EECE5644 Fall 2019 – Homework 3

Submit: Monday, 2019-November-04 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%)

Conduct the following model order selection exercise using 10-fold cross-validation procedure and report your procedure and results in a comprehensive, convincing, and rigorous fashion:

1. Select a Gaussian Mixture Model as the true probability density function for 2-dimensional real-valued data synthesis. This GMM will have 4 components with different mean vectors, different covariance matrices, and different probability for each Gaussian to be selected as the generator for each sample. Specify the true GMM that generates data.
2. Generate multiple data sets with independent identically distributed samples using this true GMM; these datasets will have respectively 10, 100, 1000, 10000 samples.
3. For each data set, using maximum likelihood parameter estimation principles (e.g. with the EM algorithm), within the framework of K(=10)-fold cross-validation, evaluate GMMs with different model orders; specifically evaluate candidate GMMs with 1, 2, 3, 4, 5, 6 Gaussian components. Note that both model parameter estimation and validation performance measures to be used is log-likelihood of data.
4. Report your results for the experiment, indicating which of the six GMM orders get selected for each of the datasets you produced. Develop a good way to describe and summarize your experiment results in the form of tables/figures.

Question 2 (50%)

Conduct the following maximum likelihood discriminative classifier training exercise on data generated from two Gaussian distributed classes:

1. Generate training set with 999 2-dimensional samples from two classes with priors $q_- = 0.3$ and $q_+ = 0.7$; the class-conditional data probability distributions are two Gaussians with different mean vectors and different covariance matrices (choose the matrices to be non-diagonal with distinct eigenvalues, so your Gaussian pdfs are tilted with respect to each other and elongated in different directions by different aspect ratios). *Hint: For more interesting results, make the Gaussians overlap with each other somewhat significantly, so that the minimum error probability achievable is not too small.*

2. Using Fisher LDA, identify a linear classifier that minimizes the error count on the training set. This classifier will have a discriminant function of the form $\mathbf{w}_{LDA}^T\mathbf{x} + b_{LDA}$ where $\mathbf{x}$ is the data vector, and the classifier decides in favor of Class $-$ if the discriminant is below 0, and decides in favor of Class $+$, if the discriminant is at least 0.
3. Train the parameters of a logistic function $y(\mathbf{x}) = 1/(1 + e^{(\mathbf{w}^T\mathbf{x}+b)})$ using the maximum likelihood estimation principle to optimize the parameters $\mathbf{w}$ and b with the training set, such that the function is trained to act as a surrogate for the posterior probability of Class $+$ given $\mathbf{x}$. In particular, your model assumes that $y(\mathbf{x}) \approx P(Label = +|\mathbf{x})$; consequently, $1 - y(\mathbf{x}) \approx P(Label = -|\mathbf{x})$. *Hint: Once you specify the optimization objective to train this logistic-linear-model for class posterior, you can solve the optimization problem using any suitable numerical optimization procedure, such as gradient ascent that you implement from scratch, or using a derivative free numerical optimization procedure like the Nelder-Mead Simplex Reflection Algorithm (e.g. in Matlab, fminsearch). Make a choice, implement correctly, perhaps consider using the LDA solution you developed earlier to provide an initial estimate for the model parameters.*
4. Report visual and numerical results that compare the following three classifiers (e.g. data scatter plots with color/shape indicators of true/decided labels), including the error counts each classifier achieve on the training set: MAP-classifier that makes use of the true data distributions and class priors, which achieves minimum probability error by design; LDA classifier you designed earlier; logistic-linear classifier you designed next.