1. Show that the within-point scatter $W(C)$ function on page 509 of the text is minimized by solving the enlarged optimization problem (14.33).

2. Show that weighted Euclidean distance

$$d_e^{(w)}(x_i, x_i') = \frac{\sum_{l=1}^{p} w_l (x_{il} - x_{i'l})^2}{\sum_{l=1}^{p} w_l}$$

satisfies

$$d_e^{(w)}(x_i, x_i') = d_e(z_i, z_i') = \sum_{l=1}^{p} (z_{il} - z_{i'l})^2$$

(1)

where

$$z_{il} = x_{il} \cdot \left( \frac{w_l}{\sum_{l=1}^{p} w_l} \right)^{1/2}.$$  

Thus, weighted Euclidean distance based on $x$ is equivalent to unweighted Euclidean distance based on $z$.

3. *Kernel k-means*: Show (18.28) in the text, and with definitions (18.30,18.31) show that $\tilde{K}_{ii'} = \langle x_i - \bar{x}, x_{i'} - \bar{x} \rangle$. Hence sketch out a K-means algorithm that requires only pairwise distances between data points.

4. *Number of clusters.*

(a) Consider a scenario with $n = 30$ data points on the real line. Generate data with two clusters of equal probability, distributed as $N(3,1)$ and $N(6,1)$ in the two clusters. Carry out $k$-means clustering for $k = 1, 2, \ldots, 20$, and plot the within cluster sum of squares versus $k$ for the training data. Comment on what you see. What value of $k$ would you select based on this plot alone?

(b) Generate a validation dataset of the same size as the training set and in the same fashion. For each $k$ assign each validation datapoint to the closest learned cluster mean from part (a) and plot the resulting validation k-means objective as a function of $k$. Comment on what you see. What value of $k$ would be selected if you were to use the minimum validation objective as a selection criterion?
(c) For each $k > 1$ compute the CH statistic from the class notes. Plot it. What value of $k$ maximizes this criterion?

(d) Write a program to compute the gap statistic for the same data, plot it and comment.

5. **Gaussian mixtures.** For the same setup as in Question 4, write a program to fit a Gaussian mixture model with $k$ components and equal variance in each cluster. Fit it to these data and plot the negative log-likelihood on the training and validation data. Comment and compare the results to those from Question 4.

6. **Clustering of non-negative data.** Suppose that we have data points $x \in \mathbb{R}^p$ with non-negative elements and consider clustering via the distance function

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} [x_{ij} \log(x_{ij}/x_{i'j}) - x_{ij} + x_{i'j}]$$

This is called a generalized KL divergence.

(a) Let $Y$ be a $n \times K$ matrix of zeroes and ones, whose rows sum to one. Consider the cost function

$$C(Y, \{\mu_k\}) = \sum_{i,k} Y_{ik} D(\mu_k, x_i)$$

where $\mu_k$ are the prototypes for each cluster $k = 1, 2, \ldots K$. Find the minimizers $\hat{\mu}_k$ with $Y$ fixed. Interpret their form.

(b) Sketch an iterative algorithm to minimize $C$ over $Y$ and $\mu$.

7. In [http://statweb.stanford.edu/~tibs/sta306bfiles/hwk1.txt](http://statweb.stanford.edu/~tibs/sta306bfiles/hwk1.txt) there is a microarray dataset, 1000 genes (rows) by 40 samples (columns). The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

(a) Apply hierarchical clustering to the samples using the correlation metric and plot the dendrogram. In R you can use hclust. Do the genes separate the samples into the two groups?

(b) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this and apply it here.