1. Computer Exercise in Vector Quantization. You will be given 100 2-dimensional points, \( \{ \mathbf{a}_i = (x_i, y_i), i = 1, 2, \ldots, 100 \} \). You are to divide them into 3 sets using vector quantization based on Euclidean distance \( d(\mathbf{a}_i, \mathbf{a}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \).

(a) Choose the three initial cluster-centers, \( \rho_k, k = 1, 2, 3 \), uniformly at random from the \( 1 \times 1 \) square in which the 100 points are located.

(b) Carry out the quantization process (cf Chapter 1) until no points change set membership.

(c) Using 3 different colors, plot the resulting sets and their cluster-centers.

Repeat the exercise several times, each with a different random choice of initial cluster-centers. Observe and report (i) the common tendencies and (ii) the occasional outlier behavior of the clustering algorithm.

2. Consider the HMM of Chapter 2, Figure 2.8, with state space \( S = \{1, 2, 3\} \) and output alphabet \( \mathcal{Y} = \{0, 1\} \). In the notation of Chapter 2, let the transition probabilities for the output-producing transitions and null transitions be given, respectively, by

\[
p(s'|s) = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & \frac{1}{3} \\ \frac{3}{4} & \frac{1}{4} & 0 \end{bmatrix} \quad \text{and} \quad q(s'|s) = \begin{bmatrix} 0 & 0 & \frac{1}{6} \\ \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix}, \quad s, s' \in S.
\]

Let the probabilities associated with the output producing transition from state \( s \) to state \( s' \) be given by

\[
q(0|s \rightarrow s') = \begin{bmatrix} 1 & \frac{1}{2} & 1 \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad q(1|s \rightarrow s') = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 1 & 1 & \frac{2}{3} \\ 1 & 1 & 1 \end{bmatrix},
\]

and note that \( q(1|s \rightarrow s') \) is simply \( 1 - q(0|s \rightarrow s') \). Let the initial state be \( s_0 = 1 \).

Perform the following calculations by hand, retaining intermediate answers in fractional form to preserve numerical accuracy.

(a) Draw a state diagram of this HMM, attaching state-labels to all nodes, probabilities \( r(y, s'|s) \) or \( q(s'|s) \) to all arcs and output-labels to all non-null arcs.
(b) Draw a 4-stage trellis for this HMM, showing only the paths which could have resulted in the output 0110.

(c) Compute the forward probabilities $\alpha_i(s)$ for the output sequence 0110, and indicate them on the trellis drawn above for (b).

(d) Compute the marginal probability of the output: $P(y_1y_2y_3y_4 = 0110|s_0 = 1)$.

(e) Redraw the trellis with paths which could have given rise to 0110, and indicate on it the Viterbi probabilities $\gamma_i(s)$. Color the most like path given $y_1y_2y_3y_4 = 0110$.

(f) Redraw the trellis, showing paths which could have given rise to 0110, and compute the backward probabilities $\beta_i(s)$. Indicate them on this redrawn trellis.

(g) Compare $\beta_0(1)$ with the marginal probability $P(y_1y_2y_3y_4 = 0110|s_0 = 1)$ of (d).

(h) Calculate the a posteriori probabilities $P(t^i = t | y_1y_2y_3y_4 = 0110, s_0 = 1)$ for each arc in the trellis of (f). Show your answers on the trellis.

(i) Based on your calculations in (h), compute the expected counts $c(t)$ of each arc, and reestimate the HMM transition probability matrices $p(s'|s)$ and $q(s'|s)$.

   Caution: the matrix $p(s'|s)$ has only one entry for a non-null transition from $s$ to $s'$, while our trellis has up to $|\mathcal{Y}|$ distinct non-null arcs from $s$ to $s'$, one arc per output symbol. Make sure you consolidate their counts appropriately.

(j) Based on your calculations in (h), compute the expected counts $c(y,t)$ of each non-null arc, and reestimate the HMM emission probability matrix $q(0|s \rightarrow s')$.

Carefully review Chapter 2 from the Jelinek book after finishing the homework, before the February 14 lecture.