Assignment 5 - report - Soft Computing

Mtech - 2nd Sem

K.Ranjith - MIT2020017 - mailID(mit2020017@iiita.ac.in, ranji.iitb@gmail.com)

Date- 15-03-2021

IIIT Allahabad

edit google collab link for code

CLICK HERE

APPROACH

1. here train and test data is created from my personal mail box(spam messages) and IIITA mail box (normal messages)
2. From each message special charectors and numbers are removed, message is converted to lower case. From this plane message bag of words are extracted
3. In GDA analysis train data is created such that each feature/column represents a unique word from entire train sample and each value in particular feature represents how many time that word is repeated and each row represents a particular message in trian data

results

1. hyperparameters = [train-data = 38, test-data = 28]
2. Naive Bayes classifier for multivariate Bernoulli model the accuracy ratio = 0.7931034482758621
3. Naive Bayes classifier for multinomial Bernoulli model the accuracy ratio = 0.6551724137931034
4. for Guassian Discriminant Analysis model the accuracy ratio = 0.689655172413793 1

comparitive analysis

1. The feautures follow discrete values and bernoulis distribution when sufficiently large train-data is taken
2. in Naive Bayes classifier for multivariate Bernoulli model the basic assumption is the features are binary valued and follow bernoulis distribution so accuracy is better in this model compared to remaining two
3. in Naive Bayes classifier for multinomial Bernoulli model the basic assumption is the features are discrete and follow multinomial distribution. however even features are discreate but dont follow multinomial distribution as our train sample less

comparitive analysis

1. for Guassian Discriminant Analysis model the basic assumption is the feaures are continous and follow normal distribution. In our implimentation the features are continuous and are not normally distributed. As a result this model produces poor results compared to multivariate Bernoulli model
2. Guassian Discriminant Analysis model produces better results compared to multinomial Bernoulli model as this model can produce better results even with less data

future work

1. Only a few train and test samples created because of time constraints.
2. Any one reading this report can add more data to csv file to get better model parameters(click the editable google collab link given at the start to add data)
3. instead of inverse in GDA model pseudo inverse is taken and det of sigma is considered a less value as sigma in GDA model is sigular because we had taken less data. when large data is added take actual inverse and real det to get precise model.