PROBABILITY MODELS FOR COMPUTER SCIENCE

PROJECT

                                              AARYA ARUN [PES1201700009]

          INDU A RALLABHANDI [PES1201700795]

        23 November, 2018

ABSTRACT:

TASK: Classify journals as national or international using minimum error classification.

Given: Dataset with journal names and seven features associated with them.

Introduction:

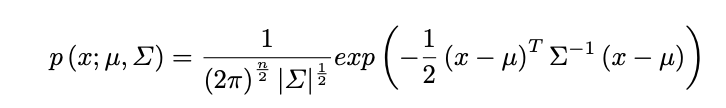
Bayes’ theorem describes the probability of an event based on the prior knowledge of conditions associated with the given event. With the given task, the events are the classification of given test points as national or international level journals. Training the dataset collects prior probability of each event. The given classifier takes a decision based on the values of the risk functions, and is a minimum-risk based model.

The given data set has journals split into two classes: national, and international level journals. The dataset records have been shuffled to make the sampling random. We have used 16 data points for training from each of these classes. 4 test points have been used for national journals and 6 for international ones.

Data points for national and international journals have numerical attributes Selfcities, Non-local count, Totalcities NLIQ, OCQ, HINDEX, and IC.

METHODOLOGY:

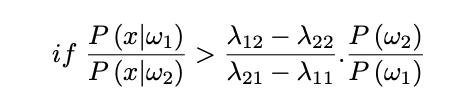
* The test points for each class are chosen from the first few datapoints, i.e the frst 4 for national and first 6 for international.
* Each data point’s feature vectors are created with attributes stored in a numpy array.
* Model:



* We calculated the mean vector(dimensions: 7x1) and covariance matrix(dimensions: 7x7) for each class. There are 7 variables in the multivariate normal distribution.

1. Mean vector calculation: Done by finding the mean of every attribute
2. Covariance matrix calculation: Done by finding the covariance of every pair of attributes.

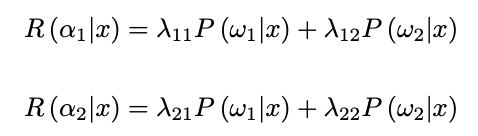
* Likelihood method used in decision rule for classification: For a data-point x, it returns      P(x|wj)
* Training the dataset involves instantiating each normal distribution oce per class.
* Each data point is then classified by repeated iterations through the training data. By comparing the model’s classification with the actual classification of the given data points, we can estimate the misclassification.
* Likelihood ratio decision rule:

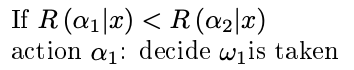


We take action α1(decide ω1)

 Otherwise, we take action α2(decide ω2)

* Risk functions:





Simple algebra is used to choose the class with minimum risk of misclassification, leading to the above decision rule.

RESULTS:

The International misclassification was found to be: 17.3333333333333332% 17.33%

The National misclassification was found to be: 9.5%

The overall misclassification was found to be:  14.2%