Курс: Методы оптимизации в машинном обучении, 2012

Оптимизация с помощью глобальных верхних оценок, зависящих от параметра

Вероятностные модели линейной регрессии

Рассмотрим вероятностную модель для задач машинного обучения в общем виде. Такая модель задается совместным распределением $p(X, T|\theta)$, где X – набор наблюдаемых величин, T – набор ненаблюдаемых (скрытых) величин и θ – набор параметров. Обучение данной модели (поиск параметров θ) с помощью метода максимального правдоподобия соответствует решению следующей задачи:

$$p(X|\theta) = \int p(X, T|\theta)dT \to \max_{\theta}. \tag{1}$$

В том случае, если T – дискретные переменные, то интеграл заменяется на сумму. Основная сложность в решении задачи (1) заключается в том, что для многих интересных с практической точки зрения вероятностных моделей интеграл по скрытым переменным T не берется аналитически и не может быть эффективно оценен численно, т.к. T, как правило, находится в существенно многомерном пространстве. В результате задача (1) относится к задачам оптимизации, в которых неизвестно ни значение оптимизируемой функции, ни значение ее производных. Следовательно, стандартные методы оптимизации, такие как градиентный спуск, ньютоновские и квази-ньютоновские методы, здесь не применимы.

Рассмотрим для примера вероятностные модели линейной регрессии с различными регуляризаторами. Пусть имеется обучающая выборка объектов $(\boldsymbol{t}, X) = \{t_n, \boldsymbol{x}_n\}_{n=1}^N$, где $\boldsymbol{x}_n \in \mathbb{R}^D$ – вектор признаков, а $t_n \in \mathbb{R}$ – регрессионная переменная. Задача заключается в предсказании на основе обучающей выборки регрессионной переменной t_{new} для нового объекта, заданного своим вектором признаков $\boldsymbol{x}_{new}$. В моделях линейной регрессии для предсказания используется линейная функция

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

где $\boldsymbol{w} \in \mathbb{R}^D$ – некоторые веса. Вероятностная модель линейной регрессии с квадратичной регуляризацией, т.е. совместное распределение на наблюдаемые переменные $\boldsymbol{t}$ и скрытые веса $\boldsymbol{w}$ с некоторыми параметрами $\alpha, \beta > 0$, задается следующим образом:

$$p(\boldsymbol{t}, \boldsymbol{w}|X, \alpha, \beta) = p(\boldsymbol{t}|X, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha), \tag{2}$$

$$p(\boldsymbol{t}|X, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{x}_n, \beta^{-1}) = \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w}, \beta^{-1}I), \tag{3}$$

$$p(\boldsymbol{w}|\alpha) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \alpha^{-1}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}I). \tag{4}$$

Здесь через $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ обозначена плотность многомерного нормального распределения с мат.ожиданием $\boldsymbol{\mu}$ и матрицей ковариации $\Sigma \succ 0$. Выражение (3) соответствует модели независимого нормального шума с некоторой дисперсией β^{-1}, а выражение (4) соответствует квадратичной регуляризации, которая уменьшает риск переобучения модели за счет предпочтения более простых регрессий (тех, у которых вектор весов $\boldsymbol{w}$ близок к нулевому вектору).

Предположим, что обучение модели (2)-(4) проведено и определены значения параметров α, β. Тогда прогнозирование регрессионной переменной для нового объекта осуществляется как $\boldsymbol{w}^T\boldsymbol{x}_{new}$, где веса $\boldsymbol{w}$ находятся путем решения задачи $p(\boldsymbol{w}|\boldsymbol{t}, X, \alpha, \beta) \to \max_{\boldsymbol{w}}$. Легко показать, что эта задача эквивалентна методу наименьших квадратов с квадратичной регуляризацией:

$$\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 + \frac{\alpha}{2}\|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}.$$

Ее решение состоит в решении следующей системы линейных уравнений:

$$(\beta X^T X + \alpha I_D)\boldsymbol{w} = \beta X^T \boldsymbol{t}. \tag{5}$$

Здесь символом I_D обозначена единичная матрица размера $D \times D$. Рассмотрим несколько способов решения системы (5):

- Разложение Холецкого. Представим матрицу $\beta X^T X + \alpha I_D$ как LL^T, где L – нижнетреугольная матрица. Это можно сделать, т.к. матрица $\beta X^T X + \alpha I_D$ для положительных α, β всегда является положительно-определенной. Тогда решение СЛАУ вида $LL^T \boldsymbol{w} = \boldsymbol{y} = \beta X^T \boldsymbol{t}$ можно представить как $L\boldsymbol{v} = \boldsymbol{y},\ L^T \boldsymbol{w} = \boldsymbol{v}$. Решение СЛАУ с нижнетреугольной (верхнетреугольной) матрицей может быть эффективно найдено методом обратного исключения неизвестных. Суммарная сложность описанного алгоритма (без учета операций, имеющих на порядки меньшую сложность) составляет ND^2 операций для вычисления матрицы СЛАУ и $D^3/3$ операций для вычисления разложения Холецкого. Требуемый объем памяти – D^2.

- Метод сопряженных градиентов. Прямое умножение матрицы СЛАУ $\beta X^T X + \alpha I_D$ на произвольный вектор длины D требует D^2 операций. Метод сопряженных градиентов сходится за K итераций, причем $K \leq D$. В результате суммарная сложность алгоритма составляет ND^2 операций для вычисления матрицы СЛАУ и D^2K операций для проведения необходимых итераций сопряженных градиентов. Требуемый объем памяти – D^2.

- Метод сопряженных градиентов с последовательным алгоритмом умножения матрицы на вектор. В последовательном алгоритме задача умножения $(\beta X^T X + \alpha I)\boldsymbol{z}$ решается как

$$\boldsymbol{z}_1 = X\boldsymbol{z},\ \boldsymbol{z}_2 = X^T \boldsymbol{z}_1,\ \boldsymbol{z}_3 = \beta \boldsymbol{z}_2 + \alpha \boldsymbol{z}.$$

  Сложность такого умножения составляет $2ND + 2D$ операций, а объем требуемой памяти – D. Учитывая K итераций метода сопряженных градиентов суммарная сложность алгоритма составляет $2NDK$ операций.

Используя тождество Вудбери, решение системы (5) можно представить как

$$\boldsymbol{w} = (\beta X^T X + \alpha I_D)^{-1} \beta X^T \boldsymbol{t} = (\alpha^{-1} I_D - \alpha^{-2} X^T (\beta^{-1} I_N + \alpha^{-1} X X^T)^{-1} X) \beta X^T \boldsymbol{t} =$$
$$\alpha^{-1} \beta X^T \boldsymbol{t} - \alpha^{-2} X^T (\beta^{-1} I_N + \alpha^{-1} X X^T)^{-1} \beta X X^T \boldsymbol{t} = \alpha^{-2} \beta X^T (\beta^{-1} I_N + \alpha^{-1} X X^T)^{-1} (\alpha(\beta^{-1} I_N + \alpha^{-1} X X^T) - X X^T)\boldsymbol{t} =$$
$$\alpha^{-1} X^T (\beta^{-1} I_N + \alpha^{-1} X X^T)^{-1} \boldsymbol{t}.$$

В результате альтернативный вариант решения системы (5) выглядит как

$$(\beta^{-1} I_N + \alpha^{-1} X X^T)\boldsymbol{u} = \boldsymbol{t}, \quad \boldsymbol{w} = \alpha^{-1} X^T \boldsymbol{u}, \tag{6}$$

где для решения СЛАУ с матрицей $\beta^{-1} I_N + \alpha^{-1} X X^T$ можно воспользоваться одним из трех алгоритмов, описанных выше. Следующая таблица сводит вместе суммарную сложность и требуемый объем памяти для всех приведенных алгоритмов:

Система	Разложение Холецкого		Метод сопр.гр. с прямым алг.		Метод сопр.гр. с послед. алг.	
	Сложность	Память	Сложность	Память	Сложность	Память
(5)	$ND^2 + D^3/3$	D^2	$ND^2 + D^2K$	D^2	$2NDK$	D
(6)	$N^2D + N^3/3$	N^2	$N^2D + N^2K$	N^2	$2NDK$	N

Из представленной таблицы можно заключить, что метод сопряженных градиентов с последовательным алгоритмом умножения является наиболее эффективным. При этом выбор между системами (5) и (6) определяется соотношением между N и D.

Задача обучения модели (2)-(4) (поиск параметров α, β) с помощью метода максимального правдоподобия выглядит как

$$p(\boldsymbol{t}|X, \alpha, \beta) = \int p(\boldsymbol{t}|X, \boldsymbol{w}, \beta) p(\boldsymbol{w}|\alpha) d\boldsymbol{w} \to \max_{\alpha,\beta}. \tag{7}$$

Эта задача относится к классу задач с интегральными функционалами вида (1), представленных выше. В данном случае значение $p(\boldsymbol{t}|X, \alpha, \beta)$ может быть вычислено аналитически, т.к. интеграл представляет собой свертку двух нормальных распределений. Тем не менее, как будет показано дальше, задача (7) может быть решена без непосредственного вычисления интеграла.

Наряду с квадратичной регуляризацией (4) можно рассмотреть другие виды регуляризации:

- L_1-регуляризация.

$$p(\boldsymbol{w}|\alpha) = \prod_{d=1}^{D} \mathcal{L}(w_d|\alpha) = \prod_{d=1}^{D} \frac{\alpha}{2} \exp(-\alpha|w_d|). \tag{8}$$

  Распределение Лапласа по сравнению с нормальным распределением (см. рис. 1) имеет более тяжелые хвосты и острый пик в нуле. Более тяжелые хвосты позволяют некоторым весам w_d иметь большие по модулю значения (нормальное распределение практически запрещает весам отклоняться от нуля больше, чем на три стандартных отклонения и, в частности, уменьшает по модулю значения всех весов, в том числе,

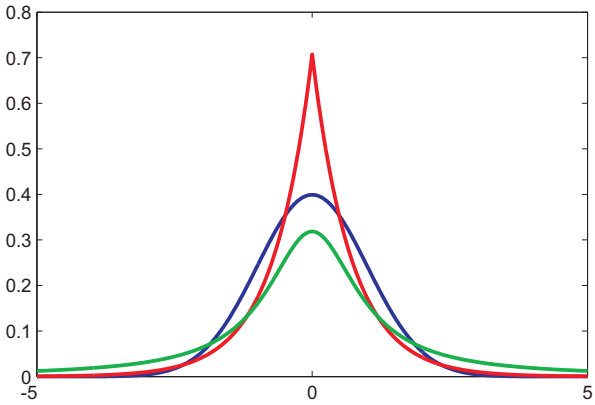

Рис. 1: Различные виды регуляризаторов. Синяя кривая – стандартное нормальное распределение (квадратичная регуляризация), красная кривая – распределение Лапласа (L_1 -регуляризация), зеленая кривая – распределение Стьюдента с одной степенью свободы.

информативных). Наличие большой массы в районе нуля и пик в нуле способствуют активному обнулению небольших по модулю весов w_d, т.е. признанию некоторых признаков неинформативными (нормальное распределение приводит к тому, что, вообще говоря, все веса отличны от нуля). Обучение модели (2)-(3) с регуляризатором (8) соответствует решению задачи

$$p(\boldsymbol{t}|X,\alpha,\beta) = \int \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w},\beta^{-1}I)\prod_{d=1}^{D}\frac{\alpha}{2}\exp(-\alpha|w_d|)d\boldsymbol{w} \to \max_{\alpha,\beta}.$$

В отличие от случая квадратичной регуляризации данный интеграл не может быть вычислен аналитически.

- Распределение Стьюдента с ν степенями свободы.

$$p(\boldsymbol{w}|\alpha) = \prod_{d=1}^{D}\mathcal{S}(w_d|\alpha,\nu) = \prod_{d=1}^{D}\frac{\sqrt{\alpha}\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{\alpha}{\nu}w_d^2\right)^{-(\nu+1)/2},\ \nu>0. \tag{9}$$

Распределение Стьюдента (см. рис. 1) также обладает более тяжелыми хвостами по сравнению с нормальным распределением и, как следствие, приводит к моделям, менее подверженным эффекту недообучения (при квадратичной регуляризации этот эффект возникает из-за уменьшения по модулю информативных весов). Кроме того, в отличие от L_1-регуляризации, распределение Стьюдента является гладким. Поэтому на этапе прогнозирования задача $p(\boldsymbol{w}|\boldsymbol{t},X,\alpha,\beta) \to \max_{\boldsymbol{w}}$ является гладкой. Тем не менее, на этапе обучения здесь возникает задача оптимизации, в которой интеграл не вычисляется аналитически:

$$p(\boldsymbol{t}|X,\alpha,\beta) = \int \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w},\beta^{-1}I)\prod_{d=1}^{D}\mathcal{S}(w_d|\alpha,\nu)d\boldsymbol{w} \to \max_{\alpha,\beta}.$$

Идея метода оптимизации с помощью верхних оценок

Рассмотрим задачу оптимизации непрерывной функции

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

Предположим, что для функции $f(\boldsymbol{w})$ известна ее оценка сверху $Q(\boldsymbol{w},\boldsymbol{\xi})$, зависящая от дополнительного параметра $\boldsymbol{\xi}$, причем данная оценка является точной при $\boldsymbol{w}=\boldsymbol{\xi}$ (см. рис. 2). Таким образом,

$$f(\boldsymbol{w}) \le Q(\boldsymbol{w},\boldsymbol{\xi}),\ \forall \boldsymbol{w},\boldsymbol{\xi},$$
$$f(\boldsymbol{\xi}) = Q(\boldsymbol{\xi},\boldsymbol{\xi}).$$

Тогда задача минимизации (10) может быть решена путем покоординатной минимизации функции Q:

$$\boldsymbol{w}^{k+1} = \arg\min_{\boldsymbol{w}} Q(\boldsymbol{w},\boldsymbol{\xi}^k), \tag{11}$$
$$\boldsymbol{\xi}^{k+1} = \boldsymbol{w}^{k+1}.$$

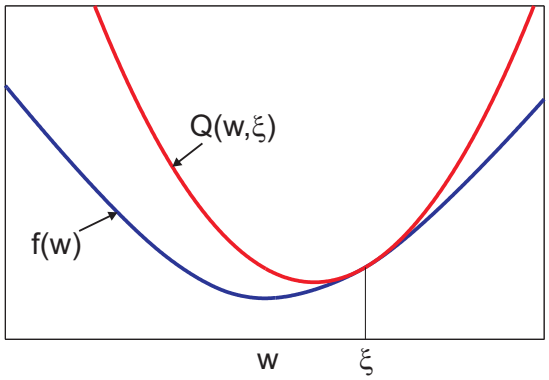

Рис. 2: Иллюстрация глобальной верхней оценки Q для функции f.

Здесь верхний индекс k обозначает номер итерации. Очевидно, что при таком подходе $f(\boldsymbol{w}^k) \geq f(\boldsymbol{w}^{k+1})$. Поэтому при дополнительном требовании об ограниченности снизу функции f итерационный процесс (11) гарантированно сходится к точке локального минимума функции f.

В общем случае параметр $\boldsymbol{\xi}$ в оценке Q может принадлежать пространству, отличному от пространства для $\boldsymbol{w}$, а сама верхняя оценка Q не обязательно должна быть точной для некоторых точек $\boldsymbol{w}$. В этом случае покоординатная минимизация Q не гарантирует нахождение локального минимума f, но может рассматриваться как приближенная процедура решения задачи (10).

Основным достоинством такой оптимизации является отсутствие необходимости вычисления функции f. Поэтому данный метод подходит для решения задачи (1) в ситуациях, когда соответствующий интеграл не вычисляется аналитически. Кроме того, по сравнению с градиентными методами оптимизация верхней оценки Q не требует дополнительной одномерной оптимизации по подбору величины шага на каждой итерации. При этом неявно предполагается, что задача минимизации функции Q по каждому из своих аргументов является значительно более простой задачей, чем минимизация исходной функции f, например, функция Q является квадратичной по $\boldsymbol{w}$ и может быть минимизирована аналитически.

Пример: LASSO

Рассмотрим в качестве примера применения оптимизации с верхней оценкой метод LASSO. Этот метод представляет собой решение задачи восстановления линейной регрессии по выборке $(\boldsymbol{t}, X)$ с использованием L_1-регуляризации:

$$f(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 + \alpha\sum_{d=1}^{D}|w_d| \to \min_{\boldsymbol{w}}. \tag{12}$$

Здесь $\boldsymbol{w} \in \mathbb{R}^D$ – веса линейной регрессии, $\alpha > 0$ – параметр регуляризации. Задача оптимизации (12) является выпуклой, но не гладкой.

Введем квадратичную верхнюю оценку для функции модуля (см. рис. 3), [2]:

$$|w_d| \leq \frac{w_d^2}{2|\xi_d|} + \frac{|\xi_d|}{2}. \tag{13}$$

Данная оценка является точной при $|w_d| = |\xi_d|$ и может быть получена непосредственно. Тогда

$$f(\boldsymbol{w}) \leq \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 + \alpha\sum_{d=1}^{D}\left(\frac{w_d^2}{2|\xi_d|} + \frac{|\xi_d|}{2}\right) \to \min_{\boldsymbol{w},\boldsymbol{\xi}}.$$

Таким образом, поиск решения исходной негладкой задачи оптимизации (12) заменяется на решение последовательности квадратичных задач оптимизации, что соответствует следующему итерационному процессу:

$$\begin{aligned} &\boldsymbol{w} = \left(X^TX + \operatorname{diag}\left(\frac{\alpha}{\xi_1}, \ldots, \frac{\alpha}{\xi_D}\right)\right)^{-1} X^T\boldsymbol{t}, \\ &\xi_d = |w_d|. \end{aligned} \tag{14}$$

Известно, что решение задачи (12) является разреженным, т.е. часть компонент оптимального вектора весов $\boldsymbol{w}$ равны нулю. Квадратичная оценка (13) становится вырожденной при $\xi_d \to 0$. Для преодоления возможных трудностей, связанных с работой со слишком большими или слишком маленькими числами, применяется

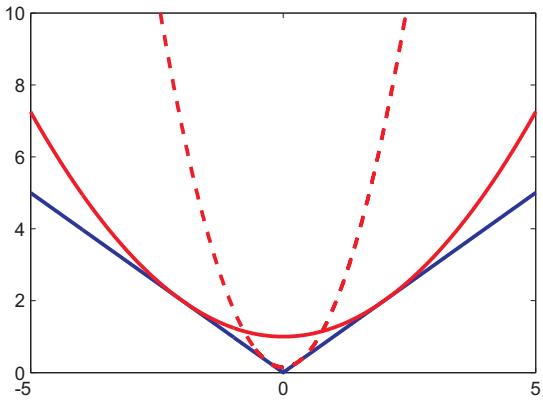

Рис. 3: Функция модуля (синяя кривая) и ее квадратичная верхняя оценка для разных значений ξ (красные кривые).

стандартная эвристика, в которой на текущей итерации все признаки, для которых $|w_d| < \varepsilon$, исключаются из рассмотрения.

В заключение заметим, что для решения СЛАУ (14) могут быть использованы алгоритмы, описанные выше, в частности, метод сопряженных градиентов с последовательным алгоритмом умножения матрицы на вектор.

Получение верхних оценок с помощью неравенства Йенсена

Пусть $f(\boldsymbol{p})$ – произвольная выпуклая функция, т.е.

$$f(\alpha \boldsymbol{p}_1 + (1-\alpha)\boldsymbol{p}_2) \le \alpha f(\boldsymbol{p}_1) + (1-\alpha) f(\boldsymbol{p}_2),\ \forall \boldsymbol{p}_1, \boldsymbol{p}_2,\ \alpha \in [0,1].$$

Это неравенство можно обобщить на случай большего числа точек $\boldsymbol{p}_i$:

$$f\left(\sum_i \alpha_i \boldsymbol{p}_i\right) \le \sum_i \alpha_i f(\boldsymbol{p}_i),\ \forall \{\boldsymbol{p}_i\}, \forall \boldsymbol{\alpha} : \alpha_i \ge 0,\ \sum_i \alpha_i = 1.$$

Полученный результат известен как неравенство Йенсена. Его можно записать и для непрерывного случая:

$$f\left(\int \alpha(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}\right) \le \int \alpha(\boldsymbol{x}) f(p(\boldsymbol{x})) d\boldsymbol{x},\ \forall \alpha(\boldsymbol{x}) \ge 0 : \int \alpha(\boldsymbol{x}) d\boldsymbol{x} = 1.$$

Рассмотрим задачу максимизации неполного правдоподобия

$$p(X|\Theta) = \int p(X, T|\Theta) dT \to \max_{\Theta}.$$

Пусть $q(T)$ – произвольное вероятностное распределение. Тогда справедлива следующая цепочка равенств и неравенств:

$$-\log p(X|\Theta) = -\log \int p(X,T|\Theta) dT = -\log \int \frac{p(X,T|\Theta)}{q(T)} q(T) dT \le \{\text{Н-во Йенсена}\} \le$$
$$-\int q(T) \log \frac{p(X,T|\Theta)}{q(T)} dT = -\mathbb{E}_q \log p(X,T|\Theta) + \mathbb{E}_q \log q \to \min_{\Theta, q}.$$

Решение данной задачи минимизации с помощью покоординатного спуска известно как EM-алгоритм:

$$\text{Минимизация по } q: \ q(T) = p(T|X,\Theta),\ \text{(E-шаг)}$$
$$\text{Минимизация по } \Theta: \ -\mathbb{E}_q \log p(X,T|\Theta) \to \min_{\Theta}.\ \text{(M-шаг)}$$

Таким образом, EM-алгоритм является примером алгоритма оптимизации с помощью верхней оценки, которая получается из неравенства Йенсена. В обобщенном EM-алгоритме минимизация по $q(T)$ может осуществляться не среди всех возможных распределений, а только в рамках некоторого семейства, например, факторизованных распределений, как в вариационном подходе [1], или параметрических. Задача оптимизации на M-шаге при этом может решаться не точно, а лишь с соблюдением условия уменьшения функционала на каждой итерации.

Рассмотрим применение EM-алгоритма для решения задачи обучения линейной регрессии с квадратичной регуляризацией:

$$p(\boldsymbol{t}|X,\alpha,\beta) = \int \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w},\beta^{-1}I)\mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha^{-1}I)d\boldsymbol{w} \to \max_{\alpha,\beta}.$$

Можно показать, что в данном случае задачи оптимизации на E и M-шаге решаются аналитически:

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\Sigma),\ \Sigma = (\beta X^T X + \alpha I)^{-1},\ \boldsymbol{\mu} = \beta\Sigma X^T\boldsymbol{t},\ \text{(E-шаг)}$$

$$\alpha = \frac{D}{\boldsymbol{\mu}^T\boldsymbol{\mu} + \operatorname{tr}\Sigma},\ \beta = \frac{N}{\|\boldsymbol{t} - X\boldsymbol{\mu}\|^2 + \operatorname{tr}\Sigma X^T X}.\ \text{(M-шаг)}$$

Заметим, что для проведения итерационного процесса по указанным формулам нет необходимости вычислять значение оптимизируемой функции $p(\boldsymbol{t}|X,\alpha,\beta)$.

Однако, при применении EM-алгоритма для решения задачи обучения модели линейной регрессии с L_1-регуляризацией или с регуляризацией по Стьюденту возникающие задачи оптимизации на E и M-шаге не решаются аналитически. Поэтому здесь возникает потребность в разработке других верхних оценок для оптимизации.

Получение верхних оценок с помощью построения касательной и замены переменных

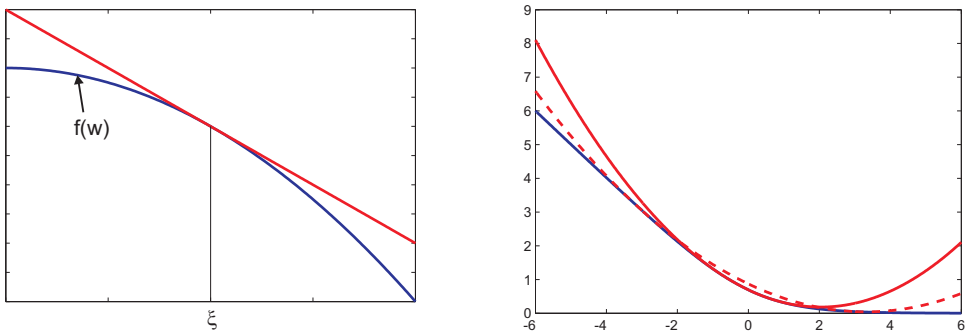

Рис. 4: Слева: касательная к вогнутой функции является верхней оценкой, справа: квадратичная верхняя оценка с различными ξ (красные кривые) для функции $\log(\exp(-w)+1)$ (синяя кривая).

Пусть $f(w)$ – произвольная вогнутая функция от скалярного аргумента. Тогда касательная к данной функции в произвольной точке ξ будет для f глобальной верхней оценкой (см. рис. 4,слева):

$$f(w) \le f(\xi) + f'(\xi)(w-\xi).$$

Пусть теперь $f(w)$ не является вогнутой функцией. Рассмотрим взаимно-однозначную замену переменной $v = h(w)$ и функцию $g(v) = f(h^{-1}(v))$. Предположим, что функция g является вогнутой. Тогда

$$g(v) = f(h^{-1}(v)) \le g(\xi) + g'(\xi)(v-\xi).$$

Возвращаясь к исходной переменной w и обозначая $\eta = h^{-1}(\xi)$, получаем верхнюю оценку для f:

$$f(w) \le g(h(\eta)) + g'(h(\eta))(h(w) - h(\eta)). \tag{15}$$

Особый интерес для нас представляют квадратичные верхние оценки, т.к. их легко минимизировать. Для получения квадратичной оценки с помощью (15) необходимо выбрать замену переменной вида $v = h(w) = w^2$, которая должна быть взаимно-однозначной. Выбранная замена будет взаимно-однозначной для функции f, определенной на положительной полуоси, а также для четной функции f.

Рассмотрим для примера получение верхней оценки для функции $f(w) = \log(1+\exp(-w))$. Эта функция может быть приведена к четной следующим образом:

$$f(w) = \log(1+\exp(-w)) = -\frac{w}{2} + \underbrace{\log\left(\exp\left(-\frac{w}{2}\right) + \exp\left(\frac{w}{2}\right)\right)}_{g(w)}.$$

Функция $g(w)$ является четной и выпуклой. Для получения вогнутой функции рассмотрим функцию $g(w)$ для $w \ge 0$ и взаимооднозначную для данной области замену переменной $v = w^2$:

$$\tilde{g}(v) = g(\sqrt{v}) = \log\left(\exp\left(-\frac{\sqrt{v}}{2}\right) + \exp\left(\frac{\sqrt{v}}{2}\right)\right).$$

Функция $\tilde{g}$ является вогнутой, поэтому мы можем построить ее верхнюю оценку с помощью касательной. Возвращаясь к исходной функции f, получаем следующую квадратичную оценку (см. рис. 4,справа):

$$f(w) = \log(1+\exp(-w)) \leq \log(1+\exp(-\xi)) - \frac{w-\xi}{2} + \frac{\tanh(\xi/2)}{4\xi}(w^2-\xi^2).$$

Здесь через tanh обозначен гиперболический тангенс. Данная оценка известна как оценка Йаакола-Джордана (Jaakkola-Jordan bound, [3]). Обычно она используется как оценка для логистической функции:

$$\sigma(w) = \frac{1}{1+\exp(-w)} \geq \sigma(\xi)\exp\left(\frac{w-\xi}{2} - \frac{\tanh(\xi/2)}{4\xi}(w^2-\xi^2)\right).$$

Рассмотрим еще два примера построения квадратичных оценок с помощью (15):

- Функция $f(w) = |w|^p$, $0 < p \leq 2$. Данная функция является четной. Для получения квадратичной оценки рассмотрим замену переменной $v = w^2$ и функцию $g(v) = \sqrt{v}^p$. Эта функция является вогнутой для $0 < p \leq 2$. Построение касательной для нее и возврат к исходной переменной w дает следующую оценку:

$$|w|^p \leq |\xi|^p + \frac{p|\xi|^p}{2\xi^2}(w^2-\xi^2).$$

  Для случая $p = 1$ данная оценка переходит в (13).

- Функция $f(w) = \log\left(1+\frac{\alpha}{\nu}w^2\right)$, $\alpha, \nu > 0$. Эта функция является четной. Действуя по аналогии с предыдущим случаем, получаем следующую оценку:

$$f(w) = \log\left(1+\frac{\alpha}{\nu}w^2\right) \leq f(\xi) + \frac{\alpha}{\alpha\xi^2+\nu}(w^2-\xi^2).$$

  Эта оценка может быть использована для получения нижней оценки для плотности распределения Стьюдента:

$$\mathcal{S}(w|\alpha,\nu) = \frac{\sqrt{\alpha}\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{\alpha}{\nu}w^2\right)^{-(\nu+1)/2} \geq \mathcal{S}(\xi|\alpha,\nu)\exp\left(-\frac{\alpha(\nu+1)}{2(\alpha\xi^2+\nu)}(w^2-\xi^2)\right). \quad (16)$$

Рассмотрим решение задачи обучения L_1-регуляризованной линейной регрессии с помощью оценки (13):

$$p(\boldsymbol{t}|X,\alpha,\beta) = \int \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w},\beta^{-1}I)\prod_{d=1}^{D}\frac{\alpha}{2}\exp(-\alpha|w_d|)d\boldsymbol{w} \geq$$

$$\geq F_{L_1}(\alpha,\beta,\boldsymbol{\xi}) = \sqrt{\frac{\beta}{2\pi}}^N\left(\frac{\alpha}{2}\right)^D\int\exp\left(-\frac{\beta}{2}\|\boldsymbol{t}-X\boldsymbol{w}\|^2 - \alpha\sum_{d=1}^{D}\left[\frac{w_d^2}{2|\xi_d|}+\frac{|\xi_d|}{2}\right]\right)d\boldsymbol{w} \to \max_{\alpha,\beta,\boldsymbol{\xi}}. \quad (17)$$

Интеграл в полученной нижней оценке вычисляется аналитически, т.к. он представляет собой интеграл от ненормированной гауссианы. Можно показать, что функция $F_{L_1}(\alpha,\beta,\boldsymbol{\xi})$ является вогнутой по всем своим переменным и максимизируется с помощью следующего итерационного процесса:

$$\Sigma = \left(\beta X^T X + \operatorname{diag}\left(\frac{\alpha}{\xi_1},\ldots,\frac{\alpha}{\xi_D}\right)\right)^{-1},\ \boldsymbol{\mu} = \beta\Sigma X^T\boldsymbol{t},$$

$$\gamma = D - \alpha\sum_{d=1}^{D}\frac{\Sigma_{dd}}{\xi_d},\ \alpha = \frac{D+\gamma}{\sum_{d=1}^{D}\left[\xi_d + \frac{\mu_d^2}{\xi_d}\right]},\ \beta = \frac{N-\gamma}{\|\boldsymbol{t}-X\boldsymbol{\mu}\|^2},$$

$$\xi_d = \sqrt{\mu_d^2+\Sigma_{dd}}.$$

Перейдем к решению задачи обучения линейной регрессии с регуляризацией по Стьюденту. Применяя оценку (16), получаем

$$p(\boldsymbol{t}|X,\alpha,\beta) = \int \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w},\beta^{-1}I)\prod_{d=1}^{D}\mathcal{S}(w_d|\alpha,\nu)d\boldsymbol{w} \geq$$

$$\geq \sqrt{\frac{\beta}{2\pi}}^N\prod_{d=1}^{D}\mathcal{S}(\xi_d|\alpha,\nu)\int\exp\left(-\frac{\beta}{2}\|\boldsymbol{t}-X\boldsymbol{w}\|^2 - \frac{\alpha(\nu+1)}{2}\sum_{d=1}^{D}\frac{w_d^2}{\alpha\xi_d^2+\nu}\right)d\boldsymbol{w}.$$

Здесь, как и раньше, под интегралом стоит ненормированная гауссиана, и поэтому интеграл вычисляется аналитически. Переходя к логарифму и отбрасывая величины, не зависящие от $\alpha, \beta, \boldsymbol{\xi}$, получаем

$$F_{student}(\boldsymbol{\mu}, \alpha, \beta, \boldsymbol{\xi}) = N \log \beta + D \log \alpha - \beta \|\boldsymbol{t} - X\boldsymbol{\mu}\|^2 - \frac{\alpha(\nu+1)}{\nu} \sum_{d=1}^{D} \xi_d^2 - \sum_{d=1}^{D} \frac{\alpha(\nu+1)(\mu_d^2 - \xi_d^2)}{\alpha\xi_d^2 + \nu} -$$
$$- \log \det \left[\beta X^T X + \operatorname{diag}\left(\frac{\alpha(\nu+1)}{\alpha\xi_1^2 + \nu}, \ldots, \frac{\alpha(\nu+1)}{\alpha\xi_D^2 + \nu} \right) \right] \to \max_{\boldsymbol{\mu}, \alpha, \beta, \boldsymbol{\xi}}. \quad (18)$$

В отличие от F_{L_1} (17) функция $F_{student}$ не является вогнутой по своим аргументам. В результате ее максимизация методом покоординатного подъема может приводить к неустойчивой процедуре.

Выпукло-вогнутая процедура (CCCP)

Рассмотрим задачу минимизации $f(\boldsymbol{w})$ и представим оптимизируемую функцию в виде суммы выпуклой $f_\cup$ и вогнутой $f_\cap$ части

$$f(\boldsymbol{w}) = f_\cup(\boldsymbol{w}) + f_\cap(\boldsymbol{w}) \to \min_{\boldsymbol{w}}.$$

Заметим, что подобное разложение является неоднозначным и всегда существует для произвольной функции с ограниченным гессианом. Ограничим сверху вогнутую часть $f_\cap$ с помощью касательной в точке $\boldsymbol{\xi}$:

$$f(\boldsymbol{w}) = f_\cup(\boldsymbol{w}) + f_\cap(\boldsymbol{w}) \le f_\cup(\boldsymbol{w}) + f_\cap(\boldsymbol{\xi}) + \nabla f_\cap(\boldsymbol{\xi})^T(\boldsymbol{w} - \boldsymbol{\xi}) \to \min_{\boldsymbol{w}, \boldsymbol{\xi}}.$$

Таким образом, получается выпуклая верхняя оценка, минимизация которой во многих случаях является существенно более простой задачей, чем минимизация исходной невыпуклой функции f. В частности, приравнивая градиент по $\boldsymbol{w}$ к нулю, получаем:

$$\nabla f_\cup(\boldsymbol{w}) + \nabla f_\cap(\boldsymbol{\xi}) = \mathbf{0} \;\Rightarrow\; \nabla f_\cup(\boldsymbol{w}) = -\nabla f_\cap(\boldsymbol{\xi}).$$

Отсюда итерационный процесс поиска точки минимума может быть записан как (см. рис. ??)

$$\boldsymbol{w}^{k+1} : \nabla f_\cup(\boldsymbol{w}^{k+1}) = -\nabla f_\cap(\boldsymbol{w}^k).$$

Различные примеры применения выпукло-вогнутой процедуры (ConCave-Convex Procedure или сокращенно CCCP) представлены в работе [4].

Рассмотрим в качестве примера применения CCCP задачу максимизации $F_{student}$ (18). Все слагаемые в выражении для $F_{student}$ являются вогнутыми по своим аргументам, кроме слагаемого $\alpha(\nu+1)\xi_d^2/(\alpha\xi_d^2 + \nu)$. Добавляя и вычитая ξ_d^2 с нужным множителем, получаем:

$$\frac{\alpha\xi_d^2}{\alpha\xi_d^2 + \nu} = \underbrace{\frac{\alpha\xi_d^2}{\alpha\xi_d^2 + \nu} - \frac{\alpha\xi_d^2}{\nu}}_{\text{вогнутая}} + \underbrace{\frac{\alpha\xi_d^2}{\nu}}_{\text{выпуклая}} \ge \frac{\alpha\xi_d^2}{\alpha\xi_d^2 + \nu} - \frac{\alpha\xi_d^2}{\nu} + \frac{\alpha\zeta_d^2}{\nu} + \frac{2\alpha\zeta_d}{\nu}(\xi_d - \zeta_d).$$

С помощью данной нижней оценки задачу (18) можно свести к решению последовательности вогнутых задач максимизации. Детали данного алгоритма представлены в работе [5].

Список литературы

[1] C.M. Bishop. Pattern Recognition and Machine Learning, Springer, 2006.

[2] D. Hunter, K. Lange. Quantile Regression via an MM Algorithm // Journal of Computational and Graphical Statistics, 2000.

[3] T. Jaakkola, M. Jordan. Bayesian parameter estimation via variational methods // Statistics and Computing, Vol. 10, 2000, pp. 25–37.

[4] A. Yuille, A. Rangarajan. The Concave-Convex Procedure (CCCP) // Neural Computation, 2003.

[5] M. Seeger, H. Nickisch. Large Scale Bayesian Inference and Experimental Design for Sparse Linear Models // SIAM Journal Imaging Sciences, Vol. 4, No. 1, 2011, pp. 166–199.