EECE5644 Fall 2019 – Homework 4

Submit: Monday, 2019-November-11 before 09:00ET

Please submit your solutions on Blackboard in a PDF file that includes all math and numerical results. Also include your code in one of the following ways: (Acceptable) upload an accompanying ZIP file containing all code files, (Preferred) keep your code in an online version control repository and provide a link to the relevant online repository in your PDF file.

Note that we will only grade the material submitted in the PDF file. Do NOT link from the PDF to online sources like Jupyter Notebook to present your numerical results. Those materials will not be considered when grading.

Make sure that you cite all resources you benefit from (books, papers, software packages). This is a graded assignment and the entirety of your submission must contain only your own work. You may benefit from literature including software (as allowed by specific restrictions in questions), as long as these sources are properly acknowledged in your submission.

Question 1 (50%)

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 value levels to each label value to 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.

Question 2 (50%)

In this exercise, you will train two support vector machine (SVM) classifiers and assess/compare their test performances. These SVMs wil respectively have linear and spherically-symmetric Gaussian (shaped radial basis function) kernels. We will refer to them as Linear-SVM and Gaussian-SVM. The data vectors are two-dimensional real-valued. The data distributions for the two classes are as follows: (1) data from class -1 are drawn from a Gaussian with zero-mean and identity-covariance-matrix; (2) data from class $+1$ are generated using a two-step procedure: a radius value is drawn from a uniform distribution over the interval $[2,3]$ and an angle value (in radians) is drawn from a uniform distribution over the interval $[-\pi,\pi]$; these radius and angle values are converted to Cartesian coordinates using the Polar-to-Cartesian coordinate transformation rule.

1. Generate a training set with 1000 independent samples from these two class distributions with priors $q_- = 0.35$ and $q_+ = 0.65$; note that this does not mean 350 samples from one class and 650 from the other – the class label needs to be randomly selected for each sample, in accordance with this prior. Visualize your training data.
2. Using 10-fold cross-validation, and minimum probability of error as the objective, select the hyper parameters for both Linear-SVM and Gaussian-SVM. For both classifiers, the constraint violation term weight (usually dwnoted by C; sometimes called the overlap penalty weight; referred to as the box constraint parameter in Matlab's fitcsvm) must be optimized. For the Gaussian kernel, the scale parameter (usually denoted by σ, corresponds to the standard deviation, if this Gaussian was a probability distribution) needs to be optimized. Visualize your cross-validation process in search of optimal hyperparameter values. Report the smallest probability of error estimate you get from cross-validation.
3. Using the best hyperparameters you identified, train your Linear-SVM and Gaussian-SVM using all of the training dataset. Visualize classification results on training data, count the erroneously classified samples and report the training dataset probability of error estimate.
4. Generate 1000 independent test samples from the same class distributions with the same priors as in the training dataset. Apply the Linear-SVM and Gaussian-SVM classifiers to the test data samples. Visualize the performance of your classifiers on the test dataset and report your test probability of error estimate.