EE 657 - Pattern Recognition and Machine Learning

Jan - May 2018

Assignment - 1

☞ Submission Deadline: **8 Apr 2018 (Sunday), 12:00 AM**

☞ Assignment submission link will be made online few days prior to the submission deadline. Submission should include:

(i) Code written in Python.

(ii) Report (pdf) explaining the summary of the experiments performed, the observations made and the conclusions. Include informative plots, figures and/or tables that justifies your findings.

☞ Usage of predefined libraries for regression/classification is not permitted. The code has to be written from scratch except for simple built-in functions (eg: function for finding matrix inverse, determinant, matrix product, ..etc).

---

1. **Dataset Description:** *Optical Recognition of Handwritten Digits Dataset* from the *UCI repository*. The original dataset consists of normalized bitmaps of handwritten digits $(0-9)$. 32x32 bitmaps are divided into non-overlapping blocks of 4x4 and the number of ON pixels are counted in each block. This generates an input matrix of 8x8 where each element is an integer in the range $0-16$. This reduces dimensionality and gives invariance to small distortions.

   The given dataset is a modified version of the above dataset, consisting of the data corresponding to the handwritten digits 5 & 6 extracted from the original dataset.

   - **Training data:** ' `P1_data_train.csv` ' consisting of 777 instances(rows) of 64 attributes(cols) corresponding to the handwritten digit value(5 or 6) given in ' `P1_labels_train.csv` '.
   - **Test data:** ' `P1_data_test.csv` ' consisting of 333 instances(rows) of 64 attributes(cols) corresponding to the handwritten digit value(5 or 6) given in ' `P1_labels_test.csv` '.

   Problem: Learn the parameters μ_5, μ_6, $\sum_5$, $\sum_6$ and π by maximizing the likelihood,

   $$\begin{array}{ll} \pi = Pr(r = C_5) & Pr(x|C_5) = \mathcal{N}(x|\mu_5,\ \sum_5) \\ 1-\pi = Pr(r = C_6) & Pr(x|C_6) = \mathcal{N}(x|\mu_6,\ \sum_6) \end{array}$$

   Use Bayes decision criterion to classify the test data. Estimate the misclassification rates of both classes and populate the 2x2 confusion matrix.
   ($Hint$: Try out different variations of covariance matrices, esp. the case of $\sum_5 = \sum_6 = \sum$)

2. The following problem is intended to illustrate alterations in performance and shape of distributions brought about by variations in covariance matrices. (*Note:* The program developed for problem 1 can be reused with a few modifications for solving problem 2.)

   - **Training data:** ' `P2_train.csv` ' consisting of 310 instances, 2 attributes +1 class label.
   - **Test data:** ' `P2_test.csv` ' consisting of 90 instances, 2 attributes +1 class label.

   Problem: Learn a binary classifier for the given data taking class conditional densities as normal density. Estimate the misclassification rates of both classes, plot the discriminant function and iso-probability contours for the following cases:

   (a) Equal diagonal $\sum_s$ of equal variances along both dimensions, $\quad \sum_0 = \sum_1 = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$

   (b) Equal diagonal $\sum_s$ with unequal variances along different dimensions, $\quad \sum_0 = \sum_1 = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$

(c) Arbitrary $\sum_s$ but shared by both classes, $\sum_0 = \sum_1 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

(d) Different arbitrary $\sum_s$ for the two classes.

3. **Dataset Description:** *Wage dataset* contains the income survey information for a group of males from Atlantic region of the United States.

   - **Data:** ‘ `Wage_dataset.csv` ′ has the numerical data from ‘ `Wage_original.csv` ′ extracted (except the 1st column). The data consist of 3000 instances, 9 attributes (including the *age* and *education* and the calendar *year*)+ 2 columns giving the natural log of the *wage*, *wage* respectively.

   Problem: We wish to understand the association between an employees *age* and *education*, as well as the calendar *year*, on his wage. Perform polynomial regression on *age* vs *wage*, *year* vs *wage*, plot *education* vs *wage*. Provide description of your observations on the variation of *wage* as a function of each these attributes. Can we get an accurate prediction of a particular man′s wage from one of these 3 attributes alone?
   Experiment with your own ideas, related to those discussed in the class and make a brief report of your findings.

****