Indian Institute of Space Science and Technology – Thiruvananthapuram

MA613 Data Mining

Assignment-III

Date: 28-09-2015

1. (a) Apply logistic regression and Gaussian discriminant analysis on the following data:
   i. Data1 (please find the attached data)
   ii. Breast Cancer Wisconsin (Diagnostic) Data Set (download the data from UCI web repository)

   (b) Report the preprocessing techniques you have done on the data, if any.

   (c) Plot ROC curve.

   (d) Compare the performance of the algorithms.

   (e) Plot the decision boundaries of the two algorithms for Data1.

   (f) Plot the logistic regression curve.

   (g) Plot multivariate Gaussian density function corresponding to the positive & negative classes and their contours for Data1.

   (h) Using GDA, plot posterior probability corresponding to positive and negative class for Data1.

   (i) Report the values of the parameters of the final model.

2. Explain how you extend logistic regression for analysing nonlinear data. Describe regularized logistic regression.

3. Describe GDA for two class classification of which the positive and negative class have different covariance matrices.

4. Find the covariance matrix of the following data manually using two formulas you

know:

$$X = \begin{bmatrix} 8 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 2 \\ 0 & 4 & 0 \end{bmatrix}$$

5. Check whether the emails having the content (a) "change your account number" (b) "send your address" are spam using Naive Bayes algorithm, on the basis of the below information:
   D1: "send us your internet banking password" spam
   D2: "send us your mobile number" ham
   D3: "change your internet banking password" ham
   D4: "change your userid " spam
   D5: "send your internet banking password" spam
   D6: "send us your account details" spam

Notes

- All the files related with the assignment should be saved in a single folder and send to sumitra@iist.ac.in.
- Last date of submission: 08-10-2015.
- **As far as assignments are concerned, students are expected to observe academic honesty and integrity. Though the students can collaborate and discuss, copying directly other students' assignment or allowing your own assignment to be copied constitute academic dishonesty and is highly discouraged.**