ECEN 689-605 Machine Learning

Home Work Assignment: Sheet 2

Due on February 28, 2019

In this assignment there are four problems. In all problems, you are given some data on which you are supposed to learn a 2-class classifier. You assess the performance of the learnt classifier using a test set.

In three problems the data is synthetic and in one problem the data may be thought of as real data. For the synthetic data problems, the details of how the data is generated is given.

Each of the problems are described in detail below. For each problem, you are asked to vary some conditions (e.g., size of training set) and also to compare different methods (e.g., Bayes classifier vs. nearest neighbour). This is the minimum exploration you are required to do. You are welcome to explore further if you can think of some other interesting things to do.

You can implement the learning algorithms on any platform you want (C, C++, MATLAB, Python, scikit-learn, PyTorch etc.). You are welcome to use codes that are freely available from any source. You are not required to submit any codes/implementation.

What you need to submit is a report summarizing your exploration of the data set. For each problem, very briefly describe what all are implemented and then present the results obtained. Discuss any points from your results that you consider surprising or interesting. In all cases, explain why such behaviour may be exhibted by the algorithm. The final submission should be in the form of a short PDF file.

The grading depends on whether or not you have done all explorations that are asked for, how you presented the results, your discussion of results and whether you have done some exploration on your own.

The problems are described below.

1. This is a 2-class classification problem with two dimensional feature vector. The class conditional densities are: f_i is normal with parame-

ters μ_i, Σ_i, $i = 1, 2$. There are three data sets here. All of them have same values for three of the parameters as given below:

$$\mu_1 = [0\ 0]^T;\ \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix};\ \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

The three data sets have three different values for μ_2:

a. $\mu_2 = [1\ 1]^T$ (Data files: P1a_train_data_2D.txt and P1a_test_data_2D.txt)

b. $\mu_2 = [3\ 3]^T$ (Data files: P1b_train_data_2D.txt and P1b_test_data_2D.txt)

c. $\mu_2 = [3\ 6]^T$ (Data files: P1c_train_data_2D.txt and P1c_test_data_2D.txt)

The training and test data files are simple text files with one feature vector and class label per line. The class labels are '1' and '-1'.

For subproblems a and b, assume class conditional densities are normal, estimate them from training data, find the accuracy of the resulting Bayes classifier on the full test set. For estimating each class conditional density, take 5, 10, 25, 75 examples by radomly sampling from the given training data. For each case, compare the accuracy of the Bayes classifier with that of nearest neighbour classifier. In each case, use the same set of training examples (as used for estimation) as prototypes in nearest neightbour classifier.

Suppose in the training data set we ignore the class labels. Then the data can be considered as sampled from a density $f(x) = p_1 f_1(x) + p_2 f_2(x)$ where p_1 and p_2 are prior probabilities and f_1 and f_2 are class conditional densities. Thus this unlabelled data is coming from a mixture density. For the data corresponding to subproblem b, ignore class labels, use the x_i in the training set to learn a two component mixture density using EM algorithm. Now use the learnt component densities as the class conditional densities, implement a Bayes classifier and compare against the Bayes classifier learnt using the class labels.

For subproblem c assume class conditinal densities to be normal, vary sample size for estimation as above and find accuracies of the implemented Bayes classifier. Then, assume class conditinal density for class-2 to be exponential (while assuming that for class-1 to be normal), estimate the exponential density and implement a Bayes classifier. Compare the two bayes classifiers. (The exponential density is given by $f(x) = \lambda e^{-\lambda x},\ x > 0$ where $\lambda > 0$ is the parameter).

2. This is also a 2-class problem with normal class conditional densities but the feature vectors are twenty dimensional. There are three sub-problems again and the training and test data files are named similarly. For all three cases, μ_1 is a vector of all zeros and μ_2 is a vector of all ones. The covariance matrices for the three cases are as given below.

   a. $\Sigma_1 = \Sigma_2 = I$ where I is the identity matrix.

   b. $\Sigma_1 = \Sigma_2 = 3I$.

   c. Σ_1 and Σ_2 have 1's on the diagonal and off-diagonal elements are randomly chosen from a uniform density over $[-1,\ 1]$. (Note that the two covariance matrices are not same now and that while off-diagonal elements are random we ensure that the matrix is symmetric).

   Vary the sample size for estimating class conditional densities as 10, 50, 100, 300. Compare the resulting Bayes classifiers with corresponding nearest neightbour classifiers.

3. This is also a 2-class problem but with one dimensional feature vector. There are two sub problems here. In both cases, class conditonal densities are mixtures of gaussian as specified below. (Notation $\mathcal{N}(a, b)$ denotes normal density with mean a and variance b).

   a. The class conditional densities for the two classes are:

   $$f_1 = 0.5\mathcal{N}(0, 4) + 0.5\mathcal{N}(4, 4)$$

   $$f_2 = 0.5\mathcal{N}(8, 5) + 0.5\mathcal{N}(12, 5)$$

   (The data files are: P3a_train_data.txt and P3a_test_data.txt)

   b. The class conditional densities for the two classes are:

   $$f_1 = 0.5\mathcal{N}(0, 4) + 0.5\mathcal{N}(8, 5)$$

   $$f_2 = 0.5\mathcal{N}(4, 4) + 0.5\mathcal{N}(12, 5)$$

   (The data files are: P3b_train_data.txt and P3b_test_data.txt)

   For each problem you implement three classifiers: (i). assume class conditional densities are mixtures of two Gaussians, estimate using

EM algorithm and implement Bayes classifier; (ii). Assume class conditional densities are single Gaussian, estimate using Maximum Likelihood method and implement Bayes classifier; (ii). using the same training data as used for estimating densities, implement nearest neighbour classifier. Compare performance of the three classifiers.

4. The final problem is about classifying documents. The data set consists of 2000 moview reviews and it is a 2-class problem. (The data is in the file sentiment_analysis.csv). For this data you explore a naive Bayes classifier using two different feature vectors: (i). 'bag-of-word' representation where each feature is binary, (ii). TF-IDF based feature vector. Discuss the performance of the classifiers. In this problem the full data is given as one file to you. part of your job is to decide how to use the data to learn the classifier and then test the classifier.

   (There are codes avaialbale for implementing such classifiers in PyTorch or scikit-learn etc. You are welcome to use any such available code.)