ECE 662 HW 17 Review

April 23, 2008

Problem 1

Comments on the methodology

For this problem, the author assessed the similarity/difference between the two linear discriminant methods by comparing their classification error rates. The author used the two criterion functions $J_1(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$ and $J_2(\mathbf{w}) = \mathbf{w}^T S_B \mathbf{w}$ and the fact that $J_2(\mathbf{w})$ can be obtained from $J_1(\mathbf{w})$ by setting S_W to scaled identity matrix. The author generated two class artificial data in two and three dimensions. The author kept the separation between the class means as constant and for both 2- and 3-dimensional case, generated 4 cases by using different sets of covariance matrices for both the classes. This procedure was used for two different class distributions, namely the multivariate normal distribution and multivariate t distribution.

The author concluded that when sum of covariance matrices of both classes differs significantly from scaled identity matrix, the classification error rates using both criteria are quite different. Another conclusion that the author drew was that difference of error rates is large for higher dimensional data since it is more likely in higher dimensions that sum of covariance matrices has off diagonal elements, taking it away from scaled identity matrix.

What the reviewer liked

The author's use of t distribution in the investigation of this problem seems novel, although some reasons justifying its choice are in order.

The two conclusions made by the author are also quite interesting and informative since they confirm the intuition that more the sum of covariance matrices deviate from scaled identity matrix, the more will be the difference in performance of the two optimization criteria.

Even though, the author used artificially generated data, he/she did not favour any one optimization criterion over another which seems to be a reasonable conclusion because in reviewer's experience, when using artificial datasets, it is easier to introduce personal biases and generate datasets which show that one method is better over another. Rather, the author stuck to a more unbiased approach, highlighting the conditions when the two methods would perform similarly or differently.

What could be improved

Although the author provides the values of covariance matrices in the appendix to the report, it would be more helpful to show the matrices for each case along with the error rate comparison results so that reviewers can convince themselves about the observation that difference in error rates between the two methods is high when sum of covariance matrices differ significantly from scaled identity.

Also the reviewer noticed that the error rates for t distribution lie in the range 30-40 % while for normal distribution, they range between 3 to 5 %. Some explanation on this difference of error rates for the two distributions would help understand the nature of t distribution better.

Problem 2

Comments on the methodology

The author compared the classification performance of neural network and SVM, by designing the neural network using MATLAB toolbox and using the simpleSVM package downloaded from www. For neural network, the author experimented with varying the number of hidden layers as 2, 3 and 5 times the feature vector size. With regards to experimental data, the author generated uncorrelated Gaussian data in 1 and 2 dimensions with varying sample sizes and separation between class means.

What the reviewer liked

The experiments on neural network demonstrate the expected conclusion that classification performance should improve with larger training data, using more hidden layers and with larger separation between the classes.

The author's discussion on strengths and weaknesses of neural networks and SVM is good.

What could be improved

One major suggestion for improvement of the report is to have some graphical illustrations of the results since they reveal the conclusions more readily than just tables of numerical results.

Somehow, the reviewer got the impression that the author's conclusion is that SVM works well for linearly separable data and may not work very well for non linearly separable data. This may be true for SVM using a linear kernel function but with SVM using more sophisticated kernels like radial basis or polynomial kernels, we can have comparable or even better performance relative to neural network. The author may design SVM using some non linear kernel functions and then compare its performance with neural network for non linearly separable data.

Problem 3

Comments on the methodology

For this problem, the author investigated the relative classification performance of 3 density estimation based techniques: Parzen window, nearest neighbour and K-nearest neighbour. For Parzen window method, the author experimented with using Gaussian and hypercube windows and also varying the window size. The author calculated the classification error rate for different cases. For K-nearest neighbour, the author used a routine called 'knn' downloaded from www. K was varied from 2 to 15 and L2 norm was used as distance metric. In addition, different feature vector sizes (1 and 2) and number of training/test samples (10, 100, 1000) were used to calculate classification error rate in different cases. For nearest neighbour technique, the author modified the 'knn' routine to incorporate Manhattan distance as an additional metric.

What the reviewer liked

The author's experiments demonstrate the expected conclusion that probability density estimation is more accurate with larger training data size.

The author's experimentation with different distance metrics for nearest neighbour is quite unique. Most authors have not considered this aspect. In addition, the reviewer found the conclusion interesting and reasonable that if feature vectors are distributed on a grid, Manhattan distance can be more accurate distance metric.

What could be improved

It seems to reviewer that the author somehow forgot to include the results on Parzen window experimentation. There are neither any figures nor tables of numerical results for this part.

The author has shown some graphical results for K-nearest neighbour method to compare the effects of varying experimental factors. The reviewer feels that instead of using multiple figures, the comparisons could be more easily highlighted by showing the results on a multi column bar chart (e.g. as generated by Microsoft Excel).

Overall Grade: A-