

Методы оптимизации для разреженных линейных моделей

Разреженные линейные модели классификации и регрессии

Рассмотрим стандартную задачу восстановления регрессии. Имеется обучающая выборка из N объектов $\{t_n, \mathbf{x}_n\}_{n=1}^N$, где $\mathbf{x}_n \in \mathbb{R}^D$ – 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 \mathbf{x} .

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

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

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

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

Здесь $\lambda \geq 0$ – параметр регуляризации, устанавливаемый пользователем. В решении задачи (2) часть компонент \mathbf{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 \mathbf{w} , как правило, оказывается равным нулю. Для этого рассмотрим следующую задачу оптимизации:

$$\begin{aligned} \frac{1}{2} \|\mathbf{t} - X\mathbf{w}\|^2 &\rightarrow \min_{\mathbf{w}}, \\ \|\mathbf{w}\|_1 &\leq \eta. \end{aligned} \quad (3)$$

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

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

Из условий (4) следует эквивалентность задач (2) и (3), где между параметрами λ и η существует неявное соответствие. Действительно, пусть $\hat{\mathbf{w}}$ является решением задачи (3) для некоторого $\hat{\eta}$. Тогда для $\hat{\mathbf{w}}$ найдется $\hat{\lambda}$ такое, что пара $(\hat{\mathbf{w}}, \hat{\lambda})$ удовлетворяет условиям (4). Это означает, что $\hat{\mathbf{w}}$ является решением задачи (2) для $\lambda = \hat{\lambda}$. Обратно, пусть $\hat{\mathbf{w}}$ является решением задачи (2) для некоторого $\hat{\lambda}$. Тогда при $\hat{\eta} = \|\hat{\mathbf{w}}\|_1$ набор $(\hat{\mathbf{w}}, \hat{\lambda})$ будет удовлетворять условиям (4), т.е. $\hat{\mathbf{w}}$ 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 $\|\mathbf{w}\|_1 \leq \eta$, представляет собой ромб с центром в нуле. Линии уровня минимизируемого функционала $\frac{1}{2} \|\mathbf{t} - X\mathbf{w}\|^2$ 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 $\mathbf{w}^+, \mathbf{w}^- \in \mathbb{R}^D$ такие, что $w_d^+ \geq 0, w_d^- \geq 0, w_d = w_d^+ - w_d^- \forall d = \overline{1, D}$, и рассмотрим задачу

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

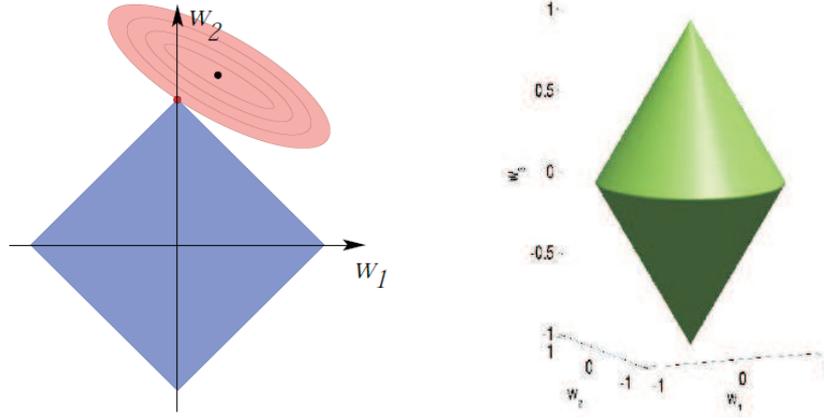


Рис. 1: Слева: иллюстрация получения разреженного решения в задаче (2), справа: ограничение вида $\sqrt{w_1^2 + w_2^2} + |w_3| \leq \eta$. Иллюстрации взяты из слайдов лекции по разреженным моделям с конференции ECML 2010 <http://www.di.ens.fr/~fbach/ecml2010tutorial/>

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 \mathbf{x} разбиты на набор непересекающихся групп. Обозначим через \mathbf{x}_g признаки, относящиеся к группе g , а через \mathbf{w}_g – соответствующие им веса в линейной регрессии. При обучении линейных регрессионных моделей со свойством групповой разреженности веса некоторых групп \mathbf{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

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

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

$$F(\mathbf{w}) = \frac{1}{2} \|\mathbf{t} - X\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|_{1/2} \rightarrow \min_{\mathbf{w}_1, \dots, \mathbf{w}_G}. \quad (6)$$

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:

$$\begin{aligned} \frac{1}{2} \|\mathbf{t} - X\mathbf{w}\|^2 &\rightarrow \min_{\mathbf{w}}, \\ \|\mathbf{w}\|_{1/2} &\leq \eta. \end{aligned} \quad (7)$$

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

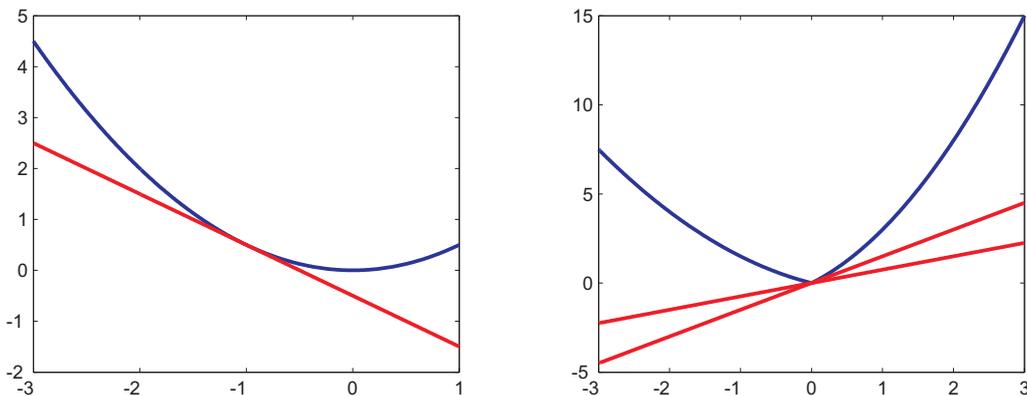


Рис. 2: Иллюстрация понятия субградиента. Слева: касательная является нижней границей для выпуклой непрерывно-дифференцируемой функции, справа: для выпуклых негладких функций в точках недифференцируемости существует множество касательных, являющихся нижней границей.

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 $\{t_n, \mathbf{x}_n\}_{n=1}^N$, где $\mathbf{x}_n \in \mathbb{R}^D$ – 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 \mathbf{x} .

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

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

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

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

где величина $\log(1 + \exp(-t_n \mathbf{w}^T \mathbf{x}_n))$ играет роль функции потерь при значении прогнозирующей функции $\mathbf{w}^T \mathbf{x}_n$ для истинного класса t_n . Как и раньше, использование L_1 -регуляризации позволяет получить разреженное решение для весов \mathbf{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 $\|\mathbf{w}\|_1$ меняется на $\|\mathbf{w}\|_{1/2}$.

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

$$F(\mathbf{w}) \rightarrow \min_{\mathbf{w}}. \quad (10)$$

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

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

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

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:

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

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

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

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

Проксимальный метод

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

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

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

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

где между L и η есть неявное соответствие. В результате добавление проксимального слагаемого означает требование поиска следующей точки итерационного процесса \mathbf{w}_{k+1} в окрестности \mathbf{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 \mathbf{w}_k .

Задача (12) может быть решена аналитически:

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

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 шага¹. Аналогично, скорость работы проксимального метода зависит от выбора L .

Рассмотрим теперь применение проксимального метода для решения задачи

$$F(\mathbf{w}) = f(\mathbf{w}) + g(\mathbf{w}) \rightarrow \min_{\mathbf{w}}. \quad (13)$$

¹при константной длине шага метод градиентного спуска имеет линейную скорость сходимости, а при длине шага, пропорциональной обратному гессиану, метод начинает сходиться квадратично

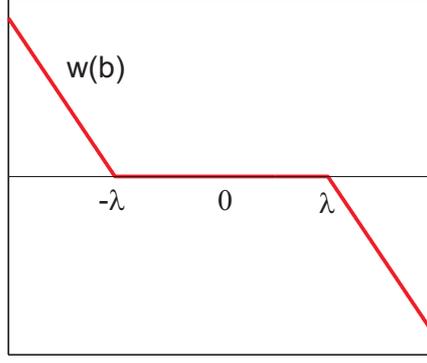


Рис. 3: Иллюстрация решения задачи (15) как функции от b .

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 $f(\mathbf{w})$ соответствует сумме функций потерь по обучающей выборке, а $g(\mathbf{w})$ отвечает за регуляризатор. В проксимальном методе для задачи (13) на шаге k решается следующая задача:

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

Здесь локальная линейная аппроксимация осуществляется только для функции f , а регуляризатор g остаётся без изменений. В минимизируемой функции (14) слагаемые $\partial f(\mathbf{w}_k)^T (\mathbf{w} - \mathbf{w}_k)$ и $\frac{L}{2} \|\mathbf{w} - \mathbf{w}_k\|^2$ являются сепарабельными по \mathbf{w} , т.е. могут быть представлены как сумма слагаемых, зависящих только от отдельных компонент w_d . Если регуляризатор $g(\mathbf{w})$ также является сепарабельным по \mathbf{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 $\|\mathbf{w}\|_1$ и $\|\mathbf{w}\|_{1/2}$.

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

$$\frac{L}{2} w_d^2 + b_d w_d + \lambda |w_d| \rightarrow \min, \quad (15)$$

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

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

Рассмотрим три случая:

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

Таким образом, решением (15) является

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

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

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

$$\frac{L}{2} \mathbf{w}_g^T \mathbf{w}_g + \mathbf{b}_g^T \mathbf{w}_g + \lambda \|\mathbf{w}_g\|_2 \rightarrow \min_{\mathbf{w}_g}. \quad (17)$$

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

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

Рассмотрим два случая:

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

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

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

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

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

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

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

$$\mathbf{b}_g + \lambda \mathbf{z} = \mathbf{0} \Rightarrow \lambda \mathbf{z} = -\mathbf{b}_g \Rightarrow \lambda \|\mathbf{z}\|_2 = \|\mathbf{b}_g\|_2 \Rightarrow \|\mathbf{z}\|_2 = \frac{\|\mathbf{b}_g\|_2}{\lambda} \leq 1 \Rightarrow \|\mathbf{b}_g\|_2 \leq \lambda.$$

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

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

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

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

где через $w_{k,d}$ обозначена d -ая компонента вектора \mathbf{w}_k , а w определяет величину сдвига вдоль d -ой координаты \mathbf{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 $f(\mathbf{w})$.

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

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

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

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

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 \mathbf{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

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

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

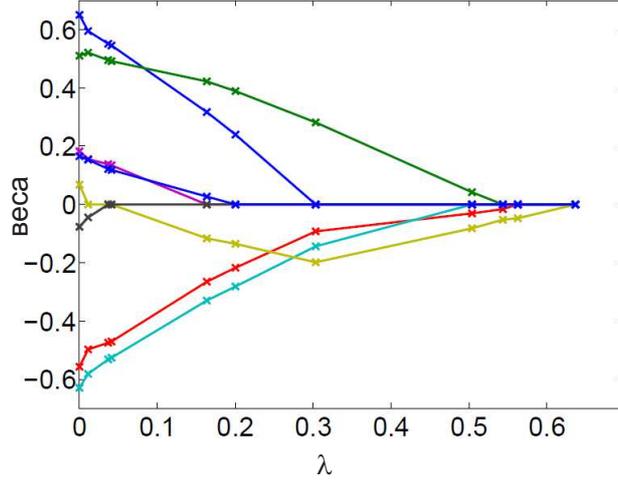


Рис. 4: Иллюстрация пути регуляризации для задачи (2).

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

$$\frac{1}{2} \mathbf{w}_g^T A_g \mathbf{w}_g + \mathbf{b}_g^T \mathbf{w}_g + \lambda \|\mathbf{w}_g\|_2 \rightarrow \min_{\mathbf{w}_g}, \quad (22)$$

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

$$\frac{1}{2} \mathbf{w}_g^T A_g \mathbf{w}_g + \mathbf{b}_g^T \mathbf{w}_g + \lambda \|\mathbf{w}_g\|_2 \rightarrow \min_{\mathbf{w}_g} \Leftrightarrow \begin{cases} \frac{1}{2} \mathbf{w}_g^T A_g \mathbf{w}_g + \mathbf{b}_g^T \mathbf{w}_g \rightarrow \min_{\mathbf{w}_g}, \\ \|\mathbf{w}_g\|_2 \leq \eta, \end{cases} \Leftrightarrow \begin{cases} \frac{1}{2} \mathbf{w}_g^T A_g \mathbf{w}_g + \mathbf{b}_g^T \mathbf{w}_g \rightarrow \min_{\mathbf{w}_g}, \\ \|\mathbf{w}_g\|_2^2 \leq \eta^2, \end{cases} \Leftrightarrow \frac{1}{2} \mathbf{w}_g^T A_g \mathbf{w}_g + \mathbf{b}_g^T \mathbf{w}_g + \lambda_2 \|\mathbf{w}_g\|_2^2 \rightarrow \min_{\mathbf{w}_g}.$$

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

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

В методах оптимизации из семейства [active-set](#) множество номеров признаков $\{1, \dots, D\}$ разбивается на два подмножества J и J^c . Предполагается, что компоненты весов из множества J строго отличаются от нуля, а $\mathbf{w}_{J^c} = \mathbf{0}$. 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. Конкретными примерами алгоритмов из семейства [active-set](#) являются SMO [1] и SVM^{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 $\mathbf{w}(\lambda)$ решение (2) для конкретного λ . Нетрудно показать, что $\mathbf{w}(\lambda)$ как функция от λ является кусочно-линейной. Действительно, пусть при текущем λ известны знаки \mathbf{s} решения $\mathbf{w}(\lambda)$:

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

Тогда задача (2) переходит в

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

Эта задача имеет аналитическое решение $\mathbf{w}(\lambda) = (X^T X)^{-1} (X^T \mathbf{t} - \lambda \mathbf{s})$. Очевидно, что при фиксированном \mathbf{s} полученное решение $\mathbf{w}(\lambda)$ является линейной функцией от λ . Из соображений непрерывности следует, что решение (2) сохраняет знаки в некоторой окрестности λ . Поэтому функция $\mathbf{w}(\lambda)$ 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 случайную модификацию значений всех признаков в обучающей выборке.

Схема метода LARS выглядит следующим образом:

1. Положить $\lambda = \|X^T \mathbf{t}\|_\infty$;
2. Положить $J = \{j : |\mathbf{x}_j^T \mathbf{t}| = \lambda\}$, $\mathbf{s}_J = \text{sign}(X_J^T \mathbf{t})$, $\mathbf{w}_J(\lambda) = (X_J^T X_J)^{-1} (X_J^T \mathbf{t} - \lambda \mathbf{s}_J)$, $\mathbf{w}_{J^c} = \mathbf{0}$;
3. Значение λ последовательно уменьшается, при этом $\mathbf{w}_J(\lambda) = (X_J^T X_J)^{-1} (X_J^T \mathbf{t} - \lambda \mathbf{s}_J)$, $\mathbf{w}_{J^c} = \mathbf{0}$. Процесс продолжается до тех пор, пока не выполняется одно из следующих условий:
 - (a) $\exists j \in J^c : |\mathbf{x}_j^T (X\mathbf{w} - \mathbf{t})| = \lambda$. Тогда добавляем новый признак j в набор активных признаков $J = J \cup \{j\}$, $s_j = -\text{sign}(\mathbf{x}_j^T (X\mathbf{w} - \mathbf{t}))$;
 - (b) $\exists j \in J : w_j = 0$. Тогда исключаем признак j из активного набора $J = J \setminus \{j\}$;
 - (c) $\lambda = 0$.

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

На шаге 1 значение λ определяется как $\|X^T \mathbf{t}\|_\infty$. Как было показано выше, для всех $\lambda \geq \|X^T \mathbf{t}\|_\infty$ оптимальным вектором весов является нулевой вектор. На шагах 2 и 3 s_j определяется как обратный знак корреляции между признаком \mathbf{x}_j и текущим остатком $X\mathbf{w} - \mathbf{t}$. 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.