Assigned: 28 August

Homework #2

EE 541: Fall 2022

Due: Wednesday, 13 September at 23:59. Submission instructions will follow separately on canvas.

1. Raman spectroscopy is a technique that uses inelastic scattering of light to identify unknown chemical substances. Spectral "peaks" indicate vibrational and rotational modes and are of special importance because they act like a chemical fingerprint. Raman spectroscopy measures photon intensity vs. Raman shift. The Raman shift relates the frequencies of the exciting laser and the scattered photons and is often reported as a wavenumber — the frequency difference in wavelengths per cm (*i.e.,* , cm^{-1}).

   - Generate a *molecular fingerprint* using the spectroscopic data in `raman.rod`. The file contains intensity vs. wavenumber data for an unknown chemical sample. A Raman Open Database (`ROD`) file includes content in addition to the raw intensity data:

     ```
     # content
     more content
     _raman_spectrum.intensity
     wavenumber1 intensity1
     wavenumber2 intensity2
     ...
     wavenumbern intensityn
     ```

     Use string matching to ignore all lines before `_raman_spectrum.intensity`. Load valid `(wavenumber, intensity)` pairs until the first invalid intensity line (or upon reaching the end of file).

     Use the method below to estimate the wavenumbers of all spectral peaks. You may use any standard `NumPy` or `SciPy` packages or experiment with your own algorithms.

   - First detect peaks in the raw spectral data. Use the peak locations to focus on regions of interest within the spectrum. For instance: if you detect peaks at x_1 cm^{-1} and x_2 cm^{-1} use regions of interest: $[x_1 - n_1, x_1 + n_1]$ and $[x_2 - n_2, x_2 + n_2]$. Experiment to find "good" widths n_1, n_2, etc. Then use a spline to interpolate intensity within each region of interest. Calculate zero-crossings of the derivative to estimate wavenumbers with maximum intensity.

     (a) Print the wavenumber estimates for the eight largest spectral peak to STDOUT sorted by magnitude (largest first).

     (b) Create a figure that shows the Raman data (intensity vs. wavenumber) and mark each of the maximum intensity values.

     (c) Produce a "zoomed-in" figure for the "regions of interest" corresponding to the four largest peaks. Plot the raw spectral data and overlay your interpolating function. Use a marker to show the wavenumber with maximal intensity.

2. Unsupervised clustering algorithms are an efficient means to identify groups of related objects within large populations. Implement the following two clustering algorithms and apply them to the data in `cluster.txt`. The file contains data as: `x`, `y`, `class`. Use a regular expression to remove lines that are empty or that are invalid data. You may safely ignore any line that fails the regular expression.

   (a) Use K-Means clustering with 3-clusters to label each (x, y) pair as `Head`, `Ear_Right`, or `Ear_Left`. You may use any standard `NumPy` or `SciPy` packages or experiment with your own implementation.

   Produce a scatter plot marking each (x, y) pair as either BLUE (class = `Head`), RED (class = `Ear_Left`) or GREEN (class = `Ear_Right`). Compare the K-means predicted labels to the true label and generate a confusion matrix showing the respective accuracies.

   (b) Gaussian Mixture Models (GMM) are a common method to cluster data from multi-modal probability densities. Expectation maximization (EM) is an iterative procedure to compute (locally) optimal GMM parameters – GMM cluster means μ_k, covariances Σ_k, and mixing weights w_k. EM consists of two-steps. The **E**[xpectation]-step uses the mixture parameters to update estimates of hidden variables. The *true* but unknown class is an example of a hidden variable. The **M**[aximization]-step then uses the new hidden variable estimates to update the mixture parameter estimates. This back-and forth update provably increases the likelihood function and the estimate eventually converges to a local likelihood maximum. The GMM update equations follow.

   E-Step: use current mixture parameters estimates to calculate membership probabilities (*a.k.a.,* the hidden variables) for each sample, $\gamma_k(x_n)$,

$$\gamma_k(x_n) = \frac{w_k f(x_n; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j f(x_n; \mu_j, \Sigma_j)}$$

   M-step: use new $\gamma_k(x_n)$ to update mixture parameter estimates,

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_k(x_n) x_n}{\sum_{n=1}^{N} \gamma_k(x_n)}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_k(x_n)(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma_k(x_n)}$$

$$w_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n)$$

   for $k \in \{1, \ldots, K\}$ where $K \in \mathbb{Z}^+$ is the (predefined) number of mixture components, x_n for $n \in \{1, \ldots, N\}$ are the data samples, and $f(x; \mu_k, \Sigma_k)$ is the d-dimensional jointly Gaussian pdf:

$$f(x; \mu_k, \Sigma_k) = \frac{\exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)}{\sqrt{(2\pi)^d |\Sigma_k|}}.$$

   - Implement Expectation Maximization and use it to estimate mixture parameters for a 3-

component GMM for the `cluster.csv` dataset. DO NOT use an EM implementation from `NumPy`, `SciPy`, or any other package. You must implement and use the above EM equations.

- Initialize $\gamma_k\left(x_n\right)$ for each sample using the K-means labels from part (a) as "one-hot" membership probabilities (*i.e.,* initialize one of the probabilities as "1" and all others are "0"). Then compute initial μ_k and Σ_k for each mixture component.

- Run EM until it has sufficiently converged. Use either the negative log-likelihood

$$\ell = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} w_k f(x_n; \mu_k, \Sigma_k)\right) \tag{1}$$

  as a convergence metric or monitor the class assignments until there are only small changes. Be aware that some points may "flip-flop" even when fully converged. Assign each datapoint to the mixture component with the largest membership probability. Produce a Blue-Red-Green scatter plot as in part (a) and generate a confusion matrix showing the respective classification accuracies.

- Generate figures showing the class assignments during the first four iterations.

- Comment on the difference between the clustering result in (a) and (b). Describe any obvious difference between the plots and indicate which performs better.