Спецсеминар «Байесовские методы машинного обучения»

Локальные вариационные оценки для решения задач оптимизации в машинном обучении

Дата: 14 марта 2012

Идея подхода

Рассмотрим задачу оптимизации непрерывной функции

$$f(\boldsymbol{w}) \to \min_{\boldsymbol{w}}. \tag{1}$$

Предположим, что для функции $f(\boldsymbol{w})$ известна ее оценка сверху $Q(\boldsymbol{w}, \boldsymbol{\xi})$, зависящая от дополнительного вариационного параметра $\boldsymbol{\xi}$, причем данная оценка является точной при $\boldsymbol{w} = \boldsymbol{\xi}$ (см. рис. 1). Таким образом,

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

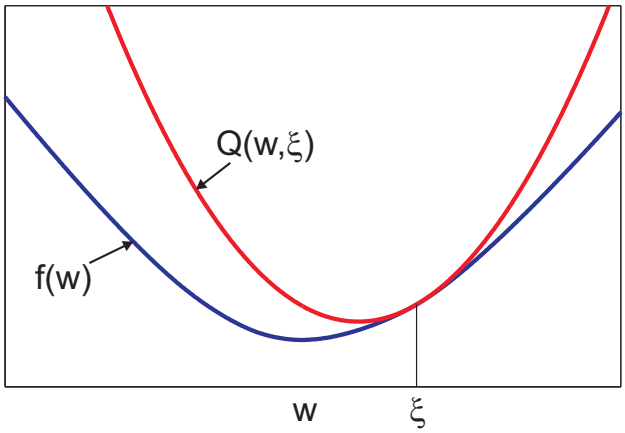

Рис. 1: Иллюстрация вариационной оценки Q для функции f.

Тогда задача минимизации (1) может быть решена путем покоординатной минимизации функции Q:

$$\boldsymbol{w}^{t+1} = \arg\min_{\boldsymbol{w}} Q(\boldsymbol{w}, \boldsymbol{\xi}^t), \tag{2}$$
$$\boldsymbol{\xi}^{t+1} = \boldsymbol{w}^{t+1}.$$

Здесь верхний индекс t обозначает номер итерации. Действительно, $f(\boldsymbol{w}^t) = \{\boldsymbol{\xi}^t = \boldsymbol{w}^t\} = Q(\boldsymbol{w}^t, \boldsymbol{\xi}^t) \ge Q(\boldsymbol{w}^{t+1}, \boldsymbol{\xi}^t) \ge Q(\boldsymbol{w}^{t+1}, \boldsymbol{\xi}^{t+1}) = f(\boldsymbol{w}^{t+1})$. При дополнительном требовании об ограниченности снизу функции f итерационный процесс (2) гарантированно сходится к точке локального минимума функции f.

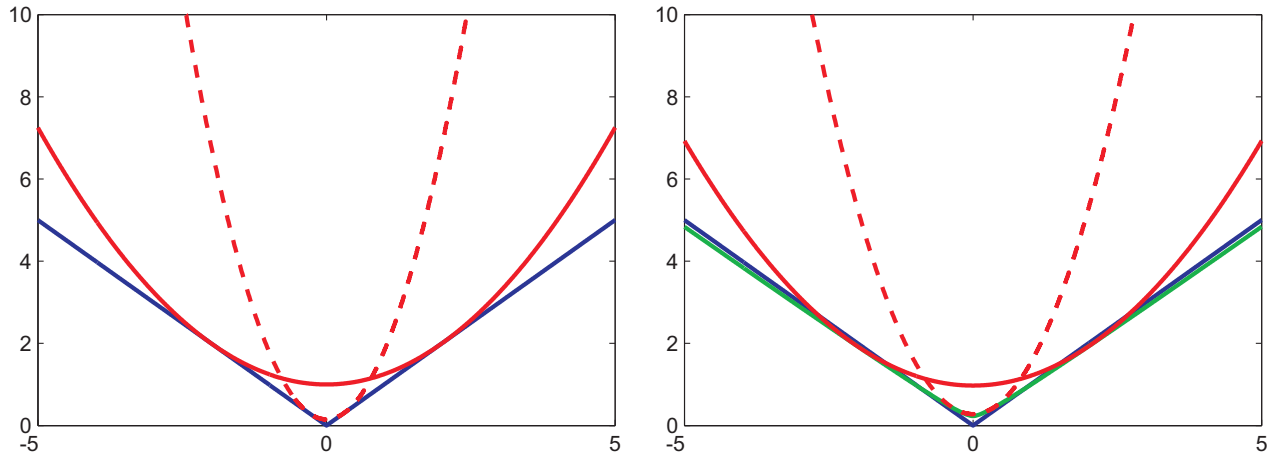

Рис. 2: Слева: функция модуля (синяя кривая), квадратичные вариационные оценки с разным значением ξ (красные кривые). Справа: функция модуля (синяя кривая), сглаженная функция модуля (зеленая кривая), квадратичные вариационные оценки для сглаженного модуля (красные кривые).

В более общем случае вариационный параметр $\boldsymbol{\xi}$ в оценке Q может принадлежать пространству, отличному от пространства для $\boldsymbol{w}$, а сама вариационная оценка Q не обязательно должна становится 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 (1).

Основным достоинством такой вариационной оптимизации по сравнению с градиентными методами является отсутствие необходимости выбора 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 вариационной оптимизации метод LASSO. Этот метод представляет собой решение задачи восстановления линейной регрессии по выборке $\{\boldsymbol{x}_n, t_n\}_{n=1}^N$, где $\boldsymbol{x}_n \in \mathbb{R}^d$ – признаки объекта, а $t_n \in \mathbb{R}$ – регрессионная переменная, с использованием L_1-регуляризации:

$$f(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 + \lambda\sum_{i=1}^{d}|w_i| \to \min_{\boldsymbol{w}}. \tag{3}$$

Здесь $\boldsymbol{w} \in \mathbb{R}^d$ – веса линейной регрессии, $\lambda > 0$ – параметр регуляризации. Задача оптимизации (3) является выпуклой, но не гладкой.

Введем квадратичную вариационную оценку для функции модуля (см. рис. 2,слева):

$$|w_i| \le \frac{w_i^2}{2|\xi_i|} + \frac{|\xi_i|}{2}.$$

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

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

Таким образом, поиск решения исходной негладкой задачи оптимизации (3) заменяется на решение последовательности квадратичных задач оптимизации, что соответствует следующему итерационному процессу:

$$\begin{aligned}\boldsymbol{w} &= \left(X^TX + \operatorname{diag}\left(\frac{\lambda}{|\xi_1|},\ldots,\frac{\lambda}{|\xi_d|}\right)\right)^{-1}X^T\boldsymbol{t},\\ \xi_i &= |w_i|.\end{aligned} \tag{4}$$

Здесь $X = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N]^T$.

Известно, что решение задачи (3) является разреженным, т.е. часть компонент оптимального вектора весов $\boldsymbol{w}$ равны нулю. Близость к нулю некоторых компонент w_i создает трудности в итерационном процессе (4), т.к. квадратичная вариационная оценка становится вырожденной при $\xi_i \to 0$. Для преодоления этих трудностей авторы статьи [1] предложили заменить функцию модуля на ее гладкую аппроксимацию (см. рис. 2,справа):

$$|w_i| \simeq |w_i| - \varepsilon\log(\varepsilon + |w_i|).$$

Здесь $\varepsilon > 0$ – некоторый параметр. Таким образом, задача оптимизации (3) заменяется на задачу оптимизации

$$f_\varepsilon(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 + \lambda\sum_{i=1}^{d}(|w_i| - \varepsilon\log(\varepsilon + |w_i|)) \to \min_{\boldsymbol{w}}. \tag{5}$$

Можно доказать (см. [1]), что для достаточно малого ε для решений исходной задачи оптимизации (3) $\hat{\boldsymbol{w}}$ и сглаженной задачи (5) $\hat{\boldsymbol{w}}_\varepsilon$ справедливы следующие утверждения:

$$\begin{aligned}&f(\hat{\boldsymbol{w}}_\varepsilon) - f(\hat{\boldsymbol{w}}) \leq -2\lambda\varepsilon N\log\varepsilon,\\ &\hat{\boldsymbol{w}}_\varepsilon \to \hat{\boldsymbol{w}} \text{ при } \varepsilon \to 0.\end{aligned}$$

Введем квадратичную вариационную оценку для сглаженной функции модуля (см. рис. 2,справа):

$$|w_i| - \varepsilon\log(\varepsilon + |w_i|) \leq |\xi_i| - \varepsilon\log(\varepsilon + |\xi_i|) + \frac{w_i^2 - \xi_i^2}{2(\varepsilon + |\xi_i|)}.$$

В результате итерационный процесс для решения задачи (5) выглядит как

$$\begin{aligned}\boldsymbol{w} &= \left(X^TX + \operatorname{diag}\left(\frac{\lambda}{|\xi_1| + \varepsilon},\ldots,\frac{\lambda}{|\xi_d| + \varepsilon}\right)\right)^{-1}X^T\boldsymbol{t},\\ \xi_i &= |w_i|.\end{aligned}$$

Теперь при $\xi_i \to 0$ матрица корректно обращается.

В заключение данного раздела заметим, что представленный метод для решения задачи (3) носит, скорее, иллюстративный характер. На практике здесь стоит применять библиотеку GLMNET[1] или LIBLINEAR[2], которые рассчитаны на использование, в том числе, в условиях больших размерностей обучающей выборки, условиях разреженности входных данных и т.д.

[1] http : //www – stat.stanford.edu/ ∼ tibs/glmnet – matlab/

[2] http : //www.csie.ntu.edu.tw/ ∼ cjlin/liblinear/

Общие методы получения вариационных оценок

Неравенство Йенсена

Пусть $f(x)$ – произвольная выпуклая функция. Тогда

$$f\left(\sum_i \alpha_i x_i\right) \leq \sum_i \alpha_i f(x_i),\ \forall \boldsymbol{x}, \forall \boldsymbol{\alpha} : \alpha_i \geq 0,\ \sum_i \alpha_i = 1.$$

Этот факт известен как неравенство Йенсена. 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:

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

Решение данной задачи максимизации с помощью покоординатного подъема известно как EM-алгоритм:

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

Таким образом, EM-алгоритм является примером алгоритма вариационной оптимизации, где соответствующая вариационная оценка получается из неравенства Йенсена.

Пусть имеется вероятностное распределение, известное с точностью до нормировочной константы:

$$p(T) = \frac{1}{Z} \tilde{p}(T).$$

Для оценки нормировочной константы Z воспользуемся неравенством Йенсена:

$$\log Z = \log \int \tilde{p}(T) dT = \log \int \frac{\tilde{p}(T)}{q(T)} q(T) dT \geq \{\text{Н-во Йенсена}\} \geq \int q(T) \log \frac{\tilde{p}(T)}{q(T)} dT \to \max_{q(T) = \prod_i q_i(T_i)}.$$

Данная задача максимизации в семействе факторизованных распределений $q(T) = \prod_i q_i(T_i)$ может быть решена с помощью покоординатного подъема:

$$q_j^*(T_j) = \frac{\exp(\int \log \tilde{p}(T) \prod_{i \neq j} q_i(T_i) dT_i)}{\int \exp(\int \log \tilde{p}(T) \prod_{i \neq j} q_i(T_i) dT_i) dT_j}.$$

Этот результат известен как вариационный подход или mean-field approximation для задачи байесовского вывода. Если интеграл от $\log \tilde{p}(T)$ по набору факторов q_j не может быть найден аналитически, то можно попробовать ввести вариационную нижнюю оценку для $\tilde{p}(T)$: $0 < F(T, \boldsymbol{\xi}) \leq \tilde{p}(T)$. Тогда

$$\log Z = \log \int \tilde{p}(T) dT \geq \log \int F(T, \boldsymbol{\xi}) dT \geq \int q(T) \log \frac{F(T, \boldsymbol{\xi})}{q(T)} dT \to \max_{q, \boldsymbol{\xi}}.$$

Решение данной задачи максимизации по распределению q в рамках факторизованного семейства аналогично результату из вариационного подхода:

$$q_j^*(T_j) = \frac{\exp(\int \log F(T, \boldsymbol{\xi}) \prod_{i \neq j} q_i(T_i) dT_i)}{\int \exp(\int \log F(T, \boldsymbol{\xi}) \prod_{i \neq j} q_i(T_i) dT_i) dT_j}.$$

Задача оптимизации по вариационному параметру $\boldsymbol{\xi}$ решается с помощью стандартных методов оптимизации. Представленная модификация вариационного подхода получила название локального вариационного подхода.

Построение касательной и замена переменных

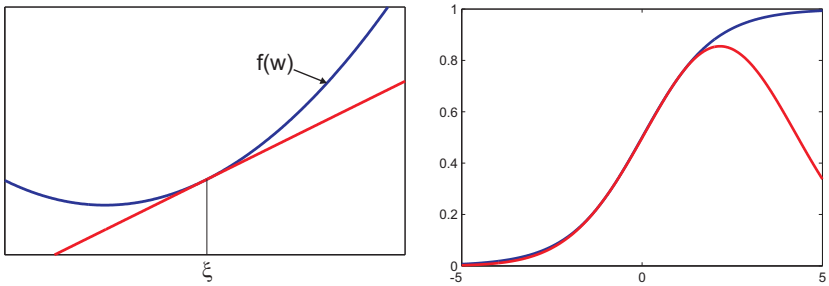

Рис. 3: Слева: касательная к выпуклой функции является вариационной оценкой, справа: нижняя вариационная оценка (красная кривая) для логистической функции (синяя кривая).

Пусть $f(w)$ – произвольная выпуклая функция от скалярного аргумента. Тогда касательная к данной функции в произвольной точке ξ будет для f вариационной нижней оценкой (см. рис. 3,слева):

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

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

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

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

$$f(w) \geq g(h(\eta)) + g'(h(\eta))(h(w) - h(\eta)).$$

Рассмотрим применение описанного способа для получения вариационной нижней оценки логистической функции $f(w) = 1/(1 + \exp(-w))$. Справедлива следующая цепочка равенств:

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

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

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

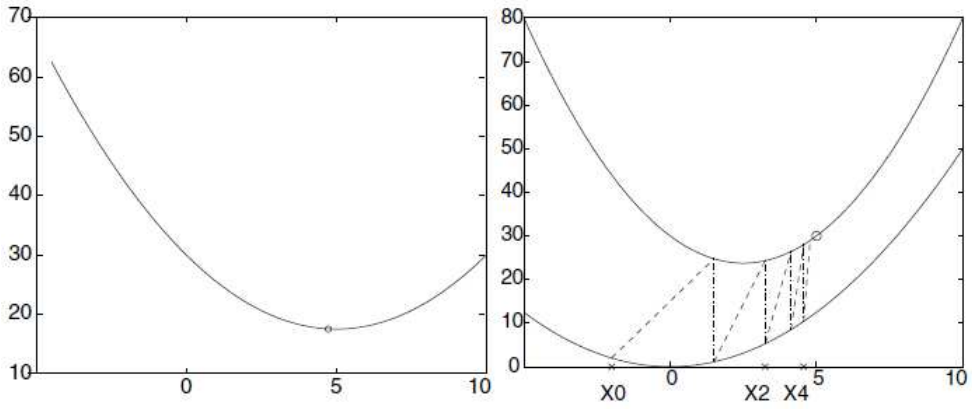

Рис. 4: Иллюстрация работы выпукло-вогнутой процедуры. Слева: исходная минимизируемая функция. Справа: выпуклая компонента (верхняя кривая) и минус вогнутая компонента (нижняя кривая), очередная точка итерационного процесса выбирается из условия совпадения производных двух кривых.

Функция $\tilde{g}$ является выпуклой и поэтому подходит для получения нижней оценки с помощью касательной. Возвращаясь к исходной функции f, получаем следующую вариационную оценку (см. рис. 3,справа):

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

Здесь через tanh обозначен гиперболический тангенс. Данная оценка известна как оценка Йаколла-Джордана (JJ bound) и лежит в основе многих применений локального вариационного подхода [2].

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 $\boldsymbol{\xi}$:

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

Таким образом, получается выпуклая вариационная оценка сверху, 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. 4)

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

Различные примеры применения выпукло-вогнутой процедуры (ConCave-Convex Procedure или сокращенно CCCP) представлены в работе [3].

Байесовские модели с супергауссовскими регуляризаторами

Рассмотрим вероятностную модель линейной регрессии с квадратичной регуляризацией:

$$\begin{aligned} &p(\boldsymbol{t}, \boldsymbol{w}|X, \alpha, \beta) = p(\boldsymbol{t}|X, \boldsymbol{w}, \beta) \prod_{i=1}^{d} p(w_i|\alpha), \\ &p(\boldsymbol{t}|X, \boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w}, \beta^{-1}) \text{ – правдоподобие данных,} \\ &p(w_i|\alpha) = \mathcal{N}(w_i|0, \alpha^{-1}),\ i = 1, \ldots, d \text{ – квадратичный регуляризатор.} \end{aligned} \tag{6}$$

Здесь $(\boldsymbol{t}, X) = \{t_n, \boldsymbol{x}_n\}_{n=1}^{N}$ – обучающая выборка, $\boldsymbol{w} \in \mathbb{R}^d$ – веса линейной регрессии, $\alpha > 0$ – параметр регуляризации, $\beta > 0$ – уровень шума.

Пусть для простоты уровень шума β известен. Найдем значение параметра регуляризации α с помощью метода максимального правдоподобия (обоснованности):

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

Справедлива следующая цепочка равенств:

$$\begin{aligned} p(\boldsymbol{t}|X, \alpha, \beta) &= \int \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w}, \beta^{-1}) \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1} I) d\boldsymbol{w} = \\ &= \sqrt{\frac{\beta}{2\pi}}^{N} \sqrt{\frac{\alpha}{2\pi}}^{d} \int \exp\left(-\frac{\beta}{2} (\boldsymbol{t} - X\boldsymbol{w})^T (\boldsymbol{t} - X\boldsymbol{w}) - \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} \right) d\boldsymbol{w} = \\ &= \sqrt{\frac{\beta}{2\pi}}^{N} \sqrt{\frac{\alpha}{2\pi}}^{d} \exp\left(\frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} \right) \int \exp\left(-\frac{1}{2} (\boldsymbol{w} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{w} - \boldsymbol{\mu}) \right) d\boldsymbol{w} = \\ &= \sqrt{\frac{\beta}{2\pi}}^{N} \sqrt{\alpha}^{d} \exp\left(\frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} \right) \sqrt{\det \Sigma}. \end{aligned}$$

Здесь $\Sigma^{-1} = \beta X^T X + \alpha I$, $\boldsymbol{\mu} = \beta \Sigma X^T \boldsymbol{t}$. Переходя к логарифму обоснованности и приравнивая производную по α к нулю, получаем следующую итерационную формулу пересчета:

$$\alpha^{new} = \frac{d - \alpha^{old} \mathrm{tr}\Sigma}{\boldsymbol{\mu}^T \boldsymbol{\mu}}.$$

Найденное после обучения значение α может быть использовано для поиска оптимальных весов $\boldsymbol{w}$ по принципу максимума апостериорного распределения:

$$p(\boldsymbol{w}|\boldsymbol{t}, X, \alpha, \beta) \to \max_{\boldsymbol{w}} \ \Leftrightarrow \ p(\boldsymbol{t}|X, \boldsymbol{w}, \beta) p(\boldsymbol{w}|\alpha) \to \max_{\boldsymbol{w}}. \tag{8}$$

Решение этой задачи может быть найдено аналитически: $\boldsymbol{w} = (\beta X^T X + \alpha I)^{-1} \beta X^T \boldsymbol{t}$.

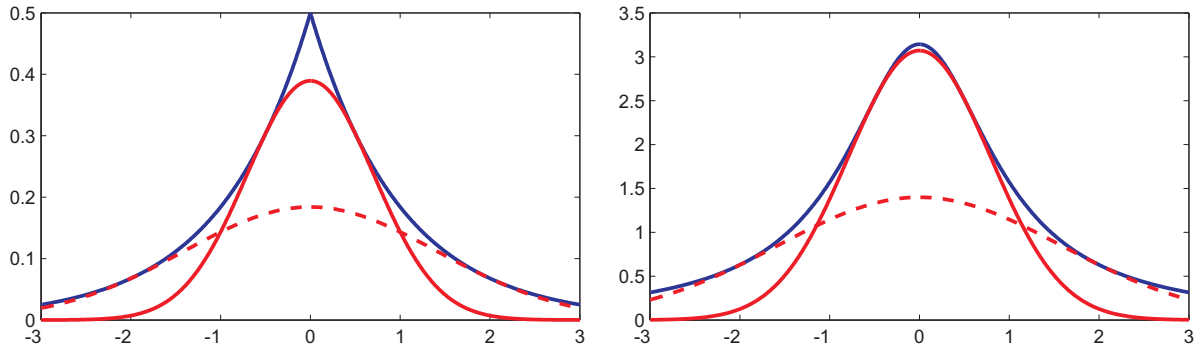

Рис. 5: Примеры супергауссовских потенциалов. Слева: распределение Лапласа (синяя кривая) и его вариационные нижние оценки с помощью ненормированной гауссианы (красные кривые), справа: распределение Стьюдента (синяя кривая) и его вариационные нижние оценки с помощью ненормированной гауссианы (красные кривые).

Рассмотрим теперь в качестве регуляризатора для отдельного веса $p(w_i|\alpha)$ распределения, ненормированная плотность (потенциал) которых принадлежит т.н. супергауссовскому семейству. Это семейство включает в себя все потенциалы (всюду неотрицательные функции), для которых может быть построена точная вариационная оценка снизу с помощью ненормированной гауссианы (см. рис. 5). Формально, $p(w)$ принадлежит супергауссовскому семейству, если найдется $b \in \mathbb{R}$ такое, что функция $\log p(w) - bw$ является четной, а функция $\log p(\sqrt{v}) - b\sqrt{v}$ является выпуклой для $v \geq 0$. Использование касательной для этой выпуклой функции дает искомую вариационную гауссовскую оценку снизу для супергауссовского потенциала.

Приведем несколько примеров супергауссовских потенциалов. Как было показано выше, логистическая функция $\sigma(w) = 1/(1 + \exp(-w))$ является супергауссовским потенциалом при $b = 1/2$ (оценка Йаколлы-Джордана). Рассмотрим потенциал, связанный с L_p-регуляризацией:

$$p(w) \propto \exp(-\alpha|w|^p).$$

Функция $\log p(w) = -\alpha|w|^p$ является четной, а функция $-\alpha\sqrt{v}^p$ является выпуклой для всех $p \in (0, 2)$. Таким образом, получаем следующую вариационную оценку:

$$\exp(-\alpha|w|^p) \geq \exp(-\alpha|\xi|^p) \exp\left(-\frac{\alpha p|\xi|^p}{2\xi^2}(w^2 - \xi^2)\right).$$

Данная оценка является точной при $|w| = |\xi|$. Случай $p = 1$ (распределение Лапласа) проиллюстрирован на рис. 5,слева. Другим примером супергауссовского потенциала является распределение Стьюдента:

$$p(w) \propto \left(1 + \frac{1}{\alpha}w^2\right)^{-\frac{\nu+1}{2}}, \ \alpha, \nu > 0.$$

Действительно, функция $\log p(w) = -\frac{\nu+1}{2}\log\left(1 + \frac{1}{\alpha}w^2\right)$ является четной (но не вогнутой), а функция $-\frac{\nu+1}{2}\log\left(1 + \frac{1}{\alpha}v\right)$ является выпуклой. В результате получаем следующую вариационную оценку:

$$\left(1 + \frac{1}{\alpha}w^2\right)^{-\frac{\nu+1}{2}} \geq \left(1 + \frac{1}{\alpha}\xi^2\right)^{-\frac{\nu+1}{2}} \exp\left(-\frac{\nu+1}{2(\alpha+\xi^2)}(w^2 - \xi^2)\right).$$

Данная оценка является точной при $|w| = |\xi|$ и проиллюстрирована на рис. 5,справа. Также можно показать (см. [4]), что линейная комбинация с положительными весами супергауссовских потенциалов является супергауссовским потенциалом.

Использование в модели линейной регрессии (6) супергауссовских регуляризаторов позволяет эффективно решить задачи обучения параметра регуляризации (7) и поиска весов (8). Представим нижнюю вариационную оценку для супергауссовских потенциалов $p(w_i|\alpha)$ в следующем виде:

$$p(w_i|\alpha) \geq \exp\left(-\frac{1}{2}a_i(\xi_i)w_i^2 + b_i(\xi_i)w_i + c_i(\xi_i)\right),$$

где ξ_i – вариационный параметр. Тогда

$$p(\boldsymbol{t}|X,\alpha,\beta) = \int \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w},\beta^{-1})\prod_{i=1}^{d} p(w_i|\alpha)d\boldsymbol{w} \geq$$

$$\sqrt{\frac{\beta}{2\pi}}^N \int \exp\left(-\frac{\beta}{2}(\boldsymbol{t}-X\boldsymbol{w})^T(\boldsymbol{t}-X\boldsymbol{w}) - \frac{1}{2}\sum_i a_i w_i^2 + b_i w_i + c_i\right) d\boldsymbol{w} \geq$$

$$\sqrt{\frac{\beta}{2\pi}}^N \exp\left(\sum_i c_i\right)\exp\left(\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right)\int \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right)d\boldsymbol{w} =$$

$$\sqrt{\frac{\beta}{2\pi}}^N \exp\left(\sum_i c_i + \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right)\sqrt{2\pi}^M\sqrt{\det\Sigma} \to \max_{\boldsymbol{w},\boldsymbol{\xi}}. \quad (9)$$

Здесь $\Sigma^{-1} = \beta X^T X + \operatorname{diag}(\boldsymbol{a})$, $\boldsymbol{\mu} = \Sigma(\beta X^T\boldsymbol{t} + \boldsymbol{b})$. Заметим, что вектора $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ зависят от вариационного параметра $\boldsymbol{\xi}$. Аналогично,

$$p(\boldsymbol{w}|X,\boldsymbol{t},\alpha,\beta) \propto \mathcal{N}(\boldsymbol{t}|X\boldsymbol{w},\beta^{-1})\prod_{i=1}^{d} p(w_i|\alpha) \geq$$

$$\sqrt{\frac{\beta}{2\pi}}^N \exp\left(-\frac{\beta}{2}\|\boldsymbol{t}-X\boldsymbol{w}\|^2 - \frac{1}{2}\sum_i a_i w_i^2 + \sum_i (b_i w_i + c_i)\right) \to \max_{\boldsymbol{w},\boldsymbol{\xi}}. \quad (10)$$

Можно показать (см. [4]), что задачи оптимизации (9),(10) являются выпуклыми тогда и только тогда, когда функция $\log p(w_i|\alpha)$ является строго вогнутой. В этом случае они могут быть решены с помощью стандартных методов выпуклой оптимизации. В частности, в задаче (10) оптимизация по весам $\boldsymbol{w}$ может быть проведена аналитически: $\boldsymbol{w} = (\beta X^T X + \operatorname{diag}(\boldsymbol{a}))^{-1}(\beta X^T\boldsymbol{t} + \boldsymbol{b})$. В случае невогнутости $\log p(w_i|\alpha)$ (например, для потенциалов Стьюдента) задачи оптимизации (9),(10) можно решать с помощью выпукло-вогнутой процедуры.

Вариационная оценка для мультиномиальной функции

Рассмотрим стандартную задачу классификации на два класса. Пусть имеется обучающая выборка $(\boldsymbol{t}, X) = \{t_n, \boldsymbol{x}_n\}_{n=1}^N$, где $\boldsymbol{x}_n \in \mathbb{R}^d$ – признаки объекта, а $t_n \in \{-1, +1\}$ – метка класса. Для решения этой задачи воспользуемся методом «логистическая регрессия». Вероятностная

модель для этого метода выглядит следующим образом:

$$p(\boldsymbol{t}, \boldsymbol{w}|X, \alpha) = p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w}|\alpha) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n, \boldsymbol{w})p(\boldsymbol{w}|\alpha),$$

$$p(t|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1+\exp(-t\boldsymbol{w}^T\boldsymbol{x})} \text{ – логистическое правдоподобие,}$$

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}I) \text{ – квадратичный регуляризатор.}$$

Здесь $\boldsymbol{w} \in \mathbb{R}^d$ – веса линейного решающего правила. Поиск весов $\boldsymbol{w}$ в данной модели с помощью максимизации апостериорного распределения приводит к следующей оптимизационной задаче:

$$f_{\text{logistic}}(\boldsymbol{w}) = \sum_{n=1}^{N} \log(1+\exp(-t_n\boldsymbol{w}^T\boldsymbol{x}_n)) + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \to \min_{\boldsymbol{w}}.$$

Эта задача является строго выпуклой и поэтому может быть решена с помощью метода Ньютона. Альтернативно, можно воспользоваться вариационной оптимизацией с применением оценки Йаколлы-Джордана для логистической функции:

$$f_{\text{logistic}}(\boldsymbol{w}) \le \sum_{n=1}^{N} \left[\log(1+\exp(-\xi_n)) + \frac{\xi_n - t_n\boldsymbol{w}^T\boldsymbol{x}_n}{2} + \frac{\tanh(\xi_n/2)}{4\xi_n}(\boldsymbol{w}^T\boldsymbol{x}_n\boldsymbol{x}_n^T\boldsymbol{w} - \xi_n^2)\right] + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \to \min_{\boldsymbol{w},\boldsymbol{\xi}}.$$

Покоординатная минимизация для $\boldsymbol{w}, \boldsymbol{\xi}$ при этом может быть выполнена аналитически:

$$\boldsymbol{w} = \left(\alpha I + 2\sum_n \frac{\tanh(\xi_n/2)}{4\xi_n}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)^{-1}\frac{1}{2}X^T\boldsymbol{t},$$

$$\xi_n = t_n\boldsymbol{w}^T\boldsymbol{x}_n.$$

Рассмотрим теперь задачу классификации на K классов. По-прежнему, имеется обучающая выборка $(\boldsymbol{t}, X) = \{t_n, \boldsymbol{x}_n\}_{n=1}^N$, где $\boldsymbol{x}_n \in \mathbb{R}^d$ – признаки объекта, $t_n \in \{1, \ldots, K\}$ – метка класса. Обобщение метода «логистическая регрессия» на многоклассовый случай известно как «мультиномиальная регрессия». Вероятностная модель в этом методе выглядит как

$$p(\boldsymbol{t}, W|X, \alpha) = p(\boldsymbol{t}|X, W)p(W|\alpha) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n, W)p(W|\alpha),$$

$$p(t|\boldsymbol{x}, W) = \frac{\exp(\boldsymbol{w}_t^T\boldsymbol{x})}{\sum_{k=1}^{K}\exp(\boldsymbol{w}_k^T\boldsymbol{x})} \text{ – мультиномиальное правдоподобие,}$$

$$p(W|\alpha) = \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{w}_k|\boldsymbol{0}, \alpha^{-1}I) \text{ – квадратичный регуляризатор.}$$

Здесь $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K]^T \in \mathbb{R}^{K\times d}$ – матрица весов для K линейных решающих функций. Прогноз метки класса $\hat{t}$ для нового объекта $\boldsymbol{x}$ находится как $\hat{t} = \arg\max_k \boldsymbol{w}_k^T\boldsymbol{x}$. Обучение весов W с помощью максимизации апостериорного распределения приводит к следующей оптимизационной задаче:

$$f_{\text{multinomial}}(W) = \sum_{n=1}^{N}\left[-\boldsymbol{w}_{t_n}^T\boldsymbol{x}_n + \log\left(\sum_{j=1}^{K}\exp(\boldsymbol{w}_j^T\boldsymbol{x}_n)\right)\right] + \frac{\alpha}{2}\sum_{k=1}^{K}\boldsymbol{w}_k^T\boldsymbol{w}_k \to \min_W. \qquad (11)$$

Решение этой задачи с помощью вариационного подхода требует вариационной верхней оценки для функции $\log(\sum_j \exp(y_j))$. Рассмотрим одну из таких оценок, предложенную в работе [5]. Верно, что

$$\prod_j (1+\exp(y_j-\alpha)) \geq \sum_j \exp(y_j-\alpha) = \exp(-\alpha)\sum_j \exp(y_j), \forall y_j, \alpha \in \mathbb{R}.$$

Отсюда

$$\sum_j \exp(y_j) \leq \exp(\alpha)\prod_j (1+\exp(y_j-\alpha)).$$

Переходя к логарифму в последнем неравенстве и используя оценку Йаколлы-Джордана, получаем искомую квадратичную вариационную оценку:

$$\log\sum_j \exp(y_j) \leq \alpha + \sum_j \log(1+\exp(y_j-\alpha)) \leq$$
$$\alpha + \sum_j \left[\log(1+\exp(\xi_j)) + \frac{y_j-\alpha-\xi_j}{2} + \frac{\tanh(\xi_j/2)}{4\xi_j}((y_j-\alpha)^2-\xi_j^2)\right]. \quad (12)$$

Подставляя данную оценку в функционал (11), получаем требуемый метод вариационной оптимизации.

Оценка на максимум и вариационный SVM

Функция максимума от двух и более переменных часто встречается в различных задачах оптимизации в машинном обучении. Рассмотрим несколько вариационных оценок для функции максимума. Справедлива следующая цепочка равенств:

$$\max(x,y) = x + \max(0, y-x) = x + \frac{y-x}{2} + \max\left(-\frac{y-x}{2}, \frac{y-x}{2}\right) = \frac{x+y}{2} + \frac{|y-x|}{2}. \quad (13)$$

Далее для функции $|y-x|$ можно применить квадратичную вариационную оценку, описанную выше. Действуя по аналогии с вариационной оценкой (12), можно получить оценку для функции максимума от набора переменных:

$$\max(y_1,\ldots,y_K) = \alpha + \max(y_1-\alpha,\ldots,y_K-\alpha) \leq \alpha + \sum_{k=1}^{K}\max(0,y_k-\alpha) = \alpha + \sum_k \left[\frac{y_k-\alpha}{2} + \frac{|y_k-\alpha|}{2}\right].$$

Аналогично, применяя квадратичную вариационную оценку для $|y_k-\alpha|$, получаем квадратичную вариационную оценку для $\max(y_1,\ldots,y_K)$.

Другой способ получения вариационных оценок для функции максимума основан на использовании следующего классического неравенства:

$$\max(y_1,\ldots,y_K) \leq \log\left(\sum_k \exp(y_k)\right).$$

Верно, что

$$\max(y_1,\ldots,y_K) = \frac{1}{a}\max(ay_1,\ldots,ay_K) \leq \frac{1}{a}\log\left(\sum_k \exp(ay_k)\right), \ \forall a > 0.$$

Далее для функции логарифма от суммы экспонент можно воспользоваться оценкой (12). Для случая двух переменных справедливо

$$\max(x,y) = x + \max(0, y-x) \le x + \frac{1}{a}\log(1+\exp(a(y-x))) \le$$
$$x + \frac{1}{a}\left[\log(1+\exp(\xi)) + \frac{a(y-x)-\xi}{2} + \frac{\tanh(\xi/2)}{4\xi}(a^2(y-x)^2 - \xi^2)\right]. \quad (14)$$

Здесь мы воспользовались оценкой Йаколла-Джордана.

В качестве примера применения вариационных оценок для функции максимума рассмотрим обучение классического метода опорных векторов:

$$f(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\max(0, 1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)) \to \min_{\boldsymbol{w},b}.$$

При использовании эквивалентного преобразования (13), получаем

$$f(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\left[\frac{1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)}{2} + \frac{|1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b)|}{2}\right] \to \min_{\boldsymbol{w},b}.$$

Таким образом, здесь возникает задача оптимизации типа LASSO, т.к. функционал f является комбинацией квадратичной функции и модулей линейных функций. Эту задачу можно решать с помощью методов, заложенных в упомянутые выше библиотеки GLMNET и LIBSVM, а также с помощью метода split Bregman [6], который хорошо себя зарекомендовал для решения задач с L_1-регуляризацией. Альтернативно, можно воспользоваться вариационным подходом с оценкой (14):

$$f(\boldsymbol{w}, b) \le \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\frac{1}{a_n}\left[\log(1+\exp(\xi_n)) + \frac{a_n - a_n t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) - \xi_n}{2} + \right.$$
$$\left. + \lambda(\xi_n)(a_n^2(1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b))^2 - \xi_n^2)\right] \to \min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{a}}.$$

Здесь $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$. Покоординатная оптимизация параметров $\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{a}$ может быть осуществлена аналитически:

$$[\boldsymbol{w},\ b] = \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} + 2C\sum_{n=1}^{N}\lambda(\xi_n)a_n\tilde{\boldsymbol{x}}_n\tilde{\boldsymbol{x}}_n^T\right)^{-1} C\sum_{n=1}^{N}\left(\frac{1}{2} + 2\lambda(\xi_n)a_n\right)t_n\tilde{\boldsymbol{x}}_n,\ \ \tilde{\boldsymbol{x}}_n = [\boldsymbol{x}_n,\ 1],$$
$$\xi_n^2 = a_n^2(1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b))^2,$$
$$a_n^2 = \frac{\log(1+\exp(\xi_n)) - \xi_n/2 - \lambda(\xi_n)\xi_n^2}{\lambda(\xi_n)(1 - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b))^2}.$$

Список вариационных оценок

В заключение для удобства запишем вариационные оценки, упомянутые выше, одним списком:

1. $|x|^p \le |\xi|^p + \frac{p|\xi|^p}{2\xi^2}(x^2 - \xi^2),\ 0 < p < 2;$

2. $|x| - \varepsilon \log(\varepsilon + |x|) \leq |\xi| - \varepsilon \log(\varepsilon + |\xi|) + \frac{x^2 - \xi^2}{2(\varepsilon + |\eta|)}$;

3. $\log(1 + \exp(x)) \leq \log(1 + \exp(\xi)) + \frac{x - \xi}{2} + \frac{\tanh(\xi/2)}{4\xi}(x^2 - \xi^2)$;

4. $\log\left(1 + \frac{x^2}{\alpha}\right) \leq \log\left(1 + \frac{\xi^2}{\alpha}\right) + \frac{x^2 - \xi^2}{\alpha + \xi^2}$;

5. $\log \sum_n \exp(x_n) \leq \alpha + \sum_{n=1}^N \log(1 + \exp(x_n - \alpha))$;

6. $\max(x_1, \ldots, x_N) \leq \frac{1}{a} \log\left[\sum_{n=1}^N \exp(a x_n)\right]$;

7. $\max(x_1, \ldots, x_N) \leq \alpha + \sum_{n=1}^N \max(0, x_n - \alpha)$.

Список литературы

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

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

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

[4] 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.

[5] G. Bouchard. Efficient Bounds for the Softmax Function and Applications to Approximate Inference in Hybrid Models // NIPS 2007 Workshop on Approximate Inference in Hybrid Models, 2007.

[6] T Goldstein, S Osher. The Split Bregman Method for L1-Regularized Problems // SIAM Journal Imaging Sciences, Vol. 2, 2009, pp. 323–343.