Лекция 15
Условная оптимизация и ядровой SVM

Е. А. Соколов
ФКН ВШЭ

19 января 2018 г.

1 Оптимизационные задачи и теорема Куна-Таккера

Рассмотрим задачу минимизации

$$\begin{cases} f_0(x) \to \min\limits_{x \in \mathbb{R}^d} \\ f_i(x) \leqslant 0, \quad i = 1, \dots, m, \\ h_i(x) = 0, \quad i = 1, \dots, p. \end{cases} \tag{1.1}$$

Если ограничения в этой задаче отсутствуют, то имеет место *необходимое условие экстремума*: если в точке x функция f_0 достигает своего минимума, то ее градиент в этой точке равен нулю. Значит, для решения задачи безусловной оптимизации

$$f_0(x) \to \min$$

достаточно найти все решения уравнения

$$\nabla f_0(x) = 0,$$

и выбрать то, в котором достигается наименьшее значение. Для решения условных задач оптимизации требуется более сложный подход, который мы сейчас и рассмотрим.

§1.1 Лагранжиан

Задача условной оптимизации (1.1) эквивалентна следующей безусловной задаче:

$$f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) + \sum_{i=1}^{p} I_0(h_i(x)) \to \min_x,$$

где $I_-(x)$ — индикаторная функция для неположительных чисел:

$$I_-(x) = \begin{cases} 0, & x \leqslant 0 \\ \infty, & x > 0, \end{cases}$$

а $I_0(x)$ — индикаторная функция для нуля:

$$I_0(x) = \begin{cases} 0, & x = 0 \\ \infty, & x \neq 0, \end{cases}$$

Такая переформулировка, однако, не упрощает задачу — индикаторные функции являются кусочно-постоянными и могут быть оптимизированы лишь путем полного перебора решений.

Заменим теперь индикаторные функции на их линейные аппроксимации:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

где $\lambda_i \geqslant 0$. Полученная функция называется *лагранжианом* задачи (1.1). Числа λ_i и ν_i называются *множителями Лагранжа* или *двойственными переменными.*

Конечно, линейные аппроксимации являются крайне грубыми, однако их оказывается достаточно, чтобы получить необходимые условия на решение исходной задачи.

§1.2 Двойственная функция

Двойственной функцией для задачи (1.1) называется функция, получающаяся при взятии минимума лагранжиана по x:

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu).$$

Можно показать, что данная функция всегда является вогнутой.

Зачем нужна двойственная функция? Оказывается, она дает нижнюю оценку на минимум в исходной оптимизационной задаче. Обозначим решение задачи (1.1) через x_*. Пусть x' — *допустимая* точка, т.е. $f_i(x') \leqslant 0$, $h_i(x') = 0$. Пусть также $\lambda_i > 0$. Тогда

$$L(x', \lambda, \nu) = f_0(x') + \sum_{i=1}^{m} \lambda_i f_i(x') + \sum_{i=1}^{p} \nu_i h_i(x') \leqslant f_0(x').$$

Если взять в левой части минимум по всем допустимым x, то неравенство останется верным; оно останется верным и в случае, если мы возьмем минимум по всем возможным x:

$$\inf_x L(x, \lambda, \nu) \leqslant \inf_{x \text{ — допуст.}} L(x, \lambda, \nu) \leqslant L(x', \lambda, \nu).$$

Итак, получаем

$$\inf_x L(x, \lambda, \nu) \leqslant f_0(x').$$

Поскольку решение задачи x_* также является допустимой точкой, получаем, что при $\lambda \geqslant 0$ двойственная функция дает нижнюю оценку на минимум:

$$g(\lambda, \nu) \leqslant f_0(x_*).$$

§1.3 Двойственная задача

Итак, двойственная функция для любой пары (λ, ν) с $\lambda > 0$ дает нижнюю оценку на минимум в оптимизационной задаче. Попробуем теперь найти наилучшую нижнюю оценку:

$$\begin{cases} g(\lambda, \nu) \to \max\limits_{\lambda,\nu} \\ \lambda_i \geqslant 0, \quad i = 1, \dots, m. \end{cases} \tag{1.2}$$

Данная задача называется *двойственной* к задаче (1.1). Заметим, что функционал в двойственной задаче всегда является вогнутым.

§1.4 Сильная и слабая двойственность

Пусть (λ^*, ν^*) — решение двойственной задачи. Значение двойственной функции всегда не превосходит условный минимум исходной задачи:

$$g(\lambda^*, \nu^*) \leqslant f_0(x_*).$$

Это свойство называется *слабой двойственностью.* Разность $f_0(x_*) - g(\lambda^*, \nu^*)$ называется *зазором* между решениями прямой и двойственной задач.

Если имеет место равенство

$$g(\lambda^*, \nu^*) = f_0(x_*),$$

то говорят о *сильной двойственности.* Существует много достаточных условий сильной двойственности. Одним из таких условий для выпуклых задач является условие Слейтера. *Выпуклой* задачей оптимизации называется задача

$$\begin{cases} f_0(x) \to \min\limits_{x \in \mathbb{R}^d} \\ f_i(x) \leqslant 0, \quad i = 1, \dots, m, \\ Ax = b. \end{cases}$$

где функции $f_0, f_1, \dots, f_m$ являются выпуклыми. Условие Слейтера требует, чтобы существовала такая допустимая точка x', в которой ограничения-неравенства выполнены строго:

$$\begin{cases} f_i(x) < 0, \quad i = 1, \dots, m, \\ Ax = b. \end{cases}$$

Условие Слейтера можно ослабить: достаточно, чтобы ограничения-неравенства были строгими только в том случае, если они не являются линейными (т.е. не имеют вид $Ax - b$).

§1.5 Условия Куна-Таккера

Пусть x_* и (λ^*, ν^*) — решения прямой и двойственной задач. Будем считать, что имеет место сильная двойственность. Можно показать (и это будет сделано на семинарах), что в этом случае выполнено несколько утверждений про связь между прямой и двойственной задачами:

- Если подставить в лагранжиан решение двойственной задачи (λ^*, ν^*), то его минимум будет достигаться на решении прямой задачи x_*. Иными словами, решение исходной задачи (1.1) эквивалентно минимизации лагранжиана $L(x, \lambda^*, \nu^*)$ с подставленным решением двойственной задачи.

- Имеют место *условия дополняющей нежёсткости*:

  $$\lambda_i^* f_i(x_*) = 0, \quad i = 1, \ldots, m.$$

  Они означают, что множитель Лагранжа при i-м ограничении может быть не равен нулю лишь в том случае, если ограничение выполнено с равенством (в этом случае говорят, что оно является *активным*).

Итак, мы можем записать условия, которые выполнены для решений прямой и двойственной задач x_* и (λ^*, ν^*):

$$\begin{cases} \nabla f_0(x_*) + \sum\limits_{i=1}^{m} \lambda_i^* \nabla f_i(x_*) + \sum\limits_{i=1}^{p} \nu_i^* \nabla h_i(x_*) = 0 \\ f_i(x_*) \leqslant 0, \quad i = 1, \ldots m \\ h_i(x_*) = 0, \quad i = 1, \ldots p \\ \lambda_i^* \geqslant 0, \quad i = 1, \ldots m \\ \lambda_i^* f_i(x_*) = 0, \quad i = 1, \ldots m \end{cases} \tag{KKT}$$

Данные условия называются *условиями Куна-Таккера* (в зарубежной литературе их принято называть условиями Каруша-Куна-Таккера) и являются необходимыми условиями экстремума. Их можно сформулировать несколько иначе:

Теорема 1.1. *Пусть x_* — решение задачи (1.1). Тогда найдутся такие векторы λ^* и ν^*, что выполнены условия (KKT).*

Если задача (1.1) является выпуклой и удовлетворяет условию Слейтера, то условия Куна-Таккера становятся *необходимыми и достаточными*.

§1.6 Экономическая интерпретация двойственной задачи

Предположим, что мы хотим открыть фирму. В нее мы можем нанимать программистов и менеджеров — обозначим их количество через x_1 и x_2 соответственно. При этом каждый программист будет приносить c_1 рублей в месяц, а каждый менеджер — c_2 рублей. Труд каждого сотрудника должен оплачиваться. Наша фирма может платить в двух формах — акциями и картошкой, причем в месяц каждому программисту нужно выдать a_{11} акций и a_{21} килограммов картошки; для менеджеров эти числа обозначим через a_{12} и a_{22}. Разумеется, наши возможности ограничены: мы можем тратить не больше b_1 акций и b_2 килограммов картошки в месяц. Запишем формально все эти соотношения:

$$\begin{cases} c_1 x_1 + c_2 x_2 \to \max\limits_{x_1, x_2} \\ a_{11} x_1 + a_{12} x_2 \leqslant b_1 \\ a_{21} x_1 + a_{22} x_2 \leqslant b_2 \\ x_1 \geqslant 0, x_2 \geqslant 0 \end{cases}$$

Это задача линейного программирования, для которой легко найти двойственную:

$$\begin{cases} b_1 y_1 + b_2 y_2 \to \min\limits_{y_1, y_2} \\ a_{11} y_1 + a_{21} y_2 \geqslant c_1 \\ a_{12} y_1 + a_{22} y_2 \geqslant c_2 \\ y_1 \geqslant 0, y_2 \geqslant 0 \end{cases}$$

Двойственную задачу можно проинтерпретировать следующим образом. Допустим, что у нас появились другие дела, и вместо открытия фирмы мы решили продать все ресурсы (т.е. акции и картошку). Разумеется, наши покупатели будут стремиться установить максимально низкую цену — иными словами, они будут минимизировать общую сумму сделки $b_1 y_1 + b_2 y_2$, где через y_1 и y_2 обозначены цены на одну акцию и на один килограмм картошки соответственно. При этом у нас есть ограничение: мы не хотим продавать ресурсы дешевле, чем могли бы на них заработать, если бы все же открыли фирму. Это означает, что суммарная стоимость a_{11} акций и a_{21} килограммов картошки (т.е. размер оплаты одного программиста) не должна быть меньше, чем доход от одного программиста c_1. Это требование, вкупе с аналогичным требованием к размеру оплаты менеджера, как раз соответствует ограничениям в двойственной задаче.

Поскольку для данных задач имеет место сильная двойственность, их решения будут совпадать. Это означает, что оптимальная прибыль, которую можно получить при открытии фирмы, совпадает с оптимальной выгодой от продажи всех ресурсов.

2 Ядровой SVM

Вспомним, что метод опорных векторов сводится к решению задачи оптимизации

$$\begin{cases} \frac{1}{2} \|w\|^2 + C \sum\limits_{i=1}^{\ell} \xi_i \to \min\limits_{w, b, \xi} \\ y_i \left(\langle w, x_i \rangle + b \right) \geqslant 1 - \xi_i, \quad i = 1, \dots, \ell, \\ \xi_i \geqslant 0, \quad i = 1, \dots, \ell. \end{cases} \tag{2.1}$$

Построим двойственную к ней. Запишем лагранжиан:

$$L(w, b, \xi, \lambda, \mu) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i - \sum_{i=1}^{\ell} \lambda_i \left[y_i \left(\langle w, x_i \rangle + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{\ell} \mu_i \xi_i.$$

Выпишем условия Куна-Таккера:

$$\nabla_w L = w - \sum_{i=1}^{\ell} \lambda_i y_i x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{\ell} \lambda_i y_i x_i \tag{2.2}$$

$$\nabla_b L = -\sum_{i=1}^{\ell} \lambda_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0 \tag{2.3}$$

$$\nabla_{\xi_i} L = C - \lambda_i - \mu_i \quad \Longrightarrow \quad \lambda_i + \mu_i = C \tag{2.4}$$

$$\lambda_i \left[y_i \left(\langle w, x_i \rangle + b \right) - 1 + \xi_i \right] = 0 \quad \Longrightarrow \quad (\lambda_i = 0) \text{ или } \left(y_i \left(\langle w, x_i \rangle + b \right) = 1 - \xi_i \right) \tag{2.5}$$

$$\mu_i \xi_i = 0 \quad \Longrightarrow \quad (\mu_i = 0) \text{ или } (\xi_i = 0) \tag{2.6}$$

$$\xi_i \geqslant 0, \lambda_i \geqslant 0, \mu_i \geqslant 0. \tag{2.7}$$

Проанализируем полученные условия. Из (2.2) следует, что вектор весов, полученный в результате настройки SVM, можно записать как линейную комбинацию объектов, причем веса в этой линейной комбинации можно найти как решение двойственной задачи. В зависимости от значений ξ_i и λ_i объекты x_i разбиваются на три категории:

1. $\xi_i = 0$, $\lambda_i = 0$.
   Такие объекты не влияют решение w (входят в него с нулевым весом λ_i), правильно классифицируются ($\xi_i = 0$) и лежат вне разделяющей полосы. Объекты этой категории называются *периферийными*.

2. $\xi_i = 0$, $0 < \lambda_i < C$.
   Из условия (2.5) следует, что $y_i \left(\langle w, x_i \rangle + b \right) = 1$, то есть объект лежит строго на границе разделяющей полосы. Поскольку $\lambda_i > 0$, объект влияет на решение w. Объекты этой категории называются *опорными граничными*.

3. $\xi_i > 0$, $\lambda_i = C$.
   Такие объекты могут лежать внутри разделяющей полосы ($0 < \xi_i < 2$) или выходить за ее пределы ($\xi_i \geqslant 2$). При этом если $0 < \xi_i < 1$, то объект классифицируется правильно, в противном случае — неправильно. Объекты этой категории называются *опорными нарушителями*.

Отметим, что варианта $\xi_i > 0$, $\lambda_i < C$ быть не может, поскольку при $\xi_i > 0$ из условия дополняющей нежесткости (2.6) следует, что $\mu_i = 0$, и отсюда из уравнения (2.4) получаем, что $\lambda_i = C$.

Итак, итоговый классификатор зависит только от объектов, лежащих на границе разделяющей полосы, и от объектов-нарушителей (с $\xi_i > 0$).

Построим двойственную функцию. Для этого подставим выражение (2.2) в лагранжиан, и воспользуемся уравнениями (2.3) и (2.4) (данные три уравнения выполнены для точки минимума лагранжиана при любых фиксированных λ и μ):

$$L = \frac{1}{2}\left\|\sum_{i=1}^{\ell}\lambda_i y_i x_i\right\|^2 - \sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle - b\underbrace{\sum_{i=1}^{\ell}\lambda_i y_i}_{0} + \sum_{i=1}^{\ell}\lambda_i + \sum_{i=1}^{\ell}\xi_i\underbrace{(C - \lambda_i - \mu_i)}_{0}$$

$$= \sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle.$$

Мы должны потребовать выполнения условий (2.3) и (2.4) (если они не выполнены, то двойственная функция обращается в минус бесконечность), а также неотрицательность двойственных переменных $\lambda_i \geqslant 0$, $\mu_i \geqslant 0$. Ограничение на μ_i и условие (2.4), можно объединить, получив $\lambda_i \leqslant C$. Приходим к следующей двойственной задаче:

$$\begin{cases} \sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle \to \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \dots, \ell, \\ \sum_{i=1}^{\ell}\lambda_i y_i = 0. \end{cases} \tag{2.8}$$

Она также является вогнутой, квадратичной и имеет единственный максимум.

Двойственная задача SVM зависит только от скалярных произведений объектов — отдельные признаковые описания никак не входят в неё. Значит, можно легко сделать ядровой переход:

$$\begin{cases} \sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j K(x_i, x_j) \to \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \dots, \ell, \\ \sum_{i=1}^{\ell}\lambda_i y_i = 0. \end{cases} \tag{2.9}$$

Вернемся к тому, какое представление классификатора дает двойственная задача. Из уравнения (2.2) следует, что вектор весов w можно представить как линейную комбинацию объектов из обучающей выборки. Подставляя это представление w в классификатор, получаем

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{\ell}\lambda_i y_i\langle x_i, x\rangle + b\right). \tag{2.10}$$

Таким образом, классификатор измеряет сходство нового объекта с объектами из обучения, вычисляя скалярное произведение между ними. Это выражение также зависит только от скалярных произведений, поэтому в нём тоже можно перейти к ядру.

В представлении (2.10) фигурирует переменная b, которая не находится непосредственно в двойственной задаче. Однако ее легко восстановить по любому граничному опорному объекту x_i, для которого выполнено $\xi_i = 0, 0 < \lambda_i < C$. Для него выполнено $y_i\left(\langle w, x_i\rangle + b\right) = 1$, откуда получаем

$$b = y_i - \langle w, x_i\rangle.$$

Как правило, для численной устойчивости берут медиану данной величины по всем граничным опорным объектам:

$$b = \operatorname{med}\{y_i - \langle w, x_i\rangle \,|\, \xi_i = 0, 0 < \lambda_i < C\}.$$

Связь с kNN. Если использовать гауссовское ядро (или, как его еще называют, RBF-ядро) в методе опорных векторов, то получится следующее решающее правило:

$$a(x) = \operatorname{sign} \sum_{i=1}^{\ell} y_i \lambda_i \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right).$$

Вспомним теперь, что решающее правило в методе k ближайших соседей выглядит как

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x, X^\ell); \quad \Gamma_y(x, X^\ell) = \sum_{i=1}^{\ell} [y_x^{(i)} = y] w(i, x),$$

где $w(i, x)$ — оценка важности i-го соседа для классификации объекта x, а $y_x^{(i)}$ — метка i-го ближайшего соседа. Для случая двух классов $\{+1, -1\}$ решающее правило можно записать как знак разности оценок за эти классы:

$$\begin{aligned}
a(x) &= \operatorname{sign}\left(\Gamma_{+1}(x, X^\ell) - \Gamma_{-1}(x, X^\ell)\right) = \\
&= \operatorname{sign}\left(\sum_{i=1}^{\ell} [y_x^{(i)} = +1] w(i, x) - \sum_{i=1}^{\ell} [y_x^{(i)} = -1] w(i, x)\right) = \\
&= \operatorname{sign} \sum_{i=1}^{\ell} ([y_x^{(i)} = +1] - [y_x^{(i)} = -1]) w(i, x) = \\
&= \operatorname{sign} \sum_{i=1}^{\ell} y_x^{(i)} w(i, x).
\end{aligned}$$

Заметим, что решающие правила метода опорных векторов с RBF-ядром и метода k ближайших соседей совпадут, если положить

$$w(i, x) = \lambda_{(i)} \exp\left(-\frac{\|x - x_{(i)}\|^2}{2\sigma^2}\right).$$

То есть SVM-RBF — это метод ℓ ближайших соседей, использующий гауссово ядро в качестве функции расстояния, и настраивающий веса объектов путем максимизации отступов.

Список литературы

[1] *Boyd, S., Vandenberghe, L.* Convex Optimization. // Cambridge University Press, 2004.