ECE 512 – Topics in Data Science
Homework 7

Dror Baron, Hangjin Liu; Fall 2023

Due: October 25, 2023

Administrative instructions:

1. For any clarification or doubts, the TA Hangjin Liu (hliu25 AT ncsu DOT edu) is in charge of homework and projects. She should be your first point of contact on homework- and project-related issues.
2. The homework can be submitted individually, in pairs, or triples.
3. You should submit electronically through Moodle by midnight the day that the homework is due.
4. Please justify your answers carefully.

1. **Decision boundaries in Bayesian classification:** Recall our two simple classification examples (nearest neighbors and linear classifiers in LecturesPPT2.pdf; slides 2-12). In this question, you will construct a Bayesian classifier for the two-dimensional mixture Gaussian source that was used to generate the data. Following the MATLAB implementation of these classifiers in classification.m, which is available on the course webpage (you can also look at the Python translation), the two classes (blue and red) are each generated using num_clusters Gaussian components. That is, the multivariate (in our case, p=2) probability density function (pdf) is a sum of num_clusters Gaussian components, where component k has pdf $N(\mu_k,\sigma^2)$, $\mu_k \in \Re^2$ is the p-dimensional mean of cluster k, and σ^2 is the variance, which in the code was equal to cluster_variance for all clusters in our implementation.
    a. Express the pdf of a mixture Gaussian source. Hint: $X \sim N(\mu,\sigma^2)$ means that the pdf satisfies $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
    b. Recall that we have two mixture Gaussian classes, a red one and blue one. Bayes' rule informs us that Pr(Class=red|X)=Pr(Class=red,X)/[Pr(Class=red,X)+ Pr(Class=blue,X)]. Using the results of part (a), express the posterior probability.
    c. Construct code that sketches the optimal Bayesian decision regions. Feel free to compare the Bayesian decision regions to results obtained with nearest neighbors (you'll need to generate training data for nearest neighbors).
    d. Quadratic discriminant analysis (QDA) considers mixture Gaussian components whose covariance matrices are different. Using material discussed in class and explored in the slides, use a single Gaussian component per class, and revisit parts a-c above. You should be able to show that the decision boundaries of QDA are quadratic in nature.

2. **Nearest neighbors using a simple kernel.** Recall that the nearest neighbors (NN) classifier takes a "plurality vote" among the K training points nearest to our test point. In this problem, you will modify the implementation of NN such that the nearest neighbor receives weight (K)/[K(K+1)/2], the second nearest weight (K-1)/[K(K+1)/2], down to the K'th nearest neighbor whose weight is (1)/[K(K+1)/2]. (Note that the weights sum to 1, because the values in the numerators of the weights, i.e., 1+2+…+K, sum to K(K+1)/2.) Compare the original NN to the modified version with different weights. You may want to compare them on the class pdf's defined in Problem 1.