Preprocessing with SVM

Ryan Del Bel

February 22, 2014

1 Simulations with uncorrelated features

We first consider the simple case with uncorrelated features. We let n=500, simulate 10 truly associated features and 190 noisy features. The effect of the features on the outcome was achieved by setting the odds ratio in the underlying logistic model. For simplicity and efficiency when using the linear SVM we simply set the tuning parameter C to be one. Since the number of support vectors tends to be very large we only remove the support vectors that have maximal weights (i.e. with absolute value 1). For each feature we obtain the p-value from the t-test on the full sample and the sample with the support vectors with maximal weight removed. We then calculate the AUC, number of truly associated features that are in the top 10, 20 smallest p-values, and the number of support vectors removed. We replicate these simulations 100 times and report the average of each of the above statistics.

We first simulate data with the 10 associated features having the same OR.

OR	**AUC**		**TOP10**		**TOP20**		
	SVM	Normal	SVM	Normal	SVM	Normal	#SV
1	0.482	0.486	0.44	0.43	0.84	0.87	118.85
1.1	0.652	0.656	1.59	1.60	2.59	2.52	108.66
1.2	0.823	0.835	3.89	4.13	5.27	5.53	95.02
1.3	0.923	0.930	5.72	6.00	7.39	7.65	80.33
1.4	0.968	0.971	7.18	7.48	8.79	8.96	59.69
1.5	0.982	0.985	7.93	8.10	9.32	9.40	46.33
1.6	0.993	0.994	8.43	8.56	9.70	9.75	33.74
1.7	0.994	0.995	8.89	8.99	9.75	9.77	24.86
1.8	0.998	0.998	9.27	9.34	9.91	9.92	17.33
1.9	0.998	0.998	9.30	9.37	9.93	9.94	13.25
2	0.998	0.999	9.45	9.49	9.92	9.94	8.47

Next we simulate 10 associated features with OR evenly spaced between a specified minimum and maximum OR.

OR2	**AUC**		**TOP10**		**TOP20**		
	SVM	Normal	SVM	Normal	SVM	Normal	#SV
1.1-1.5	0.863	0.872	5.43	5.63	6.66	6.76	75.16
1.1-2	0.900	0.903	6.83	6.85	7.63	7.63	36.59
1.5-2	0.991	0.991	8.78	8.74	9.61	9.72	17.94
2-2.5	0.999	0.999	9.53	9.56	9.92	9.95	2.81

In both cases the results are slightly worse when using SVM as a preprocessing step. We notice that even when only removing the support vectors with maximal weight, we still usually remove a very large number of support vectors. We will now limit the number of support vectors we remove to k. If there are more than k support vectors with maximal weight we then randomly remove $k/2$ support vectors with weight 1 and $k/2$ support vectors with weight -1.

	AUC		TOP10		TOP20		
OR	SVM	Normal	SVM	Normal	SVM	Normal	#SV
1	0.502	0.502	0.59	0.62	1.13	1.16	4.00
1.1	0.624	0.624	1.64	1.63	2.57	2.58	4.00
1.2	0.821	0.823	3.83	3.87	5.50	5.45	4.00
1.3	0.939	0.940	5.95	5.98	7.66	7.77	4.00
1.4	0.971	0.972	7.25	7.30	8.94	8.96	4.00
1.5	0.984	0.985	8.02	8.12	9.32	9.38	4.00
1.6	0.993	0.993	8.72	8.68	9.68	9.70	3.97
1.7	0.994	0.994	8.88	8.92	9.72	9.77	3.85
1.8	0.997	0.997	9.16	9.22	9.90	9.92	3.85
1.9	0.998	0.998	9.34	9.35	9.91	9.94	3.44
2	0.998	0.998	9.34	9.38	9.92	9.92	3.19
1.1-1.5	0.869	0.870	5.60	5.52	6.71	6.74	4.00
1.1-2	0.916	0.917	6.82	6.84	7.79	7.76	4.00
1.5-2	0.992	0.993	8.77	8.78	9.65	9.67	3.79
1	0.518	0.514	0.52	0.52	1.02	1.01	10.00
1.1	0.649	0.648	1.68	1.71	2.60	2.65	10.00
1.2	0.830	0.833	3.99	3.94	5.49	5.55	10.00
1.3	0.919	0.923	5.84	6.01	7.54	7.59	10.00
1.4	0.970	0.971	7.31	7.40	8.69	8.78	10.00
1.5	0.985	0.986	8.14	8.13	9.35	9.38	10.00
1.6	0.994	0.994	8.73	8.87	9.69	9.74	9.98
1.7	0.995	0.996	8.98	9.01	9.80	9.82	9.64
1.8	0.998	0.998	9.20	9.22	9.93	9.92	8.77
1.9	0.998	0.998	9.40	9.44	9.89	9.89	7.57
2	0.998	0.998	9.44	9.46	9.94	9.95	5.74
1.1-1.5	0.874	0.877	5.72	5.76	6.87	6.89	10.00
1.1-2	0.915	0.917	6.93	6.98	7.83	7.88	9.96
1.5-2	0.991	0.991	8.72	8.70	9.60	9.61	9.03
1	0.493	0.490	0.42	0.45	0.95	0.93	24.00
1.1	0.641	0.648	1.65	1.71	2.71	2.74	24.00
1.2	0.830	0.842	3.93	4.03	5.32	5.47	24.00
1.3	0.929	0.932	5.93	6.08	7.62	7.65	24.00
1.4	0.965	0.968	7.05	7.23	8.65	8.75	23.95
1.5	0.983	0.985	7.91	8.14	9.35	9.41	23.52
1.6	0.992	0.992	8.54	8.64	9.66	9.71	22.61
1.1-1.5	0.873	0.874	5.56	5.68	6.77	6.82	24.00
1	0.494	0.491	0.51	0.58	1.08	1.03	50.00
1.1	0.638	0.652	1.68	1.82	2.56	2.65	50.00
1.2	0.823	0.834	4.11	4.30	5.51	5.60	50.00
1.3	0.926	0.935	6.03	6.30	7.56	7.67	49.87
1.4	0.968	0.972	7.19	7.32	8.73	8.80	46.49
1.1-1.5	0.869	0.881	5.40	5.58	6.60	6.79	49.02

Even when limiting the number of support vectors we remove, the normal approach still performs better.