Лекция 2

Линейная регрессия, Оценки регрессионых параметров,

Лектор – *Сенько Олег Валентинович*

Курс «Машинное обучение и интеллектуальный анализ данных»

Содержание лекции

Линейная регрессия

Распространённым средством решения задач прогнозирования непрерывной величины Y по переменным $X_1, \ldots, X_n$ является использование метода множественной линейной регрессии. В данном методе связь переменной Y с переменными $X_1, \ldots, X_n$ задаётся с помощью линейной модели

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon, \tag{1}$$

где $\beta_0, \beta_1, \ldots, \beta_n$ - вещественные регрессионные коэффициенты, ε - случайная величина, являющаяся ошибкой прогнозирования. Регрессионные коэффициенты ищутся по обучающей выборке

$$\tilde{S}_t = \{s_1 = (y_1, \mathbf{x}_1), \ldots, s_m = (y_m, \mathbf{x}_m)\}, \tag{2}$$

где $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})$ вектор значений переменных $X_1, \ldots, X_n$ для объекта s_j.

Линейная регрессия. Метод наименьших квадратов

Традиционным способом поиска регрессионных коэффициентов является метод наименьших квадратов (МНК). МНК заключается в минимизации функционала эмпирического риска с квадратичными потерями

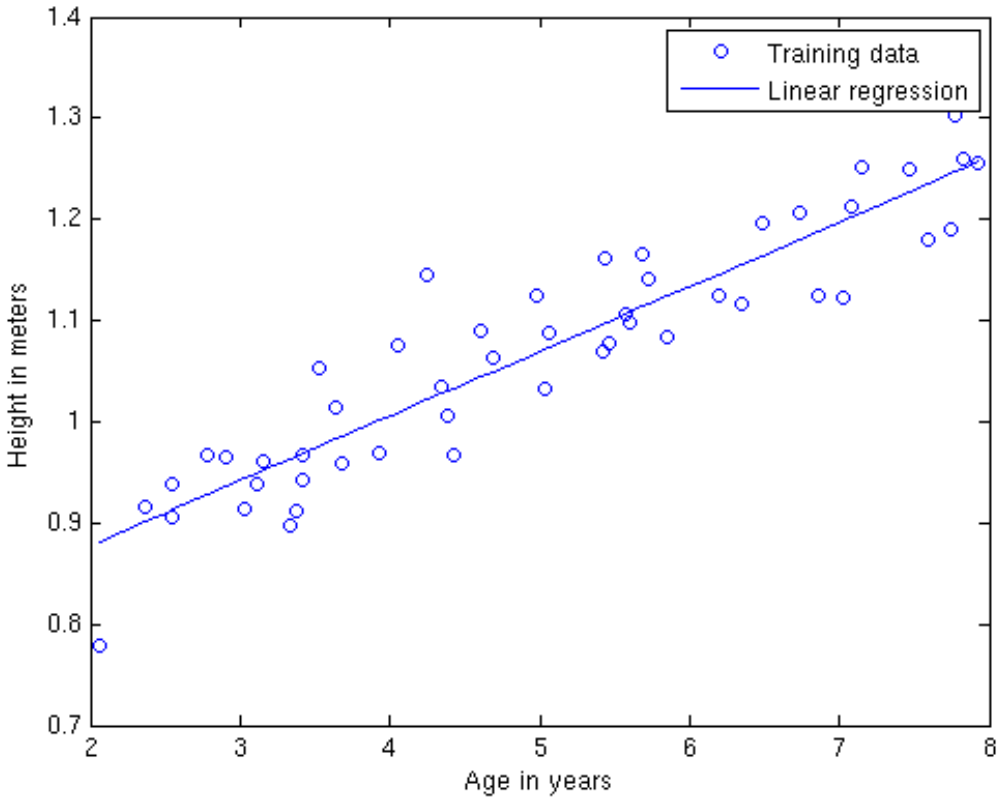
$$Q(\tilde{S}_t, \beta_0, \beta_1, \ldots, \beta_n) = \sum_{j=1}^{m} [y_j - \beta_0 - \sum_{i=1}^{n} x_i \beta_{ij}]^2 \tag{3}$$

То есть оценки $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n$ регрессионных коэффициентов $\beta_0, \beta_1, \ldots, \beta_n$ по методу МНК удовлетворяют условию минимума функционала эмпирического риска с квадратичными потерями

$$(\hat{\beta}_0, \ldots, \hat{\beta}_n) = \arg\min[Q(\tilde{S}_t, \beta_0, \beta_1, \ldots, \beta_n)]. \tag{4}$$

Линейная регрессия. Одномерная модель.

Рассмотрим простейший вариант линейной регрессии, описывающей связь между переменной Y и единственной переменной X :
$Y = \beta_0 + \beta_1 X + \varepsilon$.



Функционал эмпирического риска на выборке $\tilde{S}_t = \{(y_1, x_1), \ldots, (y_m, x_m)\}$ принимает вид

$$Q(\tilde{S}_t, \beta_0, \beta_1) = \frac{1}{m} \sum_{j=1}^{m} [y_j - \beta_0 - x_j \beta_1]^2. \tag{5}$$

Необходимым условием минимума функционала $Q(\tilde{S}_t, \beta_0, \beta_1)$ является выполнение системы из двух уравнений

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \beta_1)}{\partial \beta_0} = -\frac{2}{m} \sum_{j=1}^{m} y_j + 2\beta_0 + \frac{2\beta_1}{m} \sum_{j=1}^{m} x_j = 0 \tag{6}$$

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \beta_1)}{\partial \beta_1} = -\frac{2}{m} \sum_{j=1}^{m} x_j y_j + \frac{2\beta_0}{m} \sum_{j=1}^{m} x_j + \frac{2\beta_1}{m} \sum_{j=1}^{m} x_j^2 = 0$$

Оценки $\hat{\beta}_0, \hat{\beta}_1$ являются решением системы (11) относительно параметров соответственно β_0, β_1 .
Оценки регрессионных коэффициентов могут быть записаны в виде

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{m} x_j y_j - \frac{1}{m}\sum_{j=1}^{m} y_j \sum_{j=1}^{m} x_j}{\sum_{j=1}^{m} x_j^2 - \frac{1}{m}(\sum_{j=1}^{m} x_j)^2}, \tag{7}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

, где $\overline{y} = \frac{1}{m}\sum_{j=1}^{m} y_j$, $\overline{x} = \frac{1}{m}\sum_{j=1}^{m} x_j$. Выражение для $\hat{\beta}_1$ может быть переписано в виде

$$\hat{\beta}_1 = \frac{Cov(Y, X \mid \tilde{S}_t)}{D(X \mid \tilde{S}_t}, \tag{8}$$

где $Cov(Y, X \mid \tilde{S}_t)$ является выборочной ковариацией переменных Y и X, $D(X \mid \tilde{S}_t)$ является выборочной дисперсией переменной X.

То есть

$$Cov(Y, X \mid \tilde{S}_t) = \frac{1}{m} \sum_{j=1}^{m} (y_j - \overline{y})(x_j - \overline{x})$$

$$D(X \mid \tilde{S}_t) = \frac{1}{m} \sum_{j=1}^{m} (x_j - \overline{x})^2$$

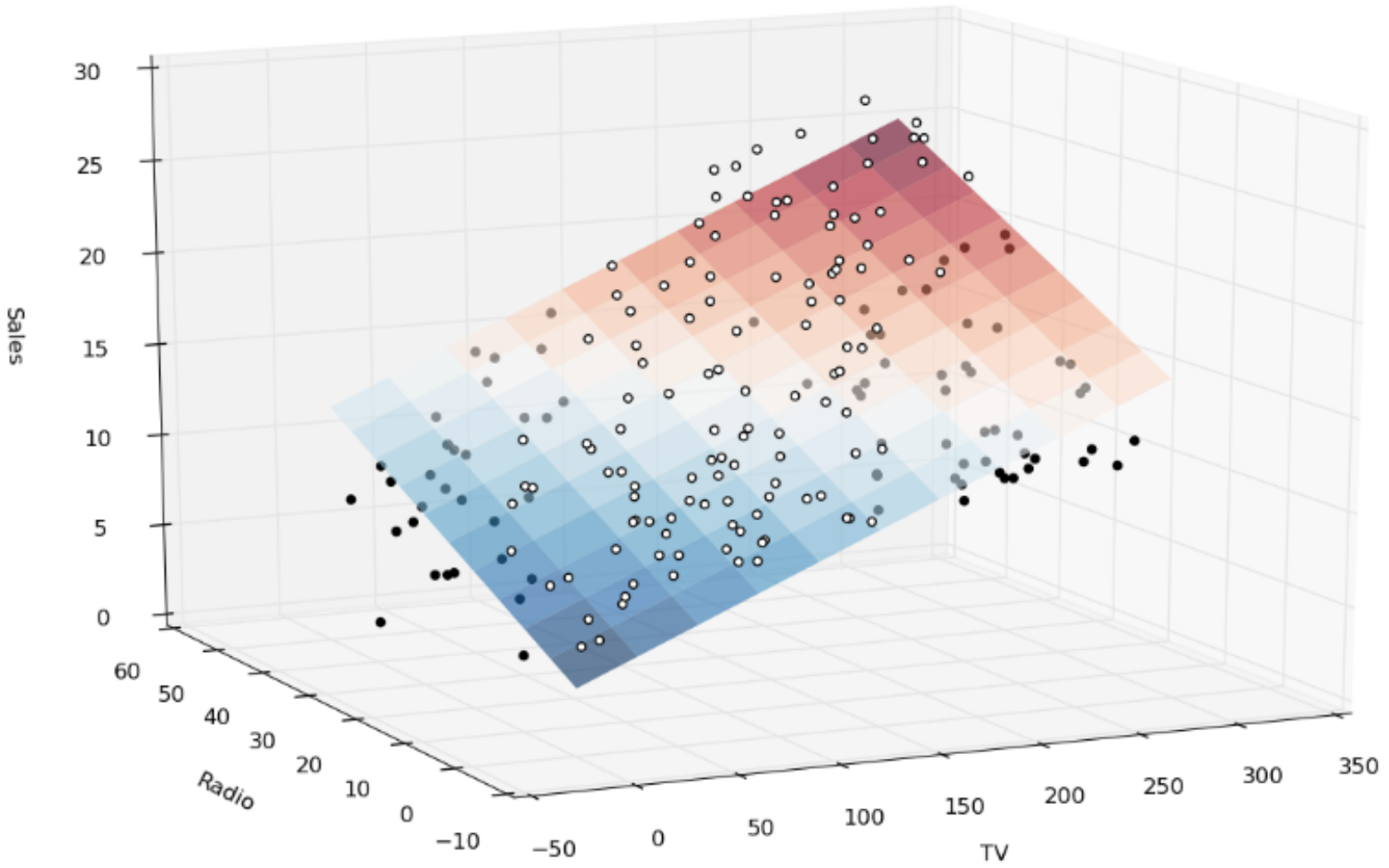
Связь коэффициента β с коэффициентом корреляции
Коэффициент корреляции

$$\hat{\rho}(Y, X) = \frac{Cov(Y, X \mid \tilde{S}_t)}{\sqrt{D(X \mid \tilde{S}_t) D(Y \mid \tilde{S}_t)}}, \tag{9}$$

То есть

$$\beta = \frac{\hat{\rho}(Y, X) \sqrt{D(X \mid \tilde{S}_t)}}{\sqrt{D(Y \mid \tilde{S}_t)}}, \tag{10}$$

Многомерная регрессия



Многомерная регрессия

При вычислении оценки вектора параметров $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_n)$ в случае многомерной линейной регрессии удобно использовать матрицу плана $\mathbf{X}$ размера $m \times (n+1)$, которая строится по обучающей выборке $\tilde{S}_t$. Матрица плана имеет вид

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{j1} & \ldots & x_{jn} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{m1} & \ldots & x_{mn} \end{pmatrix}$$

То есть j-я строка матрицы плана представляет собой вектор значений переменных $X_1, \ldots, X_n$ для объекта s_j с одной добавленной слева компонентой, содержащей 1.

Многомерная регрессия

Пусть $\mathbf{y} = (y_1, \ldots, y_m)$ - вектор значений переменной Y . Связь Y с переменными $X_1, \ldots, X_n$ на объектах обучающей выборки может быть описана с помощью матричного уравнения

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X}^t + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)$ - вектор ошибок прогнозирования для объектов $\tilde{S}_t$. Функционал $Q(\tilde{S}_t, \beta_0, \beta_1, \ldots, \beta_n)$ может быть записан в виде

$$Q(\tilde{S}_t, \beta_0, \beta_1, \ldots, \beta_n) = \sum_{j=1}^{m} [y_j - \beta_0 - \sum_{i=1}^{n} \beta_i \breve{x}_{ji}]^2, \qquad (11)$$

где $\breve{x}_{ji}$ - элементы матрицы плана $\mathbf{X}$, определяемые равенствами $\breve{x}_{j1} = 1$, $\breve{x}_{ji} = x_{j(i-1)}$ при $i > 1$.

Необходимым условием минимума функционала $Q(\tilde{S}_t, \beta_0, \beta_1, \ldots, \beta_n)$ является выполнение системы из $n+1$ уравнений

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)}{\partial \beta_0} = -\frac{2}{m}[\sum_{j=1}^{m} y_j \breve{x}_{j1} - \sum_{j=1}^{m}\sum_{i=1}^{n+1} \beta_i \breve{x}_{ji} \breve{x}_{j1}] = 0 \qquad (12)$$

$$\ldots, \ldots, \ldots, \ldots, \ldots, \ldots, \ldots, \ldots$$

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)}{\partial \beta_n} = -\frac{2}{m}[\sum_{j=1}^{m} y_j \breve{x}_{jn} - \sum_{j=1}^{m}\sum_{i=1}^{n+1} \beta_i \breve{x}_{ji} \breve{x}_{jm}] = 0$$

Вектор оценок значений регрессионных коэффициентов $\hat{\beta}_0, \ldots, \hat{\beta}_n$ является решением системы уравнений (15) .

В матричной форме система (15) может быть записана в виде

$$-2\mathbf{X}^t\boldsymbol{y}^t + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}^t = 0 \tag{13}$$

Решение системы (16) существует, если $\det(\mathbf{X}^t\mathbf{X}) \neq 0$. В этом случае для $\mathbf{X}^t\mathbf{X}$ существует обратная матрица и решение (16) относительно вектора $\boldsymbol{\beta}$ может быть записано в виде:

$$\hat{\boldsymbol{\beta}}^t = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{y}^t. \tag{14}$$

Из теории матриц следует, что $\det(\mathbf{X}^t\mathbf{X}) = 0$ если ранг матрицы $\mathbf{X}$ по строкам менее $(n+1)$, что происходит, если m-мерный вектор значений одной из переменных $X_{i'} \in \{X_1, \ldots, X_n\}$ на выборке $\tilde{S}_t$ является линейной комбинаций m-мерных векторов значений на $\tilde{S}_t$ других переменных из $\{X_1, \ldots, X_n\}$.

Многомерная регрессия

При сильной коррелированности одной из переменных $X_{i'} \in \{X_1, \ldots, X_n\}$ на выборке с какой-либо линейной комбинацией других переменных значение $\det(\mathbf{X}^t\mathbf{X})$ оказывается близким к 0. При этом вычисленный вектор оценок $\hat{\boldsymbol{\beta}}^t$ может сильно изменяться при относительно небольших чисто случайных изменениях вектора $\boldsymbol{y} = (y_1, \ldots, y_m)$. Данное явление называется мультиколлинеарностью. Оценивание регрессионных коэффициентов с использованием МНК при наличии мультиколлинеарности оказывается неустойчивым. Отметим также, что $\det(\mathbf{X}^t\mathbf{X}) = 0$ при $n + 1 > m$.Поэтому МНК не может использоваться для оценивания регрессионных коэффициентов, когда число переменных превышает число объектов в обучающей выборке. На практике высокая устойчивость достигается только, когда число объектов в выборках по крайней мере в 3-5 раз превышает число переменных.

Оценка точности

Оценка точности может производится по контрольной выборке

$$\tilde{S}_c = \{(y_1^c, \boldsymbol{x}_1^c), \ldots, (y_{m_c}^c, \boldsymbol{x}_{m_c}^c)\}$$

Величина относительной невязки

$$\Delta_{rel}(\tilde{S}_c) = \frac{\sum_{j=1}^{m_c} (y_j^c - \beta_0 - \sum_{j=1}^{m_c} \beta_i x_{ji})^2}{D(Y \mid \tilde{S}_c)}, \tag{15}$$

$R^2(\tilde{S}_c) = 1 - \Delta_{rel}(\tilde{S}_c)$ Относительная невязка и R^2 моут вычисляться в с помощью $k-$фолдовой кросс-валидации.

Отбор признаков- способ повышения

Проблемы мультиколлинеарности
малости отношения $\frac{m}{n}$
Могут быть решены через отбор информативных X переменных (регрессоров).
Пошаговые методы
Прямая пошаговая регрессия - на каждом шаге добавляется наиболее информативный по выбранному критерию регрессор.
Обратная пошаговая регрессия - на каждом шаге удаляется наименее информативный по выбранному критерию регрессор.

Структура ошибки прогнозирования. Обобщённая ошибка.

Напомним, что обобщающая способность алгоритма прогнозирования A, обученного по выборке $\widetilde{S}_t$ с помощью некоторого метода **M** измеряется величиной потерь на генеральной совокупности Ω:

$$E_\Omega \lambda[Y, A(\boldsymbol{x}, \widetilde{S}_t)].$$

Для оценки эффективности использования метода прогнозирования **M** для прогнозирования случайного процесса, связанного с генеральной совокупностью Ω, при фиксированном размере обучающей выборки m естественно использовать математическое ожидание потерь по пространству всевозможных обучающих выборок длины m:

$$\Delta_G = E_{\Omega_m} E_\Omega \lambda[Y, A(\boldsymbol{x}, \widetilde{S}_t)],$$

где Ω_m - рассмотренное ранее пространство обучающих выборок длины m.

Структура ошибки прогнозирования. Обобщённая ошибка.

Величину Δ_G будем называть обобщёнными потерями. При использовании в качестве функции потерь квадрата ошибки обобщённые потери становятся обобщённой квадратичной ошибкой и принимают вид

$$\Delta_G = E_{\Omega_m} E_{\Omega}[Y - A(\boldsymbol{x}, \widetilde{S}_t)]^2.$$

Справедливо трёхкомпонентное разложение обобщённой квадратичной ошибки Δ_G:

$$\Delta_G = \Delta_N + \Delta_B + \Delta_V \tag{16}$$

Структура обобщённой ошибки.

Шумовая компонента.

$$\Delta_N = E_\Omega[Y - E(Y \mid \boldsymbol{x})]^2$$

является минимально достижимой квадратичной ошибкой прогноза, которая не может быть устранена с использованием только математических средств. Составляющая сдвига (Bias).

$$\Delta_B = E_\Omega[E_\Omega(Y \mid \boldsymbol{x}) - \hat{A}(\boldsymbol{x})]^2$$

Высокое значение компоненты сдвига соответствует отсутствию в модели $\widetilde{M} = \{A : \widetilde{X} \to \widetilde{Y}$, внутри которой осуществляется поиск, алгоритмов, достаточно хорошо аппроксимирующих объективно существующую зависимость Y от переменных $X_1, \ldots, X_n$. Составляющая сдвига может быть снижена, например, путём расширения модели за счёт включения в него дополнительных более сложных алгоритмов, что обычно позволяет повысить точность аппроксимации данных.

Структура обобщённой ошибки.

Дисперсионная составляющая (Variance).

$$\Delta_V = E_{\Omega_m} E_{\Omega}[\hat{A}(\boldsymbol{x}) - A(\boldsymbol{x}, \widetilde{S}_t)]^2$$

характеризует неустойчивость обученных прогнозирующих алгоритмов при статистически возможных изменениях в обучающих выборках. Дисперсионная составляющая возрастает при небольших размерах обучающей выборки. Дисперсионная составляющая может быть снижена путём выбора сложности модели, соответствующей размеру обучающих данных.

Структура обобщённой ошибки.

Таким образом существует Bias-Variance дилемма Составляющая сдвига может быть снижена путём увеличения разнообразия модели. Однако увеличение разнообразия модели при недостаточном объёме обучающих данных ведёт к росту компоненты сдвига. Наиболее высокая точность прогноза достигается, при поддержании правильного баланса между разнообразием используемой модели и объёмом обучающих данных

Методы регуляризации

Одним из возможных способов борьбы с неустойчивостью является использование методов, основанных на включение в исходный оптимизируемый функционал $Q(\tilde{S}_t, \beta_0, \beta_1, \ldots, \beta_n)$ дополнительной штрафной компоненты. Введение такой компоненты позволяет получить решение, на котором $Q(\tilde{S}_t, \beta_0, \beta_1, \ldots, \beta_n)$ достаточно близок к своему глобальному минимуму. Однако данное решение оказывается значительно более устойчивым и благодаря устойчивости позволяет достигать существенно более высокой обобщающей способности. Подход к получению более эффективных решений с помощью включения штрафного слагаемого в оптимизируемый функционал принято называть регуляризацией по Тихонову.

На первом этапе переходим от исходных переменных $\{X_1, \ldots, X_n\}$ к стандартизированным $\{X_1^s, \ldots, X_n^s\}$, где $X_i^s = \frac{X_i - \hat{X}_i}{\hat{\sigma}_i}$, $\hat{X}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ji}$, $\hat{\sigma}_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{X}_i - x_{ji}^2)}$, а также от исходной прогнозируемой переменной Y к стандартизованной прогнозируемой переменной $Y^s = Y - \frac{1}{m}\sum_{j=1}^{m} y_j$. Пусть $\breve{x}_{j1}^s = 1$, $\breve{x}_{ji}^s = x_{j(i-1)}^s$ при $i > 1$, где $x_{j(i-1)}^s$ - значение признака X_i^s для j-го объекта. Пусть также

$$\mathbf{X}_s = \begin{pmatrix} 1 & \mathbf{x}_{11}^s & \ldots & \mathbf{x}_{1n}^s \\ \ldots & \ldots & \ldots & \ldots \\ 1 & \mathbf{x}_{j1}^s & \ldots & \mathbf{x}_{jn}^s \\ \ldots & \ldots & \ldots & \ldots \\ 1 & \mathbf{x}_{m1}^s & \ldots & \mathbf{x}_{mn}^s \end{pmatrix}$$

- матрица плана для стандартизированных переменных, $\boldsymbol{y}^s = (y_1^s, \ldots, y_m^s)$ - вектор значений стандартизованной переменной Y_s .

Одним из первых методов регрессии, использующих принцип регуляризации, является метод гребневой регрессии (ridge regression). В гребневой регрессии в оптимизируемый функционал дополнительно включается сумма квадратов регрессионных коэффициентов при переменных $\{X_1^s, \ldots, X_n^s\}$. В результате функционал имеет вид

$$Q_{ridge}(\tilde{S}_t, \beta_0, \ldots, \beta_n) = \frac{1}{m}\sum_{j=1}^{m}[y_j - \beta_0 - \sum_{i=1}^{n}\beta_i \breve{x}_{ji}^s]^2 + \gamma\sum_{i=1}^{n}\beta_i^2, \quad (17)$$

где γ - положительный вещественный параметр. Пусть $\boldsymbol{\beta}_r$ является вектором оценок регрессионных коэффициентов, полученным в результате минимизации $Q_{ridge}(\tilde{S}_t, \beta_0, \ldots, \beta_n)$.

Отметим, что увеличение регрессионных коэффициентов приводит к увеличению $Q_{ridge}(\tilde{S}_t, \beta_0, \ldots, \beta_n)$. Таким образом использование гребневой регрессии приводит к снижению длины вектора регрессионных коэффициентов при переменных $\{X_1^s, \ldots, X_n^s\}$. Рассмотрим конкретный вид вектора регрессионных коэффициентов $\boldsymbol{\beta}_r$ в гребневой регрессии. Необходимым условием минимума функционала $Q_{ridge}(\tilde{S}_t, \beta_0, \ldots, \beta_n)$ является выполнение системы из $n+1$ уравнений:

$$\frac{\partial Q_{ridge}(\tilde{S}_t, \beta_0, \ldots, \beta_n)}{\partial \beta_0} = -\frac{2}{m}[\sum_{j=1}^{m} y_j \breve{x}_{j1}^s - \sum_{j=1}^{m}\sum_{i=1}^{n+1} \beta_i \breve{x}_{ji}^s \breve{x}_{j1}^s] + 2\gamma\beta_0 = 0 \tag{18}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\frac{\partial Q_{ridge}(\tilde{S}_t, \beta_0, \ldots, \beta_n)}{\partial \beta_n} = -\frac{2}{m}[\sum_{j=1}^{m} y_j \breve{x}_{jn}^s - \sum_{j=1}^{m}\sum_{i=1}^{n+1} \beta_i \breve{x}_{ji}^s \breve{x}_{jm}^s] + 2\gamma\beta_n = 0$$

Поэтому вектор оценок регрессионных коэффициентов в методе гребневая регрессия является решением системы (25).

В матричной форме система (25) может быть записана в виде

$$-2\mathbf{X}_s^t\boldsymbol{y}_s^t + (2\mathbf{X}_s^t\mathbf{X}_s + 2\gamma\mathbf{I})\boldsymbol{\beta}_r^t = 0 \tag{19}$$

или в виде

$$\boldsymbol{\beta}_r^t = (\mathbf{X}_s^t\mathbf{X}_s + \gamma\mathbf{I})^{-1}\mathbf{X}_s^t\boldsymbol{y}_s^t \tag{20}$$

где $\mathbf{I}$ - единичная матрица. Отметим, что произведение $\mathbf{X}_s^t\mathbf{X}_s$ представляет собой симметрическую неотрицательно определённую матрицу. Матрица $\mathbf{X}_s^t\mathbf{X}_s + \gamma\mathbf{I}$ также является симметрической матрицей. Каждому собственному значению λ_k матрицы $\mathbf{X}_s^t\mathbf{X}_s$ соответствует собственное значение $\lambda_k + \gamma$ матрицы $\mathbf{X}_s^t\mathbf{X}_s + \gamma\mathbf{I}$. Пусть λ_{min}^{γ} минимальное собственное значение матрицы $\mathbf{X}_s^t\mathbf{X}_s + \gamma\mathbf{I}$ удовлетворяет неравенству $\lambda_{min}^{\gamma} > \gamma$. Откуда следует, что всегда $\det(\mathbf{X}_s^t\mathbf{X}_s + \gamma\mathbf{I}) > 0$, а обратная матрица $(\mathbf{X}_s^t\mathbf{X}_s + \gamma\mathbf{I})^{-1}$ всегда существует. Достаточно большая величина $\det(\mathbf{X}_s^t\mathbf{X}_s + \gamma\mathbf{I})$ приводит к относительно небольшим изменениям оценок регрессионных коэффициентов при небольших изменениях в обучающих выборках.

Методы регуляризации

Наряду с гребневой регрессией в последние годы получил распространение метод Лассо, основанный на минимизации функционала

$$Q_{lasso}(\tilde{S}_t, \beta_0, \ldots, \beta_n) = \sum_{j=1}^{m}[y_j - \beta_0 - \sum_{i=1}^{n}\beta_i \breve{x}_{ji}^s]^2 + \gamma \sum_{i=1}^{n} \mid \beta_i \mid. \quad (21)$$

Интересной особенностью метода Лассо является равенство 0 части из регрессионных коэффициентов . Однако равенство 0 коэффициента на самом деле означает исключение из модели соответствующей ему переменной. Поэтому метод Лассо не только строит оптимальную регрессионную модель, но и производит отбор переменных. Метод может быть использован для отбора переменных в условиях, когда размерность данных превышает размер выборки. Отметим, что общее число отобранных переменных не может превышать размера обучающей выборки . Эксперименты показали, что эффективность отбора переменных методом Лассо снижается, при высокой взаимной корреляции некоторых из них.

Данными недостатками не обладает другой метод построения регрессионной модели, основанный на регуляризации по Тихонову, который называется эластичная сеть. Метод эластичная сеть основан на минимизации функционала

$$Q_{elnet}(\tilde{S}_t, \beta_0, \ldots, \beta_n) = \sum_{j=1}^{m}[y_j - \beta_0 - \sum_{i=1}^{n} \beta_i \breve{x}_{ji}^s]^2 + \gamma \sum_{i=1}^{n}[\beta_i^2 \theta + (1-\theta) \mid \beta_i \mid], \tag{22}$$

где $\theta \in [0, 1]$. Метод эластичная сеть включает в себя метод гребневая регрессия и Лассо как частные случаи.