EE708: Fundamentals of Data Science and Machine Intelligence

Assignment 4

Based on Module 4B: Decision Trees and Module 5: Gaussian Mixture Modeling

1. A dataset contains 200 samples classified into two classes: 120 positive and 80 negatives.
   a. Compute the Gini index before splitting.
   b. If a split results in subsets:
      Left: (50 positive, 10 negative)
      Right: (70 positive, 70 negative)
      Compute the weighted Gini index and determine whether the split improves purity.
2. Consider the given dataset with two independent variables (x_1, x_2) and one dependent variable (y):
   a. Use the sum of squared errors (SSE) to determine the best splitting point for x_1.
   b. Construct the first split of a regression tree using SSE as the impurity measure.

x_1	x_2	y
1	5	10
2	6	12
3	8	15
4	10	18
5	12	21
6	15	25
7	18	28
8	20	30

3. Consider a 2-dimensional feature space with a dataset of $N = 10$ points. A vector quantization (VQ) system maps these points into $K = 3$ clusters using a codebook. The distortion function is the squared Euclidean distance between the original points and their assigned cluster centroids. Given the following initial cluster centroids:
   $$C_1 = (2,3), \qquad C_2 = (5,8), \qquad C_3 = (9,4)$$
   Assign the following data points to their closest centroid using squared Euclidean distance:
   $$(1,2), \qquad (3,4), \qquad (6,7), \qquad (8,3), \qquad (5,5)$$
   a. Compute the new centroids after one iteration of vector quantization.
   b. Show whether the distortion decreases after this iteration.
4. Show that if we maximize the first equation with respect to Σ_k and π_k while keeping the responsibilities $\gamma(z_{nk})$ fixed, we obtain the closed-form solutions given by the following equations:
   $$E_Z[\ln p\,(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\big(\ln \pi_k + \ln \mathcal{N}\,(x_n|\mu_k, \Sigma_k)\big)$$
   $$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^{\mathrm{T}}$$
   $$\pi_k = \frac{N_k}{N}$$
5. Consider a density model given by a mixture distribution
   $$p(x) = \sum_{k=1}^{K} \pi_k p(\, x \mid k\,)$$
   and suppose that we partition the vector **x** into two parts so that **x** = (**x**$_a$, **x**$_b$). Show that the conditional density $p(\, x_b \mid x_a\,)$ is itself a mixture distribution and find expressions for the mixing coefficients and component densities.
6. Consider a mixture of Gaussian distributions given by
   $$p(x|\Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where:
K: number of Gaussian components
π_k: mixing coefficients such that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$
$\mathcal{N}(x|\mu_k, \Sigma_k)$: Gaussian density with mean μ_k and covariance Σ_k
$\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{\mathrm{k=1}}^{\mathrm{K}}$ represents the parameters of the model.

a. Write down the complete log-likelihood function for a dataset $\{x_1, x_2, \dots, x_N\}$ assuming that the data points are drawn independently from the mixture model.
b. Derive the Maximum Likelihood Estimation (MLE) update rules for π_k, μ_k and Σ_k assuming that the component that generated each data point is known.

Programming Questions:

7. Write a code to obtain a fully grown regression tree for the data given in Q2 and visualize the regression tree.
8. Binary classification tree:
   a. Train a fully grown binary classification tree based on Gini impurity using the dataset *A4_train.csv* and visualize it.
   b. Compute the Sum of Squared Errors (SSE) on the test dataset (*A4_test.csv*) at each depth and plot the variation of SSE with depth.
   c. Determine the optimal pruning depth by selecting the depth where SSE change is minimal.
   d. Visualize the pruned tree.