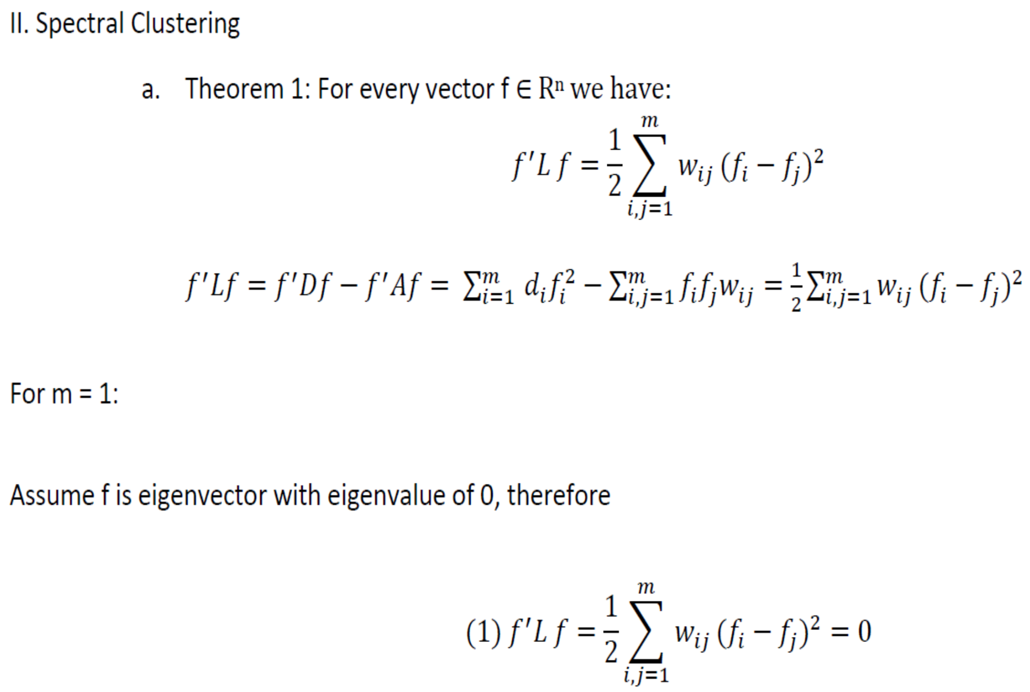
Kien Duc Vu

GTID: 903552166

Homework 2

Q1a. Spectral Clustering



II. Spectral Clustering

a. Theorem 1: For every vector f ∈ Rⁿ we have:

$$f'L\,f = \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}\,(f_i - f_j)^2$$

$$f'Lf = f'Df - f'Af = \sum_{i=1}^{m} d_i f_i^2 - \sum_{i,j=1}^{m} f_i f_j w_{ij} = \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}\,(f_i - f_j)^2$$

For m = 1:

Assume f is eigenvector with eigenvalue of 0, therefore

$$(1)\; f'L\,f = \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}\,(f_i - f_j)^2 = 0$$

(1) Happens if and only if $f_i = f_j$ ($since\ w_{ij} > 0$). Thus, f is an indicator vector of the connected component. Now, considering Laplacian matrix L with m > 1,

$$L = \begin{pmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_k \end{pmatrix}$$

Each of the L_i is a proper graph Laplacian on its own corresponding to the subgraph of the i-th connected component, with m=1, that has an indicator vector of eigenvalue 0. Therefore, L has m eigenvectors corresponding to eigenvalue and the indicator vectors of these components span the zero eigenspace.

Q2. PCA

1. Matrix data has shape of (13, 20) with 13 rows representing 13 countries, and 20 column features representing 20 kinds of food being analyzed in this study. Step by step PCA:
   - Rescale and normalize dataset by mean and standard deviation
   - Calculate mean, and covariance matrix
   - Calculate eigenvectors and eigenvalues
   - Choose k eigenvectors that have largest eigenvalues
   - Find k principal component directions corresponding to k eigenvectos
   - Project data onto k-principal component directions
   - Plot PCA
2. $X \in R^{nxd}$

$\{u_i\}$ is basic orthonormal vector representing d-dimensional space. X_n can be represented as linear combination of u_i.

$$X_n = \sum_{i=1}^{d} (X_n^T u_i) u_i$$

Equivalently,

$$X_n = \sum_{i=1}^{k}(z_i)u_i + \sum_{i=k+1}^{d}(b_i)u_i$$

We approximate each data point X_n by

$$\overline{X_n} = \sum_{i=1}^{k}(z_i)u_i + \sum_{i=k+1}^{d}(\overline{b_\imath})u_i$$

Where $\overline{\{b_\imath\}}$ is fixed for all data point and $\{z_i\}$ $\{u_i\}$ are datapoint-dependent. Distortion function J:

$$J = \frac{1}{N}\sum_{i=1}^{n}|X_n - \overline{X_n}|^2$$

Taking derivative of J with respect of z and b gives

$$z_{nj} = X_n^T u_j$$

$$b_j = \overline{X_n}^T u_j$$

Therefore, J can be expressed as

$$J = \frac{1}{N}\sum_{i=1}^{n}|X_n - \overline{X_n}|^2 = \sum_{i=k+1}^{d}(u_i^T S u_i)$$

General solution is obtained by choosing the $\{u_i\}$ to be eigenvectos of the covariance matrix given by $Su_i = \boldsymbol{\lambda}_i u_i$

And $J = \sum_{i=k+1}^{d}\boldsymbol{\lambda}_i$