L_1-регуляризация линейной регрессии. Регрессия наименьших углов (алгоритм LARS).

L_p-регуляризация линейной регрессии

Рассмотрим классическую модель линейной регрессии:

$$t = \sum_{j=1}^{d} w_j x(j) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Поиск весов $\boldsymbol{w}$ с помощью максимизации правдоподобия выборки в этой модели эквивалентен методу наименьших квадратов:

$$\sum_{n=1}^{N} \left[t_n - \sum_{j=1}^{d} w_j x_n(j) \right]^2 = \|\boldsymbol{t} - w_1 \boldsymbol{x}_1 - \cdots - w_d \boldsymbol{x}_d\|^2 = \|\boldsymbol{t} - X\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}. \quad (1)$$

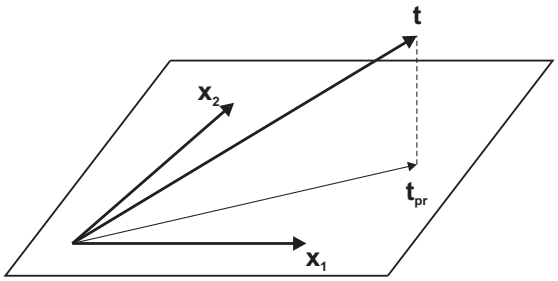

Рис. 1: Графическая иллюстрация линейной регрессии.

Здесь $\boldsymbol{x}_i \in \mathbb{R}^N$ — значения i-ого признака для всех объектов в выборке, $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d]$. Заметим, что введенное обозначение $\boldsymbol{x}_i$ отличается от стандартного, когда под $\boldsymbol{x}_i$ имеется ввиду i-ый объект выборки. Здесь и далее будем предполагать, что выборка является нормализованной, т.е. $\sum_{n=1}^{N} t_n = 0$, $\sum_{n=1}^{N} x_i(n) = 0$, $\frac{1}{N}\boldsymbol{t}^T\boldsymbol{t} = 1$, $\frac{1}{N}\boldsymbol{x}_i^T\boldsymbol{x}_i = 1$.

Задача (1) имеет простую геометрическую интерпретацию — поиск проекции вектора $\boldsymbol{t}$ на гиперплоскость с направляющими векторами $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ (см. рис. 1). Эта задача может быть решена аналитически:

$$\boldsymbol{w} = (X^T X)^{-1} X^T \boldsymbol{t}, \quad \boldsymbol{t}_{pr} = X\boldsymbol{w} = X(X^T X)^{-1} X^T \boldsymbol{t}. \quad (2)$$

Данное решение для $\boldsymbol{w}$ соответствует псевдорешению системы линейных уравнений $X\boldsymbol{w} = \boldsymbol{t}$.

Для того, чтобы избежать переобучения линейной регрессии, необходимо наложить ограничения на вариабельность решающего правила. Это можно сделать путем ограничения нормы вектора весов $\boldsymbol{w}$:

$$\begin{aligned} &\|\boldsymbol{t} - X\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}, \\ &\|\boldsymbol{w}\|_{L_p}^p \le b. \end{aligned} \quad (3)$$

Здесь под символом $\|\boldsymbol{w}\|_{L_p}$ понимается L_p норма вида $\sqrt[p]{\sum_{j=1}^{d} |w_j|^p}$. Традиционно вместо задачи оптимизации (3) рассматривают задачу оптимизации регуляризованного функционала:

$$\|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|_{L_p}^p \to \min_{\boldsymbol{w}},\ \lambda \ge 0. \quad (4)$$

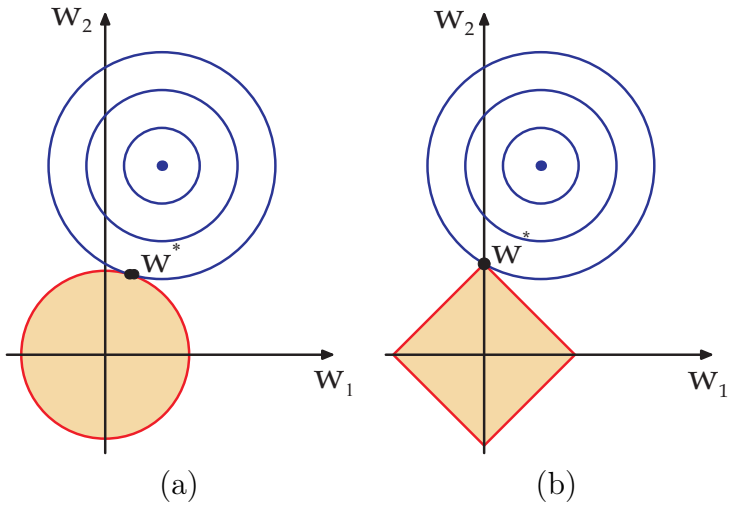

Рис. 2: Контуры минимизируемого функционала (синие кривые) и ограничения на норму весов. Случай (a) соответствует квадратичной норме, случай (b) — L_1 норме.

Нетрудно показать, что задачи оптимизации (3) и (4) эквивалентны при $p \geq 1$, т.е. когда все рассматриваемые функции являются выпуклыми. Введем функцию Лагранжа $L(\boldsymbol{w}, \lambda) = \|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \lambda(\|\boldsymbol{w}\|^p_{L_p} - b)$. Тогда согласно выпуклому варианту теоремы Куна-Таккера необходимым и достаточным условием существования решения $\hat{\boldsymbol{w}}$ в задаче (3) является наличие $\lambda \geq 0$:

1. *Принцип минимума*: $L(\hat{\boldsymbol{w}}, \lambda) = \min_{\boldsymbol{w}} L(\boldsymbol{w}, \lambda)$,
2. *Условие дополняющей нежесткости*: $\lambda(\|\hat{\boldsymbol{w}}\|^p_{L_p} - b) = 0$.

Заметим, что для достаточности условий 1 и 2 требуется также выполнение условия Слейтера, т.е. существования $\boldsymbol{w} : \|\boldsymbol{w}\|^p_{L_p} < b$. Очевидно, что это условие выполнено при $b > 0$. Задача оптимизации (4) эквивалентна условию 1. Рассмотрим выполнение условия 2. Это условие эквивалентно наступлению одного из двух событий: $\lambda = 0$ или $\|\boldsymbol{w}\|^p_{L_p} = b$. Если $\lambda = 0$, то оптимальная точка $\hat{\boldsymbol{w}}$ лежит внутри области $\|\boldsymbol{w}\|^p_{L_p} < b$. Следовательно, ограничение $\|\boldsymbol{w}\|^p_{L_p} \leq b$ становится излишним, и задача оптимизации (3) переходит в задачу оптимизации без ограничений, что эквивалентно задаче (4) при $\lambda = 0$. Пусть $\lambda > 0$, $\|\boldsymbol{w}\|^p_{L_p} = b$. Выполнения этого условия легко добиться в задаче (4), просто обозначив через b значение нормы весов, оптимальных с точки зрения задачи (4).

Рассмотрим оптимальные решения задачи (3) при различных p. Можно показать, что в случае $p \leq 1$ оптимальное решение $\hat{\boldsymbol{w}}$ обладает свойством разреженности, т.е. часть весов в точности равна нулю. В случае $p > 1$ строго нулевые веса в оптимальном решении практически невозможны. Это свойство иллюстрируется на рис. 2. Заметим, что ситуация $p = 1$ является выделенной, т.к. в этом случае оптимизируемый функционал (4) является выпуклым, а оптимальное решение обладает свойством разреженности. Метод настройки весов линейной регрессии с помощью решения задачи (3) или (4) с L_1 нормой получил название LASSO (сокращение от Least Absolute Shrinkage and Selection Operator) [1]. Именно этот метод мы и будет рассматривать ниже.

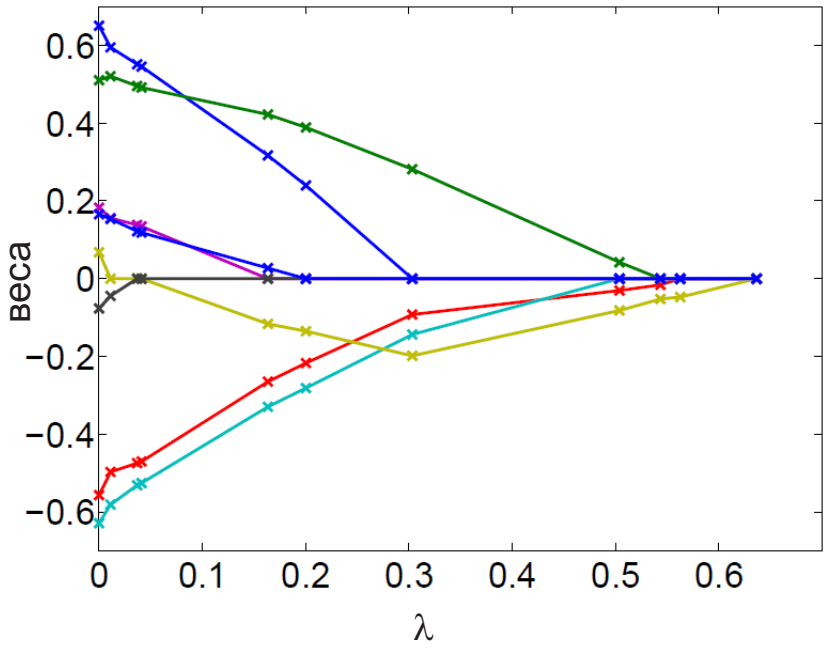

Рис. 3: «Путь регуляризации» для LASSO.

Теоретические свойства LASSO

Рассмотрим т.н. «путь регуляризации», т.е. совокупность решений $\hat{\boldsymbol{w}}$ задачи (4) для всех значений $\lambda \geq 0$. Можно показать, что этот путь является кусочно-линейным. Пусть при фиксированном λ нам известны знаки весов $\boldsymbol{s}$ в оптимальном решении $\hat{\boldsymbol{w}}$:

$$
\begin{aligned}
&s_j = 0, \text{ если } \hat{w}_j = 0,\\
&s_j = 1, \text{ если } \hat{w}_j > 0,\\
&s_j = -1, \text{ если } \hat{w}_j < 0.
\end{aligned}
$$

В этом случае задача оптимизации (4) переходит в следующую:

$$\|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \lambda \boldsymbol{s}^T \boldsymbol{w} \to \min_{\boldsymbol{w}}.$$

Данная задача может быть решена аналитически:

$$
\begin{aligned}
&\hat{\boldsymbol{w}}_J(\lambda, \boldsymbol{s}) = (X_J^T X_J)^{-1}(X_J^T \boldsymbol{t} - \frac{1}{2}\lambda \boldsymbol{s}),\\
&\hat{\boldsymbol{w}}_{J^c}(\lambda, \boldsymbol{s}) = \boldsymbol{0}.
\end{aligned}
\tag{5}
$$

Здесь $J = \{j : s_j \neq 0\}$, $J^c = \{j : s_j = 0\}$, $X_J = X_{:,J}$. Можно показать, что необходимыми и достаточными условиями оптимальности решения (5) являются следующие:

$$
\begin{aligned}
&\operatorname{sign}(\hat{\boldsymbol{w}}_J(\lambda, \boldsymbol{s})) = \boldsymbol{s}_J,\\
&\|X_{J^c}^T(X_J \hat{\boldsymbol{w}}_J(\lambda, \boldsymbol{s}) - \boldsymbol{t})\|_\infty \leq \lambda.
\end{aligned}
\tag{6}
$$

Решение (5) является линейной функцией относительно λ. Это решение в совокупности с условиями (6) показывают, что итоговый путь регуляризации является кусочно-линейным (см.

рис. 3). Таким образом, возможный алгоритм решения задачи (4) состоит в поиске точек изменения направления на этом пути.

Существует ряд теоретических результатов относительно способности метода LASSO находить истинные релевантные веса (свойство состоятельности). Пусть наблюдаемая выборка данных сгенерирована из вероятностной модели линейной регрессии с весами $\boldsymbol{w}^{true}$, где часть весов равна нулю. Назовем метод LASSO состоятельным по знакам, если при стремлении объема выборки N к бесконечности верно

$$\lim_{N\to+\infty} \mathrm{P}(\exists\lambda \geq 0 : \mathrm{sign}(\hat{\boldsymbol{w}}(\lambda)) = \mathrm{sign}(\boldsymbol{w}^{true})) = 1.$$

Здесь через $\hat{\boldsymbol{w}}(\lambda)$ обозначены веса, полученные с помощью LASSO с коэффициентом регуляризации λ. Можно показать [3], что метод LASSO является состоятельным по знакам тогда и только тогда, когда при стремлении объема выборки к бесконечности выполняется следующее условие:

$$\|Q_{J^cJ}Q_{JJ}^{-1}\mathrm{sign}(\boldsymbol{w}_J^{true})\|_\infty \leq 1. \tag{7}$$

Здесь $Q = \lim_{N\to+\infty} \frac{1}{N}X_N^T X_N$ – выборочная матрица ковариации при стремлении объема выборки к бесконечности, а $J = \{j : w_j^{true} \neq 0\}$. На практике множество J, а также знаки $\boldsymbol{w}^{true}$ являются неизвестными. Поэтому условие (7) приходится поверять для всех возможных знаков весов и всех возможных множеств J.

Рассмотрим условие, при котором требование (7) гарантированно выполнено. Обозначим через q мощность множества J, т.е. количество истинных информативных весов. Пусть матрица ковариации Q_N для выборки объема N имеет единицы на диагонали и внедиагональные элементы q_{ij}. Пусть найдется константа $0 < c \leq 1$: $|q_{ij}| \leq \frac{c}{2q-1}$. Тогда условие (7) выполнено [3].

Результат (7) является теоретическим. Экспериментальные исследования показывают, что на практике LASSO не является состоятельным по знакам [4] и, как правило, отбирает больше информативных признаков, чем нужно [5]. Для более устойчивого поиска множества информативных признаков обычно генерируется набор бутстрапированных выборок, для каждой из которых находится набор информативных признаков с помощью LASSO. Затем итоговое множество информативных признаков образуется в результате пересечения всех множеств информативных признаков [4] или путем усреднения рангов полученных признаков [6] (так как LASSO строит кусочно-линейную траекторию в пространстве весов, то помимо набора информативных признаков данный метод выдает еще и порядок вхождения признаков в информативное множество).

Алгоритм LARS (регрессия наименьших углов)

Изначательно автор LASSO предлагал решать задачу оптимизации (4) путем сведения к выпуклой задачи квадратичного программирования [1]. Для этого представим каждый вес w_j в следующем виде: $w_j = w_j^+ - w_j^-$, $w_j^+, w_j^- \geq 0$. Это представление не единственно. Рассмотрим следующую задачу оптимизации:

$$\|\boldsymbol{t} - X(\boldsymbol{w}^+ - \boldsymbol{w}^-)\|^2 + \lambda\sum_{j=1}^{d}(w_j^+ + w_j^-) \to \min_{\boldsymbol{w}^+,\boldsymbol{w}^-},$$
$$w_j^+, w_j^- \geq 0,\ j = 1,\ldots,d.$$

Легко показать, что решение данной задачи квадратичного программирования совпадает с решением LASSO. Однако, применение общих методов решения задачи квадратичного

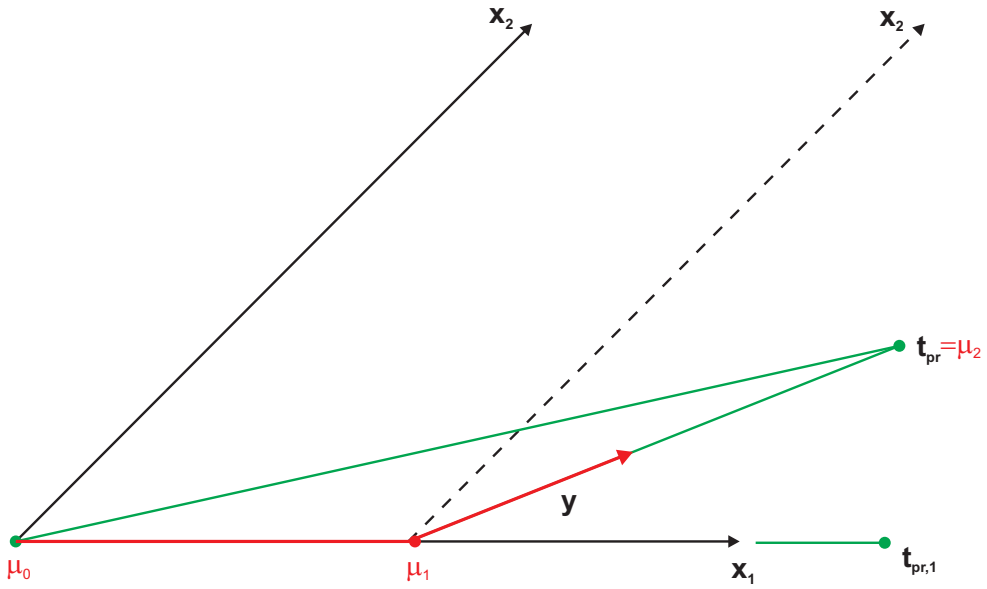

Рис. 4: Иллюстрация алгоритма LARS для двухмерного случая. В начале $\boldsymbol{\mu}_0 = \mathbf{0}$ и наибольшая корреляция между $\boldsymbol{t}-\boldsymbol{\mu}_0$ и векторами признаков достигается для $\boldsymbol{x}_1$. На первом шаге алгоритма $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \gamma_1\boldsymbol{x}_1$, где γ_1 выбрано таким образом, чтобы корреляция между $\boldsymbol{t}-\boldsymbol{\mu}_1$ и векторами признаков $\boldsymbol{x}_1, \boldsymbol{x}_2$ была бы одинакова. Далее выбирается вектор движения $\boldsymbol{y}$ такой, вдоль которого корреляция между $\boldsymbol{t}-\boldsymbol{\mu}$ и векторами признаков $\boldsymbol{x}_1, \boldsymbol{x}_2$ была бы все время одинаковой. Этот вектор соответствует биссектрисе угла, образуемый векторами $\boldsymbol{x}_1, \boldsymbol{x}_2$ с центром в $\boldsymbol{\mu}_1$. На втором шаге $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1+\gamma_2\boldsymbol{y}$. Так как все имеющиеся признаки уже включены в текущий прогноз, то величина шага γ_2 выбирается в соответствии с величиной проекции $\boldsymbol{t}-\boldsymbol{\mu}_1$ на вектор движения $\boldsymbol{y}$. Таким образом, $\boldsymbol{\mu}_2$ совпадает с оценкой наименьших квадратов $\boldsymbol{t}_{pr}$.

программирования требует значительного времени. Для метода LASSO был предложен специальный метод оптимизации [2], скорость работы которого сопоставима с поиском решения (2) (т.е. очень быстрый метод). Этот метод получил название LARS (сокращение от Least Angle Regression for laSso/Stepwise regression).

Алгоритм LARS направлен на построение оптимальной кусочно-линейной траектории в пространстве весов $\boldsymbol{w} \in \mathbb{R}^d$ (рис. 3) или прогнозов $\boldsymbol{\mu} = X\boldsymbol{w} \in \mathbb{R}^N$. Таким образом, алгоритм LARS находит решения LASSO *сразу для всех* возможных значений коэффициента регуляризации λ.

Идея алгоритма LARS состоит в следующем (см. рис. 4). Будем обозначать через $\boldsymbol{w}_j$ значение весов после j-ой итерации, а через $\boldsymbol{\mu}_j = X\boldsymbol{w}_j$ значение прозноза для вектора $\boldsymbol{t}$ после j-ой итерации. В начале $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{w}_0 = \mathbf{0}$. Найдем среди всех признаков $\boldsymbol{x}_i$ наиболее коррелированный (по модулю) с $\boldsymbol{t}$. Обозначим его через $\boldsymbol{x}_m$, а знак корреляции через s_m. На первом шаге алгоритма модифицируем текущее значение прогноза $\boldsymbol{\mu}_0$ путем движения вдоль наиболее коррелированного вектора с учетом знака корреляции $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \gamma s_m\boldsymbol{x}_m$, $\gamma > 0$. В результате такого движения величина корреляции между текущим остатком $\boldsymbol{t}-\boldsymbol{\mu}_1$ и признаком $\boldsymbol{x}_m$ будет уменьшаться по мере увеличения шага γ. Обозначим эту корреляцию через r_γ. Будем двигаться в выбранном направлении до тех пор, пока корреляция r_γ не совпадет с корреляцией между $\boldsymbol{t}-\boldsymbol{\mu}_1$ и каким-нибудь еще признаком $\boldsymbol{x}_k$. Затем, на втором шаге алгоритма, будем модифицировать текущий прогноз $\boldsymbol{\mu}_1$ путем движения вдоль такого направления $\boldsymbol{y}_1$, что корреляция между текущим остатком $\boldsymbol{t}-\boldsymbol{\mu}_2$ и каждым из множества активных признаков $\boldsymbol{x}_m, \boldsymbol{x}_k$

была бы одинакова. Будем двигаться вдоль выбранного направления, пока еще один признак не окажется равно коррелированным с текущим остатком, что и признаки из активного множества. Тогда добавим этот признак в активное множество и найдем новое эквикоррелированное направление $\boldsymbol{y}_2$, вдоль которого все признаки из активного множества являются одинаково коррелированными с текущим остатком. Алгоритм заканчивает работу, когда все имеющиеся (линейно независимые) признаки входят в активное множество. На последнем шаге алгоритма величина шага определяется значением корреляции между текущим остатком $\boldsymbol{t} - \boldsymbol{\mu}_{d-1}$ и последним эквикоррелированным направлением $\boldsymbol{y}_{d-1}$. При этом $\boldsymbol{\mu}_d = \boldsymbol{t}_{pr}$, где $\boldsymbol{t}_{pr}$ является оценкой наименьших квадратов. Количество шагов алгоритма LARS определяется количеством признаков d.

Рассмотрим теперь шаги алгоритма LARS более подробно. Обозначим через $R = \frac{1}{N}X^TX$ выборочную матрицу ковариации выборки. Так как все вектора признаков $\boldsymbol{x}_i$ считаются нормализованными, то значения в этой матрице R_{ij} определяют корреляцию между векторами $\boldsymbol{x}_i$ и $\boldsymbol{x}_j$, а $R_{ii} = 1$. Обозначим через $\boldsymbol{r} = \frac{1}{N}X^T\boldsymbol{t}$ корреляцию между векторами признаков и целевым вектором $\boldsymbol{t}$.

Поиск оптимальной величины шага γ на первой итерации. Пусть на первом шаге алгоритма LARS $m = \arg\max_j |r_j|$, $s_m = \operatorname{sign}(r_m)$. Найдем значение шага γ вдоль направления $s_m\boldsymbol{x}_m$. Корреляция между текущим остатком $\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m$ и вектором $s_m\boldsymbol{x}_m$ составляет

$$\operatorname{corr}(\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m, s_m\boldsymbol{x}_m) = \frac{\frac{1}{N}(\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m)^T s_m\boldsymbol{x}_m}{\sqrt{\mathbb{D}(\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m)}} = \frac{s_m r_m - \gamma}{\sqrt{\mathbb{D}(\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m)}}.$$

Здесь через $\mathbb{D}(\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m)$ обозначена выборочная дисперсия случайной величины $t - \gamma s_m x_m$. Найдем корреляцию между текущим остатком и вектором $\boldsymbol{x}_j$, $j \neq m$:

$$\operatorname{corr}(\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m, \boldsymbol{x}_j) = \frac{r_j - \gamma s_m R_{mj}}{\sqrt{\mathbb{D}(\boldsymbol{t} - \gamma s_m\boldsymbol{x}_m)}}.$$

Приравнивая корреляции между собой, получаем:

$$\gamma_j^+ = \frac{s_m r_m - r_j}{1 - s_m R_{mj}}.$$

Вычисляя корреляцию между текущим остатком и вектором $-\boldsymbol{x}_j$, $j \neq m$, получаем:

$$\gamma_j^- = \frac{s_m r_m + r_j}{1 + s_m R_{mj}}.$$

Таким образом, оптимальная величина шага $\gamma = \min_{j\neq m}\gamma_j$, где $\gamma_j = \min^+(\gamma_j^+, \gamma_j^-)$. Под символом $\min^+$ понимается минимум среди положительных аргументов. После выполнения первого шага в активное множество признаков $\{s_m\boldsymbol{x}_m\}$ добавляется признак $s_k\boldsymbol{x}_k$, где $k = \arg\min_{j\neq m}\gamma_j$, $s_k = 1$, если $\gamma_k = \gamma_k^+$, $s_k = -1$ в противном случае.

Поиск эквикоррелированного направления $\boldsymbol{y}$. Пусть J – текущее активное множество признаков, $X_J^{\boldsymbol{s}} = [\dots s_j\boldsymbol{x}_j \dots]$ – признаки из активного множества с учетом знаков $\boldsymbol{s}$, $\boldsymbol{\mu}$ – текущий прогноз. Тогда текущий эквикоррелированный вектор $\boldsymbol{y}$ можно найти из следующих трех соображений. Во-первых, вектор $\boldsymbol{y}$ является линейной комбинацией признаков из активного множества:

$$\boldsymbol{y} = X_J^{\boldsymbol{s}}\boldsymbol{v}_J. \tag{8}$$

Здесь через $\boldsymbol{v}_J$ обозначены веса этой линейной комбинации. Во-вторых, можно считать, что y имеет единичную дисперсию, т.к. нас интересует только эквикоррелированное направление:

$$\frac{1}{N}\boldsymbol{y}^T\boldsymbol{y} = 1. \tag{9}$$

В-третьих, по определению эквикоррелированного направления $\boldsymbol{y}$ должен иметь одинаковую корреляцию (обозначим ее через a) со всеми векторами признаков из активного множества, т.е.

$$\frac{1}{N}(X_J^{\boldsymbol{s}})^T\boldsymbol{y} = a\mathbf{1}_J. \tag{10}$$

Здесь через $\mathbf{1}_J$ обозначен вектор, состоящий из единиц. Подставляя (8) в (10), получаем

$$\frac{1}{N}(X_J^{\boldsymbol{s}})^T X_J^{\boldsymbol{s}}\boldsymbol{v}_J = a\mathbf{1}_J. \tag{11}$$

Матрица $R_J^{\boldsymbol{s}} = \frac{1}{N}(X_J^{\boldsymbol{s}})^T X_J^{\boldsymbol{s}}$ является матрицей корреляций между векторами признаков из активного множества с учетом их знаков. Эта матрица может быть получена из подматрицы $R_{J,J}$ следующим образом: $R_J^{\boldsymbol{s}} = D_J R_J D_J$, где $D_J = \mathrm{diag}(\boldsymbol{s}_J)$. Теперь мы можем найти $\boldsymbol{v}_J$ из условия (11):

$$\boldsymbol{v}_J = a(D_J R_J D_J)^{-1}\mathbf{1}_J. \tag{12}$$

Подставляя (8) в (9), получаем

$$\boldsymbol{v}_J^T\left[\frac{1}{N}(X_J^{\boldsymbol{s}})^T X_J^{\boldsymbol{s}}\right]\boldsymbol{v}_J = 1.$$

Объединяя этот результат с (12), находим a:

$$a = \frac{1}{\sqrt{\mathbf{1}_J^T(D_J R_J D_J)^{-1}\mathbf{1}_J}}.$$

Подставляя найденное значение a в (12), находим $\boldsymbol{v}_J$, по которому в свою очередь из условия (8) находим $\boldsymbol{y}$. Обозначим через a_j корреляцию между $\boldsymbol{y}$ и $\boldsymbol{x}_j$, $j \notin J$. Она может быть вычислена как

$$a_j = \frac{1}{N}\boldsymbol{x}_j^T\boldsymbol{y} = \boldsymbol{x}_j^T\frac{1}{N}X_J^{\boldsymbol{s}}\boldsymbol{v}_J = (D_J R_{Jj})^T\boldsymbol{v}_J.$$

Поиск оптимальной величины шага γ вдоль эквикоррелированного направления $\boldsymbol{y}$. Пусть $\boldsymbol{\mu}$ – текущий прогноз, $\boldsymbol{y}$ – текущее эквикоррелированное направление. Нам необходимо вычислить величину шага γ вдоль эквикоррелированного направления для получения нового прогноза $\boldsymbol{\mu}^{new} = \boldsymbol{\mu} + \gamma\boldsymbol{y}$. Обозначим через r корреляцию признаков из активного множества с текущим остатком $\boldsymbol{t} - \boldsymbol{\mu}$, а через r_j – корреляцию остальных признаков с текущим остатком. Тогда

$$\forall m \in J:\ \mathrm{corr}(\boldsymbol{t} - \boldsymbol{\mu} - \gamma\boldsymbol{y}, \boldsymbol{x}_m) = \frac{r - \gamma a}{\sqrt{D(\boldsymbol{t} - \boldsymbol{\mu} - \gamma\boldsymbol{y})}},$$
$$\forall j \notin J:\ \mathrm{corr}(\boldsymbol{t} - \boldsymbol{\mu} - \gamma\boldsymbol{y}, \boldsymbol{x}_j) = \frac{r_j - \gamma a_j}{\sqrt{D(\boldsymbol{t} - \boldsymbol{\mu} - \gamma\boldsymbol{y})}},$$
$$\forall j \notin J:\ \mathrm{corr}(\boldsymbol{t} - \boldsymbol{\mu} - \gamma\boldsymbol{y}, -\boldsymbol{x}_j) = \frac{-r_j + \gamma a_j}{\sqrt{D(\boldsymbol{t} - \boldsymbol{\mu} - \gamma\boldsymbol{y})}}.$$

Алгоритм 1 Алгоритм LARS

Вход: Коэффициенты корреляций между нормированными признаками $R \in \mathbb{R}^{d\times d}$, коэффициенты корреляций между нормированными признаками и нормированным целевым вектором $\boldsymbol{r} \in \mathbb{R}^d$.

Выход: Набор информативных признаков J в порядке убывания информативности, их знаки $\boldsymbol{s}$ и веса (точки излома в пути регуляризации) $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K$.

1: Найти максимальное количество линейно независимых признаков $d_{max} = \text{rank}(R)$.
2: Инициализировать активное множество $J = \emptyset$, знаки признаков $s_J = \emptyset$ и веса $\boldsymbol{w} = \boldsymbol{0}$.
3: Найти $m = \arg\max_j |r_j|$, $s_m = \text{sign}(r_m)$. Найти корреляцию признака m с текущим остатком $r_0 = s_m r_m$.
4: **пока** $|J| < d_{max}$
5: Добавить признак m в активное множество $J = J \cup \{m\}$, $s_J = s_J \cup \{s_m\}$.
6: Найти $a = [\boldsymbol{1}_J^T (D_J R_J D_J)^{-1} \boldsymbol{1}_J]^{-1/2}$, где $R_J = R_{J,J}$, $D_J = \text{diag}(s_J)$, $\boldsymbol{1}_J$ – вектор из единиц.
7: Найти $\boldsymbol{v}_J = a(D_J R_J D_J)^{-1}\boldsymbol{1}_J$, а также корреляцию между неактивными признаками и текущим эквикоррелированным направлением $a_j = (D_J R_{jJ})^T \boldsymbol{v}_J,\ j \in J^c$. // В том случае, если множество J состоит только из одного признака, то $a = 1, a_j = s_m R_{jm}$.
8: Для $j \in J^c$ вычислить $\gamma_j^+ = (r_0 - r_j)/(a - a_j)$, $\gamma_j^- = (r_0 + r_j)/(a + a_j)$, $\gamma_j = \min^+\{\gamma_j^+, \gamma_j^-\}$. Найти $m = \arg\min_{j\in J^c} \gamma_j$. Если $\gamma_m = \gamma_m^+$, то $s_m = 1$, иначе $s_m = -1$.
9: Найти новые веса $\boldsymbol{w}^{new} = \boldsymbol{w}^{old} + \gamma_m D_J \boldsymbol{v}_J$.
10: Найти корреляцию признаков из активного множества с текущим остатком $r_0 = r_0 - \gamma_m a$ и корреляцию неактивных признаков с текущим остатком $r_j = r_j - \gamma_m a_j$.
11: На последнем шаге веса совпадают с оценкой наименьших квадратов $\boldsymbol{w} = R^{-1}\boldsymbol{r}$.

Приравнивая соответствующие корреляции между собой, получаем:

$$\gamma = \min_{j\notin J} \gamma_j, \quad \gamma_j = \min^+(\gamma_j^+, \gamma_j^-),$$
$$\gamma_j^+ = \frac{r - r_j}{a - a_j},$$
$$\gamma_j^- = \frac{r + r_j}{a + a_j}.$$

Помимо величины шага γ мы находим также признак $\boldsymbol{x}_k$, который присоединяется к активному множеству, и его знак s_k: $k = \arg\min_{j\notin J} \gamma_j$, $s_k = 1$, если $\gamma_k = \gamma_k^+$, $s_k = -1$ в противном случае.

Остановка процесса оптимизации в алгоритме LARS. Очевидно, что в процессе оптимизации алгоритма LARS имеет смысл рассматривать только линейно независимые признаки. Только в этом случае матрица R_J будет невырожденной, а вычисления для a и $\boldsymbol{v}_J$ будут корректными. Поэтому в алгоритме LARS множество активных признаков J наращивается до мощности d_{max}, где $d_{max} = \text{rank}(R)$. При этом на последнем шаге алгоритма оптимальные веса определяются через оценку наименьших квадратов $\boldsymbol{w} = (X_J^T X_J)^{-1} X_J^T \boldsymbol{t}$. Общая схема алгоритма LARS представлена в алгоритме 1.

Модификация LARS/LASSO. Алгоритм LARS, описанный выше, не решает задачу оптимизации для LASSO. Можно показать, что для LASSO справедливо следующее утверждение. Пусть $\boldsymbol{\mu} = X\boldsymbol{w}$ является решением LASSO для некоторого параметра

регуляризации λ. Тогда

$$\text{sign}(w_j) = s_j = \text{sign}(\boldsymbol{t} - \boldsymbol{\mu})^T \boldsymbol{x}_j,\ j \in J,$$

т.е. знаки весов оптимального решения совпадают со знаками корреляций между текущим остатком и активными признаками. Метод LARS, вообще говоря, не гарантирует выполнения данного условия. Однако, можно предложить небольшую модификацию алгоритма LARS, которая будет решать задачу оптимизации для LASSO.

Пусть на текущем шаге алгоритма LARS вычислено $\boldsymbol{v}_J$ и новые веса $\boldsymbol{w}^{new}(\gamma)$ определяются как $\boldsymbol{w}^{old} + \gamma D_J \boldsymbol{v}_J$. Таким образом, в точке $\gamma_j = -w_j^{old}/(s_j v_j)$ новый вес w_j^{new} меняет знак относительно w_j^{old}. Самое первое подобное изменение происходит для веса $w_{\tilde{j}}^{new}$ для величины шага

$$\tilde{\gamma} = \min_j^+ \gamma_j,\ \tilde{j} = \arg\min_j^+ \gamma_j.$$

Если все $\gamma_j < 0$, то положим по определению $\tilde{\gamma} = +\infty$. Тогда сделаем следующую модификацию алгоритма LARS:

$$\text{Если } \tilde{\gamma} < \gamma_m,\ \text{то } \boldsymbol{w}^{new} = \boldsymbol{w}^{old} + \tilde{\gamma} D_J \boldsymbol{v}_J,\ J = J \backslash \{\tilde{j}\}.$$

Здесь γ_m – оптимальная величина шага по эквикоррелированному направлению из алгоритма LARS. Можно показать [2], что с учетом данной модификации алгоритм LARS решает задачу оптимизации для LASSO сразу для всех значений коэффициента регуляризации λ.

Заметим, что в описанной модификации LARS/LASSO в некоторых ситуациях происходит уменьшение множества активных признаков. Подобная флуктуация мощности множества активных признаков увеличивает время работы алгоритма LARS/LASSO. В том случае, если нас интересует только само множество информативных признаков, а не значения весов оптимальной линейной комбинации, то лучше использовать исходный алгоритм LARS, который гарантированно сходится за d_{max} шагов.

Список литературы

[1] *R. Tibshirani.* Regression Shrinkage and Selection via LASSO // Journal of the Royal Statistical Society, V. 58, I. 1, 1996, pp. 267–288.

[2] *B. Efron, T. Hastie, I. Johnstone and R. Tibshirani.* Least Angle Regression // Annals of Statistics, V. 32, No. 2, 2004, pp. 407–499.

[3] *P. Zhao, B. Yu.* On Model Selection Consistency of Lasso // Journal of Machine Learning Research, V. 7, 2006, pp. 2541–2563.

[4] *F. Bach.* Bolasso: model consistent lasso estimation through the bootstrap // ICML, 2008.

[5] *J. Lv, Y. Fan.* A unified approach to model selection and sparse recovery using regularized least squares // Annals of Statistics, V. 37, No. 6A, 2009, pp. 3498–3528.

[6] *S. van Aelst, J.A. Khan and R.H. Zamar.* Robust Linear Model Selection Based on Least Angle Regression // Journal of the American Statistical Association, V. 102, 2007, pp. 1289–1299.