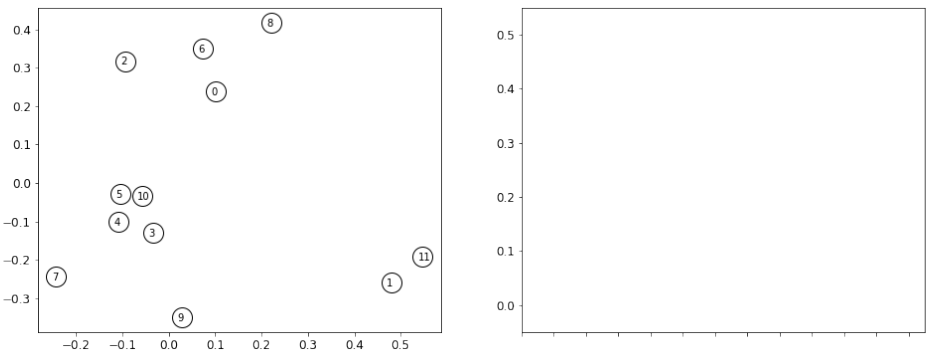

Figure 1: Plot of `hw2-toy-dendro.dat` from Problem 2.

STAT/BIOST 527
Homework 3
Out Wednesday May 10, 2023
Due Wednesday May 19, 2023

©Marina Meilă
mmp@stat.washington.edu

Problem 1 – Dendrograms and distances between clusterings (from Homework 2)
a. Figure 1 below displays $n = 12$ data points found in file `hw2-toy-dendro.dat`. The first two columns contain the x and y coordinates and the third column contains the label. Compute all the distances between these points (you do not need to submit code or output). Using the calculated distance matrix, draw the dendrogram of this dataset obtained by the Single Linkage Algorithm. Implemementation is not required. The algorithm can be "run" manually for these data and the dendrogram can be copied by hand.

Use the plot on the right to display this dendrogram; the height of a dendrogram node should be equal to the distance between the two clusters merged at this node. Recall that the distance between two clusters C_1 and C_2 in the single-linkage framework is given by

$$\text{distance}(C_1, C_2) = \min_{x \in C_1, y \in C_2} ||x - y||_2. \tag{1}$$

b. The dendrogram in Figure 2 displays the output of a different hierarchical clustering algorithm on the same data as above. On the plots below, draw the first 5 stages of this algorithm (from the top down). Stages are denoted by the number of clusters K; for example level $K = 2$ is the clustering with 2 clusters, resulting after the first split, level $K = 3$ results after the second split and has 3 clusters, etc.

c. Denote by Δ_1 the clustering at level $K_1 = 3$ in the dendrogram obtained in **a**, and denote by Δ_2 the clustering at level $K_2 = 4$ in the dendrogram in **b**. Compute the confusion matrix M of these clusterings. *Cluster labels are arbitrary, hence any permutation of rows or columns of a correct M is equally correct*

d. From the confusion matrix M obtained in **c**, calculate the Misclassification Error distance $d_{ME}(\Delta_1, \Delta_2)$.

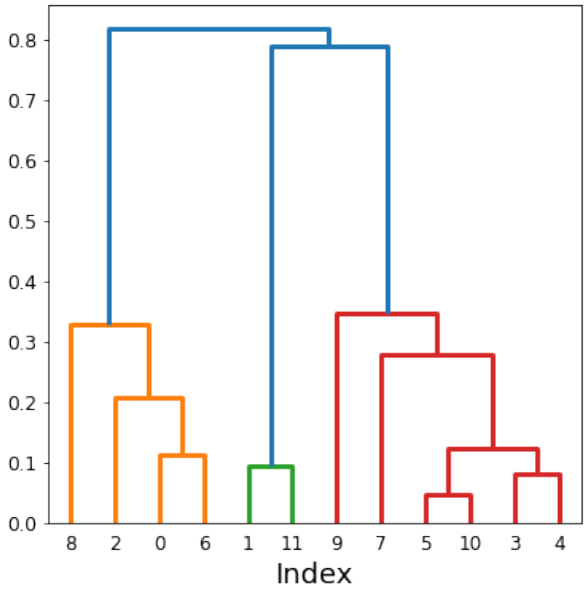

Figure 2: Dendrogram on `hw2-toy-dendro.dat`.

e. From the confusion matrix M obtained in **c**, calculate N_{22}, N_{12}, N_{21} and the Jaccard index $J(\Delta_1, \Delta_2)$.

f. Calculate the $n \times K$ matrix representation $\tilde{X}_{1,2}$ for clusterings Δ_1, Δ_2 (see Lecture II, part 3). Verify that $M = \tilde{X}_1^T \tilde{X}_2$. Provide a code snippet that prints M, $\tilde{X}_1$, and $\tilde{X}_2$, and computes and prints

$$||M - \tilde{X}_1^T \tilde{X}_2||_F,$$

where $||\cdot||_F$ denotes the Frobenius norm of a matrix.

g. Prove that $M = \tilde{X}_1^T \tilde{X}_2$ for any two (arbitrary) clusterings represented by $\tilde{X}_1, \tilde{X}_2$.

h. Denote $\tilde{Z}(\Delta) = \tilde{X}\tilde{X}^T$. Show that $Z \in \{0,1\}^{n \times n}$. Find a simple expression for the Jaccard index of clusterings $\Delta_{1,2}$ as a function of $\tilde{Z}_{1,2}$ and n. Both matrix and elementwise arithmetic or Boolean operations with $\tilde{Z}_{1,2}$ are allowed as long as they are on all elements. E.g. $\tilde{Z}_1 + 3\tilde{Z}_2$, $\max(\tilde{Z}_1, \tilde{Z}_2)$; $\tilde{Z}_1 + C$ (where C is a constant matrix).

Problem 2 – Leave one out CV and support vectors
Assume the data set $\mathcal{D}$ contains n samples. You perform *leave-one-out cross-validation* i.e, for $i = 1 : n$ you compute a linear support vector machine classifier f_{-i} on $n-1$ points, leaving out (x^i, y^i). More precisely, f_{-i} is a SVM trained on $\mathcal{D}_{-i} = \mathcal{D} \setminus \{(x^i, y^i)\}$.

a. Assume that the original data set is linearly separable. Prove that each of the n support vector problems is also linearly separable.

b. Is it possible that $f_{-i}(x) \equiv f_{-j}(x)$ for $i \neq j$ two points in the training set $\mathcal{D}$? Give a short motivation or proof.

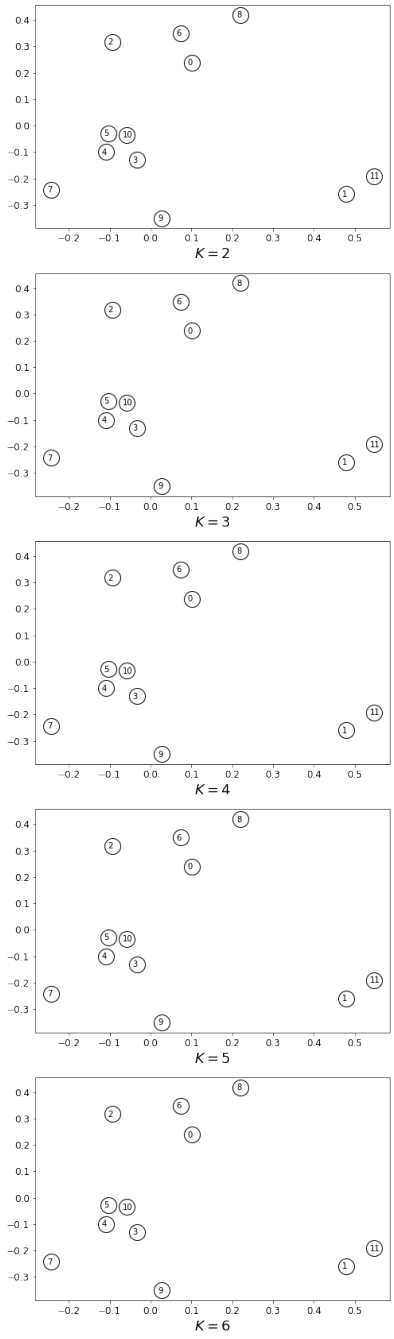

Figure 3: Plots on which to draw cluster assignments at each stage of the dendrogram from Figure 2.

c. Denote by $\hat{L}_{01}^{loo}$ the error rate in leave-one-out CV, i.e

$$\hat{L}_{01}^{loo} \;=\; \frac{|\{i,\, f_{-i}(x^i) \neq y^i\}|}{n}$$

Prove that $\hat{L}_{01}^{loo} \leq \frac{\#\text{support vectors of } f}{n}$, where f is the linear support vector classifier trained on all the data.

[Problem 3 – Quadratic kernel – NOT GRADED]

In this problem, the points lie on the real line, there are two classes and we use the polynomial degree 2 kernel $K(x, x') = (1 + xx')^2$.

1. What is the mapping $\phi : R \longrightarrow R^d$ satisfying

   $$K(x, x') \;=\; \phi(x)^T \phi(x')$$

   and what is its dimension d?

2. Let the data be $\mathcal{D} = \{\, (-1, +1), (0, -1), (1, +1) \,\}$.

   Compute $\phi(x_i)$ and the Gram matrix for this dataset.

3. Write the expression of the primal SVM problem for this data set. Be specific, give numerical values.

4. Write the expression of the dual SVM problem for this data set. Be specific, give numerical values.

5. This dual problem is small enough that it can be solved "manually". [Hint: you can notice that due to symmetry, $\alpha_1 = \alpha_3$ and turn it into a 2 variable problem.] Show that the solution is $\alpha_1 = \alpha_3 = 1,\ \alpha_2 = 2$.

6. What are the values of w and b? Write the expression of the discriminant function $f(x) = w^T \phi(x) + b$. Write the same function now using the kernel K. What are the decision regions of this classifier?

   Make a sketch of the data and the decision regions.

Problem 4 – SVM solution

a. Let $f = \sum_{i=1}^{n} \alpha_i K_{x^i}()$, where $x^{1:n} \in \mathbb{R}^d$ are vectors, $K(x, x')$ is a Mercer kernel defining a scalar product, and $K_x(u) \equiv K_(x, u)$. Give an expression for

$$\langle f, f \rangle_{\mathcal{H}} = \|f\|_{\mathcal{H}}^2.$$

b. Assume now that you have a non-linear SVM with $C = 0$, defined by a kernel $K(\,,\,)$ and by the feature map $\phi : \mathbb{R}^n \to \mathcal{H}$. In this case, $w \in \mathcal{H}$, hence it cannot be represented explicitly. However, its norm in $\mathcal{H}$ can be computed as $\|w\|_{\mathcal{H}}^2 =< w, w >_{\mathcal{H}}$.

Assume that you have solved the dual problem and that the dual variables $\alpha_{1:N}$, as well as b are known. It is a fact from convex optimization that (under generic conditions) the optimal value of the Primal SVM problem is equal to the optimal value of the Dual SVM problem (this is known as *strong duality*). Show that

$$\|w\|_{\mathcal{H}}^2 \;=\; \sum_{i=1}^{n} \alpha_i \tag{2}$$