Tutorial on Constant Shift Embeddings (SLT 2019)

May 7, 2019

These are **informal** notes whose purpose is to help you to follow the tutorial class.

1 Setting (pairwise dissimilarity and clustering)

We have N objects

$$O = \{o_1 \dots, o_i, \dots, o_N\}$$

which are **not** necessarily vectors. We want to cluster them based on their **pairwise dissimilarity** represented by a matrix D where

$$D_{ij} = \text{dissimilarity between object } o_i \text{ and object } o_j$$

We want to map these objects into K clusters so that each cluster contains "similar" objects. Our cost function will sum all pairwise dissimilarities **inside** each cluster α:

Pairwise clustering:

$$\mathcal{R}^{pc}(c, D) := \sum_{\alpha} \sum_{(i,j)\in\mathcal{E}_{\alpha\alpha}} \frac{D_{ij}}{|\mathcal{G}_\alpha|} \tag{1}$$

where

$$\begin{aligned} \mathcal{G}_\alpha &= \{o \in O :\ c(o) = \alpha\} && \text{(objects in cluster } \alpha) \\ \mathcal{E}_{\alpha\beta} &= \{(i,j) :\ o_i \in \mathcal{G}_\alpha \text{ and } o_j \in \mathcal{G}_\beta\} && (\mathcal{E}_{\alpha\alpha} = \text{ all pairs in cluster } \alpha) \end{aligned}$$

Return the clustering c_D minimizing $\mathcal{R}^{pc}$ above ($c_D = \arg\min_c \mathcal{R}^{pc}(c, D)$).

Remark 1. *Throughout these notes, we shall use* ***greek letters*** *to denote* ***clusters****.*

At this point we make **no assumption** on the dissimilarity matrix D which may also contain negative numbers. We only assume that the "self-dissimilarities" terms are all zero ($D_{ii} = 0$ for all i).

2 Invariance (symmetrization + off-diagonal shift)

Symmetrization. If D_{ij} is **not** symmetric

$$D_{ij} \neq D_{ji}$$

then we can make it symmetric by considering

$$D^S_{ij} := \frac{D_{ij} + D_{ji}}{2}$$

This will not change the cost function and thus the clustering:

$$\mathcal{R}^{pc}(c, D^S) = \mathcal{R}^{pc}(c, D) \tag{2}$$

Proof of (2). Intuitively, in $\mathcal{R}^{pc}(c, D)$ in each cluster α containing objects o_i and o_j, we are summing "$\cdots + D_{ij} + \cdots + D_{ji} + \cdots$", while in $\mathcal{R}^{pc}(c, D^S)$ we are summing "$\cdots + \frac{D_{ij}+D_{ji}}{2} + \cdots + \frac{D_{ij}+D_{ji}}{2} + \cdots$", which is the same. In formulas:

$$\begin{aligned}\mathcal{R}^{pc}(c, D^S) &= \sum_{\alpha} \sum_{(i,j)\in\mathcal{E}_{\alpha\alpha}} \frac{D_{ij} + D_{ij}}{2|\mathcal{G}_\alpha|} \\ &= \underbrace{\frac{1}{2}\sum_{\alpha} \sum_{(i,j)\in\mathcal{E}_{\alpha\alpha}} \frac{D_{ij}}{|\mathcal{G}_\alpha|}}_{\frac{1}{2}\mathcal{R}^{pc}(c,D)} + \underbrace{\frac{1}{2}\sum_{\alpha} \sum_{(i,j)\in\mathcal{E}_{\alpha\alpha}} \frac{D_{ji}}{|\mathcal{G}_\alpha|}}_{\frac{1}{2}\mathcal{R}^{pc}(c,D)}\end{aligned}$$

□

Off-diagonal shift. Consider adding the same quantity d_0 to **all** elements **not** in the diagonal:

$$\tilde{D}_{ij} = D_{ij} + \begin{cases} d_0 & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

Now the cost function changes, but the computed **clustering does not change**:

$$\mathcal{R}^{pc}(c, \tilde{D}) = \mathcal{R}^{pc}(c, D) + \underbrace{(N - K)d_0}_{\text{constant shift}} \tag{3}$$

Proof of (3). Let us fix a cluster α and see how the corresponding contribution in $\mathcal{R}^{pc}(\cdot, \cdot)$ changes. All pairs $(i, j) \in \mathcal{E}_{\alpha\alpha}$ with $i \neq j$ are increased by d_0, while those with $i = j$ do not change:

$$\underbrace{\sum_{(i,j)\in\mathcal{E}_{\alpha\alpha}} \frac{\tilde{D}_{ij}}{|\mathcal{G}_\alpha|}}_{|\mathcal{G}_\alpha||\mathcal{G}_\alpha| \text{ terms}} = \sum_{\substack{(i,j)\in\mathcal{E}_{\alpha\alpha} \\ i\neq j}} \frac{\tilde{D}_{ij}}{|\mathcal{G}_\alpha|} + \underbrace{\sum_{\substack{(i,j)\in\mathcal{E}_{\alpha\alpha} \\ i=j}} \frac{\tilde{D}_{ij}}{|\mathcal{G}_\alpha|}}_{|\mathcal{G}_\alpha| \text{ terms}}$$

as the latter summation contains exactly $|\mathcal{G}_\alpha|$ terms (all pairs (i,i) with $o_i \in \mathcal{G}_\alpha$) and thus

$$
\begin{aligned}
&= \sum_{\substack{(i,j)\in\mathcal{E}_{\alpha\alpha}\\ i\neq j}} \frac{D_{ij}+d_0}{|\mathcal{G}_\alpha|} + \sum_{\substack{(i,j)\in\mathcal{E}_{\alpha\alpha}\\ i=j}} \frac{D_{ij}}{|\mathcal{G}_\alpha|} \\
&= (|\mathcal{G}_\alpha|-1)d_0 + \sum_{\substack{(i,j)\in\mathcal{E}_{\alpha\alpha}\\ i\neq j}} \frac{D_{ij}}{|\mathcal{G}_\alpha|} + \sum_{\substack{(i,j)\in\mathcal{E}_{\alpha\alpha}\\ i=j}} \frac{D_{ij}}{|\mathcal{G}_\alpha|}
\end{aligned}
$$

By summing over all K clusters α we have

$$
\mathcal{R}^{pc}(c,\tilde{D}) = \sum_\alpha \sum_{(i,j)\in\mathcal{E}_{\alpha\alpha}} \frac{\tilde{D}_{ij}}{|\mathcal{G}_\alpha|} = \underbrace{\sum_\alpha (|\mathcal{G}_\alpha|-1)d_0}_{(N-K)d_0} + \underbrace{\sum_\alpha \sum_{(i,j)\in\mathcal{E}_{\alpha\alpha}} \frac{D_{ij}}{|\mathcal{G}_\alpha|}}_{\mathcal{R}^{pc}(c,D)}
$$

where we used $\sum_\alpha |\mathcal{G}_\alpha| = N$ (each object is in one cluster). □

3 One special case (Dissimilarity ≡ Euclidean Distance)

Consider the case

$$D_{ij} = ||x_i - x_j||^2$$

so we can think of our objects as points $x_1, \ldots, x_N \in R^d$. Now we have two clustering methods (cost functions):

- K-means: $\mathcal{R}^{km}(c,x) = \sum_i ||x_i - y_{c(i)}||$ where $y_\alpha = \frac{1}{|\mathcal{G}_\alpha|}\sum_{i:c(i)=\alpha} x_i$.
- pairwise clustering: $\mathcal{R}^{pc}(c,D)$ for $D_{ij} = ||x_i - x_j||^2$.

Which method should we choose?

They are (essentially) the same:

$$\mathcal{R}^{km}(c,x) = \frac{1}{2}\mathcal{R}^{pc}(c,D) \qquad \text{for } D_{ij} = ||x_i - x_j||^2 \tag{4}$$

Proof of (4). We show that

$$
\underbrace{\sum_i ||x_i - y_{c(i)}||^2}_{\mathcal{R}^{km}(c,x)} = \underbrace{\frac{1}{2}\sum_{i,j}\sum_\nu \frac{\mathbf{1}_{\{c(i)=\nu\}}\mathbf{1}_{\{c(j)=\nu\}}}{|\mathcal{G}_\nu|}||x_i - x_j||^2}_{\mathcal{R}^{pc}(c,D)}
$$

Start from the left hand side,

$$
\sum_i ||x_i - y_{c(i)}||^2 = \sum_i \sum_\alpha \mathbf{1}_{\{c(i)=\alpha\}}\big(||x_i||^2 + ||y_\alpha||^2 - \underbrace{2x_i y_\alpha}_{\text{vector product}}\big) \tag{5}
$$

we want to rewrite last term in (5) as

$$= \sum_i \sum_\ell \sum_\alpha \frac{\mathbf{1}_{\{c(i)=\alpha\}} \mathbf{1}_{\{c(\ell)=\alpha\}}}{|\mathcal{G}_\alpha|} (\cdots\cdots) \tag{6}$$

The first trick is $\sum_\ell \mathbf{1}_{\{c(\ell)=\alpha\}} = |\mathcal{G}_\alpha|$ and thus

$$||x_i||^2 = ||x_i||^2 \cdot \frac{|\mathcal{G}_\alpha|}{|\mathcal{G}_\alpha|} = ||x_i||^2 \cdot \frac{\sum_\ell \mathbf{1}_{\{c(\ell)=\alpha\}}}{|\mathcal{G}_\alpha|}$$

The second trick is that for K-means clustering

$$y_\alpha = \frac{1}{|\mathcal{G}_\alpha|} \sum_j x_j \mathbf{1}_{\{c(j)=\alpha\}}$$

and thus

$$||y_\alpha||^2 = \frac{1}{|\mathcal{G}_\alpha|^2} \sum_{j,\ell} x_j x_\ell \mathbf{1}_{\{c(j)=\alpha\}} \mathbf{1}_{\{c(\ell)=\alpha\}}$$

By plugging these quantities into the three quantities in (5) we get:

$$\begin{aligned}
\sum_\alpha \sum_i \mathbf{1}_{\{c(i)=\alpha\}} ||x_i||^2 &= \sum_\alpha \sum_i \sum_j \frac{\mathbf{1}_{\{c(i)=\alpha\}} \mathbf{1}_{\{c(j)=\alpha\}}}{|\mathcal{G}_\alpha|} ||x_i||^2 \\
\sum_\alpha \sum_i \mathbf{1}_{\{c(i)=\alpha\}} (-2 x_i y_\alpha) &= \sum_\alpha \sum_i \sum_j \frac{\mathbf{1}_{\{c(i)=\alpha\}} \mathbf{1}_{\{c(j)=\alpha\}}}{|\mathcal{G}_\alpha|} (-2 x_i x_j) \\
\sum_\alpha \sum_i \mathbf{1}_{\{c(i)=\alpha\}} ||y_\alpha||^2 &= \sum_\alpha \sum_i \sum_j \sum_\ell \frac{\mathbf{1}_{\{c(j)=\alpha\}} \mathbf{1}_{\{c(\ell)=\alpha\}}}{|\mathcal{G}_\alpha|^2} x_j x_\ell \\
&= \sum_\alpha \sum_j \sum_\ell \frac{\mathbf{1}_{\{c(j)=\alpha\}} \mathbf{1}_{\{c(\ell)=\alpha\}}}{|\mathcal{G}_\alpha|} x_j x_\ell
\end{aligned}$$

Back to (5) we get

$$\begin{aligned}
\sum_i ||x_i - y_{c(i)}||^2 &= \sum_\alpha \sum_i \sum_j \frac{\mathbf{1}_{\{c(i)=\alpha\}} \mathbf{1}_{\{c(j)=\alpha\}}}{|\mathcal{G}_\alpha|} \left(||x_i||^2 - x_i x_j \right) \\
&= \sum_\alpha \sum_i \sum_j \frac{\mathbf{1}_{\{c(i)=\alpha\}} \mathbf{1}_{\{c(j)=\alpha\}}}{|\mathcal{G}_\alpha|} \left(||x_i||^2 + ||x_j||^2 - 2 x_i x_j \right) \frac{1}{2}
\end{aligned}$$

which concludes the proof since $||x_i - x_j||^2 = ||x_i||^2 + ||x_j||^2 - 2 x_i x_j$. □

4 Centralization and decomposition

Look back at the case of "geometric dissimilarities":

$$D_{ij} = ||x_i - x_j||^2 = \underbrace{||x_i||^2}_{S_{ii}} + \underbrace{||x_j||^2}_{S_{jj}} - 2\underbrace{x_i x_j}_{S_{ij}}$$

where $S_{ij} = x_i x_j$. This suggests the following idea:

Decomposition of D with zero diagonal: find another matrix S such that

$$D_{ij} = S_{ii} + S_{jj} - 2S_{ij} \tag{7}$$

Exercise 1. *Check the following:*

1. *The "diagonal" elements are ok, i.e., (7) can be satisfied for $i = j$.*
2. *There is always one solution of (7).*
3. *There are infinitely many solutions of (7).*
4. *Matrix S is symmetric if and only if D is symmetric.*

A **centralized** matrix is a matrix M such that the **sum** of the elements in each **row**, and the **sum** of the elements in each **column**, equals to zero:

$$\text{for all } i \text{ and } j \qquad \sum_k M_{ik} = 0 \qquad \text{and} \qquad \sum_k M_{kj} = 0 \tag{8}$$

Example 1 (centralized matrix)**.** *An example of a centralized matrix is the matrix which has $1 - 1/N$ in the diagonal, and $-1/N$ off diagonal:*

$$Q = \begin{pmatrix} 1-1/N & -1/N & \cdots & -1/N \\ -1/N & 1-1/N & \cdots & -1/N \\ & & \vdots & \\ -1/N & -1/N & \cdots & 1-1/N \end{pmatrix} = I - \frac{1}{N}O$$

where I is the identity matrix and O is the matrix with all entries equal to 1.

Centralization

$$M \Longrightarrow M^c := QMQ \textbf{ centralized} \tag{9}$$

Proof of (9). We show that, for any symmetric M, matrix M^c is indeed a centralized matrix. Using $Q = I - \frac{1}{N}O$ we can write for $X := \frac{1}{N}O$

$$\begin{aligned} M^c =& QMQ = (I - X)M(I - X) \\ =& (I - X)(M - MX) \\ =& M - MX - XM + XMX \end{aligned}$$

which gives

$$M^c_{ij} = M_{ij} - \frac{1}{N}\sum_k M_{ik} - \frac{1}{N}\sum_k M_{kj} + \frac{1}{N^2}\sum_{k,\ell} M_{k\ell} \tag{10}$$

and from this we can verify that M^c satisfies (8):

$$\begin{aligned} \sum_j M^c_{ij} =& \sum_j \left(M_{ij} - \frac{1}{N}\sum_k M_{ik} - \frac{1}{N}\sum_k M_{kj} + \frac{1}{N^2}\sum_{k,\ell} M_{k\ell} \right) \\ =& \underbrace{\left(\sum_j M_{ij}\right) - \left(N \cdot \frac{1}{N}\sum_k M_{ik}\right)}_{=0} - \underbrace{\left(\frac{1}{N}\sum_{j,k} M_{kj}\right) + \left(N \cdot \frac{1}{N^2}\sum_{k,\ell} M_{k\ell}\right)}_{=0} \end{aligned}$$

(The proof of $\sum_i M^c_{ij} = 0$ is similar.) □

Centralization + decomposition

$$S^c = -\frac{1}{2}D^c \tag{11}$$

This is interesting because different decompositions give the same centralization.

Proof of (11). It is enough to plug (7) into (10)

$$\begin{aligned} D^c_{ij} =& D_{ij} - \frac{1}{N}\sum_k D_{ik} - \frac{1}{N}\sum_k D_{kj} + \frac{1}{N^2}\sum_{k,\ell} D_{k\ell} \\ =& S_{ii} + S_{jj} - 2S_{ij} - \frac{1}{N}\sum_k (S_{ii} + S_{kk} - 2S_{ik}) - \frac{1}{N}\sum_k (S_{kk} + S_{jj} - 2S_{kj}) + \frac{1}{N^2}\sum_{k,\ell}(S_{kk} + S_{\ell\ell} - 2S_{k\ell}) \end{aligned}$$

and since $\frac{1}{N}\sum_k S_{ii} = S_{ii}$, $\frac{1}{N}\sum_k S_{jj} = S_{jj}$

$$= -2S_{ij} - \frac{1}{N}\sum_k (S_{kk} - 2S_{ik}) - \frac{1}{N}\sum_k (S_{kk} - 2S_{kj}) + \frac{1}{N^2}\sum_{k,\ell}(S_{kk} + S_{\ell\ell} - 2S_{k\ell})$$

and similarly this further simplifies to

$$\begin{aligned}
&= -2S_{ij} - \frac{1}{N}\sum_k(-2S_{ik}) - \frac{1}{N}\sum_k(-2S_{kj}) + \frac{1}{N^2}\sum_{k,\ell}(-2S_{k\ell}) \\
&= -2S^c_{ij}
\end{aligned}$$

□

Theorem 2. *Dissimilarity D comes from a Euclidean distance if and only if $S^c = -\frac{1}{2}D^c$ is positive semidefinite.*

Proof of Theorem 2. We prove only the "⇒" direction in these notes. Observe that

$$\begin{aligned}
D_{ij} =& x_i x_i + x_j x_j - 2x_i x_j \\
=& S_{ii} + S_{jj} - 2S_{ij}
\end{aligned}$$

where $S_{ij} = x_i x_j$ for all i and j. In matrix multiplication this means $S = X^T X$. By applying the centralization (9):

$$S^c = QSQ = \underbrace{QX^T}_{Y^T}\underbrace{XQ}_{Y} = Y^T Y$$

□

5 From dissimilarity to geometric distances

Given our dissimilarity matrix D, let us decompose it using the matrix S as in (7). Then our goal is to write S as

$$S = X^T X = \begin{pmatrix} -x_1- \\ -x_2- \\ \vdots \\ -x_N- \end{pmatrix} \begin{pmatrix} | & | & | & | \\ x_1 & x_2 & \cdots & x_N \\ | & | & | & | \end{pmatrix}$$

that is $S_{ij} = x_i x_j$ which then implies $D_{ij} = ||x_i - x_j||^2$. We need two conditions (symmetry and positive semidefinite):

- If S is **symmetric** then can be decomposed as

  $$U^T D U$$

  where D contains the eigenvalues in the main diagonal and zeros off diagonal.

- If S is also **positive semidefinite** then all eigenvalues are positive and D can be further decomposed as
$$D^{1/2}D^{1/2}$$
where $D^{1/2}$ contains the roots $\sqrt{\lambda_i}$ of the eigenvalues in the main diagonal (these roots exist because $\lambda_i \geq 0$).

Then we get $S = \underbrace{U^T D^{1/2}}_{X^T}\underbrace{D^{1/2}U}_{X}$.

Make S positive semidefinite:

$$S \Rightarrow S^{ps} := S - \lambda_{\min} I \tag{12}$$

(subtract the minimum eigenvalue $\lambda_{\min} = \lambda_{\min}(S)$ of S from the diagonal – here I denotes the identity matrix)

Proof of (12). For any vector v we have

$$\frac{v^T S v}{v^T v} \geq \lambda_{\min}$$

and therefore

$$v^T \tilde{S} v = v^T (S - \lambda_{\min} I) v^T = v^T S v - \lambda_{\min} v^T I v \geq 0$$

□

This transformation (12) affects D which changes into D^{ps} as

$$D^{ps}_{ij} = S^{ps}_{ii} + S^{ps}_{jj} - 2S^{ps}_{ij}$$

which however is only an **off-diagonal** shift of D:

1. For $i = j$ we have $D^{ps}_{ii} = S_{ii} - \lambda_{\min} + S_{jj} - \lambda_{\min} - 2(S_{ii} - \lambda_{\min}) = D_{ii}$
2. For $i \neq j$ we have $D^{ps}_{ii} = S_{ii} - \lambda_{\min} + S_{jj} - \lambda_{\min} - 2(S_{ij}) = D_{ij} - 2\lambda_{\min}$

Pipeline: (1) symmetrization, (2) centralization, (3) decomposition, (4) positive semidefinite

$$D \text{ “generic”} \Longrightarrow D^S \text{ symmetric} \Longrightarrow D^c \text{ centered} \overset{(3)}{\Longleftrightarrow} S^c$$
$$\Downarrow$$
$$D^{ps} \overset{(3)}{\Longleftrightarrow} S^{ps}$$

Note that D^c and D^{ps} still give the **same clustering** since (as we argued above) D^{ps} is obtained from D^c via an off-diagonal shift (we apply (12) to the decomposition S^c of centralized matrix D^c). Since $S^p s$ is positive semidefinite, D^{ps} comes from a Euclidean distance, i.e.,

$$D^{ps}_{ij} = ||x_i - x_j||^2$$

for some $x_1, \ldots, x_N \in R^d$. In particular, these vectors can be obtained by the eigenvalues of S^{ps} and the decomposition described above.

Constant-shift embedding: For any D with zero diagonal we can map our objects into points of the Euclidean space

$$o_i \to x_i \in R^d \qquad \text{(embedding)}$$

such that

$$\mathcal{R}^{pc}(c, D) = \sigma \mathcal{R}^{km}(c, x) + \sigma' \qquad \text{(constant shift)}$$

and thus K-mean clustering produces the same result.