EECE5644 Fall 2018 – Exam 2

This assignment is due on Blackboard by 10:00am ET on Wednesday, December 5, 2018. Please submit your solutions on Blackboard in a single PDF file that includes all math, visual and quantitative results (plots, tables, etc), as well as your code (appended after your answers/solutions for each question). Do not submit a *.zip file! In this PDF file, start your answer for each question on a new page and clearly indicate on this new page which question is being answered. Clearly label axes of plots (use equal scales for axis where appropriate; e.g., using axis equal in Matlab), table rows/columns, and use descriptive captions for results displayed in this fashion. Whether you use Matlab, Python, or another computing language for your codes, please make use of built-in functions to the maximum possible extent.

1 Image segmentation with KNN and GMM (50 points)

Using the K-Means clustering algorithm with minimum Euclidean-distance-based assignments of samples to cluster centroids, segment the two attached color images into $K \in \{2, 3, 4, 5\}$ segments. As the feature vector for each pixel use a 5-dimensional feature vector consisting of normalized vertical and horizontal coordinates of the pixel relative to the top-left corner of the image, as well as normalized red, green, and blue values of the image color at that pixel. Normalize each feature by linearly shifting and scaling the values to the interval $[0, 1]$, such that the set of 5-dimensional normalized feature vectors representing each pixel are in the unit-hypercube $[0, 1]^5$.

For each $K \in \{2, 3, 4, 5\}$, let the algorithm assign labels to each pixel; specifically, label $l_{rc} \in \{1, ..., K\}$ to the pixel located at row r and column c. Present your clustering results in the form of an image of these label values. Make sure you improve this segmentation outcome visualization by using a contrast enhancement method; for instance, assign a unique color value to each label and make your label image colored, or assign visually distinct grayscale values to each label value or make best use of the range of gray values at your disposal for visualization.

Repeat this segmentation exercise using GMM-based clustering. For each specific K, use the EM algorithm to fit a GMM with K components, and then use that GMM to do MAP-classification style cluster label assignments to pixels. Display results similarly for this alternative clustering method. Briefly comment on the reasons of any differences if any.

2 Classification with SVM and MLP (50 points)

Train and test two classifiers on data that is described below. One of the classifiers should be a Support Vector Machine that uses Gaussian (radial-basis function) kernels, and the other one should be a single-hidden layer Multi-layer Perceptron (MLP) with sigmoid nonlinearities in all perceptron units. Optimize the following to minimize the empirical probability of error estimates obtained with 10-fold cross-validation on the available training data: (a) two hyperparameters of the SVM – overlap penalty weight factor in the objective and kernel width; (b) the number of perceptron units in the hidden layer – i.e. model order.

Generate 1000 independent and identically distributed (iid) samples for training and 10000 iid samples for test performance estimation, using the following data distribution for each class label $l \in \{0, 1\}$: $x = r_l \begin{bmatrix} cos(\theta) \\ sin(\theta) \end{bmatrix} + n$, where $\theta \sim Uniform[-\pi, \pi]$ and $n \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. Use $r_0 = 2$, $r_1 = 4$, $\sigma = 1$ as the parameters of the true pdf when generating data samples. You can see the attached Matlab script called *Exam2Q2_GenerateData.m* for a sample implementation of the data generation process.

Present your results that describe the entire process of training with cross-validation indicating clearly how/why hyperparameters and model order were selected. Also demonstrate the test performance of your classifiers (both visually and numerically by indicating their respective test probability of error estimates).