实验报告

一. 编写一个 R 脚本，分析 wmc，dit，noc，cbo，rfc 和 lcom 的缺陷预测能力。

通过 fivenum，mean，cor 以及 moments 包中的 skewness/kurtosis 函数得到如下统计信息：

	wmc	dit	noc	cbo	rfc	lcom
Min	0	1	0	0	0	0
1st Qu.	3	1	0	4	8	0
Median	6	2	0	8	19	3
3rd Qu.	12.5	4	0	18	41	22.5
Max	123	8	29	171	355	6589
Mean	11.44952	2.565698	0.608575	14.49793	30.16183	130.0816
Skewness	0.478202	-2.34313	4.332323	0.472056	0.014723	4.684377
Kurtosis	18.08622	2.700852	66.16812	19.41106	17.6192	70.13383
Pearson	0.378792	-0.00186	0.054916	0.223544	0.459294	0.307576
Spearman	0.314245	-0.02612	0.090944	0.217624	0.356342	0.259252

二. 编写一个 R 脚本，不进行特征选择，利用十种机器学习方法建立多变量的缺陷预测模型 (在这里预测的是有无缺陷，而非缺陷数目)，并利用十折交叉验证评价上述模型。

所选机器学习方法与所使用的函数见下表：

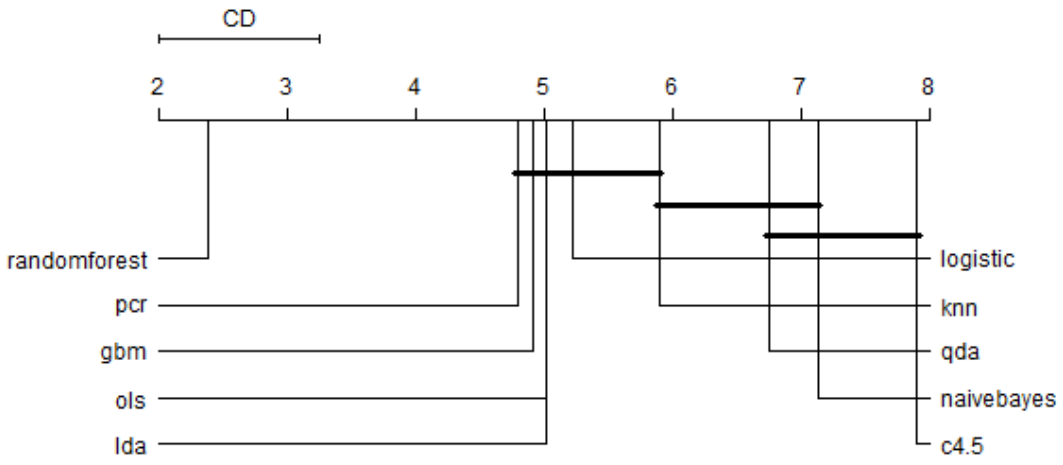
方法	函数
ols	lm()
logistic	lm()
lda	lda()
qda	qda()
knn	knn()
c4.5	tree()
randomforest	randomForest()
gbm	gbm()

pcr	pcr()
naivebayes	NaiveBayes()

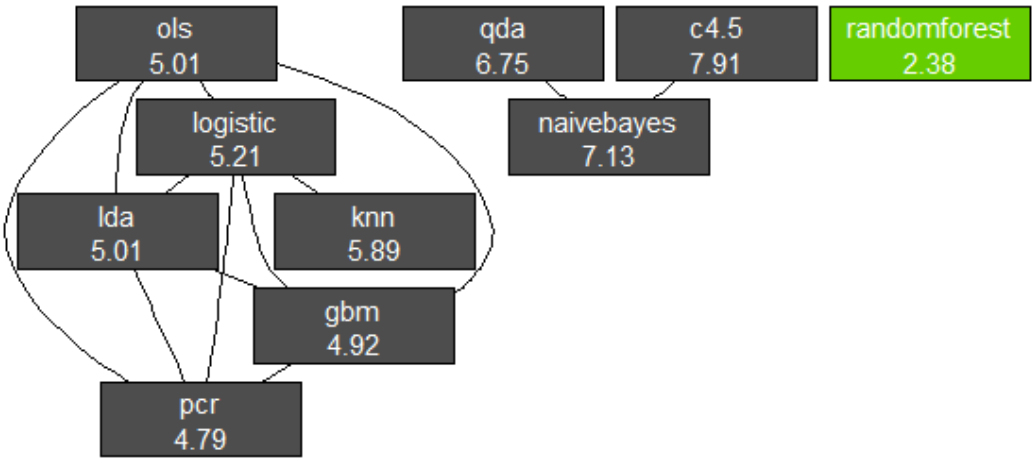
十折交叉验证所得 auc 矩阵（100×10，即 10 中方法各训练 10*10 次）如下所示：

	ols	logistic	lda	qda	knn	c4.5	randomfor	gbm	pcr	naivebayes
1	0.877419	0.904839	0.877419	0.897581	0.807258	0.818548	0.875806	0.833871	0.862903	0.868548
2	0.671182	0.69335	0.671182	0.592365	0.700123	0.522783	0.706897	0.687192	0.641626	0.642857
3	0.866319	0.866319	0.866319	0.770833	0.876736	0.651042	0.857639	0.904514	0.847222	0.829861
4	0.668056	0.676389	0.668056	0.676389	0.659028	0.704167	0.820833	0.748611	0.697222	0.601389
5	0.796791	0.773262	0.796791	0.809626	0.740642	0.782353	0.826738	0.797861	0.826738	0.787166
6	0.747801	0.728739	0.747801	0.68695	0.627566	0.54912	0.730205	0.683284	0.766862	0.733871
7	0.663139	0.694885	0.663139	0.632275	0.872134	0.832451	0.846561	0.814815	0.758377	0.749559
8	0.855379	0.857143	0.855379	0.82716	0.707231	0.805115	0.830688	0.747795	0.84127	0.783951
9	0.698413	0.66843	0.698413	0.640212	0.704586	0.671076	0.751323	0.733686	0.673721	0.626102
10	0.920635	0.895238	0.920635	0.849206	0.919841	0.780952	0.957143	0.930159	0.952381	0.88254
11	0.717813	0.767196	0.717813	0.829806	0.591711	0.739859	0.844797	0.740741	0.694885	0.795414
12	0.748276	0.75977	0.748276	0.745977	0.683333	0.545977	0.77069	0.786207	0.724138	0.68908
13	0.770492	0.786885	0.770492	0.733234	0.754098	0.770492	0.794337	0.780924	0.771982	0.676602
14	0.846491	0.809211	0.846491	0.726974	0.816338	0.743421	0.882675	0.773026	0.83443	0.783991
15	0.853846	0.855128	0.853846	0.839744	0.869872	0.778205	0.901282	0.825641	0.878205	0.853846
16	0.703125	0.705078	0.703125	0.65918	0.738281	0.616211	0.785156	0.828125	0.705078	0.691406
17	0.731922	0.716049	0.731922	0.564374	0.685185	0.693122	0.804233	0.767196	0.756614	0.641093
18	0.770149	0.81194	0.770149	0.78806	0.898507	0.462687	0.758209	0.629851	0.814925	0.740299
19	0.83867	0.828818	0.83867	0.799261	0.716133	0.628079	0.828818	0.849754	0.847291	0.791256
20	0.741935	0.722581	0.741935	0.822581	0.754839	0.766935	0.859677	0.76129	0.785484	0.798387
21	0.888889	0.866667	0.888889	0.758333	0.781944	0.743056	0.884722	0.875	0.851389	0.765972
22	0.783871	0.762903	0.783871	0.799194	0.629032	0.751613	0.804839	0.859677	0.762903	0.737097
23	0.75	0.711111	0.75	0.834722	0.749306	0.534028	0.834722	0.804167	0.770833	0.779167

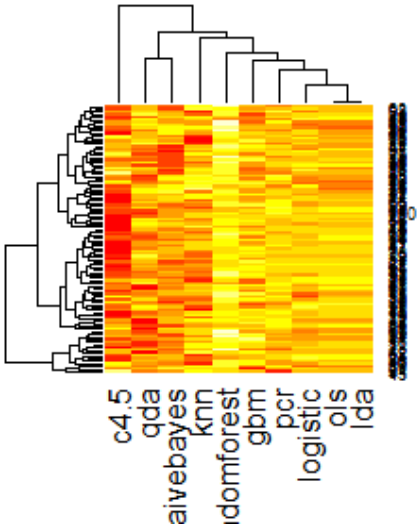
CD 图



算法图


ols
5.01
qda
6.75
c4.5
7.91
randomforest
2.38
logistic
5.21
naivebayes
7.13
lda
5.01
knn
5.89
gbm
4.92
pcr
4.79

热力图


c4.5
qda
aivebayes
knn
idomforest
gbm
pcr
logistic
ols
lda