CS678 Assignment 2

Pranav Maneriker, 12497

October 26, 2015

1   Info

For question 1, subsets of the datasets are chosen:

1. Covtypes : 100 pts train, 300 pts test

2. rcv1: 40 pts train, 120 pts test

   For SMO, the implementation follows the pseudocode given in Platt's paper

2   Liblinear

The training times as well as the testing time are of the orders of a few seconds, even with the full datasets. Detailed parameter tuning tables and related data can be found in the ipython notebook `liblinear_tests.ipynb`

   Note: The lower accuracy on rcv1 with scaling as compared with without is probably because when scaling the test data, many columns were not available in the original data hence the scaling was imperfect.

Dataset	Train time(sec)	Stopping Criteria	Iterations	Accuracy
leukemia	(2.37, 2.37, 2.37)	(1, 2*, 3*)	(14, 16, 13)	(73.52, 73.52, 73.52) %
leukemia, scaled	2.53	1	13	64.7%
leukemia, scaled	3.14	2	14	64.8%
leukemia, scaled	3.56	3*	16	64.7%
rcv1	2.91	1	13	83.19%
rcv1	3.38	2*	14	84.03%
rcv1-scaled	(2.99, 3.76, 3.61)	(1, 2*, 3*)	(14, 15, 13)	(86.55, 86.55, 86.55)
covtype	9.04	1	18	44.81 %
covtype	12.58	2*	24	57.19 %
covtype	12.01	3*	18	60.86%

Table 1: Chunking

Dataset	Train time(sec)	Stopping Criteria	Iterations	Accuracy
leukemia	15.40	1	38	82.35 %
leukemia, scaled	25.81	1	13	88.23%
rcv1	10.19	1	23	85.71%
rcv1-scaled	(2.99, 3.76, 3.61)	1	23	86.55
covtype	90.56	1	594	66.88 %

Table 2: SMO

Dataset	C	Reported Accuracy
leukemia	0.25	79.41% (27/34)
leukemia with scaling	0.5	88.23% (30/34)
rcv1	1	96.15% (651374/677399)
rcv1 with scaling	0.0625	96.05% (650654/677399)
covtype (already scaled)	0.015625	63.36% (73632/116202)

Table 3: Liblinear results

3 LibSVM

1. The training times reported are with 5 fold cross validation.

2. Apart from these, subsets of the data(as indicated in the ratios in the tables) are used since each run (with one set of parameters) on rcv and covtypes takes 28 mins

3. Training set size: rcv: 300, covtypes: 1000

4. Only the c parameter is reported in these tables, the full data is available in `libsvm_tests.ipynb`

Further details about parameters can be found in `libsvm_tests.ipynb`

Dataset	Final C	Train time(sec)	Reported Accuracy
leukemia	0.25	1.10	82.35% (28/34)
leukemia with scaling	0.25	1.19	88.23% (30/34)
rcv1	1	3.14	77.56% (930/1199)
rcv1 with scaling	0.625	3.04	77.64% (931/1199)
covtype (already scaled)	1	1.30	19.7991% (690/3485)

Table 4: LibSVM linear kernel

Dataset	Final C	Train time(sec)	Reported Accuracy
leukemia	0.25	16.88	70.58% (24/34)
leukemia with scaling	0.25	17.03	88.23% (30/34)
rcv1	1	41.51	79.89% (958/1199)
rcv1 with scaling	0.625	45.22	85.40% (1024/1199)
covtype (already scaled)	1	290.08	23.64% (824/3485)

Table 5: LibSVM polynomial kernel degree 2

Dataset	Final C	Train time(sec)	Reported Accuracy
leukemia	4	16.88	61.76% (21/34)
leukemia with scaling	4	17.03	91.17% (31/34)
rcv1	1	41.51	56.71% (680/1199)
rcv1 with scaling	4	45.22	81.65% (979/1199)
covtype (already scaled)	1	290.08	29.84% (895/2999)

Table 6: LibSVM polynomial kernel degree 3

Dataset	Final C	Train time(sec)	Reported Accuracy
leukemia	4	7.79	58.82% (20/34)
leukemia with scaling	4	9.77	58.82% (20/34)
rcv1	16	22.40	79.23% (950/1199)
rcv1 with scaling	16	22.94	82.23% (986/1199)
covtype (already scaled)	16	11.40	31.54% (946/2999)

Table 7: LibSVM RBF Kernel