Machine Learning- Homework-1

Results for HW1 Dataset-

	Score values	Naïve Bayes	Multinomial Naïve Bayes	Logistic Regression	Stochastic Gradient Descent
Bag of Words Model	Accuracy	-	0.9435	0.9351	0.9351
	F1 Score	-	0.9618	0.9555	0.9556
	Precision	-	0.9470	0.9568	0.9515
	Recall	-	0.9770	0.9541	0.9597
Bernoulli Model	Accuracy	0.9205	-	0.9539	0.9435
	F1 Score	0.9487	-	0.9685	0.9577
	Precision	0.9078	-	0.9741	0.9392
	Recall	0.9913	-	0.9630	0.9770

Results for Enron 1 Dataset-

	Score values	Naïve Bayes	Multinomial Naïve Bayes	Logistic Regression	Stochastic Gradient Descent
Bag of Words Model	Accuracy	-	0.9276	0.9122	0.9320
	F1 Score	-	0.9478	0.9350	0.9496
	Precision	-	0.9202	0.9381	0.9481
	Recall	-	0.9771	0.9320	0.9511
Bernoulli Model	Accuracy	0.8815	-	0.9451	0.9572
	F1 Score	0.9189	-	0.9591	0.9655
	Precision	0.8523	-	0.9543	0.9735
	Recall	0.9967	-	0.9638	0.9576

Results for Enron 4 dataset-

	Score values	Naïve Bayes	Multinomial Naïve Bayes	Logistic Regression	Stochastic Gradient Descent
Bag of Words Model	**Accuracy**	**-**	**0.9447**	**0.9650**	**0.9613**
	F1 Score	**-**	**0.8973**	**0.9377**	**0.9311**
	Precision	**-**	**0.9357**	**0.9407**	**0.9281**
	Recall	**-**	**0.8618**	**0.9346**	**0.9342**
Bernoulli Model	**Accuracy**	**0.9465**	**-**	**0.9761**	**0.9629**
	F1 Score	**0.8961**	**-**	**0.9565**	**0.9143**
	Precision	**0.9843**	**-**	**0.9407**	**1.0**
	Recall	**0.8224**	**-**	**0.9728**	**0.8421**

Hyper-Parameters for Logistic Regression and SGD Classifier-

Logistic Regression-

In case of Logistic Regression, a list of lambda values was given, and the validation set (i.e. 30% of the train set) was used to choose a lambda parameter such that it gives the maximum accuracy as compared to other lambda values. Then the chosen lambda value was used to train on the whole training set and then tested on the test set.

Following is the list of the Lambda values used-

[0.001, 0.01, 0.05, 0.075,0.3, 0.45, 0.5, 0.75]

Alpha (learning rate) of 0.05 was used.

Number of iterations – 100

HW1 dataset-

Lambda selected for Bag of Words Model= 0.45

Lambda selected for Bernoulli Model= 0.01

Enron1 Dataset-

Lambda selected for Bag of Words Model= 0.001

Lambda selected for Bernoulli Model= 0.45

Enron4 Dataset-

Lambda selected for Bag of Words Model= 0.45

Lambda selected for Bernoulli Model= 0.01

The Accuracy drops drastically with lambda values >3.

(*Lambda values change slightly on each new run as the shuffling of data changes)

SGD Classifier-

HW1 Dataset-

Best Parameters for Bag of Words Model are--> {'alpha': 0.1, 'loss': 'hinge', 'max_iter': 350, 'penalty': 'l2'}

Best Parameters for Bernoulli Model are--> {'alpha': 0.01, 'loss': 'log', 'max_iter': 350, 'penalty': 'l2'}

Enron1 Dataset-

Best Parameters for Bag of Words Model are--> {'alpha': 0.05, 'loss': 'hinge', 'max_iter': 350, 'penalty': 'l2'}

Best Parameters for Bernoulli Model are--> {'alpha': 0.01, 'loss': 'hinge', 'max_iter': 50, 'penalty': 'l2'}

Enron4 Dataset-

Best Parameters for Bag of Words Model are--> {'alpha': 0.0001, 'loss': 'log', 'max_iter': 50, 'penalty': 'l2'}

Best Parameters for Bernoulli Model are--> {'alpha': 0.01, 'loss': 'hinge', 'max_iter': 100, 'penalty': 'l2'}

Questions-

1. Which data representation and algorithm combination yields the best performance (measured in terms of the accuracy, precision, recall and F1 score) and why?

- The Bernoulli Model with Logistic Regression yields the best performance.
- It gives Avg. accuracy=0.9583
  Avg. F1 score= 0.9614
  Avg. Precision= 0.9563
  Avg. Recall score= 0.9665

2. Does Multinomial Naive Bayes perform better (again performance is measured in terms of the accuracy, precision, recall and F1 score) than LR and SGDClassifier on the Bag of words representation? Explain your yes/no answer.

- No, Logistic Regression and SGD classifier perform better than the Multinomial Naïve Bayes when compared with the Avg. values of the performance scores.

- Multinomial Naïve Bayes gives the following Avg. values-

  Avg. accuracy=0.9386

  Avg. F1 score=0.9356

  Avg. Precision=0.9343

  Avg. Recall=0.9386

- While Logistic regression for Bag of Words Model gives-
  Avg. Accuracy=0.9374
  Avg. F1 Score=0.9427
  Avg. Precision=0.9452
  Avg. Recall=0.9402
- While SGD classifier for Bag of words Model gives-
  Avg. Accuracy=0.9428
  Avg. F1 score=0.9454
  Avg. Precision=0.9426
  Avg. Recall=0.9483

3. Does Discrete Naive Bayes perform better (again performance is measured in terms of the accuracy, precision, recall and F1 score) than LR and SGDClassifier on the Bernoulli representation? Explain your yes/no answer.

- No, Logistic Regression and SGD classifier perform better than the Discrete Naïve Bayes when compared with the Avg. values of the performance scores.
- Discrete Naïve Bayes gives the following Avg. values-

  Avg. accuracy=0.9162

  Avg. F1 score=0.9212

  Avg. Precision=0.9148

  Avg. Recall=0.9386

- While Logistic regression for Bernoulli Model gives-
  Avg. Accuracy=0.9583
  Avg. F1 Score=0.9614
  Avg. Precision=0.9563
  Avg. Recall=0.9665
- While SGD classifier for Bernoulli Model gives-
  Avg. Accuracy=0.9545
  Avg. F1 score=0.9458
  Avg. Precision=0.9709
  Avg. Recall=0.9255

4. Does your LR implementation outperform the SGDClassifier (again performance is measured in terms of the accuracy, precision, recall and F1 score) or is the difference in performance minor? Explain your yes/no answer.

The difference in performance between the two is minor.

For Bernoulli Model-

- Logistic regression for Bernoulli Model gives-
  Avg. Accuracy=0.9583
  Avg. F1 Score=0.9614
  Avg. Precision=0.9563
  Avg. Recall=0.9665
- While SGD classifier for Bernoulli Model gives-
  Avg. Accuracy=0.9545
  Avg. F1 score=0.9458
  Avg. Precision=0.9709
  Avg. Recall=0.9255

For Bag of Words Model-

- While Logistic regression for Bag of Words Model gives-
  Avg. Accuracy=0.9374
  Avg. F1 Score=0.9427
  Avg. Precision=0.9452
  Avg. Recall=0.9402
- While SGD classifier for Bag of words Model gives-
  Avg. Accuracy=0.9428
  Avg. F1 score=0.9454
  Avg. Precision=0.9426
  Avg. Recall=0.9483

As seen in both cases there is not much difference in the score values.