ЛЕКЦИЯ 11. Множественная регрессия

В простой линейной регрессии устанавливается линейная зависимость между одной входной переменной и одной выходной. Но на практике в задачах анализа часто возникает необходимость привлечь для решения больше исходной информации. Поэтому представляет интерес исследование зависимости выходной переменной от нескольких входных, Многие задачи Data Mining имеют достаточно большую размерность и содержат десятки и сотни переменных. Для установления таких связей служит множественная (иногда ее называют многомерной линейной регрессией). Как правило, она обеспечивает более высокую точность оценки, чем простая.

Если простая линейная регрессия использует приближение в виде прямой линии, то множественная линейная регрессия для моделирования линейных связей между выходной переменной и набором входных плоскость (если входных переменных две) или гиперплоскость (если входных переменных более, чем две).

Обычно входные переменные являются непрерывными, но в некоторых задачах могут содержаться и категориальные входные переменные. В этом случае применяется специальный подход, основанный на введении так называемых фиктивных переменных.

Фиктивные переменные – это такие переменные, которые принимают одно из двух значений — 0 или 1. Их также называют бинарными или дамми-переменными. Например, фиктивная переменная может быть, равна единице, если i-ый работник — женщина, и равна нулю, если мужчина.

Как и в случае простой линейной регрессии, в множественной требуется определить, можно ли распространить линейную зависимость между набором входных переменных и выходной переменной, построенную на основе выборочных наблюдений, на всю имеющуюся совокупность данных. Необходимо не только вычислить коэффициенты уравнения множественной регрессии, но и построить соответствующую регрессионную модель и определить ее значимость.

При построении такой модели можно использовать те же подходы, что и для простой линейной регрессионной модели. Множественная регрессионная модель представляет собой прямое расширение простой линейной регрессионной модели для заданного числа переменных:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad (11.1)$$

Для оценки параметров уравнения линейной множественной регрессии применяют метод наименьших квадратов – строится система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии:

$$\begin{cases} \sum y = na + b_1 \sum x_1 + b_2 \sum x_2 + \cdots + b_p \sum x_p \\ \sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + b_1 \sum x_2 x_1 + \cdots + b_p \sum x_p x_1 \\ \cdots \\ \sum yx_1 = a \sum x_p + b_1 \sum x_1 x_p + b_2 \sum x_2 x_p + \cdots + b_p \sum x_p^2 \end{cases} \quad (11.2)$$

Другой вид уравнения множественной регрессии – уравнение регрессии в стандартизированном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \cdots + \beta_p t_{x_p} \quad (11.3)$$

где $t_y = (y - \bar{y})/\sigma_y, t_x = (x_i - \bar{x}_i)/\sigma_{x_i}$ – стандартизированные переменные; β_i – стандартизированные коэффициенты регрессии.

К уравнению множественной регрессии в стандартизированном масштабе применим МНК (метод наименьших квадратов), что приводит к решению системы уравнений:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + \cdots + \beta_p r_{x_px_1} \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_3x_2} + \cdots + \beta_p r_{x_px_2} \\ \cdots \\ r_{yx_p} = \beta_1 r_{x_1x_p} + \beta_2 r_{x_2x_p} + \beta_3 r_{x_3x_p} + \cdots + \beta_p \end{cases} \quad (11.4)$$

Для двухфакторной модели линейной регрессии $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2}$расчет β-коэффициентов можно выполнить по формулам (следуют из решения системы (11.4)):

$$\beta_1 = (r_{yx_1} - r_{yx_2} r_{x_1x_2})/1 - r_{x_1x_2}^2, \beta_2 = (r_{yx_2} - r_{yx_1} r_{x_1x_2})/1 - r_{x_1x_2}^2 \quad (11.5)$$

Связь коэффициентов множественной регрессии b_i со стандартизированными коэффициентами β_i описывается соотношением:

$$b_i = \beta_i(\sigma_y/\sigma_{x_i}), \beta_i = b_i(\sigma_{x_i}/\sigma_y) \quad (11.6)$$

При этом: $a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$

Тесноту совместного влияния факторов на результат оценивает коэффициент множественной корреляции, который можно определить по формуле:

$$R_{yx_1x_2\ldots x_p} = \sqrt{\sum \beta_i r_{yx_i}} \quad (11.7)$$

где β_i – стандартизированные коэффициенты регрессии, r_{yx_i} – парные коэффициенты корреляции между переменными y и x_i. Качество построенной модели в целом оценивает коэффициент (индекс) детерминации. Коэффициент множественной детерминации рассчитывается как квадрат индекса множественной корреляции:

$$R_{yx_1x_2\ldots x_p}^2 \quad (11.8)$$

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния (при

закреплении их влияния на постоянном уровне) других факторов, включенных в уравнение регрессии. Для двухфакторной модели их можно определить по формулам:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}}; \; r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}};$$

$$r_{x_1x_2 \cdot y} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{yx_2}^2)}}. \quad (11.9)$$

При построении уравнения множественной регрессии может возникнуть проблема мультиколлениарности факторов (тесная линейная зависимость более двух факторов). Считается, что две переменные явно коллинеарны, если $r_{x_ix_j} > 0{,}7$. Статистическая значимость уравнения множественной регрессии в целом оценивается с помощью общего F-критерия Фишера:

$$F = \frac{R^2_{yx_1x_2\ldots x_p}}{1-R^2_{yx_1x_2\ldots x_p}} \times \frac{n-m-1}{m} \quad (11.10)$$

где m – число факторов в линейном уравнении регрессии; n – число наблюдений.

Вывод о статистической значимости уравнения множественной регрессии в целом и коэффициента множественной детерминации можно сделать, если наблюдаемое значение критерия больше табличного, найденного для заданного уровня значимости (например, $\alpha = 0{,}05$) и степенях свободы $k_1 = m\,, k_2 = m - n - 1$.

Частный F-критерий оценивает статистическую значимость присутствия каждого из факторов в уравнении множественной регрессии. Для двухфакторной модели F_{x_1} оценивает целесообразность включения в уравнение фактора x_1 после того, как в него был включен фактор x_2; F_{x_2} оценивает целесообразность включения в уравнение фактора x_2 после того, как в него был включен фактор x_2

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1-R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1}, \; F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1-R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1}, \quad (11.11)$$

где m – число факторов в линейном уравнении регрессии; n – число наблюдений.

Фактическое значение частного F-критерия сравнивается с табличным при 5%-ном или 1%-ном уровне значимости и числе степеней свободы: $k_1 = m\,, k_2 = m - \; n - 1$. Если фактическое значение превышает табличное, то дополнительное

включение соответствующего фактора в модель статистически оправдано, в противном случае фактор в модель включать нецелесообразно.

Решение примера

Имеются данные о стоимости автомобилей ВАЗ 2110 (результативная переменная y, тыс. руб.) в Краснодарском крае, о годе выпуска (возраст автомобиля – фактор x_1, лет) и о пробеге (фактор x_2, тыс. км) (табл. 11.1):

Таблица 11.1. Данные

№	Возраст, лет	Пробег, тыс. км	Цена, тыс. руб.
1	5	50	167
2	5	70	175
3	8	110	146
4	8	120	120 143
5	10	175	120
6	4	62	220
7	5	87,5	150
8	5	84	172
9	7	77	170
10	4	83	190
11	4	65	210
12	8	120	143
13	6	88	167
14	7	89	150
15	4	83	195

Требуется:

1) Найти уравнение линейной множественной регрессии в стандартизированной ($t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2}$) и естественной форме ($y = a + b_1 x_1 + b_2 x_2$).

2) Найти коэффициенты множественной и частной корреляции, множественной детерминации; дать их характеристику.

3) Рассчитать общий и частные F-критерии Фишера; оценить статистическую надежность уравнения регрессии и коэффициента множественной детерминации; оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после фактора x_2 и целесообразность включения фактора x_2 после фактора x_1.

4) При необходимости найти уравнение парной регрессии (исключив статистически незначимый фактор).

Решение

1. Рассчитаем параметры уравнения линейной множественной регрессии в стандартизированной форме $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2}$ и естественной форме $y = a + b_1 x_1 + b_2 x_2$ методом наименьших квадратов. Составим расчетную таблицу 2.

Таблица 11.2.

№	y	x_1	x_2	yx_1	yx_2	x_1x_2	y^2	x_1^2	x_2^2
1	167	5	50	835	8350	250	27889	25	2500
2	175	5	70	875	12250	350	30625	25	4900
3	146	8	110	1168	16060	880	21316	64	12100
4	143	8	120	1144	17160	960	20449	64	14400
5	120	10	175	1200	21000	1750	14400	100	30625
6	220	4	62	880	13640	248	48400	16	3844
7	150	5	87,5	750	13125	437,5	22500	25	7656,25
8	172	5	84	860	14448	420	29584	25	7056
9	170	7	77	1190	13090	539	28900	49	5929
10	190	4	83	760	15770	332	36100	16	6889
11	210	4	65	840	13650	260	44100	16	4225
12	143	8	120	1144	17160	960	20449	64	14400
13	167	6	88	1002	14696	528	27889	36	7744
14	150	7	89	1050	13350	623	22500	49	7921
15	195	4	83	780	16185	332	38025	16	6889
Сумма	2518	90	1363,5	14478	219934	8869,5	433126	590	137078,3
Среднее	**167,87**	**6,00**	**90,90**	**965,20**	**14662,27**	**591,30**	**28875,07**	**39,33**	**9138,55**

Найдем средние квадратические отклонения переменных:

$$\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2} = \sqrt{28875{,}07 - 167{,}87^2} = 26{,}38$$

$$\sigma_{x_1} = \sqrt{\overline{x_1}^2 - \overline{x_1}^2} = \sqrt{39{,}33 - 6{,}0^2} = 1{,}83$$

$$\sigma_{x_2} = \sqrt{\overline{x_2}^2 - \overline{x_2}^2} = \sqrt{9138{,}55 - 90{,}9^2} = 29{,}59$$

Найдем коэффициенты парной корреляции (vepa, используемая для представления, насколько сильно связаны две случайные величины, известная как корреляция):

$$r_{yx_1} = \frac{cov(y, x_1)}{\sigma_y \sigma_{x_1}} = \frac{\overline{y.x_1} - \bar{y}.\overline{x_1}}{\sigma_y \sigma_{x_1}} = \frac{965{,}2 - 167{,}87 \times 6{,}0}{26{,}38 \times 1{,}83} = -0{,}87$$

$$r_{yx_2} = \frac{cov(y, x_2)}{\sigma_y \sigma_{x_2}} = \frac{\overline{y.x_2} - \bar{y}.\overline{x_2}}{\sigma_y \sigma_{x_2}} = \frac{14662{,}27 - 167{,}87 \times 90{,}9}{26{,}38 \times 29{,}59} = -0{,}76$$

$$r_{x_1x_2} = \frac{cov(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}} = \frac{\overline{x_1.x_2} - \overline{x_1}.\overline{x_2}}{\sigma_{x_1} \sigma_{x_2}} = \frac{591{,}3 - 6{,}0 \times 90{,}9}{1{,}83 \times 29{,}59} = 0{,}85$$

Стандартизированные β-коэффициенты определим по формулам (11.5):

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{-0{,}87 - (-0{,}76) \times 0{,}85}{1 - 0{,}85^2} = -0{,}8$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{-0{,}76 - (-0{,}87) \times 0{,}85}{1 - 0{,}85^2} = -0{,}09$$

Таким образом, уравнение регрессии в стандартизированной форме имеет вид: $t_y = -0{,}8t_{x_1} - 0{,}09t_{x_2}$.

Вывод: Сравнение модулей значений стандартизированных коэффициентов регрессии ($|\beta_1| = 0{,}8 > |\beta_2| = 0{,}09$) говорит о том, что на цену автомобиля возраст (фактор x_1) оказывает значительно большее влияние, нежели пробег (фактор x_2).

Рассчитаем естественные коэффициенты регрессии:

$$b_1 = \beta_1 \frac{\sigma_y}{\sigma_{x_1}} = -0{,}8 \frac{26{,}38}{1{,}83} = -11{,}56$$

$$b_2 = \beta_1 \frac{\sigma_y}{\sigma_{x_1}} = -0{,}09 \frac{26{,}38}{29{,}59} = -0{,}08$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = 167{,}87 - (-11{,}56) \times 6{,}0 - (-0{,}08) \times 90{,}9 = 244{,}09$$

Получаем уравнение линейной множественной (двухфакторной) регрессии в естественной форме: $y = 244{,}09 - 11{,}56x_1 - 0{,}08x_2$

Вывод: с увеличением возраста машины на 1 год ее цена уменьшается в среднем на 11,56 тыс. рублей, а с увеличением пробега на 1 тыс. км цена уменьшается в среднем на 0,08 тыс. рублей (80 рублей).

2. Найдем коэффициенты множественной и частной корреляции, а также множественной детерминации. Коэффициент множественной корреляции находится по формуле:

$$R_{yx_1x_2} = \sqrt{\beta_1 r_{yx_1} + \beta_2 r_{yx_2}} = \sqrt{-0{,}8 \times (-0{,}87) - 0{,}09 \times (-0{,}76)} = \sqrt{0{,}76} = 0{,}87$$

$$R_{yx_1x_2}^2 = \left(\sqrt{0{,}76}\right)^2 = 0{,}76$$

Вывод: величина коэффициента множественной корреляции показывает, что связь между y, x_1, x_2 – высокая (при качественной интерпретации коэффициента корреляции используется шкала Чеддока), причем 76,3% вариации цены на автомобиль объясняется вариацией возраста машины и пробега.

Шкала Чеддока: слабая — от 0,1 до 0,3; умеренная — от 0,3 до 0,5; заметная — от 0,5 до 0,7; высокая — от 0,7 до 0,9; весьма высокая (сильная) — от 0,9 до 1,0.

Коэффициенты частной корреляции определяются через парные коэффициенты корреляции по формулам:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}} = \frac{-0{,}87 - (-0{,}76) \cdot 0{,}85}{\sqrt{(1-(-0{,}76)^2)(1-0{,}85^2)}} = -0{,}65;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}} = \frac{-0{,}76 - (-0{,}87) \cdot 0{,}85}{\sqrt{(1-(-0{,}87)^2)(1-0{,}85^2)}} = -0{,}09;$$

$$r_{x_1x_2 \cdot y} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{yx_2}^2)}} = \frac{0{,}85 - (-0{,}76) \cdot (-0{,}87)}{\sqrt{(1-(-0{,}76)^2)(1-(-0{,}87)^2)}} = 0{,}32.$$

Вывод: коэффициенты частной корреляции характеризуют тесноту связи между двумя переменными, исключив влияние третьей переменной. Значит, связь между ценой на ВАЗ 2110 и годом выпуска при исключении влияния величины пробега обратная и заметная; между ценой автомобиля и пробегом без учета возраста машины – обратная, но слабая; связь между факторами x_1 и x_2 – умеренная. Сравним соответствующие коэффициенты парной и частной корреляции: $r_{yx_1} = -0{,}87$, $r_{yx_2} = -0{,}76$, $r_{x_1x_2} = 0{,}85$.

$r_{yx_1 \times x_2} = -0{,}65$, $r_{yx_2 \times x_1} = -0{,}09$, $r_{x_1x_2 \times y} = 0{,}32$

Вывод:

1) при закреплении фактора x_2 на постоянном уровне влияние на y фактора x_1 оказалось несколько менее сильным (– 0,65 против – 0,87), но все равно остается заметным;

2) при закреплении фактора x_1 на постоянном уровне влияние на y фактора x_2 стало весьма слабым (– 0,09 против – 0,76);

3) межфакторная связь ($r_{x_1x_2} = 0{,}85$) говорит о высокой коллинеарности (наличие линейной зависимости) факторов, причем исключив влияние результативной переменной y эта связь становится умеренной.

3. Оценим значимость уравнения регрессии и коэффициента множественной детерминации с помощью F-критерия Фишера. Наблюдаемое значение критерия находится по формуле:

$$F_{\text{набл}} = \frac{R_{yx_1x_2}^2}{1 - R_{yx_1x_2}^2} \times \frac{n-m-1}{m} = \frac{0{,}76}{1-0{,}76} \times \frac{15-2-1}{2} = 19{,}27$$

Табличное значение критерия при уровне значимости $\propto = 0{,}05$ и $k_1 = m = 2, k_2 = n - m - 1 = 15 - 2 - 1 = 12$

$$F_{\text{табл}} = \text{F}(0{,}05; 2; 12) = 3{,}88$$

Вывод: т.к. $F_{\text{табл}} < F_{\text{набл}}$, то с вероятностью $1-\propto = 0{,}95$ делаем заключение о статистической значимости уравнения регрессии и коэффициента множественной детерминации, которые сформировались под неслучайным воздействием факторов x_1 и x_2.

Оценим целесообразность включения в уравнение множественной регрессии фактора x_1 после фактора x_2 и целесообразность включения фактора x_2 после фактора x_1 с помощью частных F-критериев F_{x_1} и F_{x_2}.

$$F_{x_1\ набл} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1-R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1} = \frac{0{,}76-(-0{,}76)^2}{1-0{,}76} \cdot \frac{15-2-1}{1} = 9{,}00;$$

$$F_{x_2\ набл} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1-R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1} = \frac{0{,}76-(-0{,}87)^2}{1-0{,}76} \cdot \frac{15-2-1}{1} = 0{,}10.$$

Найдем табличные значения критерия на уровне значимости $\propto = 0{,}05$ и $k_1 = 1,\ k_2 = n-m-1 = 15-2-1 = 12$: $F_{\text{табл}} = \text{F}(0{,}05; 1; 12) = 4{,}75$

Вывод:

1) Поскольку $F_{x_1\text{набл}} > F_{\text{табл}}$, то включение в модель фактора x_1 (возраста автомобиля) после фактора x_2 статистически оправдано и коэффициент b_1 при факторе x_1 статистически значим.

2) Поскольку $F_{x_2\text{набл}} < F_{\text{табл}}$, то нецелесообразно включать в модель фактор x_2 (пробег) после фактора x_1. Это означает, что парная регрессия зависимости цены ВАЗ 2110 от возраста машины является достаточно статистически значимой, надежной и что нет необходимости улучшать ее, включая дополнительный фактор x_2.

Найдем уравнение парной регрессии $y = a + bx_1$, где y – цена автомобиля (тыс. руб), x_1 – возраст машины (лет):

$$b = \frac{\text{cov}(x_1; y)}{\sigma^2_{x_1}} = \frac{\overline{x_1 \cdot y} - \overline{x_1} \cdot \overline{y}}{\sigma^2_{x_1}} = \frac{965{,}2 - 6 \cdot 167{,}87}{1{,}83^2} = -12{,}6;$$

$$a = \overline{y} - b \cdot \overline{x_1} = 167{,}87 - (-12{,}6) \cdot 6 = 243{,}47.$$

Получаем: $y = 243{,}47 - 12{,}6x_1$

Методы отбора переменных (признаков) регрессионные модели

Большинство реальных анализируемых процессов и объектов являются сложными, для их описания требуется много признаков и показателей. Поэтому типична ситуация, когда аналитику при построении регрессионных моделей приходится иметь дело с десятками переменных и, соответственно, производить

отбор переменных для построения модели. Данная задача совсем не так проста, как кажется. На первый взгляд, нужно отобрать только те переменные, которые непосредственно связаны с решаемой задачей. Однако даже после того, как посторонние переменные будут отсеяны, нет гарантии успешного решения.

Действительно, из ранее рассмотренного материала по регрессии мы увидели, что:

- входные переменные могут иметь низкую значимость, то есть линейная зависимость между ними и выходной переменной может либо отсутствовать, либо быть очень слабой. Такие переменные не способствуют повышению точности полученных оценок, а только усложняют модель;
- входные переменные могут коррелировать между собой, что приводит к мультиколлинеарности и, как следствие, к снижению точности и устойчивости модели, к противоречивости результатов и т. д.

Как показывает практика, визуально выявить эти проблемы в исходных данных и результатах регрессии практически невозможно. Чтобы оптимизировать процесс отбора переменных для использования в регрессионной модели, разработаны различные методы, которые позволяют учесть указанные факторы и выбрать наилучшую модель.

Существует общая рекомендация, которая в первом приближении дает возможность построить хорошую регрессионную модель: необходимо включить в рассмотрение все переменные, которые позволяют повысить точность оценок, получаемых с помощью регрессии. Но возникает дилемма: как реализовать модель с приемлемыми точностью и затратами? На практике приходится соблюдать два противоречивых требования.

- В регрессионной модели нужно использовать как можно больше входных переменных, содержащих новую информацию о выходной переменной.
- Поскольку включение в модель каждой новой переменной увеличивает временные и вычислительные затраты на ее реализацию, нужно стремиться, чтобы модель содержала как можно меньше входных переменных.

Выбор лучшей регрессионной модели заключается в поиске компромисса между данными требованиями.

Таким образом, процедура отбора признаков решает следующие задачи:

1. Упрощение моделей с целью улучшения их интерпретируемости.
2. Сокращение размерности пространства признаков.
3. Уменьшение временных и вычислительных затрат на построение и эксплуатацию модели.

4. Повышение обобщающей способности модели и борьба с переобучением.

Обобщающая способность – это способность аналитической модели, построенной на основе машинного обучения выдавать правильные результаты не только для примеров, участвовавших в процессе обучения, но и для любых новых, которые не участвовали в нем. Обобщающая способность является важнейшим свойством аналитической модели, приобретаемым в процессе обучения.

Переобучение – это явление, когда обучаемая модель хорошо распознает примеры из обучающего множества, но при этом не распознает или плохо распознает любые другие примеры, не участвовавшие в процессе обучения (т.е. предъявляемые ей в процессе практического использования).

В основе идеи отбора признаков лежит понимание того, что не все обучающие данные являются полезными: они могут содержать избыточные и незначащие (нерелевантные) признаки, которые могут быть удалены без существенной потери информации и ухудшения качества модели. При этом даже значимый признак может оказаться избыточным, если коррелирует с другим значимым признаком.

Технология отбора признаков основана на формировании подмножеств из общего числа признаков и вычисления для каждого из них некоторой оценки качества. Простейшей из таких оценок является ошибка модели: выбирается тот набор признаков, который минимизирует ошибку. Однако на практике этот подход реализуем только для задач небольшой размерности, поскольку для большого числа признаков формируется огромное число подмножеств, которое требуется проверить.

Выделяют четыре класса методов отбора признаков:

1. Обёрточные (wrapper) методы — используют предсказательное моделирование для оценивания подмножеств признаков. Каждое подмножество используется для обучения модели, а затем модель проверяется на тестовом множестве. Лучшим принимается то подмножество признаков, для которого количество ошибок минимально. Несмотря на то, что как отмечалось выше, данный метод требователен к вычислительным ресурсом, он позволяет получить наилучший результат для конкретного вида задачи и аналитической модели. Кроме этого, обёрточные методы склонны к переобучению. Рассмотренные в методы будут относятся именно к этой категории.

2. Методы фильтрации используют косвенные меры качества модели вместо ошибки, например корреляцию между входными переменными и выходной. В простейшем случае для каждой входной переменной вычисляется

коэффициент корреляции с выходной, и исключаются те переменные, для которых он ниже заданного порога. Таким образом формируется своего рода фильтр, которые пропускает переменные с сильной корреляцией относительно выходной, и "подавляют" со слабой. Методы фильтрации менее требовательны к вычислительным ресурсам, чем обёрточные методы, но являются общими и не ориентированы на конкретный вид модели, поэтому обычно показывают несколько худшие результаты.

3. Встроенные (embedded) методы. Представляют наиболее универсальную группу методов, в которых отбор признаков рассматривается как часть процесса построения модели. Встроенные методы специфичны для конкретной модели.

4. Рекурсивные методы отбора (Recursive Feature Elimination — RFE). В этом случае ищутся не подмножества признаков, а каждому признаку присваиваются веса, по которым они ранжируются. Затем исключаются признаки с малыми весами. Присвоение весов производится с помощью специальной модели-оценщика, которая сначала обучается на начальном наборе признаков. Затем признаки с малыми весами исключаются и обучение производится снова, в результате чего веса оставшихся признаков вновь меняются. И так рекурсивно производится до тех пор, пока не будет получен оптимальный набор признаков.

Постановка задачи

Зададим признаковое описание объекта с использованием следующих обозначений. Каждая независимая переменная представлена вектором-столбцом $x_j = (x_{j1}, \dots x_{jm})$, а зависимая $y_i = (y_{i1}, \dots y_{im})$. Тогда

$$y = b_1 x_1 + \cdots + b_n x_n$$

или в матричном представлении

$$y = Xb$$

где X – матрица признаков со столбцами $x_1, \dots, x_n$, $b = (b_1, \dots, b_n)$ – вектор параметров модели.

Пусть задана выборка $D = \{x_i, y_i\}, i = 1, \dots, m$ состоящая из m пар, включающих векторы значений зависимых переменных $x_i = x_{ij}$, $j = 1, \dots, n$ и значений единственной независимой переменной y_i. Индексы наблюдений i и индекс независимых переменных j, будем рассматривать как элементы множеств $i \in I = \{1, \dots, m\}, j \in J = \{1, \dots, n\}$.

Также пусть задано разбиение на обучающее и тестовое множества L и T, $I = L \cup T$.

Зададим модель линейной регрессии в виде:

$$y_i = f_s(b_s x_i) = \sum_{j=1}^{n} b_j x_{ij}$$

где $s = \{1, \dots, 2^n\}$ – индекс модели, $b_s = (b_j)$ – вектор параметров модели.

Алгоритм выбора модели задаёт метод оптимизации, доставляющий оптимальное значение параметрам $\hat{b}$ модели на обучающей выборке. Минимизируемый функционал качества модели определим как сумму квадратов остатков регрессии:

$$S = \sum_{j=1}^{n} \left(y_i - f(b_s x_i)\right)^2 \quad (1)$$

Требуется найти такую модель, которая обеспечит минимум данному функционалу качества. В литературе величину S часто обозначают RSS — Residual Sum of Squares (сумма квадратов остатков).

Принудительное (полное) включение – включение в аналитическую модель всех доступных в обучающем наборе признаков. Этот подход целесообразно использовать в следующих случаях:

1. Количество признаков относительно невелико и их полное включение не приводит к излишней сложности модели как в плане интерпретируемости, так и в плане вычислительной сложности.
2. Исключение любого признака приводит к критичному уменьшению количества информации, используемой для обучения модели. Иными словами, когда незначимые и избыточные признаки просто отсутствуют.

Прямое включение (Forward selection) – метод, который базируется на принципе: начать с пустой модели, в которой признаки отсутствуют и постепенно добавляя признаки найти самые «лучшие».

Обратное исключение (Backward elimination) – исходная модель содержит все признаки, которые поочерёдно исключаются с целью найти «худшие» и не применять их в модели.

Пошаговое включение/исключение (Stepwise) — модификация метода прямого включения с тем отличием, что на каждом шаге после включения новой переменной в модель, осуществляется проверка на значимость остальных переменных, которые уже были введены в нее ранее.

Гребневая регрессия (Ridge regression) — использует процедуру регуляризации для ограничения пространства решений с целью сделать модель более устойчивой в случае высокой коррелированности входных признаков. Подразумевает введение штрафов для уменьшения значений коэффициентов регрессии. При этом значения параметров модели не обращаются в ноль, т.е. отбора переменных не происходит.

LASSO-регрессия — также использует регуляризацию для повышения устойчивости модели. Но отличается от гребневой регрессии тем, что допускает обнуление параметров модели (т.е. реализует процедуру отбора).

Регрессия «Эластичная сеть» — также использует регуляризацию, но в отличии от гребневой регрессии в ней применяет два регуляризующих члена.