Elements of Statistical Learning
Homework Sheet 2

Optional. Only if you want. You can hand in solutions to question 3 for marking for formative FEEDBACK. Question 3 is marked below in blue and with $*$. Hand in solutions on Blackboard on Monday Feb 21st. Ensure your name is CLEARLY written on them. If you submit more than one page, ensure that the pages uploaded as a single document and in the right order, please!

1. In lectures we showed that

$$\begin{aligned} b_{m,\ell} &= -\tfrac{1}{2}\left(e_{m,\ell} - \tfrac{e_{m,\bullet}}{n} - \tfrac{e_{\bullet,\ell}}{n} + \tfrac{e_{\bullet,\bullet}}{n^2}\right) \qquad (1) \\ &= -\tfrac{1}{2}(\text{entry} - \text{row av.} - \text{col av.} + \text{grand av.}), \qquad (2) \end{aligned}$$

where $E = (e_{m,\ell})$ is the Euclidean distance matrix and $B = (b_{m,\ell})$ is the inner product matrix. Show that

$$B = -\frac{1}{2}(I_n - \mathbf{1}_n\mathbf{1}_n^T/n)\, E\, (I_n - \mathbf{1}_n\mathbf{1}_n^T/n), \qquad (3)$$

where $\mathbf{1}_n$ is the n-vector consisting of ones. Show that $\mathbf{1}_n$ is an eigenvector of B with eigenvalue of 0.

2. Repeat the classical scaling analyses on the `eurodist` and `UScitiesD` datasets that are built into R.

3. $*$ You can calculate the quantities here by hand, or using a calculator or computer.

   (a) Let the two-dimensional data matrix $X = \begin{pmatrix} 4 & 1 \\ 2 & 6 \\ 1 & 9 \end{pmatrix}$. Starting from X, calculate the inner product matrix, B; the Euclidean distance matrix, E; mean of the data matrix X; the centred data matrix X_c, and the inner product matrix of the centred data matrix B_c.

   (b) Compute the eigendecomposition of the centred matrix inner product matrix B_c, then form the recovered configuration Y, and then check that the inner product matrix from Y is the same as B_c).

4. Read the Wikipedia page on ‘Seriation (archaeology)’ and see how scaling helps with the problem of putting objects into chronological order.

5. *Lemma:* If $\{d_\alpha\}_{\alpha \in \mathcal{A}}$ is a family of metrics, where $\mathcal{A}$ is a discrete set, then $\sum_{\alpha \in \mathcal{A}} d_\alpha$ is a metric. Prove the lemma.

6. An ultrametric on a set M is a real-valued function $d : M \times M \to \mathbb{R}^+$, such that for all $x, y, z \in M$:

   (a) $d(x, y) \geq 0$;

   (b) $d(x, y) = d(y, x)$;

(c) $d(x, x) = 0$;

(d) if $d(x, y) = 0$, then $x = y$;

(e) if $d(x, z) \leq \max\{d(x, y), d(y, z)\}$,

where property (5) is known as the ultrametric inequality (or the strong triangle inequality). An ultrametric space (M, d) is a set together with an ultrametric d on M.

Suppose x, y, z are three points in an ultrametric space (M, d). Show that every such triple forms an isosceles triangle — that is, at least one of the three equalities $d(x, y) = d(y, z)$ or $d(x, z) = d(y, z)$ or $d(x, y) = d(z, x)$ holds.

Suppose M is a set. Define the *discrete metric* ρ on X to be

$$\rho(x, y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{if } x = y, \end{cases}$$

for $x, y \in M$. Show that (M, ρ) is an ultrametric space.

[Updated: Feb 13th 2023]