Lab 2. Task 2.2

Лагутин Евгений

Рассмотрим задачу предсказание числа заболевших некоторой болезнью от некоторых экологических анализов. Гарантируется, что предсказание описывается линейной моделью.

Так как проведение анализов не является бесплатным, то стоит вопрос о том какие из анализов являются лишними (на уровне значимости $\alpha = 0.05$) для предсказания линейной модели.

Требуется:

1. Записать задачу формально;
2. Провести отбор приизнаков линейной модели.

Все выкладки должны быть сделаны аналитически, без использования компьютера.

Решение

Вдохновение для решения я черпал с этого сайта https://basegroup.ru/community/articles/feature-selection

Частный F-тест

Критерий позволяет оценить необходимость ввода дополнительной переменной в линейную модель.

Множественная регрессия:

$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_i^j + \varepsilon_i,\ i = 1..n$

$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_i^j,\ \ i = 1..n$

$\overline{y} = \frac{1}{n} \sum_{i=1} y_i$

Вводится понятие суммы квадратов регрессии $SSR = \sum_{i=1}^{n} (\overline{y}_i - \hat{y}_i)^2$ - то же самое что объясненная моделью дисперсия.

Сумма квадратов ошибок: $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Пусть на основе $X^1, \ldots, X^k$ построена линейная регрессионная модель. Для нее объясненная дисперсия s_1.

Хотим добавить новый признак X^{k+1}. Для новых признаков объясненная дисперсия s_2.

Очевидно, что $s_2 > s_1$.

$s_\Delta = s_2 - s_1$

Частный F-test помогает ответить на вопрос, как выбрать порог для s_Δ, чтобы принять решение добавить признак.

Формальная задача:

H_0 : s_Δ недостаточно велико, чтобы добавить новый признак.

Переходим к статистике $T(X) = \frac{(n-k-2)s_\Delta}{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \sim F(1, n-k-2)$

По таблице можно определить t_α - α квантиль распределения Фишера со степенями свободы $1,\ n-k-2$.

Если статистика меньше этого значения - принимается решение о нецелесообразности добавления этого нового признака, иначе включаем признак в модель.

Метод прямого отбора

Данный алгоритм включает в себя следующие шаги:

1. Из списка всех возможных входных переменных выбирается та, которая имеет наибольшую корреляцию с Y, после чего модель, содержащая лишь одну выбранную независимую переменную, проверяется на значимость при помощи частного F-критерия. Если значимость модели не подтверждается, то алгоритм на этом заканчивается за неимением существенных входных переменных. В противном случае эта переменная вводится в модель и осуществляется переход к следующему пункту алгоритма.
2. По всем оставшимся переменным рассчитывается значение статистики T, которая представляет собой отношение прироста суммы квадратов регрессии.
3. Из всех переменных-претендентов на включение в модель выбирается та, которая имеет наибольшее значение критерия, рассчитанного в пункте 2.
4. Проводится проверка на значимость выбранной в пункте 3 независимой переменной. Если ее значимость подтверждается, то она включается в модель, и осуществляется переход к пункту 2 (но уже с новой независимой переменной в составе модели). В противном случае алгоритм останавливается.

Перейдем к выборке

Найдем корреляцию каждого признака с целевой переменной. В качестве корреляции естественно взять корреляцию Пирсона.

feature	**corr(x_j, y)**
x1	-0.244527
x2	0.0857819
x3	0.785655
x4	0.185746
x5	0.0104914
x6	-0.354543
x7	0.0195487
x8	0.110317
x9	-0.13446
x10	0.0961255

Самую большую корреляцию имеет признак x_3. Рассматриваем модель $\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3$, параметры оценим МНК.

$\hat{\beta}_{OLS} = \left(X^T X\right)^{-1} X^T Y$

В одномерном случае $\hat{\beta_3} = \frac{\sum_{t=1}^{n} x_t^3 y_t}{\sum_{t=1}^{n} (x_t^3)^2}$, $\hat{\beta_0} = \frac{\sum_{t=1}^{n} y_t - \tilde{\beta_3} \sum_{t=1}^{n} x_t^3}{n}$

$\hat{\beta}_3 = 1.78$, $\hat{\beta}_0 = 3.16$

Считаем для него статистику $T = \frac{(30-1-2)(\sum_{i=1}^{n}(\overline{y}_i - \hat{y}_i)^2 - 0)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 45.15$

Табличное значение для $F(1, 30-1-2)$: $t_\alpha \approx 4,2$.

Добавляем признак.

Далее мы должны посчитать критерий для всех остальных признаков (модель для 2 признаков и константы).

Коэффициенты придется считать МНК.

Посчитаем получившиеся объясненные дисперсии, суммы квадратов ошибок и статистики.

	feature	sse	ssr	T
0	x1	95.85	22.19	6.2507
1	x2	116.49	1.52	0.352305
2	x4	114.27	3.75	0.886059
3	x5	116.47	1.54	0.357002
4	x6	115.31	2.67	0.625184
5	x7	117.06	0.95	0.219118
6	x8	117.59	0.38	0.0872523
7	x9	117.95	0.09	0.0206019
8	x10	117.77	0.26	0.0596077

Наибольшее значение статистики имеет признак x_1.

Проверим его значимость.

Табличное значение для $F(1, 30-2-2)$: $t_\alpha \approx 4,2 < 6.25$

Добавляем признак в модель.

Снова ищем статистики для оставшихся признаков, но уже для 3 признаков + константы.

	feature	sse	ssr	T
0	x2	94.34	1.48	0.423574
1	x4	90.27	5.58	1.66899
2	x5	95.58	0.25	0.0706215
3	x6	92.71	3.09	0.899903
4	x7	94.66	1.18	0.336573
5	x8	95.07	0.77	0.218681
6	x9	95.76	0.08	0.0225564
7	x10	95.86	-0.03	-0.00844982

Все значения статистики меньше уровня значимости, значит никакой признак добавлять не стоит.

Итак, отбираем только признаки x_1, x_3