

# Пошаговая дискриминация, кросс-валидация и бутстрап в задаче классификации пострадавших с сочетанной травмой груди.

Феофанов Василий

Санкт-Петербургский государственный университет,  
факультет прикладной математики - процессов управления,  
кафедра МТИСР

5 апреля, 2016

- 1 Постановка задачи
- 2 Линейный дискриминатный анализ
- 3 Пошаговая дискриминация
- 4 Оценка вероятности ошибочной классификации
- 5 Классификация пострадавших с сочетанной травмой груди
- 6 Выводы
- 7 Список литературы

В исследовании рассматриваются 52 пациента с сочетанной травмой грудной клетки, госпитализированные в экстренном порядке. База содержит общую информацию о пострадавших, результаты лабораторных и инструментальных исследований, проведенных в течение первых 12 часов с момента поступления в больницу пострадавшего, а также исход получения травмы. По имеющимся 160 признакам, описывающих каждого пациента, необходимо построить классификационное правило, которое позволит спрогнозировать летальность исхода для будущих пациентов с сочетанной травмой груди.

# Фрагмент базы данных

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z			
2	Общие данные										Объективный статус при поступлении														Шкалы				
3	№	Возраст	Пол	Датность травмы, минуты	Сроки госпитализации, минуты	Причина травмы	Алкогольное опьянение	Исход	Ковыльный порок	Характер ранения/длина	Аускультативные изменения в легких	Речевой контакт	Реакция на боль	Зрачковый или роговичный рефлекс	Валикулярный транкус	Характер пульса	Частота пульса	САД	Ориентировочная величина кровяного давления	Шумы выслушиваемой перистальтики	ВРХ - СП	Сознание (шкала Глазго)	GSC - M	ВРХ - П (МТ)	ISS	ВРХ - СТ			
4																										0 - 124			
5	1	45	1	120	90	1	1	1	1	3	3	6	6	6	6	6	1	4	1	1	32	3	1	15.9	44	35			
6	2	43	1	90	60	4	1	2	1	8	5	6	6	6	6	1	4	1	1	3	48	5	3	5.3	21	39			
7	3	56	1	60	45	1	1	1	1	1	3	6	1	1	2	1	3	3	3	1	25	7	4	34	43	29			
8	4	53	1	120	60	2	2	2	1	1	3	3	3	1	2	1	3	1	1	3	21	14	6	7.2	33	24			
9	5	28	1	120	50	4	1	2	1	1	1	1	1	1	2	1	1	3	1	1	15	15	6	4.65	18	27			
10	6	60	1	120	60	4	2	2	1	1	3	1	1	1	2	1	3	1	1	1	19	15	6	3.25	19	21			
11	7	35	1	80	60	4	2	2	1	1	3	6	6	1	2	1	1	1	1	3	27	3	1	1.6	17	23			
12	8	82	1	70	60	4	1	1	1	1	3	6	6	1	4	1	3	3	4	1	37	3	1	5.55	20				
13	9	80	1	45	30	1	1	1	1	1	3	6	6	6	8	2	1	1	3	1	36	3	1	19	34	41			
14	10	42	1	80	60	6	1	2	4	1	3	3	1	1	2	1	3	4	4	3	26	11	4	5	34	40			
15	11	21	1	45	30	2	1	2	1	1	3	1	1	1	2	1	4	1	3	1	20	14	6	3.6	25	29			
16	12	45	1	30	20	1	2	1	4	1	3	6	6	6	8	2	1	1	4	3	3	42	3	1	13.3	33	36		
17	13	27	1	40	30	4	2	1	4	1	3	6	6	6	8	2	1	3	1	3	41	7	5	11.9	38	38			
18	14	27	1	60	30	4	1	1	1	1	3	6	6	6	8	2	1	4	6	6	3	48	3	1	20.3	57	40		
19	15	34	1	80	80	1	1	1	2	4	5	1	1	1	2	1	1	1	4	1	23	15	6	2.6	19	34			
20	16	21	1	120	120	1	1	2	1	1	3	6	6	6	8	2	1	3	3	3	3	40	3	1	3.5	23	29		
21	17	57	1	70	60	4	1	2	2	1	3	1	1	1	2	1	3	4	4	3	26	15	6	3.3	29	34			
22	18	34	1	2100	30	4	2	1	4	1	3	1	1	1	2	1	3	3	9	3	32	15	6	7.2	33	44			
23	19	28	1	120	70	2	1	2	1	1	3	3	3	1	1	2	1	4	3	4	3	27	12	6	6	27	35		
24	20	25	1	60	40	2	2	2	4	1	3	3	3	1	1	2	1	3	1	1	22	10	4	4.2	17	26			
25	21	68	2	60	40	1	1	1	4	1	3	6	1	1	2	1	3	3	4	3	32	4	2	4	25	36			
26	22	49	1	30	60	6	1	2	4	1	1	1	1	1	1	1	18	15	6	4.25	17	1	1	34	33				
27	23	31	1			2	1	1	1	1	3	6	6	6	8	2	1	4	1	4	1	38	6	4	11.4	36			
28	24	54	1	60	40	2	1	2	4	1	3	6	1	8	2	1	3	1	1	1	32	5	3	14.2	43	38			
29	25	57	2	90	35	1	1	2	4	1	3	3	1	1	2	1	7	4	4	1	32	9	4	2.4	21	42			
30	26	68	1			2	1	2	1	5	3	1	1	1	1	1	3	1	3	1	22	15	6	2.2	17				
31	27	30	1	60	30	2	1	2	4	1	3	3	1	1	2	1	3	1	1	1	22	13	5	5	34	33			
32	28	51	2	60	30	4	2	2	4	1	1	1	1	1	2	1	1	1	1	1	26	15	6	6.4	24	33			

# Линейный дискриминатный анализ (ЛДА)

Один из классических методов классификации. Если  $\mu_1, \mu_2$  — математические ожидания классов  $W_1$  и  $W_2$  соответственно, а  $\Sigma$  — общая ковариационная матрица, тогда для произвольного наблюдения  $x$  классификационное правило будет выглядеть следующим образом:

$$W_1 : (\mu_1 - \mu_2)^T \Sigma^{-1} x \geq \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) + \ln \left( \frac{p_2}{p_1} \right)$$

$$W_2 : (\mu_1 - \mu_2)^T \Sigma^{-1} x < \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) + \ln \left( \frac{p_2}{p_1} \right),$$

где  $p_1, p_2$  — априорные вероятности принадлежности к классам  $W_1$  и  $W_2$ .

- наблюдений меньше, чем оцениваемых параметров.
- риск получения плохо обусловленной ковариационной матрицы
- ошибка на обучении:  $\overline{err} = \frac{1}{n} \sum_{i=1}^n |y_i - m(x_i)|$  не является корректной оценкой вероятности ошибочной классификации

# Пошаговая дискриминация

В основе пошагового дискриминатного анализа лежит тест на добавочную информацию, который определяет значимость вклада новых включенных переменных по отношению к старым при проверки гипотезы  $H_0 : \mu_1 = \mu_2$ . Для этого вычисляется частная лямбда Уилкса:

$$\Lambda(x|t) = \frac{\Lambda(t, x)}{\Lambda(t)},$$

где  $t$  — изначальный набор признаков,  $x$  — набор добавленных переменных, а  $\Lambda(z)$ :

$$\Lambda = \frac{|W|}{|W + B|},$$

где  $B$  и  $W$  меж- и внутри- групповые ковариационные матрицы.

# Алгоритм Forward Selection

1. Вначале из модели удаляются все рассматриваемые признаки, число которых  $p$ .
2. Задаемся значением  $\Lambda$ -включения.
3. Для каждого  $x_j$  высчитывается  $\Lambda(x_j)$  и затем включается в модель та переменная, значение соответствующей статистики которой наименьшее среди рассматриваемых. Включенный в модель признак обозначим за  $t_1$ .



4. Среди остальных  $p - 1$  переменных ищется признак с наименьшей частной лямбдой Вилкса:

$$\Lambda(x_j|t_1) = \frac{\Lambda(t_1, x_j)}{t_1},$$

Вместе с тем, статистика должна удовлетворять условию:  
 $\Lambda \leq \Lambda$ -включения.

5. Процесс продолжается аналогичным образом до тех пор, пока ни одна из переменных не будет удовлетворять условию или все переменные не войдут в модель.

## 1. Cross-validation leave-one-out:

$$\widehat{Err}^{(CV)} = \frac{1}{n} \sum_{i=1}^n |y^{(i)} - m^{-(i)}(x^{(i)})|$$

## 2. Bootstrap leave-one-out:

$$\widehat{Err}^{(LOOB)} = \frac{1}{n} \sum_{i=1}^n \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} |y^{(i)} - m^{(b)}(x^{(i)})|,$$

где  $C^{-i}$  — набор индексов, идентифицирующие те бустрап-выборки, которые не содержат объект  $i$ .

## 3. Bootstrap 0.632:

$$\widehat{Err}^{(0.632)} = 0.368 \cdot \overline{err} + 0.632 \cdot \widehat{Err}^{(LOOB)},$$

## 4. Bootstrap 0.632+:

$$\widehat{Err}^{(0.632+)} = (1 - \alpha) \cdot \overline{err} + \alpha \cdot \widehat{Err}^{(LOOB)},$$

где  $\alpha = \frac{0.632}{1 - 0.368\widehat{R}}$ , а  $\widehat{R}$  — относительная частота переобучения.

# Классификация пострадавших с сочетанной травмой груди. Этап I.

Все 160 признаков были разбиты на следующие группы:

Группа 1: Общие данные

Группа 2: Объективный статус при поступлении + шкалы

Группа 3: Шкалы

Группа 4: Структура повреждений внутренних органов

Группа 5: Параметры ИВЛ

Группа 6: ВСП + ВСАД + ВДАД

Группа 7: ВД + общие данные спироартериокардиоритмографа (САКР)

Группа 8: Электрокардиография (ЭКГ)

Группа 9: Общий анализ крови (ОАК)

Группа 10: Биохимия

Группа 11: Маркеры повреждения сердца + газы крови

# Результаты пошагового дискриминантного анализа

Группа	Признак 1	Λ	Признак 2	Λ	Признак 3	Λ	Признак 4	Λ	Признак 5	Λ
Группа 1	Возраст	0.6	Сроки госпитализации	0.73						
Группа 2	ВПХ - П (МТ)	0.59	Речевой контакт	0.55	Величина кровопотери	0.51	САД при поступлении	0.44	Частота пульса При поступлении	0.4
Группа 3	ВПХ - голова	0.76	ВПХ - грудь	0.63	ВПХ - таз	0.53	АIS - грудь	0.45	АIS - таз	0.41
Группа 4	Повреждение ЦНС САК	0.85								
Группа 5	ИВЛ АДД	0.92								
Группа 6	BCP LF (п. у.)	0.8	Вар. САД ТР	0.65	Вар. САД HF	0.57	Вар. САД LF (п. у.)	0.5	Вар. САД VLF	0.43
Группа 7	САКР PQ	0.84	САКР ЧССср	0.73	САКР АДсмакс	0.57	САКР АДдср	0.46	САКР Вар. Дых. HF п.	0.36
Группа 8	ЭКГ RR	0.92								
Группа 9	ОАК Гемоглобин	0.77								
Группа 10	Биохимия Натрий	0.63								
Группа 11	Газы крови FiO2	0.79	Газы крови RI	0.67						

# Оценка вероятности ошибочной классификации.

Группа	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+
Группа 1	0.289	0.333	0.399	0.358	0.367
Группа 2	0.118	0.157	0.177	0.155	0.157
Группа 3	0.178	0.222	0.242	0.219	0.221
Группа 4	0.28	0.28	0.305	0.296	0.296
Группа 5	0.306	0.347	0.345	0.331	0.332
Группа 6	0.114	<b>0.114</b>	<b>0.164</b>	<b>0.146</b>	<b>0.147</b>
Группа 7	<b>0.074</b>	0.185	0.222	0.168	0.179
Группа 8	0.304	0.326	0.375	0.349	0.353
Группа 9	0.362	0.426	0.385	0.376	0.377
Группа 10	0.267	0.267	0.291	0.282	0.282
Группа 11	0.25	0.296	0.309	0.287	0.289

# Классификация пострадавших с сочетанной травмой груди. Этап II.

Группа А: Отобранные признаки из групп 1-3

Группа В: Отобранные признаки из групп 4-6

Группа С: Отобранные признаки из групп 7-11

Группа F: Отобранные признаки из групп А-С

Группа	Признак 1	Λ	Признак 2	Λ	Признак 3	Λ	Признак 4	Λ	Признак 5	Λ	Признак 6	Λ	Признак 7	Λ
Группа А	ВРХ - П (MT)	0.61	Речевой контакт	0.53	Возраст	0.45	Величина кровопотери	0.39	САД при поступлении	0.33	Сроки госпитализации	0.3		
Группа В	ВСР L F (п. у.)	0.83	Вар. САД TP	0.67	Вар. САД HF	0.58	Вар. САД L F (п. у.)	0.51	ИВЛ АДД	0.42	Вар. САД VLF	0.38		
Группа С	Газы крови FIO2	0.85	Газы крови RI	0.77										
Группа F	ВРХ - П (MT)	0.63	Возраст	0.53	Вар. САД L F (п. у.)	0.44	Речевой контакт	0.4	Величина кровопотери	0.34	Сроки госпитализации	0.3	САД при поступлении	0.26

Группа	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+
Группа А	<b>0.044</b>	<b>0.044</b>	<b>0.099</b>	<b>0.079</b>	<b>0.08</b>
Группа В	0.116	0.163	0.161	0.144	0.145
Группа С	0.25	0.296	0.312	0.289	0.291
Группа F	0.075	0.1	0.124	0.106	0.107

В задаче прогнозирования летальности исхода постравдшего с сочетанной травмой груди стоит принимать во внимание такие параметры, как:

- Шкала военной-полевой хирургии П(МТ) (П–повреждение, МТ–механическая травма)
- Систолическое артериальное давление при поступлении в больницу
- Величина кровопотери и сроки госпитализации
- Вариабельность сердечного ритма (мощность низкочастотного компонента)
- Вариабельность систолического артериального давления
- Диастолическое артериальное давление
- Речевой контакт



- Наилучшая точность была достигнута на группе признаков A, но не на группе F. Это подтверждает, что пошаговый дискриминантный анализ не выдает гарантированно оптимальный набор признаков
- Несмотря на наличие большого числа признаков, превышающее количество наблюдений, удалось добиться высокой точности классификации (90.1–95.6%)
- Ошибка на обучении в среднем давала оптимистически заниженную оценку в сравнении с методами кросс-валидации и бутстрап
- 0.632 и 0.632+ показали примерно одинаковые результаты, что может говорить о том, что в исследовании наблюдался незначимый эффект переобучения

- [1] Буре В. М., Парилина Е. М., Рубша А. И., Свиркина Л. А. Анализ выживаемости по медицинской базе данных больных раком предстательной железы // Вестн. С.-Петерб. ун-та. Сер. 10: Прикладная математика, информатика и процессы управления. 2014. N. 2. С. 27–35.
- [2] Буре В. М., Щербакова А. А. Применение дискриминантного анализа и метода деревьев принятия решений для диагностики офтальмологических заболеваний // Вестн. С.-Петерб. ун-та. Сер. 10: Прикладная математика, информатика и процессы управления. 2013. N. 1. С. 70–76.

- [3] Rencher A. C. Methods of Multivariate Analysis. 2nd Ed. New York: John Wiley & Sons, Inc. 2002. P. 738.
- [4] Рао С. Р. Линейные статистические методы и их применения. / науч ред. Линник Ю. В.; пер. с англ. Калинина В. М. и др. М.: Наука, 1968. 548 с.
- [5] Fu W. J., Carroll R. J., Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. // Bioinformatics. 2005. Vol. 21(9). P. 1979–1986.
- [6] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed. New York: Springer-Verlag. 2009. P. 745.

- [7] Molinaro A. M., Simon R., Pfeiffer R. M. Prediction error estimation: a comparison of resampling methods. // Bioinformatics. 2005. Vol. 21(15). P. 3301–3307.
- [8] Zavorka S., Perrett J. J. Minimum sample size considerations for two-group linear and quadratic discriminant analysis with rare populations // Communications in Statistics - Simulation and Computation. 2014. Vol. 43(7). P. 1726–1739.