Московский государственный университет имени М.В.Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра Методов математического прогнозирования

Пакет lasso2: L1 Constrained Estimation Routines

Обзор пакета системы R

Куракин Александр, 317-я группа

1. Введение

В данном отчете рассмотривается пакет `lasso2` системы `GNU R`[1]. Данный пакет[2] предоставляет процедуры, стастические данные и документацию для решения задач регрессии в случае использования L1-ограничений оценки. Основа алгоритма предложена Тибширани (Tibshirani, 1996) и доработана Озборном (Osborne, 1998)[3].

2. Описание алгоритма

Рассмотрим регрессию с целевой функцией $y = a(x)$ и решающую функцию положим как линейную комбинацию признаков: $\hat{y} = \sum_{j=1}^{m} b_j x_j^i$, где $(x^i, y^i), i = \overline{1,n}$ — обучающая выборка. Оптимизируя коэффициенты $B = (b_1, \ldots, b_m)$ по разности квадратов:

$$\sum_{i=1}^{n} \left(y^i - \sum_{j=1}^{m} b_j x_i^j \right)^2 \to \min_B,$$

при достаточно больших m или высокой коррелированности выборки можно столкнуться с разбросом оценки наименьшими квадратами.

Суть предлагаемого метода состоит в добавлении к функционалу оценки дополнительного условия в виде

$$\sum_{j=1}^{m} |b_j| \le \kappa, \quad \kappa > 0.$$

Тибширани в 1996 году назвал этот метод "lasso least absolute shrinkage and selection operator (LASSO)". Детальнее, Тибширани предложил ???. Тишбирани показал, что данная модель имеет определенные свойства, в частности, устремляется к нулю разность квадратов, мало того, иногда она равна нулю.

Пусть $x = (g_1, \ldots, g_p, h_1, \ldots, h_q), p + q = m$, и пусть G и H — соответствующие матрицы признаков, предлагается рассматривать технику LASSO применительно только к H-признакам. Обозначим матрицу зависимых переменных (коэффициентов) для них через соответственно $\mathcal{B} = (\beta_1, \ldots, \beta_p)$ и $\mathcal{Y} = (\gamma_1, \ldots, \gamma_q)$. Пусть $W \in \text{diag}[m \times n]$. Основная функция пакета `l1ce` работает по следующему алгоритму:

1. Вычисляем $\hat{\mathcal{B}} = (G^T W G)^{-1} G^T W B$ и проецируем G и H ортгонально пространству столбцов $W^{1/2}X$:

$$Z^* = \{E - W^{1/2} X (G^T W G)^{-1} G^T W^{1/2}\} W^{1/2} Z,$$
$$B^* = \{E - W^{1/2} X (G^T W G)^{-1} G^T W^{1/2}\} W^{1/2} B.$$

[1] R is a language and environment for statistical computing and graphics, `http://www.r-project.org/`.

[2] Package `lasso2` provides routines and documentation for solving regression problems while imposing an L1 constraint on the estimates, based on the algorithm of Osborne et al. (1998), `http://cran.r-project.org/web/packages/lasso2/`

[3] См. список литературы к `http://cran.r-project.org/web/packages/lasso2/vignettes/Manual-vignette.pdf`.

2. Опционально, столбцы Z^* приводятся к виду с отклонением 1.

3. Решается задача минимизации

$$(B^* - Z^*\mathcal{Y})^T(B^* - Z^*\mathcal{Y}) \to \min_{\mathcal{Y}}, \quad |\mathcal{Y}| \leq \kappa.$$

4. ???

3. Спецификация

3.1. Класс `l1ce`

Основной объект библиотеки — класс, хранящий модель задачи — класс `l1ce`. Основные его компоненты:

`coefficients`	Итоговые коэффициенты. Их имена-индексы совпадают с именами-индексами столбцов модели	
`residuals`	Остаточные векторы ???; веса не применяются	
`fitted.values`	Оптимизированные значения модели; веса не применяются	
`bound`	Ограничитель	
`xtx`	Матрица зависимых параметров; после применения весов и преобразований	
`xtr`	Матрица `xtx` и остаточные векторы	
`constrained.coefficients`	Коэффициенты в начальной шкале	
`sweep.out`	Формула, задающая переменные $\mathcal{B}$	

3.2. Функция `l1ce`

Основная функция пакета, работающая по описанному выше алгоритму и возвращает объект типа `l1ce`. Параметры функции:

`formula`	???	
`data`	Объект класса `data.frame`, в котором хранятся переменные выборки. Имена этих переменных можно использовать в `formula`, `weights`, `subset` и `sweep.out`	
`weights`	Вектор весов	
`subset`	Вектор-индикатор, указывающий, какие перенные использовать в модели, а какие — нет	
`na.action`	Функция, применяемяя для фильтрации `NA`-значений в процессе работы алгоритма	
`sweep.out`	Формула, задающая переменные $\mathcal{B}$	
`standardize`	Флаг: применять ли шаг 2 алгоритма?	
`trace`	Флаг: выводить отчет о работе алгоритма?	
`coefficients`	Начальное значение вектора весов	
`bound`	Ограничитель	

3.3. Статистическая информация

`data(Iowa)`	Данные о погоде и урожае пшеницы в штате Iowa, USA, в 1930-1962гг.
`data(Prostate)`	Данные, полученные при изучении корреляции антигена простаты и другими клиническими показателями мужчин с показаниями к радикальной простатэктомии

3.4. Другие возможности пакета

`gl1ce`	Обобщенные модели LASSO — generalized lasso least absolute shrinkage and selection operator
`l1celist`	Совокупности моделей `l1ce`

4. Пример использования

Приведем пример на данных, которые предоставляются самим пакетом.

4.1. Prostate Cancer Data

Данные, полученные при изучении корреляции антигена простаты и другими клиническими показателями мужчин с показаниями к радикальной простатэктомии. Имеется выборка длины 97 со следующими данными[4]:

`lcavol`	log(cancer volume)
`lweight`	log(prostate weight)
`age`	age
`lbph`	log(benign prostatic hyperplasia amount)
`svi`	seminal vesicle invasion
`lcp`	log(capsular penetration)
`gleason`	Gleason score
`pgg45`	percentage Gleason scores 4 or 5
`lpsa`	log(prostate specific antigen)

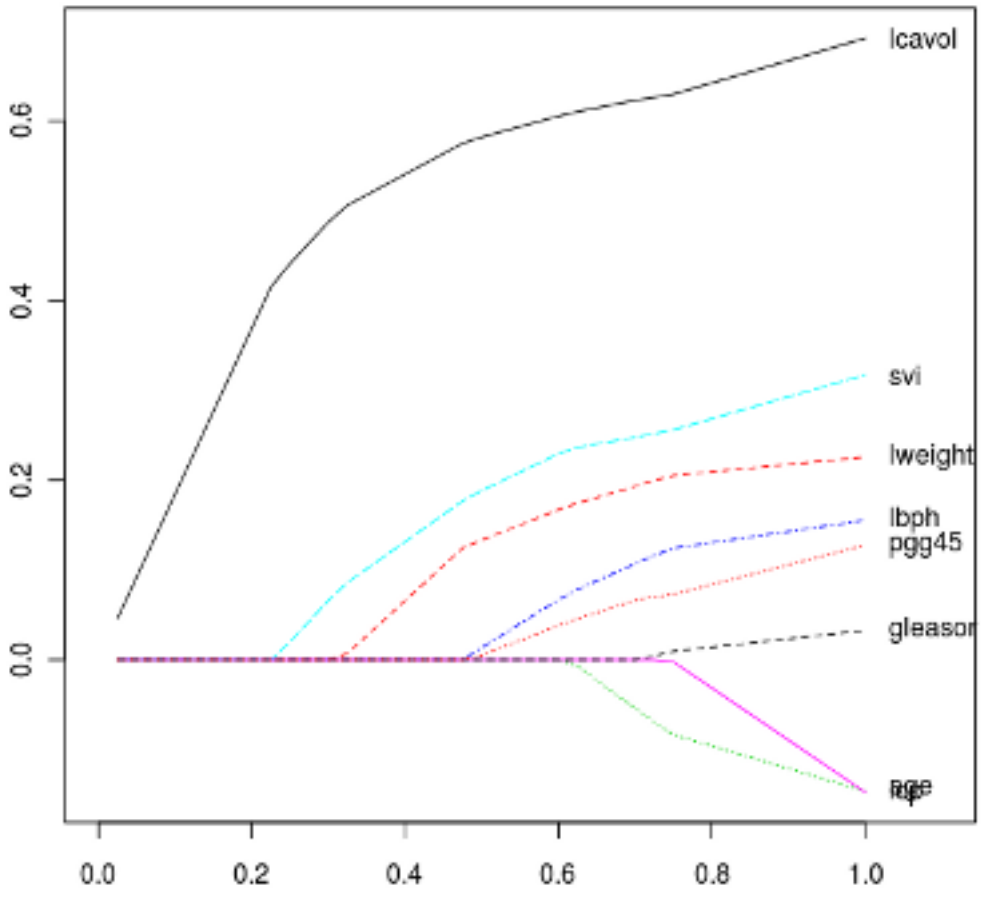
Тогда для получения коэффициентов модели LASSO запустим:

```
#            'lasso2'
library("lasso2")
#
data(Prostate)
#
p.mean <- apply(Prostate, 2, mean)
pros <- sweep(Prostate, 2, p.mean, "-")
p.std <- apply(pros, 2, var)
pros <- sweep(pros, 2, sqrt(p.std), "/")
pros[, "lpsa"] <- Prostate[, "lpsa"]
pros <- as.data.frame(pros)
#
res <- l1ce(lpsa ~ ., pros, bound=(1:100)/100)
#
plres <- plot(res, plot=F)
matplot(plres$bound[,"rel.bound"], plres$mat[,-1], type="l", xlim=c(0,1.1))
text(cbind(1.03, coef(res[40])[-1]), labels(res), adj=0)
```

[4]Приводится дословное описание из документации, во избежание искажений при переводе терминов, http://cran.r-project.org/web/packages/lasso2/lasso2.pdf

(bound)	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
0.01	0,018	0	0	0	0	0	0	0
0.02	0,037	0	0	0	0	0	0	0
0.03	0,055	0	0	0	0	0	0	0
0.04	0,074	0	0	0	0	0	0	0
0.05	0,092	0	0	0	0	0	0	0
0.06	0,111	0	0	0	0	0	0	0
0.07	0,129	0	0	0	0	0	0	0
0.08	0,148	0	0	0	0	0	0	0
0.09	0,166	0	0	0	0	0	0	0
0.10	0,184	0	0	0	0	0	0	0
0.11	0,203	0	0	0	0	0	0	0
0.12	0,221	0	0	0	0	0	0	0
0.13	0,240	0	0	0	0	0	0	0
0.14	0,258	0	0	0	0	0	0	0
0.15	0,277	0	0	0	0	0	0	0
0.16	0,295	0	0	0	0	0	0	0
0.17	0,313	0	0	0	0	0	0	0
0.18	0,332	0	0	0	0	0	0	0
0.19	0,350	0	0	0	0	0	0	0
0.20	0,369	0	0	0	0	0	0	0
0.21	0,387	0	0	0	0	0	0	0
0.22	0,406	0	0	0	0	0	0	0
0.23	0,423	0	0	0	0,001	0	0	0
0.24	0,432	0	0	0	0,011	0	0	0
0.25	0,441	0	0	0	0,020	0	0	0
0.26	0,450	0	0	0	0,029	0	0	0
0.27	0,460	0	0	0	0,038	0	0	0
0.28	0,469	0	0	0	0,048	0	0	0
0.29	0,478	0	0	0	0,057	0	0	0
0.30	0,487	0	0	0	0,066	0	0	0
0.31	0,496	0	0	0	0,075	0	0	0
0.32	0,504	0,003	0	0	0,083	0	0	0
0.33	0,508	0,011	0	0	0,089	0	0	0
0.34	0,513	0,019	0	0	0,095	0	0	0
0.35	0,517	0,027	0	0	0,101	0	0	0
0.36	0,522	0,035	0	0	0,107	0	0	0
0.37	0,527	0,042	0	0	0,113	0	0	0
0.38	0,531	0,050	0	0	0,119	0	0	0
0.39	0,536	0,058	0	0	0,125	0	0	0
0.40	0,540	0,066	0	0	0,131	0	0	0
0.41	0,545	0,074	0	0	0,137	0	0	0
0.42	0,550	0,081	0	0	0,144	0	0	0
0.43	0,554	0,089	0	0	0,150	0	0	0
0.44	0,559	0,097	0	0	0,156	0	0	0
0.45	0,563	0,105	0	0	0,162	0	0	0
0.46	0,568	0,113	0	0	0,168	0	0	0
0.47	0,573	0,120	0	0	0,174	0	0	0
0.48	0,577	0,126	0	0,002	0,180	0	0	0
0.49	0,580	0,130	0	0,008	0,184	0	0	0,002
0.50	0,582	0,133	0	0,013	0,188	0	0	0,005

(bound)	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
0.51	0,584	0,137	0	0,019	0,192	0	0	0,008
0.52	0,586	0,140	0	0,024	0,197	0	0	0,012
0.53	0,589	0,143	0	0,029	0,201	0	0	0,015
0.54	0,591	0,147	0	0,035	0,205	0	0	0,018
0.55	0,593	0,150	0	0,040	0,209	0	0	0,022
0.56	0,596	0,154	0	0,045	0,213	0	0	0,025
0.57	0,598	0,157	0	0,051	0,217	0	0	0,028
0.58	0,600	0,160	0	0,056	0,221	0	0	0,031
0.59	0,603	0,164	0	0,061	0,225	0	0	0,035
0.60	0,605	0,167	0	0,067	0,229	0	0	0,038
0.61	0,607	0,170	0	0,072	0,233	0	0	0,041
0.62	0,609	0,173	−0,004	0,076	0,236	0	0	0,044
0.63	0,611	0,176	−0,010	0,080	0,237	0	0	0,047
0.64	0,613	0,178	−0,017	0,084	0,239	0	0	0,050
0.65	0,614	0,181	−0,023	0,088	0,240	0	0	0,053
0.66	0,616	0,183	−0,029	0,092	0,242	0	0	0,055
0.67	0,618	0,186	−0,035	0,096	0,243	0	0	0,058
0.68	0,619	0,188	−0,042	0,099	0,245	0	0	0,061
0.69	0,621	0,191	−0,048	0,103	0,246	0	0	0,063
0.70	0,623	0,193	−0,054	0,107	0,248	0	0	0,066
0.71	0,624	0,196	−0,060	0,111	0,249	0	0,001	0,068
0.72	0,626	0,198	−0,066	0,114	0,251	0	0,003	0,069
0.73	0,627	0,201	−0,073	0,118	0,252	0	0,005	0,070
0.74	0,628	0,203	−0,079	0,121	0,254	0	0,008	0,071
0.75	0,630	0,205	−0,083	0,124	0,256	−0,002	0,009	0,073
0.76	0,632	0,206	−0,086	0,125	0,258	−0,008	0,010	0,075
0.77	0,635	0,207	−0,088	0,127	0,261	−0,014	0,011	0,077
0.78	0,637	0,208	−0,091	0,128	0,263	−0,020	0,012	0,079
0.79	0,640	0,209	−0,093	0,129	0,266	−0,026	0,013	0,082
0.80	0,642	0,209	−0,096	0,130	0,268	−0,031	0,014	0,084
0.81	0,645	0,210	−0,098	0,132	0,271	−0,037	0,015	0,086
0.82	0,647	0,211	−0,101	0,133	0,273	−0,043	0,016	0,088
0.83	0,650	0,212	−0,103	0,134	0,276	−0,049	0,017	0,090
0.84	0,652	0,213	−0,106	0,135	0,278	−0,055	0,018	0,093
0.85	0,655	0,213	−0,108	0,137	0,280	−0,060	0,019	0,095
0.86	0,657	0,214	−0,111	0,138	0,283	−0,066	0,020	0,097
0.87	0,660	0,215	−0,114	0,139	0,285	−0,072	0,021	0,099
0.88	0,662	0,216	−0,116	0,140	0,288	−0,078	0,021	0,101
0.89	0,664	0,217	−0,119	0,142	0,290	−0,084	0,022	0,104
0.90	0,667	0,218	−0,121	0,143	0,293	−0,089	0,023	0,106
0.91	0,669	0,218	−0,124	0,144	0,295	−0,095	0,024	0,108
0.92	0,672	0,219	−0,126	0,145	0,298	−0,101	0,025	0,110
0.93	0,674	0,220	−0,129	0,147	0,300	−0,107	0,026	0,112
0.94	0,677	0,221	−0,131	0,148	0,303	−0,113	0,027	0,114
0.95	0,679	0,222	−0,134	0,149	0,305	−0,118	0,028	0,117
0.96	0,682	0,222	−0,136	0,150	0,307	−0,124	0,029	0,119
0.97	0,684	0,223	−0,139	0,152	0,310	−0,130	0,030	0,121
0.98	0,687	0,224	−0,141	0,153	0,312	−0,136	0,031	0,123
0.99	0,689	0,225	−0,144	0,154	0,315	−0,142	0,032	0,125
1.00	0,692	0,226	−0,146	0,155	0,317	−0,147	0,033	0,128



5. Литература

1. The R Project for Statistical Computing,
   `http://r-project.org`.
2. CRAN - Package `lasso2`: L1 constrained estimation aka 'lasso',
   `http://cran.r-project.org/web/packages/lasso2`
3. School of Mathematics and Statistics: `lasso2` (L1 Constrained Estimation Routines),
   `http://school.maths.uwa.edu.au/ berwin/software/lasso.html`.
4. Manual of `lasso2` package,
   `http://cran.r-project.org/web/packages/lasso2/vignettes/Manual-vignette.pdf`.
5. Reference manual of `lasso2` package,
   `http://cran.r-project.org/web/packages/lasso2/lasso2.pdf`.
6. Источники и литература предыдущих трех пунктов.
7. Воронцов К. В., Лекции по статистическим (байесовским) алгоритмам классификации, часть 1, 2009, стр. 84-89,
   `http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf`.
8. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных
   `http://www.machinelearning.ru/`.