Coursework **Fall 2012**
Statistical Pattern Recognition (COMP136) **Roni Khardon**
Department of Computer Science, Tufts University

Assignment 5

This assignment is due by Thursday November 1 in class.

Experiments with Classification Algorithms

In this assignment you will implement and evaluate several algorithms for the classification problem. In particular, we will look at 4 algorithms discussed in class:

- 2 class generative model with shared covariance matrix as developed in section 4.2 of the textbook.
- 2 class generative model with a separate covariance matrix for each class. In this case, learning is done by estimating a Normal distribution on the examples from each class separately.
- Fisher's linear discriminant as in Eq (4.38) in the textbook with threshold given by $w^T \cdot m$ where m is the mean of the examples.
- Bayesian logistic regression where we use the simple prior $w \sim \mathcal{N}(0, \frac{1}{\alpha}I)$. In this case the update formula for w is: $w_{n+1} \leftarrow w_n - (\alpha I + \Phi^T R \Phi)^{-1}[\Phi^T(y - t) + \alpha w_n]$ where R is as in Eq (4.98) in the textbook.

  Note that unlike the other algorithms, logistic regression relies on a free parameter (w_0) to capture an appropriate separating hyperplane. Therefore, you will need to add a feature fixed at one to the data for this algorithm. For the implementation: initialize $w = 0$ and use $\alpha = 10$. The value $\alpha = 10$ yields reasonable results across our datasets but is not optimal for each dataset. It is not required for the assignment but you can use additional train-set only cross validation to pick a value of α in each dataset.

We will use 4 datasets, available through the course web page, to evaluate the algorithms:

- The first dataset (marked as A) has uniformly distributed data with an arbitrary weight vector separating positive from negative examples. Therefore data does not conform to the Gaussian generative model but the data is linearly separable.
- For the second dataset (marked as B) data is generated using the Gaussian generative model where the two classes have different covariance matrices and there is some overlap between the distributions. Recall that linear models are not expected to do as well in this case.
- In the third dataset (marked as C) each class is generated from multiple Gaussians with differing covariance structure. Here none of the algorithms uses the correct model for the data, yet we designed it to be somewhat linearly separable.

We also designed it so that a quadratic separator is likely to perform much better than a linear separator, but this point is not directly tested in the experiments of this assignment. If you are interested and have time you can test this point in addition to the ones requested below.

- The forth dataset, *ionosphere*, is taken from the UCI repository.

Each dataset is given in two files with the data in one and the labels in the other file. Unlike the previous assignment we did not split the data into a training set and test set and you will use cross validation to report results.

Your Task

Implement the 4 algorithms, and evaluate them as follows. For each dataset, perform 10-fold cross validation and report the obtained accuracies and standard deviations.

Repeat this procedure for different training set sizes to obtain learning curves (with error bars) for the algorithms. To reduce variance in the results, you can keep the test set for each fold fixed and subsample the training set for the fold to obtain error rates as a function of training set size.

Submitting Your Assignment

Please submit printouts of your code, and a short report on the experiments, their results, and your conclusions from them.