Курс: Методы оптимизации в машинном обучении, 2012

Методы оптимизации для разреженных линейных моделей

Разреженные линейные модели классификации и регрессии

Рассмотрим стандартную задачу восстановления регрессии. Имеется обучающая выборка из N объектов $\{t_n, \boldsymbol{x}_n\}_{n=1}^N$, где $\boldsymbol{x}_n \in \mathbb{R}^D$ – признаки объекта, а $t_n \in \mathbb{R}$ – целевая переменная. Требуется на основе данной информации спрогнозировать значение целевой переменной $\hat{t}$ для нового объекта, представленного своим вектором признаков $\boldsymbol{x}$.

В линейных регрессионных моделях прогноз целевой переменной осуществляется с помощью линейной функции

$$\hat{t}(\boldsymbol{x}) = \sum_{d=1}^{D} w_d x_d = \boldsymbol{w}^T \boldsymbol{x}, \tag{1}$$

где $\boldsymbol{w} \in \mathbb{R}^D$ – набор некоторых весов. Обучение весов $\boldsymbol{w}$ в модели L_1-регуляризованной линейной регрессии происходит путем решения задачи

$$F(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 + \lambda\sum_{d=1}^{D}|w_d| = \frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_1 \to \min_{\boldsymbol{w}}. \tag{2}$$

Здесь $\lambda \geq 0$ – параметр регуляризации, устанавливаемый пользователем. В решении задачи (2) часть компонент $\boldsymbol{w}$, как правило, равны нулю. С точки зрения решающего правила (1) нулевые веса отвечают исключению соответствующего признака из прогнозирующей модели. Линейные модели, в которых в результате обучения часть компонент могут иметь нулевой вес, получили название разреженных.

Рассмотрим подробнее причины того, почему в решении задачи (2) часть компонент $\boldsymbol{w}$, как правило, оказывается равным нулю. Для этого рассмотрим следующую задачу оптимизации:

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}, \\ &\|\boldsymbol{w}\|_1 \leq \eta. \end{aligned} \tag{3}$$

Здесь $\eta \geq 0$ – некоторый параметр. Задача (3) является выпуклой задачей оптимизации, т.к. минимизируемый функционал является выпуклым, а допустимое множество $\{\boldsymbol{w} \mid \|\boldsymbol{w}\|_1 \leq \eta\}$ также является выпуклым, т.к. любая норма является выпуклой функцией. Отсюда $\hat{\boldsymbol{w}}$ является решением задачи (3) тогда и только тогда, когда найдется $\hat{\lambda} \geq 0$ такое, что пара $(\hat{\boldsymbol{w}}, \hat{\lambda})$ удовлетворяет условиям Куна-Таккера:

$$\begin{aligned} &\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \hat{\lambda}\|\boldsymbol{w}\|_1, \\ &\|\hat{\boldsymbol{w}}\|_1 \leq \eta, \\ &\hat{\lambda}(\|\hat{\boldsymbol{w}}\|_1 - \eta) = 0. \end{aligned} \tag{4}$$

Из условий (4) следует эквивалентность задач (2) и (3), где между параметрами λ и η существует неявное соответствие. Действительно, пусть $\hat{\boldsymbol{w}}$ является решением задачи (3) для некоторого $\hat{\eta}$. Тогда для $\hat{\boldsymbol{w}}$ найдется $\hat{\lambda}$ такое, что пара $(\hat{\boldsymbol{w}}, \hat{\lambda})$ удовлетворяет условиям (4). Это означает, что $\hat{\boldsymbol{w}}$ является решением задачи (2) для $\lambda = \hat{\lambda}$. Обратно, пусть $\hat{\boldsymbol{w}}$ является решением задачи (2) для некоторого $\hat{\lambda}$. Тогда при $\hat{\eta} = \|\hat{\boldsymbol{w}}\|_1$ набор $(\hat{\boldsymbol{w}}, \hat{\lambda})$ будет удовлетворять условиям (4), т.е. $\hat{\boldsymbol{w}}$ будет решением задачи (3) для $\eta = \hat{\eta}$. Очевидно, что соответствие между λ и η в задачах (2) и (3) носит характер обратной пропорциональности.

Рассмотрим графическую иллюстрацию решения задачи (3) с двухмерном пространстве (см. рис. 1, слева). Допустимое множество, определяемое ограничением $\|\boldsymbol{w}\|_1 \leq \eta$, представляет собой ромб с центром в нуле. Линии уровня минимизируемого функционала $\frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2$ представляют собой эллипсы. В результате решение задачи (3) определяется точкой касания соответствующей линии уровня и ромба. Для большинства вариантов расположения эллипсов в пространстве точка касания будет попадать на одну из вершин ромба, что соответствует обнулению одного из двух весов. В многомерном случае рассуждения аналогичны.

В задаче оптимизации (2) минимизируемый функционал является выпуклым, но негладким. Введем новые переменные $\boldsymbol{w}^+, \boldsymbol{w}^- \in \mathbb{R}^D$ такие, что $w_d^+ \geq 0, w_d^- \geq 0$, $w_d = w_d^+ - w_d^-$ $\forall d = \overline{1, D}$, и рассмотрим задачу

$$\begin{aligned} &\frac{1}{2}\sum_{n=1}^{N}(t_n - (\boldsymbol{w}^+ - \boldsymbol{w}^-)^T\boldsymbol{x}_n)^2 + \lambda\sum_{d=1}^{D}(w_d^+ + w_d^-) \to \min_{\boldsymbol{w}^+, \boldsymbol{w}^-}, \\ &w_d^+ \geq 0,\ w_d^- \geq 0,\ d = \overline{1, D}. \end{aligned} \tag{5}$$

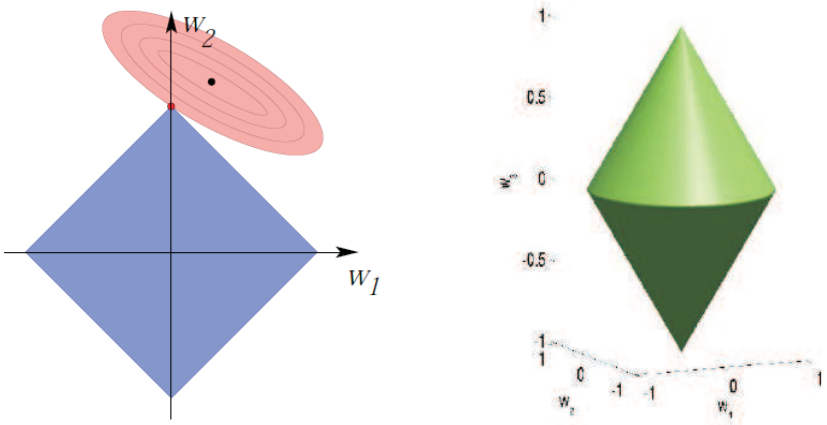

Рис. 1: Слева: иллюстрация получения разреженного решения в задаче (2), справа: ограничение вида $\sqrt{w_1^2 + w_2^2} + |w_3| \leq \eta$. Иллюстрации взяты из слайдов лекции по разреженным моделям с конференции ECML 2010 http://www.di.ens.fr/~fbach/ecml2010tutorial/

Можно показать, что задача (5) эквивалентна задаче (2), но, в отличие от последней, является выпуклой и гладкой. Для решения (5) можно воспользоваться методом внутренней точки.

Как известно, при решении задач условной оптимизации в точке оптимума часть ограничений вида неравенства переходит в равенство (из-за условий дополняющей нежесткости). Для задачи (5) это соответствует тому, что, возможно, для некоторых d $w_d^+ = w_d^- = 0$, т.е. $w_d = 0$. Таким образом, мы получили еще одно объяснение тому факту, что задача (2) соответствует получению разреженной линейной модели.

В некоторых практических задачах актуальным является свойство т.н. групповой разреженности. Пусть все признаки $\boldsymbol{x}$ разбиты на набор непересекающихся групп. Обозначим через $\boldsymbol{x}_g$ признаки, относящиеся к группе g, а через $\boldsymbol{w}_g$ – соответствующие им веса в линейной регрессии. При обучении линейных регрессионных моделей со свойством групповой разреженности веса некоторых групп $\boldsymbol{w}_g$ становятся равными нулю. Таким образом, здесь происходит не исключение отдельных признаков из прогнозирующей модели, а исключение целых групп признаков.

Одним из примеров ситуации, в которой может использоваться групповая разреженность, является наличие номинальных признаков и кодирование их значений в виде векторов бинарных признаков. Тогда бинарные признаки, относящиеся к кодировке одного номинального признака, имеет смысл объединить в одну группу. Другим примером ситуации с групповой разреженностью является часто используемый в анализе изображений подход к описанию изображения в виде набора дескрипторов, вычисляемых для каждого прямоугольного блока изображения. Тогда в одну группу могут объединяться все дескрипторы, вычисленные для фиксированного блока изображения, или все дескрипторы определенного типа, вычисленные для всех блоков изображения.

Обобщением подхода L_1-регуляризации на случай групповой разреженности является т.н. L_1/L_2-регуляризация, определяемая как

$$\|\boldsymbol{w}\|_{1/2} = \sum_{g=1}^{G} \|\boldsymbol{w}_g\|_2.$$

Здесь через $\|\boldsymbol{w}_g\|_2$ обозначена обычная евклидова норма вектора весов, относящихся к группе g. Теперь при обучении линейной регрессии со свойством групповой разреженности вместо задачи (2) решается задача

$$F(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_{1/2} \to \min_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_G}. \tag{6}$$

Заметим, что если все группы состоят из одного признака, то задача (6) переходит в задачу (2). Покажем, что решение задачи (6) действительно позволяет обнулять векторы весов для некоторых групп. Для этого, как и в случае задачи (2), рассмотрим эквивалентную (6) задачу условной оптимизации:

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}, \\ &\|\boldsymbol{w}\|_{1/2} \leq \eta. \end{aligned} \tag{7}$$

Пусть имеется три признака, первые два из которых образуют первую группу, а третий признак – вторую группу. Тогда ограничение вида $\|\boldsymbol{w}\|_{1/2} \leq \eta$ соответствует «волчку», показанному на рис. 1, справа. Линии уровня минимизируемого функционала $\frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2$ представляют собой трехмерные эллипсоиды, а решение задачи (7) определяется точкой касания между соответствующим эллипсоидом и волчком. Если точка касания

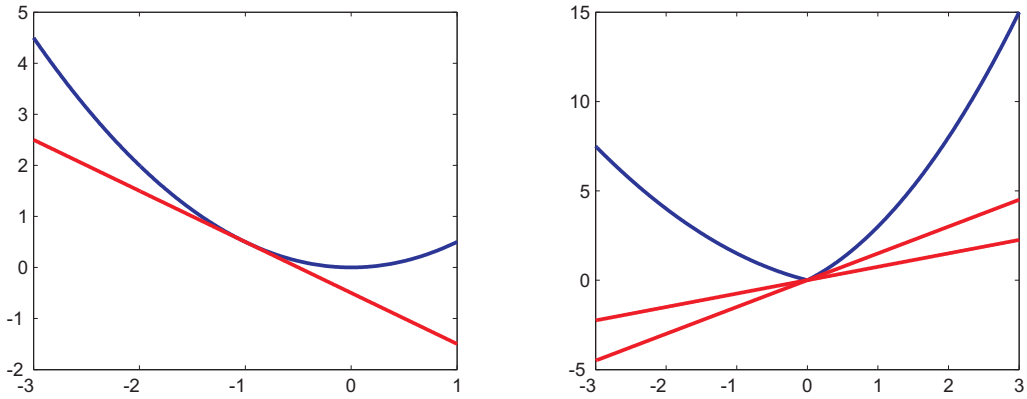

Рис. 2: Иллюстрация понятия субградиента. Слева: касательная является нижней границей для выпуклой непрерывно-дифференцируемой функции, справа: для выпуклых негладких функций в точках недифференцируемости существует множество касательных, являющихся нижней границей.

приходится на одну из вершин волчка, то первая группа признаков признается неинформативной и исключается из линейной модели. Если точка касания приходится на обод волчка, то вторая группа признаков признается неинформативной.

Заметим, что при использовании L_1/L_2-регуляризации требуется, чтобы набор групп признаков был непересекающимся. В практических задачах признаки могут образовывать более сложные группы с пересечениями, с наличием иерархии групп и т.д. О способах учета таких структурных разреженностей, соответствующих приложениях и методах оптимизации можно прочитать здесь.

Рассмотрим теперь стандартную задачу классификации на два класса. Имеется обучающая выборка из N объектов $\{t_n, \boldsymbol{x}_n\}_{n=1}^N$, где $\boldsymbol{x}_n \in \mathbb{R}^D$ – признаки объекта, а $t_n \in \{-1, 1\}$ – метка класса. Требуется на основе данной выборки спрогнозировать метку класса $\hat{t}$ для нового объекта, представленного своим вектором признаков $\boldsymbol{x}$.

В линейных моделях классификации прогноз метки класса осуществляется по знаку линейной функции

$$\hat{t}(\boldsymbol{x}) = \begin{cases} +1, \text{ если } \boldsymbol{w}^T\boldsymbol{x} > 0, \\ -1, \text{ иначе.} \end{cases} \tag{8}$$

Здесь $\boldsymbol{w} \in \mathbb{R}^D$ – набор некоторых весов. Для обучения L_1-регуляризованной логистической регрессии решается задача

$$F(\boldsymbol{w}) = \sum_{n=1}^{N} \log(1 + \exp(-t_n \boldsymbol{w}^T \boldsymbol{x}_n)) + \lambda \|\boldsymbol{w}\|_1 \to \min_{\boldsymbol{w}}, \tag{9}$$

где величина $\log(1+\exp(-t_n\boldsymbol{w}^T\boldsymbol{x}_n))$ играет роль функции потерь при значении прогнозирующей функции $\boldsymbol{w}^T\boldsymbol{x}_n$ для истинного класса t_n. Как и раньше, использование L_1-регуляризации позволяет получить разреженное решение для весов $\boldsymbol{w}$. С точки зрения решающего правила (8) нулевые веса означают исключение соответствующего признака из модели. Аналогично случаю линейной регрессии, для получения модели классификации со свойством групповой разреженности в задаче (9) регуляризатор $\|\boldsymbol{w}\|_1$ меняется на $\|\boldsymbol{w}\|_{1/2}$.

Субградиент выпуклой функции

В задачах оптимизации (2),(6),(9), возникающих при обучении разреженных линейных моделей классификации и регрессии, минимизируемый функционал является выпуклым, но негладким. Рассмотрим необходимое и достаточное условие экстремума для задач безусловной выпуклой негладкой оптимизации вида

$$F(\boldsymbol{w}) \to \min_{\boldsymbol{w}}. \tag{10}$$

Если функция F является дифференцируемой в точке $\boldsymbol{w}_0$, то касательная прямая вида $F(\boldsymbol{w}_0)+\nabla F(\boldsymbol{w}_0)^T(\boldsymbol{w}-\boldsymbol{w}_0)$ является глобальной нижней границей для функции $F(\boldsymbol{w})$ (см. рис. 2, слева). Угол наклона этой касательной определяется значением градиента $\nabla F(\boldsymbol{w}_0)$. Если выпуклая функция не является дифференцируемой в точке $\boldsymbol{w}_0$, то найдется множество линейных функций вида $F(\boldsymbol{w}_0) + \boldsymbol{z}^T(\boldsymbol{w} - \boldsymbol{w}_0)$, являющихся глобальными нижними границами для $F(\boldsymbol{w})$ (см. рис. 2, справа). Назовем множество таких векторов $\boldsymbol{z}$ субградиентом:

$$\partial F(\boldsymbol{w}_0) = \{\boldsymbol{z} \in \mathbb{R}^D \mid F(\boldsymbol{w}) \geq F(\boldsymbol{w}_0) + \boldsymbol{z}^T(\boldsymbol{w} - \boldsymbol{w}_0)\ \forall \boldsymbol{w}\}. \tag{11}$$

Заметим, что понятие субградиента определяется только для выпуклой функции, а в точках дифференцируемости функции F субградиент совпадает с градиентом. Вектор $\boldsymbol{w}_0$ является решением задачи (10) тогда и только тогда, когда $\mathbf{0} \in \partial F(\boldsymbol{w}_0)$[1].

Рассмотрим для примера вычисление субградиента одномерной функции $F(w) = |w|$. Если $w < 0$, то $F(w)$ является дифференцируемой функцией, и значение субградиента определяется производной $F'(w) = -1$. Аналогичные рассуждения верны для случая $w > 0$, где $F'(w) = 1$. Если $w = 0$, то субградиентом является любое $z \in \mathbb{R}$ такое, что $|w| \geq zw$. Решая данное неравенство, получаем, что $|z| \leq 1$. Таким образом,

$$\partial|w| = \begin{cases} 1, \text{ если } w > 0, \\ -1, \text{ если } w < 0, \\ [-1, 1], \text{ если } w = 0. \end{cases}$$

Рассмотрим теперь необходимое и достаточное условие оптимальности для задачи обучения разреженной линейной регрессии (2). Для этого найдем каждую компоненту субградиента минимизируемой функции:

$$\partial\left(\frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_1\right)_d = \boldsymbol{x}_d^T(X\boldsymbol{w} - \boldsymbol{t}) + \lambda\partial|w_d| = \begin{cases} \boldsymbol{x}_d^T(X\boldsymbol{w} - \boldsymbol{t}) + \lambda, \text{ если } w_d > 0, \\ \boldsymbol{x}_d^T(X\boldsymbol{w} - \boldsymbol{t}) - \lambda, \text{ если } w_d < 0, \\ \boldsymbol{x}_d^T(X\boldsymbol{w} - \boldsymbol{t}) + [-\lambda, \lambda], \text{ если } w_d = 0. \end{cases}$$

Здесь через $\boldsymbol{x}_d \in \mathbb{R}^N$ обозначен вектор значений d-го признака для всех объектов обучения. Теперь легко записать условие того, что $\hat{\boldsymbol{w}}$ является решением задачи (2):

$$\begin{aligned} &\boldsymbol{x}_d^T(X\hat{\boldsymbol{w}} - \boldsymbol{t}) = -\lambda, \ \hat{w}_d > 0, \\ &\boldsymbol{x}_d^T(X\hat{\boldsymbol{w}} - \boldsymbol{t}) = \lambda, \ \hat{w}_d < 0, \\ &|\boldsymbol{x}_d^T(X\hat{\boldsymbol{w}} - \boldsymbol{t})| \leq \lambda, \ \hat{w}_d = 0. \end{aligned}$$

В частности, нулевой вектор $\hat{\boldsymbol{w}} = \mathbf{0}$ будет решением (2), если $|\boldsymbol{x}_d^T\boldsymbol{t}| \leq \lambda \ \forall d$, что эквивалентно $\|X^T\boldsymbol{t}\|_\infty \leq \lambda$.

Проксимальный метод

Рассмотрим задачу (10), где $F(\boldsymbol{w})$ является выпуклой, но, возможно, негладкой функцией. В проксимальном методе оптимизации на шаге k предлагается решать следующую задачу:

$$\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w}}\left(F(\boldsymbol{w}_k) + \partial F(\boldsymbol{w}_k)^T(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{L}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2\right), \tag{12}$$

где $\partial F(\boldsymbol{w}_k)$ обозначает субградиент. Таким образом, здесь рассматривается локальная линеаризация функции F в окрестности $\boldsymbol{w}_k$ с добавлением т.н. проксимального слагаемого $\frac{L}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2$. Задача (12) эквивалентна задаче условной оптимизации

$$\begin{aligned} &\partial F(\boldsymbol{w}_k)^T(\boldsymbol{w} - \boldsymbol{w}_k) \to \min_{\boldsymbol{w}}, \\ &\|\boldsymbol{w} - \boldsymbol{w}_k\|^2 \leq \eta, \end{aligned}$$

где между L и η есть неявное соответствие. В результате добавление проксимального слагаемого означает требование поиска следующей точки итерационного процесса $\boldsymbol{w}_{k+1}$ в окрестности $\boldsymbol{w}_k$. Это соответствует ограничению на большие шаги в оптимизационном процессе, которые, как правило, являются нежелательными. Величина L при этом задает неявно размер допустимой окрестности $\boldsymbol{w}_k$.

Задача (12) может быть решена аналитически:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{1}{L}\partial F(\boldsymbol{w}_k).$$

Таким образом, проксимальный метод для задачи (10) совпадает с методом субградиентного спуска с фиксированной величиной шага $1/L$. Как известно, скорость сходимости метода градиентного спуска существенно зависит от стратегии выбора длины шага[2]. Аналогично, скорость работы проксимального метода зависит от выбора L.

Рассмотрим теперь применение проксимального метода для решения задачи

$$F(\boldsymbol{w}) = f(\boldsymbol{w}) + g(\boldsymbol{w}) \to \min_{\boldsymbol{w}}. \tag{13}$$

[1] достаточность данного условия следует из выпуклости функции F

[2] при константной длине шага метод градиентного спуска имеет линейную скорость сходимости, а при длине шага, пропорциональной обратному гессиану, метод начинает сходиться квадратично

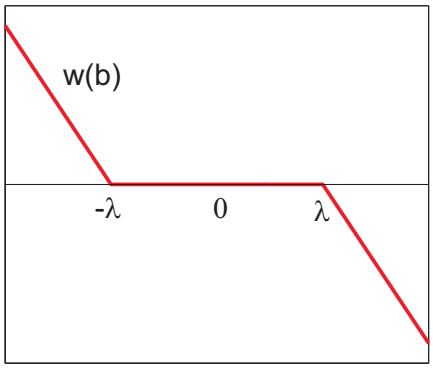

Рис. 3: Иллюстрация решения задачи (15) как функции от b.

Здесь функции f и g по-прежнему являются выпуклыми, но, возможно, негладкими. Задачи обучения разреженных линейных моделей классификации и регрессии (2),(6),(9) являются задачами типа (13), где $f(\boldsymbol{w})$ соответствует сумме функций потерь по обучающей выборке, а $g(\boldsymbol{w})$ отвечает за регуляризатор. В проксимальном методе для задачи (13) на шаге k решается следующая задача:

$$\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w}} \left(f(\boldsymbol{w}_k) + \partial f(\boldsymbol{w}_k)^T(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{L}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2 + g(\boldsymbol{w}) \right). \tag{14}$$

Здесь локальная линейная аппроксимация осуществляется только для функции f, а регуляризатор g остаётся без изменений. В минимизируемой функции (14) слагаемые $\partial f(\boldsymbol{w}_k)^T(\boldsymbol{w} - \boldsymbol{w}_k)$ и $\frac{L}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2$ являются сепарабельными по $\boldsymbol{w}$, т.е. могут быть представлены как сумма слагаемых, зависящих только от отдельных компонент w_d. Если регуляризатор $g(\boldsymbol{w})$ также является сепарабельным по $\boldsymbol{w}$, то задача многомерной оптимизации (14) разбивается на набор задач одномерной оптимизации относительно каждой из компонент w_d. В некоторых случаях эти одномерные задачи решаются аналитически. Рассмотрим подробнее случай регуляризатора $\|\boldsymbol{w}\|_1$ и $\|\boldsymbol{w}\|_{1/2}$.

При использовании L_1-регуляризации $g(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$ одномерные задачи оптимизации, образованные из (14), могут быть записаны как

$$\frac{L}{2}w_d^2 + b_d w_d + \lambda|w_d| \to \min_{w_d}, \tag{15}$$

где $b_d = (\partial f(\boldsymbol{w}_k))_d - Lw_{k,d}$. Для решения данной одномерной задачи вычислим субградиент минимизируемой функции:

$$\partial\left(\frac{L}{2}w_d^2 + b_d w_d + \lambda|w_d|\right) = Lw_d + b_d + \begin{cases} \lambda, & w_d > 0, \\ -\lambda, & w_d < 0, \\ [-\lambda, \lambda], & w_d = 0. \end{cases}$$

Рассмотрим три случая:

$$\begin{aligned} &w_d > 0: \ Lw_d + b_d + \lambda = 0 \ \Rightarrow \ w_d = -\frac{b_d + \lambda}{L} > 0 \ \Rightarrow \ b_d < -\lambda, \\ &w_d < 0: \ Lw_d + b_d - \lambda = 0 \ \Rightarrow \ w_d = \frac{\lambda - b_d}{L} < 0 \ \Rightarrow \ b_d > \lambda, \\ &w_d = 0: \ 0 \in b_d + [-\lambda, \lambda] \ \Rightarrow \ |b_d| \le \lambda. \end{aligned}$$

Таким образом, решением (15) является

$$w_d = \begin{cases} \frac{-b_d + \text{sign}(b_d)\lambda}{L}, & \text{если } |b_d| > \lambda, \\ 0, & \text{если } |b_d| \le \lambda. \end{cases} \tag{16}$$

Данное решение проиллюстрировано на рис. 3. Оно наглядно отражает свойство разреженности для $\boldsymbol{w}$.

Рассмотрим теперь решение задачи (14) для регуляризатора $g(\boldsymbol{w}) = \|\boldsymbol{w}\|_{1/2}$. Этот регуляризатор является сепарабельным по группам признаков $\boldsymbol{w}_g$. В результате приходим к задаче

$$\frac{L}{2}\boldsymbol{w}_g^T\boldsymbol{w}_g + \boldsymbol{b}_g^T\boldsymbol{w}_g + \lambda\|\boldsymbol{w}_g\|_2 \to \min_{\boldsymbol{w}_g}. \tag{17}$$

Для ее решения вычислим сначала субградиент функции $\|\boldsymbol{w}_g\|_2$:

$$\partial\|\boldsymbol{w}_g\|_2 = \begin{cases} \frac{\boldsymbol{w}_g}{\|\boldsymbol{w}_g\|_2}, & \text{если } \|\boldsymbol{w}_g\|_2 > 0, \\ \boldsymbol{z} : \|\boldsymbol{z}\|_2 \le 1, & \text{если } \|\boldsymbol{w}\|_2 = 0. \end{cases}$$

Рассмотрим два случая:

1. $\|\boldsymbol{w}_g\|_2 > 0$. Приравнивая к нулю градиент минимизируемой функции в задаче (17), получаем

$$\partial\left(\frac{L}{2}\boldsymbol{w}_g^T\boldsymbol{w}_g + \boldsymbol{b}_g^T\boldsymbol{w}_g + \lambda\|\boldsymbol{w}_g\|_2\right) = L\boldsymbol{w}_g + \boldsymbol{b}_g + \lambda\frac{\boldsymbol{w}_g}{\|\boldsymbol{w}_g\|_2} = \boldsymbol{0} \;\Rightarrow\; \left(L + \frac{\lambda}{\|\boldsymbol{w}_g\|_2}\right)\boldsymbol{w}_g = -\boldsymbol{b}_g. \tag{18}$$

   Вычисляя норму от левой и правой части в последнем уравнении, находим

$$\left(L + \frac{\lambda}{\|\boldsymbol{w}_g\|_2}\right)\|\boldsymbol{w}_g\|_2 = \|\boldsymbol{b}_g\|_2 \;\Rightarrow\; L\|\boldsymbol{w}_g\|_2 + \lambda = \|\boldsymbol{b}_g\|_2 \;\Rightarrow\; \|\boldsymbol{w}_g\|_2 = \frac{\|\boldsymbol{b}_g\|_2 - \lambda}{L} > 0 \;\Rightarrow\; \|\boldsymbol{b}_g\|_2 > \lambda.$$

   Поставляя результат для $\|\boldsymbol{w}_g\|_2$ в (18), имеем

$$\boldsymbol{w}_g = -\frac{1}{L}\left(1 - \frac{\lambda}{\|\boldsymbol{b}_g\|_2}\right)\boldsymbol{b}_g.$$

2. $\|\boldsymbol{w}_g\|_2 = 0$. Приравнивая к нулю субградиент минимизируемой функции в задаче (17), получаем

$$\boldsymbol{b}_g + \lambda\boldsymbol{z} = \boldsymbol{0} \;\Rightarrow\; \lambda\boldsymbol{z} = -\boldsymbol{b}_g \;\Rightarrow\; \lambda\|\boldsymbol{z}\|_2 = \|\boldsymbol{b}_g\|_2 \;\Rightarrow\; \|\boldsymbol{z}\|_2 = \frac{\|\boldsymbol{b}_g\|_2}{\lambda} \le 1 \;\Rightarrow\; \|\boldsymbol{b}_g\|_2 \le \lambda.$$

Объединяя эти два случая, находим окончательно, что решением (17) является

$$\boldsymbol{w}_g = -\frac{1}{L}\left(1 - \frac{\lambda}{\|\boldsymbol{b}_g\|_2}\right)_+\boldsymbol{b}_g, \tag{19}$$

где через u_+ обозначен оператор $\max(u, 0)$. Заметим, что задача (17) переходит в задачу (15), если группа g состоит из одного элемента. Легко убедиться, что в этом случае решение (19) действительно переходит в (16).

Метод покоординатного спуска

Другим подходом к решению задач (2), (6) и (9) является метод покоординатного спуска. Рассмотрим сначала применение этого метода для случая L_1-регуляризации. Здесь на очередном шаге оптимизационного процесса решается одномерная задача

$$(\nabla f(\boldsymbol{w}_k))_d(w - w_{k,d}) + \frac{1}{2}(\nabla^2 f(\boldsymbol{w}_k))_{dd}(w - w_{k,d})^2 + \lambda|w| \to \min_{w\in\mathbb{R}}, \tag{20}$$

где через $w_{k,d}$ обозначена d-ая компонента вектора $\boldsymbol{w}_k$, а w определяет величину сдвига вдоль d-ой координаты $\boldsymbol{e}_d$. Для случая линейной регрессии (2) задача (20) соответствует точной оптимизации функции F вдоль очередной координаты, а для случая логистической регрессии (9) задача (20) соответствует оптимизации функции F с использованием локального квадратичного приближения для $f(\boldsymbol{w})$.

Задача (20) может быть решена аналитически. Ее решение вычисляется по формуле (16), где

$$L = (\nabla^2 f(\boldsymbol{w}_k))_{dd},\;\; b_d = (\nabla f(\boldsymbol{w}_k))_d - (\nabla^2 f(\boldsymbol{w}_k))_{dd}w_{k,d}.$$

В связи с тем, что для логистической регрессии задача (20) является приближением оптимизации вдоль $\boldsymbol{e}_d$, то для стабильности оптимизационного процесса применяется дополнительная неточная одномерная оптимизация по $\alpha \in \mathbb{R}$, приводящая к выполнению критерия

$$F(\boldsymbol{w}_k + \alpha w\boldsymbol{e}_d) \le F(\boldsymbol{w}_k) + \alpha\rho((\nabla f(\boldsymbol{w}_k)_d w + \lambda(|w_{k,d} + w| - |w_{k,d}|)).$$

Здесь w – решение задачи (20), а $\rho \in (0, 1)$ – параметр, определяемый пользователем. Предлагаемый критерий является первым условием Флетчера для неточной одномерной оптимизации, где вместо производной по направлению $\boldsymbol{e}_d$ недифференцируемой функции $g(w) = |w|$ используется ее разностная аппроксимация.

Рассмотрим теперь случай L_1/L_2-регуляризации. Применение здесь блочного покоординатного спуска приводит к решению задачи

$$(\nabla f(\boldsymbol{w}_k))_g^T(\boldsymbol{w}_g - \boldsymbol{w}_{k,g}) + \frac{1}{2}(\boldsymbol{w}_g - \boldsymbol{w}_{k,g})^T(\nabla^2 f(\boldsymbol{w}_k))_{gg}(\boldsymbol{w}_g - \boldsymbol{w}_{k,g}) + \lambda\|\boldsymbol{w}_g\|_2 \to \min_{\boldsymbol{w}_g}. \tag{21}$$

Как и раньше, для логистической регрессии данная задача является аппроксимацией. Поэтому после решения (21) необходимо дополнительно решать задачу неточной одномерной оптимизации вдоль направления $\boldsymbol{w}_g - \boldsymbol{w}_{k,g}$.

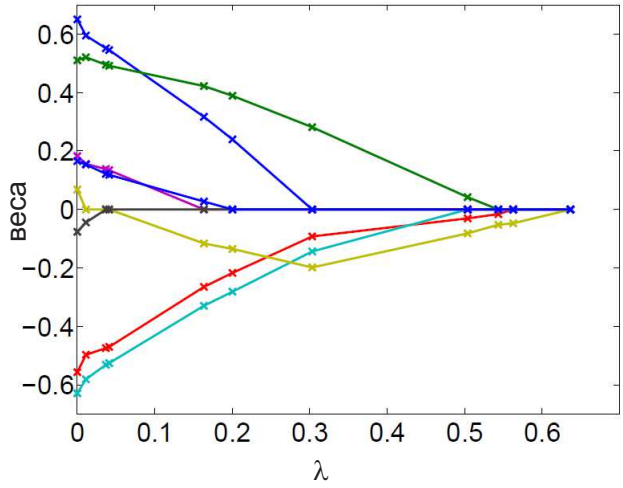

Рис. 4: Иллюстрация пути регуляризации для задачи (2).

Рассмотрим решение задачи (21), которая может быть эквивалентно представлена как

$$\frac{1}{2}\boldsymbol{w}_g^T A_g \boldsymbol{w}_g + \boldsymbol{b}_g^T \boldsymbol{w}_g + \lambda \|\boldsymbol{w}_g\|_2 \to \min_{\boldsymbol{w}_g}, \tag{22}$$

где $A_g = (\nabla^2 f(\boldsymbol{w}_k))_{gg}$, а $\boldsymbol{b}_g = (\nabla f(\boldsymbol{w}_k))_g - (\nabla^2 f(\boldsymbol{w}_k))_{gg} \boldsymbol{w}_{k,g}$. Действуя по аналогии со случаем $\|\boldsymbol{w}_g\|_2 = 0$ для задачи (17), можно показать, что задача (22) имеет нулевое решение при $\lambda \geq \|\boldsymbol{b}_g\|_2$. Однако, для случая $\lambda < \|\boldsymbol{b}_g\|_2$ решение (22) не выписывается аналитически. Проведем последовательность эквивалентных преобразований:

$$\frac{1}{2}\boldsymbol{w}_g^T A_g \boldsymbol{w}_g + \boldsymbol{b}_g^T \boldsymbol{w}_g + \lambda \|\boldsymbol{w}_g\|_2 \to \min_{\boldsymbol{w}_g} \Leftrightarrow \begin{cases} \frac{1}{2}\boldsymbol{w}_g^T A_g \boldsymbol{w}_g + \boldsymbol{b}_g^T \boldsymbol{w}_g \to \min\limits_{\boldsymbol{w}_g}, \\ \|\boldsymbol{w}_g\|_2 \leq \eta, \end{cases} \Leftrightarrow$$

$$\begin{cases} \frac{1}{2}\boldsymbol{w}_g^T A_g \boldsymbol{w}_g + \boldsymbol{b}_g^T \boldsymbol{w}_g \to \min\limits_{\boldsymbol{w}_g}, \\ \|\boldsymbol{w}_g\|_2^2 \leq \eta^2, \end{cases} \Leftrightarrow \frac{1}{2}\boldsymbol{w}_g^T A_g \boldsymbol{w}_g + \boldsymbol{b}_g^T \boldsymbol{w}_g + \lambda_2 \|\boldsymbol{w}_g\|_2^2 \to \min_{\boldsymbol{w}_g}.$$

Здесь между λ и λ_2 существует неявное соответствие. Последняя задача эквивалентна СЛАУ вида $(A_g + 2\lambda_2 I)\,\boldsymbol{w}_g = -\boldsymbol{b}_g$. В результате решение (22) для случая $\lambda < \|\boldsymbol{b}_g\|_2$ может быть найдено путем решения задачи одномерной оптимизации по λ_2, где в качестве возможных кандидатов $\boldsymbol{w}_g$ рассматриваются решения обозначенной СЛАУ.

Регрессия наименьших углов (LARS)

Рассмотрим задачу обучения L_1-регуляризованной линейной регрессии (2). Как было показано выше, необходимыми и достаточными условиями оптимальности $\hat{\boldsymbol{w}}$ в этой задаче являются

$$\begin{aligned} &\boldsymbol{x}_d^T(X\hat{\boldsymbol{w}} - \boldsymbol{t}) = -\lambda, \ \hat{w}_d > 0, \\ &\boldsymbol{x}_d^T(X\hat{\boldsymbol{w}} - \boldsymbol{t}) = \lambda, \ \hat{w}_d < 0, \\ &|\boldsymbol{x}_d^T(X\hat{\boldsymbol{w}} - \boldsymbol{t})| \leq \lambda, \ \hat{w}_d = 0. \end{aligned} \tag{23}$$

В методах оптимизации из семейства active-set множество номеров признаков $\{1, \ldots, D\}$ разбивается на два подмножества J и J^c. Предполагается, что компоненты весов из множества J $\boldsymbol{w}_J$ строго отличаются от нуля, а $\boldsymbol{w}_{J^c} = \boldsymbol{0}$. Общая схема метода выглядит следующим образом:

1. Решаем задачу оптимизации при фиксированных J и J^c;
2. Проверяем выполнимость необходимых и достаточных условий оптимальности; если выполнено, то стоп;
3. Перебрасываем часть элементов между множествами J и J^c, как правило, из учета наиболее нарушаемого условия оптимальности, переходим к шагу 1.

Применение этой общей схемы требует конкретизации действий на всех трех шагах. Примерами алгоритмов из семейства active-set являются SMO [1] и SVM$^{\text{light}}$ [2] для решения задачи обучения метода опорных векторов,

а также симплекс-метод [4] для решения задачи линейного программирования. Для задачи (2) соответствующий метод получил название регрессии наименьших углов (LARS: Least Angle Regression for laSso/Stepwise regression) [3], который рассматривается ниже. Как правило, методы из семейства active-set являются эффективными в ситуации, когда итоговое множество J является небольшим.

В методе LARS задача (2) решается сразу для всех значений параметра регуляризации λ (строится т.н. путь регуляризации). Обозначим через $\boldsymbol{w}(\lambda)$ решение (2) для конкретного λ. Нетрудно показать, что $\boldsymbol{w}(\lambda)$ как функция от λ является кусочно-линейной. Действительно, пусть при текущем λ известны знаки $\boldsymbol{s}$ решения $\boldsymbol{w}(\lambda)$:

$$s_d = \begin{cases} 1, w_d(\lambda) > 0, \\ -1, w_d(\lambda) < 0, \\ 0, w_d(\lambda) = 0. \end{cases}$$

Тогда задача (2) переходит в

$$\frac{1}{2}\|\boldsymbol{t} - X\boldsymbol{w}\|^2 + \lambda \boldsymbol{s}^T \boldsymbol{w} \to \min_{\boldsymbol{w}}.$$

Эта задача имеет аналитическое решение $\boldsymbol{w}(\lambda) = (X^T X)^{-1}(X^T \boldsymbol{t} - \lambda \boldsymbol{s})$. Очевидно, что при фиксированном $\boldsymbol{s}$ полученное решение $\boldsymbol{w}(\lambda)$ является линейной функцией от λ. Из соображений непрерывности следует, что решение (2) сохраняет знаки в некоторой окрестности λ. Поэтому функция $\boldsymbol{w}(\lambda)$ является кусочно-линейной функцией от λ (см. рис. 4), а для построения пути регуляризации достаточно найти все точки излома.

Для применения метода LARS необходимо выполнение следующих двух требований:

1. Матрица $X_J^T X_J$ является невырожденной для любого множества J, где под X_J понимается подмножество признаков J для всех объектов обучения. Если это условие не выполнено, то обычно к матрице $X_J^T X_J$ добавляют αI, где α – небольшое положительное число.

2. Вдоль пути регуляризации признаки входят и покидают множество J по-одному. На практике при использовании действительных признаков это требование всегда выполняется. Если оно не выполнено, то достаточно сделать небольшую модификацию значений всех признаков в обучающей выборке.

Схема метода LARS выглядит следующим образом:

1. Положить $\lambda = \|X^T \boldsymbol{t}\|_\infty$;

2. Положить $J = \{j : |\boldsymbol{x}_j^T \boldsymbol{t}| = \lambda\}$, $\boldsymbol{s}_J = \text{sign}(X_J^T \boldsymbol{t})$, $\boldsymbol{w}_J(\lambda) = (X_J^T X_J)^{-1}(X_J^T \boldsymbol{t} - \lambda \boldsymbol{s}_J)$, $\boldsymbol{w}_{J^c} = \boldsymbol{0}$;

3. Значение λ последовательно уменьшается, при этом $\boldsymbol{w}_J(\lambda) = (X_J^T X_J)^{-1}(X_J^T \boldsymbol{t} - \lambda \boldsymbol{s}_J)$, $\boldsymbol{w}_{J^c} = \boldsymbol{0}$. Процесс продолжается до тех пор, пока не выполняется одно из следующих условий:

   (a) $\exists j \in J^c : |\boldsymbol{x}_j^T (X\boldsymbol{w} - \boldsymbol{t})| = \lambda$. Тогда добавляем новый признак j в набор активных признаков $J = J \cup \{j\}$, $s_j = -\text{sign}(\boldsymbol{x}_j^T (X\boldsymbol{w} - \boldsymbol{t}))$;

   (b) $\exists j \in J : w_j = 0$. Тогда исключаем признак j из активного набора $J = J \backslash \{j\}$;

   (c) $\lambda = 0$.

   Если выполнено (a) или (b), то повторяем шаг 3, иначе стоп.

На шаге 1 значение λ определяется как $\|X^T \boldsymbol{t}\|_\infty$. Как было показано выше, для всех $\lambda \geq \|X^T \boldsymbol{t}\|_\infty$ оптимальным вектором весов является нулевой вектор. На шагах 2 и 3 s_j определяется как обратный знак корреляции между признаком $\boldsymbol{x}_j$ и текущим остатком $X\boldsymbol{w} - \boldsymbol{t}$. Это следует из условий оптимальности (23). Выражение для шага 3 «значение λ последовательно уменьшается» носит неформальный характер. Здесь можно аналитически определить, при каком λ будет выполнено условие (a) или (b). Конкретные формулы можно найти в [3].

На практике ситуация (b) происходит редко. Поэтому в большинстве случаев алгоритм LARS сходится за D шагов. Возможность поиска пути регуляризации для задачи (2) позволяет значительно ускорить процедуру скользящего контроля по подбору параметра регуляризации λ.

Список литературы

[1] J. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, MIT Press, 1999.

[2] T. Joachims. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, MIT Press, 1999.

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

[4] I. Maros. Computational Techniques of the Simplex Method. Kluwer Academic, Boston, 2003.