

Семинары по линейным классификаторам

Евгений Соколов
sokolov.evg@gmail.com

20 ноября 2015 г.

4 Условия Куна-Таккера и SVM

§4.1 Условия Куна-Таккера и двойственность, продолжение

Задача 4.1. Покажите, что задача минимизации регуляризованного функционала

$$Q(w) + \tau \|w\|_p \rightarrow \min_w \quad (4.1)$$

с $p \geq 1$ и $\tau \geq 0$ эквивалентна условной задаче

$$\begin{cases} Q(w) \rightarrow \min_w \\ \|w\|_p \leq C \end{cases} \quad (4.2)$$

если функционал $Q(w)$ является выпуклым. Здесь имеется ввиду, что для любого τ найдется такое C , что эти задачи эквивалентны, и наоборот — для любого C найдется такое τ .

Решение. Задача (4.2) является выпуклой и для нее выполнено условие Слейтера, поэтому вектор w_* является ее решением тогда и только тогда, когда он удовлетворяет условиям Куна-Таккера:

$$\begin{cases} \nabla_w (Q(w_*) + \lambda^* \|w_*\|_p) = 0 \\ \|w_*\|_p \leq C \\ \lambda^* \geq 0 \\ \lambda^* (\|w_*\|_p - C) = 0 \end{cases}$$

Пусть w_* — решение задачи (4.2), тогда из условий Куна-Таккера получаем, что градиент лагранжиана в данной точке равен нулю. Поскольку лагранжиан является выпуклым, то из равенства нулю градиента в точке w_* следует, что w_* является глобальным минимумом лагранжиана. Следовательно, вектор w_* является решением задачи

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

которая эквивалентна задаче (4.1) при $\tau = \lambda^*$. Значит, если w_* является решением задачи (4.2), то он является решением задачи (4.1).

Пусть теперь w_* — решение задачи (4.1). Положим $C = \|w_*\|_p$ и $\lambda^* = \tau$. Тогда пара (w_*, λ^*) удовлетворяет условиям Куна-Таккера и, следовательно, является решением задачи (4.2). ■

§4.2 Метод опорных векторов

4.2.1 Формулировка

Будем рассматривать линейные классификаторы вида

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

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 $a(x) = \text{sign}\langle w, x \rangle + b$. 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

$$\min_{x \in X^\ell} |\langle w, x \rangle + b| = 1. \quad (4.3)$$

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

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

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:

$$\begin{cases} \frac{1}{2} \|w\|^2 \rightarrow \min_{w, b} \\ y_i (\langle w, x_i \rangle + b) \geq 1, \quad i = 1, \dots, \ell. \end{cases} \quad (4.4)$$

Здесь мы воспользовались тем, что линейный классификатор дает правильный ответ на объекте x_i тогда и только тогда, когда $\langle w, x_i \rangle + b \geq 0$. Более того, из условия нормировки (4.3) следует, что $\langle w, x_i \rangle + b \geq 1$.

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 (4.4) будет нарушено:

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

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

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

Отметим, что если отступ объекта лежит между нулем и единицей ($0 \leq y_i (\langle w, x_i \rangle + b) < 1$), то объект верно классифицируется, но имеет ненулевой штраф $\xi > 0$. Таким образом, мы штрафует 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

$$\begin{cases} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min_{w, b, \xi} \\ y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad i = 1, \dots, \ell, \\ \xi_i \geq 0, \quad i = 1, \dots, \ell. \end{cases} \quad (4.5)$$

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

Данная задача также является выпуклой и имеет единственное решение.

§4.3 Вывод двойственной задачи

Построим двойственную задачу к (4.5):

$$\begin{cases} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min_{w, b, \xi} \\ y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad i = 1, \dots, \ell, \\ \xi \geq 0, \quad i = 1, \dots, \ell. \end{cases}$$

Запишем лагранжиан:

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

Выпишем условия Куна-Таккера:

$$\nabla_w L = w - \sum_{i=1}^{\ell} \lambda_i y_i x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{\ell} \lambda_i y_i x_i \quad (4.6)$$

$$\nabla_b L = - \sum_{i=1}^{\ell} \lambda_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0 \quad (4.7)$$

$$\nabla_{\xi_i} L = C - \lambda_i - \mu_i \quad \Longrightarrow \quad \lambda_i + \mu_i = C \quad (4.8)$$

$$\lambda_i [y_i (\langle w, x_i \rangle + b) - 1 + \xi_i] = 0 \quad \Longrightarrow \quad (\lambda_i = 0) \text{ или } (y_i (\langle w, x_i \rangle + b) = 1 - \xi_i) \quad (4.9)$$

$$\mu_i \xi_i = 0 \quad \Longrightarrow \quad (\mu_i = 0) \text{ или } (\xi_i = 0) \quad (4.10)$$

$$\xi_i \geq 0, \lambda_i \geq 0, \mu_i \geq 0. \quad (4.11)$$

Проанализируем полученные условия. Из (4.6) 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 на 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 (4.9) следует, что $y_i (\langle w, x_i \rangle + b) = 1$, 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 ($\xi_i \geq 2$). 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 (4.10) следует, что $\mu_i = 0$, и отсюда из уравнения (4.8) 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 (4.6) в лагранжиан, и воспользуемся уравнениями (4.7) и (4.8) (данные три уравнения выполнены для точки минимума лагранжиана при любых фиксированных λ и μ):

$$\begin{aligned} L &= \frac{1}{2} \left\| \sum_{i=1}^{\ell} \lambda_i y_i x_i \right\|^2 - \sum_{i,j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle - b \underbrace{\sum_{i=1}^{\ell} \lambda_i y_i}_0 + \sum_{i=1}^{\ell} \lambda_i + \sum_{i=1}^{\ell} \xi_i \underbrace{(C - \lambda_i - \mu_i)}_0 \\ &= \sum_{i=1}^{\ell} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle. \end{aligned}$$

Мы должны потребовать выполнения условий (4.7) и (4.8) (если они не выполнены, то двойственная функция обращается в минус бесконечность), а также неотрицательность двойственных переменных $\lambda_i \geq 0$, $\mu_i \geq 0$. Ограничение на μ_i и условие (4.8), можно объединить, получив $\lambda_i \leq C$. Приходим к следующей двойственной задаче:

$$\begin{cases} \sum_{i=1}^{\ell} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \rightarrow \max_{\lambda} \\ 0 \leq \lambda_i \leq C, \quad i = 1, \dots, \ell, \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases} \quad (4.12)$$

Она также вогнутой, квадратичной и имеет единственный максимум.

Вернемся к тому, какое представление классификатора дает двойственная задача. Из уравнения (4.6) следует, что вектор весов w можно представить как линейную комбинацию объектов из обучающей выборки. Подставляя это представление w в классификатор, получаем

$$a(x) = \text{sign} \left(\sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle + b \right). \quad (4.13)$$

Таким образом, классификатор измеряет сходство нового объекта с объектами из обучения, вычисляя скалярное произведение между ними.

В представлении (4.13) 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 $y_i (\langle w, x_i \rangle + b) = 1$, откуда получаем

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

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

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

Список литературы

- [1] *Mohri, M., Rostamizadeh, A., Talwalkar, A.* Foundations of Machine Learning. // MIT Press, 2012.
- [2] *Bishop, C.M.* Pattern Recognition and Machine Learning. // Springer, 2006.
- [3] *Crammer, K., Singer, Y.* On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. // Journal of Machine Learning Research, 2:265-292, 2001.