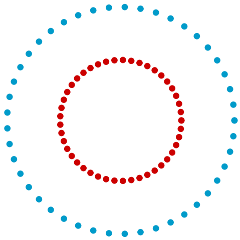
Solution

Let $\mathbf{L}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \forall i \in \{1, 2, \ldots, n\}$ be the EVD of the matrix $\mathbf{L}$. Here, we assume that the eigenvalues are such that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then, the solution to the above problem is $\mathbf{H}^{\star} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k \end{bmatrix}$.



$$\mathbf{H} = \begin{bmatrix} 1 & 0.5 \\ 1 & 0.5 \\ \vdots & \vdots \\ 1 & 0.5 \\ 1 & -0.5 \\ 1 & -0.5 \\ \vdots & \vdots \\ 1 & -0.5 \end{bmatrix}$$





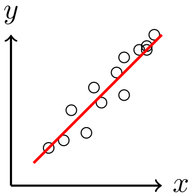
Spectral Clustering

Algorithm 1 Spectral Clustering

1: **Input:** $\mathbf{W} \in \mathbb{R}^{n \times n}$, Number of clusters k.
2: **Initialize:** Compute the graph Laplacian $\mathbf{L}$.
3: $\mathbf{H} \leftarrow$ matrix whose columns are the eigenvectors of $\mathbf{L}$ corresponding to the k-smallest eigenvalues.
4: $\mathbf{r}_1, \ldots, \mathbf{r}_n$ be the rows of $\mathbf{H}$.
5: Cluster the points $\mathbf{r}_1, \ldots, \mathbf{r}_n$ using k-means algorithm.
6: **Output:** Clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ of the k-means algorithm.

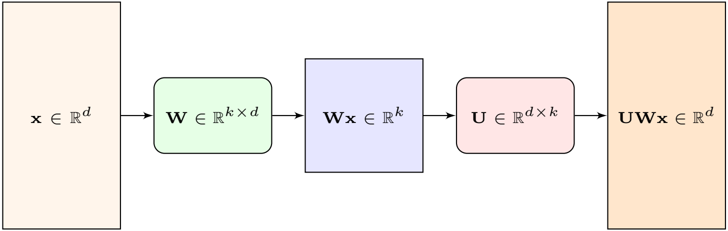
Dimensionality Reduction

- Dimensionality reduction is the process of mapping it into a new space whose dimensionality is much smaller.
- High dimensional data impose computational challenges.
- Dimensionality reduction can be used for interpretability of the data, for finding meaningful structure of the data, and for illustration purposes.



Dimensionality Reduction

- Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the input data points where each data point $\mathbf{x}_i \in \mathbb{R}^d$.
- We would like to reduce the dimensionality of these vectors using a linear transformation.
- A matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$, where $k < d$, induces a mapping $\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$, where $\mathbf{W}\mathbf{x} \in \mathbb{R}^k$ is the lower dimensionality representation of $\mathbf{x}$.
- Then, a second matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$ can be used to recover the each original vector $\mathbf{x}$ from its compressed version.



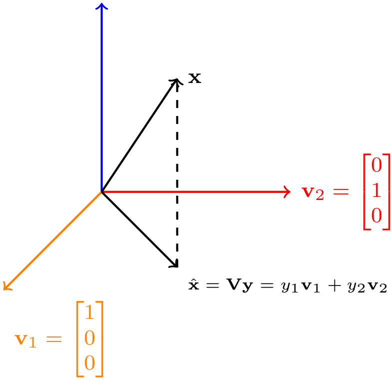
Principal Component Analysis

- In PCA, we find the compression matrix $\mathbf{W}$ and the recovering matrix $\mathbf{U}$ so that the total squared distance between the original and recovered vectors is as minimum as possible, i.e.,

$$\underset{\mathbf{W}\in\mathbb{R}^{k\times d},\mathbf{U}\in\mathbb{R}^{d\times k}}{\text{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{x}_i\|_2^2.$$

Claim:

Let $(\mathbf{U}, \mathbf{W})$ be a solution. Then the columns of $\mathbf{U}$ are orthonormal and $\mathbf{W} = \mathbf{U}^\top$.



Subspace Projection

Let $\mathcal{R}$ be a k dimensional subspace of $\mathbb{R}^d$. Let $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{bmatrix} \in \mathbb{R}^{d \times k}$ be the orthonormal matrix containing the basis vectors of $\mathcal{R}$. Let $\mathbf{x} \in \mathbb{R}^d$ a vector. Then, the closest vector $\mathbf{x}^\star \in \mathcal{R}$ to the vector $\mathbf{x}$ can be found by solving the below optimization problem.

$$\mathbf{x}^\star = \underset{\hat{\mathbf{x}} \in \mathcal{R}}{\operatorname{argmin}} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$$

Since we know that $\hat{\mathbf{x}} = \mathbf{V}\mathbf{y}$ for some $\mathbf{y} \in \mathbb{R}^k$, we have

$$\mathbf{y}^\star = \underset{\mathbf{y} \in \mathbb{R}^k}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{V}\mathbf{y}\|_2^2$$

Then, $\mathbf{y}^\star = \mathbf{V}^\top \mathbf{x} \Rightarrow \mathbf{x}^\star = \mathbf{V}\mathbf{V}^\top \mathbf{x}$.

Principal Component Analysis

$$\underset{\mathbf{U}\in\mathbb{R}^{d\times k},\mathbf{U}^\top\mathbf{U}=\mathbf{I}}{\text{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{U}\mathbf{U}^\top\mathbf{x}_i\|_2^2.$$

$$\begin{aligned}
\|\mathbf{x}_i - \mathbf{U}\mathbf{U}^\top\mathbf{x}_i\|_2^2 &= (\mathbf{x}_i - \mathbf{U}\mathbf{U}^\top\mathbf{x}_i)^\top(\mathbf{x}_i - \mathbf{U}\mathbf{U}^\top\mathbf{x}_i) \\
&= (\mathbf{x}_i^\top - \mathbf{x}_i^\top\mathbf{U}\mathbf{U}^\top)(\mathbf{x}_i - \mathbf{U}\mathbf{U}^\top\mathbf{x}_i) \\
&= \mathbf{x}_i^\top\mathbf{x}_i - 2\mathbf{x}_i^\top\mathbf{U}\mathbf{U}^\top\mathbf{x}_i + \mathbf{x}_i^\top\mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top\mathbf{x}_i \\
&= \mathbf{x}_i^\top\mathbf{x}_i - \mathbf{x}_i^\top\mathbf{U}\mathbf{U}^\top\mathbf{x}_i
\end{aligned}$$

$$\max_{\mathbf{U}\in\mathbb{R}^{d\times k},\mathbf{U}^\top\mathbf{U}=\mathbf{I}} \sum_{i=1}^{n} \mathbf{x}_i^\top\mathbf{U}\mathbf{U}^\top\mathbf{x}_i \Leftrightarrow \max_{\mathbf{U}\in\mathbb{R}^{d\times k},\mathbf{U}^\top\mathbf{U}=\mathbf{I}} \text{Trace}(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X})$$

Principal Component Analysis

Problem

$$\max_{\mathbf{U}\in\mathbb{R}^{d\times k},\mathbf{U}^\top\mathbf{U}=\mathbf{I}} \text{Trace}(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U})$$

Solution

Let $\mathbf{X}\mathbf{X}^\top\mathbf{u}_i = \lambda_i\mathbf{u}_i, \forall i \in \{1, 2, \ldots, n\}$ be the EVD of the matrix $\mathbf{X}\mathbf{X}^\top$. Here, we assume that the eigenvalues are such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then, the solution to the above problem is $\mathbf{U}^\star = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k\end{bmatrix}$.