Assignment 5 : IIT2016067

(1) Spam/ham classification using naive baysian method

(2) Identifying river and non-river points in the image using naive baysian method

QUESTION DESCRIPTIONS

(1) Using Naive Bayesian classifier predict where a given mail is spam or not. Use the data set provided for this purpose. (structured data set)

(2) Using Naive Bayesian classifier predict river non river using Satellite data set of Hooghly river (unstructured data set).

[QUESTION 1]

INTRODUCTION

We are given a dataset of email text and a label specifying that whether the email is spam or ham(non-spam). We need to build a classifier using the naive baysian concepts to classify some text as spam or ham.

CONCEPTS USED

1. Basic Formula

P(A | B) = (P(B | A) * P(A)) | P(B)

2. Formula Definition

P(H | X) = Probability that email is ham(H) given that it contains document- X (lets say X = content of email)

P(S | X) = Probability that email is spam(S) given that it contains document- X **P(H | X) = (P(H) * P(X | H)) / P(X)**

P(S | X) = (P(S) * P(X | S)) / P(X)

$P(H \mid X) = (P(H) * P(X \mid H)) / P(X)$

$P(S \mid X) = (P(S) * P(X \mid S)) / P(X)$

P(H) = (number of ham documents) / (total number of documents)

P(S) = (number of spam documents) / (total number of documents)

$P(X \mid S) = P(X_1 \mid S) * P(X_2 \mid S) * \ldots$

$P(X \mid H) = P(X_1 \mid H) * P(X_2 \mid H) * \ldots$

Where X_1, X_2, … are the words of the dictionary

$P(X_j \mid S)$ = count of word X_j belonging to category spam / total count of words belonging to category spam.

$P(X_j \mid H)$ = count of X_j belonging to category ham / total count of words belonging to category ham.

3. Results

The dataset was divided into 75% training and 25% testing.

Accuracy of the training set: 98.64102564102564%

Accuracy of the testing set: 97.72727272727273%

[QUESTION 2]

INTRODUCTION

We are given a dataset of 4 images of the hoogly river in the R, G, B and I band. We need to build a classifier to classify each of 512 X 512 pixels as river or non river using the naive baysian approach.

CONCEPTS USED

Gaussian class conditional densities

Here, $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, \ldots, K\}$. Estimate Bayes classifier via MLE:

- **Class priors**: The MLE estimate of π_y is $\hat{\pi}_y = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i = y)$.
- **Class conditional density**: Choose $p(x|Y = y) = N(x|\mu_y, \Sigma_y)$. The MLE estimate of (μ_y, Σ_y) is

$$\hat{\mu}_y = \frac{1}{n_y}\sum_{i=1}^{n} \mathbb{1}(y_i = y)x_i,$$

$$\hat{\Sigma}_y = \frac{1}{n_y}\sum_{i=1}^{n} \mathbb{1}(y_i = y)(x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T.$$

This is just the empirical mean and covariance of class y.

- **Plug-in classifier**:

$$\hat{f}(x) = \arg\max_{y \in \mathcal{Y}} \; \hat{\pi}_y |\hat{\Sigma}_y|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \hat{\mu}_y)^T \hat{\Sigma}_y^{-1}(x - \hat{\mu}_y)\right\}.$$

Taken from the slides of the edX course ColumbiaX: CSMM.102x Machine Learning

Steps Followed:

Four satellite Images of Kolkata (Rband, Gband, Bband and Iband) are given to you with equal image size (512 * 512).

• The feature vector dimension is 4

• Each pixel location we have four values.

• Two Classes are given (River and NonRiver)

• Take 50 sample points (Pixel location's corresponding pixel values) from river class for training for each band

•Take 100 sample points (Pixel location's corresponding pixel values) from non river class for training for each band.

•Take (512 * 512) sample points (Pixel location's corresponding pixel values) for testing for each band.

•Apply baye's decision rule to classify all the test sample either in river or nonriver class denoting 0 and 255 at corresponding pixel locations.

•Show the result in image form with black and white image (either 0 and 255)

•Step 1: Calculate Mean of River Class : T1 = [Mean1; Mean2; Mean3 ;Mean4]; Mean1 = mean of Rband image for 50 sample points

Mean2 = mean of Gband Image for 50 sample points

Mean3 = mean of Bband image for 50 sample points

Mean4 = mean of Iband image for 50 sample points

•Step 2: Calculate Mean of NonRiver Class : T2 = [Mean1; Mean2; Mean3; Mean4];

Mean1 = mean of Rband image for 100 sample points

Mean2 = mean of Gband Image for 100 sample points

Mean3 = mean of Bband image for 100 sample points

Mean4 = mean of Iband image for 100 sample points

•Step 3: Calculate the Covariance Matrix for River Class for 50 samples which is 4 * 4 dimensions. Basically (X – T1) deviation and (Y – T1) deviation and multiply it and summing up where X and Y represents all the sample points considered for training (R, G, B and I band image) we will get 2^4 = 16 values in the covariance matrix for possible combinations of 4 band images. We are doing the deviation of sample points from the mean vector.

(Apply covariance matrix calculation formula)

• Step 4: Calculate the Covariance Matrix for Non River Class for 100 samples which is 4 * 4 dimensions also by applying same process explained in step 3.

• Step 5: Take whole image for test data where : test_data= [Rband_img(i,j) Gband_img(i,j) Bband_img(i,j) Iband_img(i,j)]; i = 1 to 512; and j = 1 to 512;

• step 6: The dimension of test data is (4 * (512 * 512));

• Step 7: For each pixel location of test image Run the loop from i = 1 to (512*512) Do

• Step 8: For river class calculate (test_data – T1) deviation and $(\text{test\_data} - T1)^T$ Then Multiply it :

River_class = $(\text{Test\_data} - T1)^T$ * Inverse (Covariance_matrix_Riverclass) *(Test_data – T1)

• Step 9: For Non_river class calculate (test_data – T2) deviation and $(\text{test\_data} - T2)^T$. Then Multiply it :

Nonriver class = $(\text{Test\_data} - T2)^T$ * Inverse (Covariance_matrix_NonRiverclass) *(Test_data – T2)

• Step 10: Calculate density function p1 for river class where P1 = 0.3 given

p1 = (-0.5) * 1/sqrt(Determinant of Covariance_matrix_Riverclass) * exp(River_class);

(Here we apply multivariate Normal Distrubution)

• Step 11: Calculate density function p2 for nonriver class where P2 = 0.7 given

p2 = (-0.5) * 1/sqrt(Determinant of Covariance_matrix_nonRiverclass) * exp(NonRiver_class);

• Step 12: For each pixel location of test image apply baye's rule:

(P1 * p1) >= (P2 * p2) then Out_image(i) = 255 (River class)

Else Out_image(i) = 0; (Nonriver class)

• Step 13 : Goto step 7;

• Step 14: Show the three output image Image using imshow function for three cases:

Case 1 : River class (Prior Prob:) = 0.3 , Nonriver class(Prior Prob) = 0.7

Case 2 : River class (Prior Prob:) = 0.7 , Nonriver class(Prior Prob) = 0.3

Case 3 : River class (Prior Prob:) = 0.5 , Nonriver class(Prior Prob) = 0.5

3. Results

For P1 = 0.7 and P2 = 0.3 the results are :

[The color is changed for clarity purposes]