Assignment 1

Dataset Choice:

Sentiment analysis using **Bo and Pang's movie review** dataset with binary classification between positive and negative reviews.

Feature Design:

After getting the dataset some pre-processing was done on the dataset. All kinds of numbers, punctuations and special characters were removed. All the text was changed to lower case. After that I have selected bag-of-words as the feature design.

Bag-of-words (bow): A **bag-of-words** is a representation of text that describes the occurrence of words within a document. It involves two things: A vocabulary of known words and a measure of the presence of known words. In bag-of-words approach the number of occurrence and not sequence or order of words matters.

I have selected **bow** as the feature design as **n-gram** is mostly used for paragraph subjectivity classification and bow is mainly used for tweet or review classification. Since, the dataset of my choice was movie review classification so bag-of-words is justified.

A **5-fold cross validation** was performed to select the best model.

Evaluation:

The accuracy for the different classifiers used is presented below in the form of tables.

Q1) **K-NN classification.**

p \ k	**1**	**3**	**5**
1	0.69	0.72	0.74
2	0.66	0.65	0.63
infinity	0.63	0.55	0.59

Q3) **Kernelized SVM with RBF kernel.**

Gamma\C	**0.01**	**0.1**	**1**	**10**	**100**
0.01	0.55	0.55	0.55	0.55	0.55
0.1	0.55	0.55	0.55	0.55	0.71
1	0.54	0.54	0.54	0.71	0.83
10	0.54	0.55	0.73	0.83	0.84
100	0.56	0.60	0.81	0.84	0.84

Q4) **Regularized logistic regression.**

C	**0.001**	**0.01**	**0.1**	**1**	**10**	**100**	**1000**
Accuracy	0.685	0.685	0.655	0.655	0.695	0.775	0.81

Q5) **Gaussian based Bayes classifier.**

Accuracy: 0.695

All codes are present in GitHub.