Exam 3

(There will be no extensions or late submission excepted, so make sure you start early, submit often, and make sure RStudio is working. We will not accept any submissions over email.) The data for problem two can be found at https://github.com/resteorts/data-clean/tree/main/exams/data/

Instructions

- This is a taken home examination to be completed individually.
- This exam is open note, open book, and class resources (please cite as you go).
- The exam is closed to talking to students (or anyone else except your instructor or teaching assistants in the course). We will only answer clarification questions and not anything about course material for the exam.
- You may not search Google, use ChatGPT, or other resources to aide in answering the questions. You may utilize any suggested books or resources listed for the course materials.
- You may not talk to anyone about the exam until the grades are released to the entire class. - Clarification questions will be addressed and should be asked to the instructor (and teaching assistants privately on Canvas).
- **No group postings should be made regarding the exam or exam related material**.
- Your submission must be to Gradescope and to Canvas (in the same format) as the homework.

- Please make sure your name is attached and by submitting you agree to the abide by all rules of the assignment and the Duke Community Standard.

1. True/False (8.5 points total, 1/2 point each)

   (a) The initialization of the cluster centroids does not affect the final result of k-means clustering.

   (b) K-means clustering can effectively handle clusters of any shape and/or size.

   (c) The K-means algorithm updates cluster centroids by recalculating the mean of all points assigned to each cluster.

   (d) K-means clustering guarantees finding the global optimum of the clustering objective function.

   (e) K-means is guaranteed to terminate.

   (f) The Elbow Method can always be used to determine the optimal number of clusters for K-means.

   (g) Hierarchical clustering does not require the number of clusters to be specified in advance.

   (h) Single linkage in hierarchical clustering measures the distance between the closest points of two clusters.

   (i) The cutting of the dendrogram at a specific level can help determine the number of clusters.

   (j) A dendrogram is a tree-like diagram that shows the hierarchical relationships between clusters.

   (k) The Expectation-Maximization (EM) algorithm is commonly used to estimate the parameters of mixture models.

   (l) Mixture models provide probabilistic assignments of data points to clusters, unlike K-means clustering, which provides hard assignments.

   (m) Poisson Mixture Models assume that the data is generated from a mixture of several Poisson distributions.

   (n) Mixture models are sensitive to the initialization of parameters.

   (o) The Expectation-Maximization (EM) algorithm for mixture models alternates between an E-step and an M-step.

   (p) Mixture models require that all components have the same weight.

(q) In a Gaussian Mixture Model, the probability of each data point belonging to a component is fixed over all iterations of the EM algorithm (and never updated).

2. (13 points) K-means clustering and Gaussian mixture models are two common algorithms to perform clustering. In this assignment, you will explore these algorithms and their limitations. For each problem, you may use functionality in R or code up your own functions. You can always check your own functions by comparing them to existing packages!

   (a) (5 points) k-means clustering. Use a set.seed(1234).

      i. (2 points) Apply k-means to both data sets provided using the settings centers = 2, nstart = 20, algorithm ="Lloyd", iter.max=100. Plot each data set, indicating for each point which cluster it was placed it.

      ii. (2 points) How well do you think k-means did for each data set?

      iii. (1 point) Explain, intuitively, what (if anything) went badly and why.

   (b) (5 points) Two-component Gaussian mixture model.

      i. (2 points) Apply a two-component Gaussian mixture model to the two data sets provided, where you may assume the mixing proportions are the same.

      ii. (2 points) What parameters do you find for each data set?

      iii. (1 point) Plot each data set, indicating for each point which cluster it was placed in.

   (c) (3 points) Modeling the second data set as a mixture of Gaussians is unrealistic, but the EM algorithm still gives *an* answer. Is there anything regarding your answers suggesting something may not be quite correct?

3. (Bonus, 2 points) The EM algorithm is often used with Gaussian mixture models, however, they can be used with any mixture models, such as exponential distributions (or others). One type of probability density suitable for "ring-shaped data" is the von Mises distribution. Using the hint below, code up your own function to sample the data from this distribution and output a plot illustrating how it captures the original data. Briefly, comment, visually, how well it does.

Hint for Bonus Question:

Let's talk about how to sample data from a ring shaped distribution, i.e. from the von Mises distribution.

1. Radial Distance (r)

The radial distance r is sampled from a normal distribution:

$$r \sim \mathcal{N}(\mu_r, \sigma_r^2)$$

where:

- μ_r: Mean radial distance.
- σ_r: Standard deviation of the radial distance.

To ensure r is positive, the absolute value is taken:

$$r = |r|$$

2. Angular Component (θ)

The angular coordinate θ is determined as follows:

- If $\kappa > 0$: θ is sampled from the von Mises distribution:

  $$\theta \sim VM(\mu_\theta, \kappa)$$

  where:

  - $\mu_\theta = 0$: Mean direction of the distribution (centered at zero).
  - κ: Concentration parameter ($\kappa = 0$ corresponds to a uniform distribution, and larger κ values produce more concentrated distributions around μ_θ).

- If $\kappa = 0$: θ is sampled uniformly from the interval $[0, 2\pi]$:

  $$\theta \sim \mathcal{U}(0, 2\pi)$$

3. Conversion to Cartesian Coordinates

After r and θ are determined, the polar coordinates (r, θ) are converted to Cartesian coordinates (x, y) using:

$$x = r\cos(\theta)$$

$$y = r\sin(\theta)$$

4. Result

Each sampled point is a pair (x, y), where:

$$(x, y) = (r\cos(\theta), r\sin(\theta))$$

and:

$$r \sim |\mathcal{N}(\mu_r, \sigma_r^2)|$$

$$\theta \sim VM(\mu_\theta, \kappa) \quad \text{if } \kappa > 0, \quad \text{or} \quad \theta \sim \mathcal{U}(0, 2\pi) \quad \text{if } \kappa = 0.$$