HW1 Machine Learning CS6375

Naive Bayes - Bag of words and Bernoulli representation

Datasets	Representation & Algorithm	Accuracy	Precision	Recall	F1 score
HW1	Discrete Naive Bayes & Bernoulli	89.12	71.91	98.46	83.11
ENRON 1	Discrete Naive Bayes & Bernoulli	93.2	83.15	99.33	90.52
ENRON 4	Discrete Naive Bayes & Bernoulli	78.45	77.07	99.74	86.96
HW1	Multinomial Naive Bayes & Bag of words	94.76	93.67	99.08	96.31
ENRON 1	Multinomial Naive Bayes & Bag of words	94.95	93.81	98.63	96.16
ENRON 4	Multinomial Naive Bayes & Bag of words	78.82	35.53	76.05	48.43

Logistic Regression - Bag of words and Bernoulli representation

Datasets	Representation & Algorithm	Accuracy	Precision	Recall	F1 score	Lambda	Learning Rate	Iterations
HW1	Logistic Regression & Bernoulli	96.44	97.41	97.69	97.55	0.9	0.01	1000
ENRON1	Logistic Regression & Bernoulli	96.27	97.39	97.07	97.23	0.3	0.01	1000
ENRON4	Logistic Regression & Bernoulli	87.47	73.02	80.43	76.55	0.9	0.01	1000

HW1 Machine Learning CS6375

Datasets	Representation & Algorithm	Accuracy	Precision	Recall	F1 score	Lambda	Learning Rate	Iterations
HW1	Logistic Regression & Bag of words	93.93	97.13	94.67	95.88	0.9	0.01	1000
ENRON1	Logistic Regression & Bag of words	96.27	96.09	98.33	97.2	0.3	0.01	1000
ENRON4	Logistic Regression & Bag of words	86.0	67.11	79.68	72.85	0.3	0.01	1000

The limit set for the number of iterations is 1000 as the increase in number of iterations above this decrease the gradient. The difference in weights is also so small. So at this learning rate, weights are almost close to converging points.

The learning parameter is set to 0.01 and the number of iterations are 1000.

Lambda is set to values from 0.1 to 0.9 with a difference of 0.2 . Fixed the number of iterations and learning parameter and tested on the validation test data. The value which gives the highest accuracy is set as the regularization value.

SGDClassifier - Bag of words and Bernoulli representation

Datasets	Representation & Algorithm	Accuracy	Precision	Recall	F1 score
HW1	SGDClassifier& Bernoulli	95.60	93.96	100	96.88
ENRON1	SGDClassifier& Bernoulli	93.85	91.21	99.64	95.23
ENRON4	SGDClassifier& Bernoulli	86.0	52.63	95.23	67.79
HW1	SGDClassifier & Bag of words	88.91	85.05	99.66	91.78
ENRON1	SGDClassifier & Bag of words	90.78	86.3	100	92.65
ENRON4	SGDClassifier & Bag of words	80.84	32.23	98.0	48.51

HW1 Machine Learning CS6375

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

SGDClassifier with Bernoulli representation performs better in terms of the accuracy, precision, recall and F1 score when compared with the other algorithms. At some frequencies, the weights are similar to the SGDC Bag of words model. So there is no much variation between SGDC Bernoulli and Bag of words model. The logistic regression performs better in some cases. SGDC and Logistic regression performed better with the given datasets.

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.

In LR and SGDC, some weights might become zero and this results in the elimination of feature, whereas in multinomial naive bayes features are not eliminated. So, sometimes Multinomial Naive Bayes perform better. The assumptions made in Naive Bayes are sometimes leading to better F1 scores in multinomial naive bayes than LR and SGDC.

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.

In most cases of Bernoulli representation, LR and SGDC perform better than discrete naive bayes in terms of accuracy, precision, recall and F1 score. Although the representation is same in all the three cases, discrete naive bayes considers non-occurrences which might decrease the probability, but in LR and SGDC we don't consider.

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.

LR and SGDC almost perform the same in all the cases. In some cases LR outperformed SGDC might be due to efficient covergence.