Семинары по линейным классификаторам

Евгений Соколов
sokolov.evg@gmail.com

30 октября 2013 г.

3 Условия Куна-Таккера и SVM

§3.1 Условия Куна-Таккера и двойственность, продолжение

Задача 3.1. *Покажите, что задача минимизации регуляризованного функционала*

$$Q(w) + \tau\|w\|_p \to \min_w \tag{3.1}$$

с $p \geqslant 1$ *и* $\tau \geqslant 0$ *эквивалентна условной задаче*

$$\begin{cases} Q(w) \to \min\limits_w \\ \|w\|_p \leqslant C \end{cases} \tag{3.2}$$

если функционал $Q(w)$ *является выпуклым.*

Решение. Задача (3.2) является выпуклой и для нее выполнено условие Слейтера, поэтому вектор w_* является ее решением тогда и только тогда, когда он удовлетворяет условиям Куна-Таккера:

$$\begin{cases} \nabla_w \left(Q(w_*) + \lambda^*\|w_*\|_p\right) = 0 \\ \|w_*\|_p \leqslant C \\ \lambda^* \geqslant 0 \\ \lambda^*(\|w_*\|_p - C) = 0 \end{cases}$$

Пусть w_* — решение задачи (3.2), тогда из условий Куна-Таккера получаем, что градиент лагранжиана в данной точке равен нулю. Поскольку лагражиан является выпуклым, то из равенства нулю градиента в точке w_* следует, что w_* является глобальным минимумом лагранжиана. Следовательно, вектор w_* является решением задачи

$$Q(w) + \lambda^* \left(\|w\|_p - C\right) \to \min_w,$$

которая эквивалентна задаче (3.1) при $\tau = \lambda^*$. Значит, если w_* является решением задачи (3.2), то он является решением задачи (3.1).

Пусть теперь w_* — решение задачи (3.1). Положим $C = \|w\|_p$ и $\lambda^* = \tau$. Тогда пара (w_*, λ^*) удовлетворяет условиям Куна-Таккера и, следовательно, является решением задачи (3.2).

∎

§3.2 Метод опорных векторов

3.2.1 Формулировка

Будем рассматривать линейные классификаторы вида

$$a(x) = \operatorname{sign}\langle w, x\rangle + b, \qquad w \in \mathbb{R}^d, b \in \mathbb{R}.$$

Линейной разделимая выборка. Будем считать, что существуют такие параметры w_* и b_*, что соответствующий им классификатор $a(x)$ не допускает ни одной ошибки на обучающей выборке. В этом случае говорят, что выборка *линейно разделима.*

Пусть задан некоторый классификатор $a(x) = \operatorname{sign}\langle w, x\rangle + b$. Заметим, что если одновременно умножить параметры w и b на одну и ту же положительную константу, то классификатор не изменится. Распорядимся этой свободой выбора и отнормируем параметры так, что

$$\min_{x \in X^\ell} |\langle w, x\rangle + b| = 1. \tag{3.3}$$

Расстояние от произвольной точки $x_0 \in \mathbb{R}^d$ до гиперплоскости, определяемой данным классификатором, равно

$$\rho(x_0, a) = \frac{|\langle w, x\rangle + b|}{\|w\|}.$$

Тогда расстояние от гиперплоскости до ближайшего объекта обучающей выборки равно

$$\min_{x \in X^\ell} \frac{|\langle w, x\rangle + b|}{\|w\|} = \frac{1}{\|w\|} \min_{x \in X^\ell} |\langle w, x\rangle + b| = \frac{1}{\|w\|}.$$

Данная величина также называется *отступом (margin).*

Таким образом, если классификатор без ошибок разделяет обучающую выборку, то ширина его разделяющей полосы равна $\frac{2}{\|w\|}$. Известно, что максимизация ширины разделяющей полосы приводит к повышению обобщающей способности классификатора [1]. Вспомним также, что на повышение обобщающей способности направлена и регуляризация, которая штрафует большую норму весов — а чем больше норма весов, тем меньше ширина разделяющей полосы.

Итак, требуется построить классификатор, идеально разделяющий обучающую выборку, и при этом имеющий максимальный отступ. Запишем соответствующую оптимизационную задачу:

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,b} \\ y_i\left(\langle w, x_i\rangle + b\right) \geqslant 1, \quad i = 1, \dots, \ell. \end{cases} \tag{3.4}$$

Здесь мы воспользовались тем, что линейный классификатор дает правильный ответ на объекте x_i тогда и только тогда, когда $\langle w, x_i\rangle + b \geqslant 0$. Более того, из условия нормировки (3.3) следует, что $\langle w, x_i\rangle + b \geqslant 1$.

В данной задаче функционал является строго выпуклым, а ограничения линейными, поэтому сама задача является выпуклой и имеет единственное решение. Более того, задача является квадратичной и может быть решена крайне эффективно.

Неразделимый случай. Рассмотрим теперь общий случай, когда выборку невозможно идеально разделить гиперплоскостью. Это означает, что какие бы w и b мы не взяли, хотя бы одно из ограничений в задаче (3.4) будет нарушено:

$$\exists x_i \in X^\ell : \; y_i \left(\langle w, x_i \rangle + b \right) < 1.$$

Сделаем эти ограничения «мягкими», введя штраф $\xi_i \geqslant 0$ за их нарушение:

$$y_i \left(\langle w, x_i \rangle + b \right) \geqslant 1 - \xi_i, \quad i = 1, \dots, \ell.$$

Отметим, что если отступ объекта лежит между нулем и единицей ($0 \leqslant y_i \left(\langle w, x_i \rangle + b \right) < 1$), то объект верно классифицируется, но имеет ненулевой штраф $\xi > 0$. Таким образом, мы штрафуем объекты за попадание внутрь разделяющей полосы.

Величина $\frac{1}{\|w\|}$ в данном случае называется *мягким отступом (soft margin)*. С одной стороны, мы хотим максимизировать отступ, с другой — минимизировать штраф за неидеальное разделение выборки $\sum_{i=1}^{\ell} \xi_i$. Эти две задачи противоречат друг другу: как правило, излишняя подгонка под выборку приводит к маленькому отступу, и наоборот — максимизация отступа приводит к большой ошибке на обучении. В качестве компромисса будем минимизировать взвешенную сумму двух указанных величин. Приходим к оптимизационной задаче

$$\begin{cases} \frac{1}{2} \|w\|^2 + C \sum\limits_{i=1}^{\ell} \xi_i \to \min\limits_{w, b, \xi} \\ y_i \left(\langle w, x_i \rangle + b \right) \geqslant 1 - \xi_i, \quad i = 1, \dots, \ell, \\ \xi \geqslant 0, \quad i = 1, \dots, \ell. \end{cases} \tag{3.5}$$

Чем больше здесь параметр C, тем сильнее мы будем настраиваться на обучающую выборку.

Данная задача также является выпуклой и имеет единственное решение.

3.2.2 Обобщение на многоклассовый случай

Рассмотрим теперь многоклассовую задачу: $\mathbb{Y} = \{1, \dots, K\}$ и способы ее решения с помощью метода опорных векторов.

Один против всех. Обучим K линейных классификаторов $a_1(x), \dots, a_K(x)$, $a_k(x) = \operatorname{sign}(\langle w_k, x \rangle + b_k)$. Классификатор с номером k будем обучать по выборке $(x_i, [y_i = k])_{i=1}^{\ell}$; иными словами, мы учим классификатор отличать k-й класс от всех остальных.

Итоговый классификатор определим следующим образом:

$$a(x) = \operatorname*{arg\,max}_{k \in \{1, \dots, K\}} \left\{ \langle w_k, x \rangle b_k \right\}.$$

Парный подход. Обучим C_K^2 классификаторов $a_{ij}(x)$, $i, j = 1, \dots, K$, $i \neq j$. Классификатор $a_{ij}(x)$ будем настраивать по подвыборке X^ℓ, содержащей только объекты классов i и j.

Чтобы классифицировать новый объект, подадим его на вход каждого из построенных бинарных классификаторов. Каждый из них проголосует за своей класс; в качестве ответа выберем тот, за который наберется больше всего голосов:

$$a(x) = \mathop{\arg\max}_{k \in \{1, \dots, K\}} \sum_{i=1}^{\ell} \sum_{j \neq i} [a_{ij}(x) = k].$$

Многоклассовый SVM [2]. В алгоритме «один против всех» мы *независимо* строили свой классификатор за каждый класс. Попробуем теперь строить эти классификаторы одновременно, в рамках одной оптимизационной задачи.

Для простоты будем считать, что в выборке имеется константный признак, и не будет явно указывать сдвиг b. Будем настраивать K наборов параметров $w_1, \dots, w_K$, и итоговый алгоритм определим как

$$a(x) = \mathop{\arg\max}_{k \in \{1, \dots, K\}} \langle w_k, x \rangle.$$

Рассмотрим следующую функцию потерь:

$$\max_k \Bigl\{ \langle w_k, x \rangle + 1 - [k = y(x)] \Bigr\} - \langle w_{y(x)}, x \rangle. \tag{3.6}$$

Разберемся сначала с выражением, по которому берется максимум. Если $k = y(x)$, то оно равно $\langle w_k, x \rangle$; в противном же случае оно равно $\langle w_k, x \rangle + 1$. Если оценка за верный класс больше оценок за остальные классы хотя бы на единицу, то максимум будет достигаться на $k = y(x)$; в этом случае потеря будет равна нулю. Иначе же потеря будет больше нуля. Здесь можно увидеть некоторую аналогию с бинарным SVM: мы штрафуем не только за неверный ответ на объекте, но и за неуверенную классификацию (за попадание объекта в разделяющую полосу).

Рассмотрим сначала линейно разделимую выборку — т.е. такую, что существуют веса $w_{1*}, \dots, w_{K*}$, при которых потеря (3.6) равна нулю. В бинарном SVM мы строили классификатор с максимальным отступом. Известно, что аналогом отступа для многоклассового случая является норма Фробениуса матрицы W, k-я строка которой совпадает с w_k:

$$\rho = \frac{1}{\|W\|^2} = \frac{1}{\sum_{k=1}^K \sum_{j=1}^d w_{kj}}.$$

Получаем следующую задачу:

$$\begin{cases} \frac{1}{2}\|W\|^2 \to \min\limits_W \\ \langle w_{y_i}, x_i \rangle + [y_i = k] - \langle w_k, x_i \rangle \geqslant 1, \quad i = 1, \dots, \ell; k = 1, \dots, K. \end{cases} \tag{3.7}$$

Перейдем теперь к общему случаю. Как и в бинарном методе опорных векторов, перейдем к мягкой функции потерь, введя штрафы за неверную или неуверенную классификацию. Получим задачу

$$\begin{cases} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{\ell}\xi_i \to \min\limits_{W,\xi} \\ \langle w_{y_i}, x_i\rangle + [y_i = k] - \langle w_k, x_i\rangle \geqslant 1 - \xi_i, \quad i = 1, \dots, \ell; k = 1, \dots, K; \\ \xi_i \geqslant 0, \quad i = 1, \dots, \ell. \end{cases} \tag{3.8}$$

Решать задачу (3.8) можно, например, при помощи пакета SVM$^{\text{multiclass}}$.

§3.3 Вывод двойственной задачи

Вернемся к бинарному методу опорных векторов и построим двойственную задачу к (3.5):

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

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:

$$\nabla_w L = w - \sum_{i=1}^{\ell}\lambda_i y_i x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{\ell}\lambda_i y_i x_i \tag{3.9}$$

$$\nabla_b L = -\sum_{i=1}^{\ell}\lambda_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell}\lambda_i y_i = 0 \tag{3.10}$$

$$\nabla_{\xi_i} L = C - \lambda_i - \mu_i \quad \Longrightarrow \quad \lambda_i + \mu_i = C \tag{3.11}$$

$$\lambda_i\left[y_i\left(\langle w, x_i\rangle + b\right) - 1 + \xi_i\right] = 0 \quad \Longrightarrow \quad (\lambda_i = 0) \text{ или } \left(y_i\left(\langle w, x_i\rangle + b\right) = 1 - \xi_i\right) \tag{3.12}$$

$$\mu_i\xi_i = 0 \quad \Longrightarrow \quad (\mu_i = 0) \text{ или } (\xi_i = 0) \tag{3.13}$$

Проанализируем полученные условия. Из (3.9) 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. Из первого условия дополняющей нежесткости (3.12) следует, что итоговый классификатор будет зависеть от объекта x_i (т.е. вес при x_i будет отличен от нуля) тогда и только тогда, когда $\left(y_i\left(\langle w, x_i\rangle + b\right) = 1 - \xi_i\right)$. Если $\xi_i = 0$, то $y_i(\langle w, x_i\rangle + b) = 1$, и объект x_i лежит на границе разделяющей полосы. Если же $\xi \neq = 0$, то, согласно (3.13), $\mu_i = 0$; из (3.11) отсюда следует, что $\lambda_i = C$. Итак, итоговый классификатор зависит только от объектов, лежащих на границе разделяющей полосы, и от объектов-нарушителей (с $\xi_i > 0$), для которых выполнено $\lambda_i = C$.

Построим двойственную функцию. Для этого подставим выражение (3.9) в лагранжиан, и воспользуемся уравнениями (3.10) и (3.11) (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 (3.10) и (3.11) (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 (3.11), можно объединить, получив $\lambda_i \leqslant C$. Приходим к следующей двойственной задаче:

$$\begin{cases} \sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle \to \max\limits_{\alpha} \\ 0 \leqslant \alpha_i \leqslant C, \quad i = 1, \dots, \ell, \\ \sum_{i=1}^{\ell}\lambda_i y_i = 0. \end{cases} \tag{3.14}$$

Список литературы

[1] *Mohri, M., Rostamizadeh, A., Talwalkar, A.* Foundations of Machine Learning. // MIT Press, 2012.

[2] *Crammer, K., Singer, Y.* On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. // Journal of Machine Learning Research, 2:265-292, 2001.