Линейные методы классификации II

Виктор Китов
v.v.kitov@yandex.ru

МГУ им.Ломоносова, ф-т ВМиК, кафедра ММП.

I семестр 2015 г.

Содержание

Постановка задачи

- Классификация двумя классами ω_1 и ω_2
- Линейное решающее правило:

$$\widehat{c} = \begin{cases} \omega_1, & w^T x \geq -w_0 \\ \omega_2, & w^T x < w_0 \end{cases}$$

эквивалентно:

1. снижению размерности до 1-мерного подпространства (определяемого вектором w)
2. классификации в этом подпространстве, путем сравнения координаты с порогом

Постановка задачи

- Классификация двумя классами ω_1 и ω_2
- Линейное решающее правило:

$$\widehat{c} = \begin{cases} \omega_1, & w^T x \geq -w_0 \\ \omega_2, & w^T x < w_0 \end{cases}$$

эквивалентно:

1. снижению размерности до 1-мерного подпространства (определяемого вектором w)
2. классификации в этом подпространстве, путем сравнения координаты с порогом

Идея линейного дискриминанта Фишера

Определить направление, проекции на которое лучше всего разделят классы.

Возможный подход

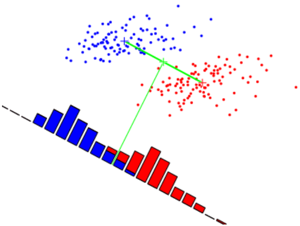
- Определим $C_1 = \{i : x_i \in \omega_1\}$, $\quad C_2 = \{i : x_i \in \omega_2\}$ и

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n, \quad m_2 = \frac{1}{N_1} \sum_{n \in C_2} x_n$$

$$\mu_1 = w^T m_1, \quad \mu_2 = w^T m_2$$

Очевидное, но не оптимальное решение:

$$\begin{cases} (\mu_1 - \mu_2)^2 \to \max_w \\ \|w\| = 1 \end{cases}$$



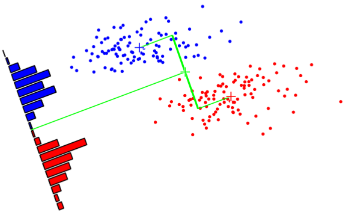
Линейный дискриминант Фишера

- Определим дисперсии проекций каждого класса:

$$s_1 = \sum_{n \in C_1} (w^T x_n - w^T m_1)^2, \quad s_2 = \sum_{n \in C_2} (w^T x_n - w^T m_2)^2$$

- Определение w в линейном дискриминанте Фишера:

$$\frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2} \to \max_w$$



Эквивалентное определение

$$\frac{(\mu_1-\mu_2)^2}{s_1^2+s_2^2} = \frac{\left(w^T m_1 - w^T m_2\right)^2}{\sum_{n\in C_1}(w^T x_n - w^T m_1)^2 + \sum_{n\in C_2}(w^T x_n - w^T m_2)^2}$$

$$= \frac{\left[w^T(m_1-m_2)\right]^2}{\sum_{n\in C_1}\left[w^T(x_n-m_1)\right]^2 + \sum_{n\in C_2}\left[w^T(x_n-m_1)\right]^2}$$

$$= \frac{w^T(m_1-m_2)(m_1-m_2)^T w}{w^T\left[\sum_{n\in C_1}(x_n-m_1)(x_n-m_1)^T + \sum_{n\in C_2}(x_n-m_2)(x_n-m_2)^T\right]w}$$

$$= \frac{w^T S_B w}{w^T S_W w} \to \max_w$$

$$\begin{aligned} S_B &= (m_1-m_2)(m_1-m_2)^T, \\ S_W &= \sum_{n\in C_1}(x_n-m_1)(x_n-m_1)^T + \sum_{n\in C_2}(x_n-m_2)(x_n-m_2)^T \end{aligned}$$

Решение

$$Q(w) = \frac{w^T S_B w}{w^T S_W w} \to \max_w$$

Используя свойство $\frac{d}{dw}\left(w^T A w\right) = 2Aw$ для каждой $A \in \mathbb{R}^{KxK}$, $A^T = A$, получим

$$\frac{dQ(w)}{dw} \propto 2S_B w \left[w^T S_W w\right] - 2\left[w^T S_B w\right] S_W w = 0,$$

что эквивалентно

$$\left[w^T S_W w\right] S_B w = \left[w^T S_B w\right] S_W w.$$

Таким образом,

$$w \propto S_W^{-1} S_B w \propto S_W^{-1}(m_1 - m_2)$$

Содержание

Логистическая регрессия

- Добавим в x константный признак и w_0 к w.
- Сигмоидная функция активации $\sigma(z) = \frac{1}{1+e^{-z}}$.
- Двухклассовая классификация:

$$score(\omega_1|x) = w^T x$$

$$p(\omega_1|x) = \sigma(w^T x)$$

- Многоклассовая классификация:

$$\begin{cases} score(\omega_1|x) = w_1^T x \\ score(\omega_2|x) = w_2^T x \\ \dots \\ score(\omega_C|x) = w_C^T x \end{cases}$$

Логистическая регрессия

Вероятности классов аппроксимируются через soft-max функцию:

$$p(\omega_c|x) = \frac{exp(w_c^T x)}{\sum_i exp(w_i^T x)}$$

w_c, $c = 1, 2, ...C$ определены с точностью до сдвига на произвольный вектор v:

$$\frac{exp((w_c - v)^T x)}{\sum_i exp((w_i - v)^T x)} = \frac{exp(-v^T x)exp(w_c^T x)}{\sum_i exp(-v^T x)exp(w_i^T x)} = \frac{exp(w_c^T x)}{\sum_i exp(w_i^T x)}$$

Обычно сдвигают все w_c на $v = w_C$.

Замечание: нелинейное преобразование score в вероятность могло быть определено и по-другому - получили бы другой метод.

Логистическая регрессия

- Пусть γ_1, γ_2 - цены неправильной классификации классов ω_1 и ω_2.
- Предположим

$$\ln\left(\frac{\gamma_1 p(\omega_1|x)}{\gamma_2 p(\omega_2|x)}\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$$

- это эквивалентно

$$\begin{aligned} p(\omega_2|x) &= \frac{1}{1 + exp(\beta_0' + \boldsymbol{\beta}^T \boldsymbol{x})} \\ p(\omega_1|x) &= \frac{exp(\beta_0' + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + exp(\beta_0' + \boldsymbol{\beta}^T \boldsymbol{x})} \end{aligned}$$

- где $\beta_0' = \beta_0 - \ln(\gamma_1/\gamma_2)$

Логистическая регрессия

Решающее правило (следуя Байесовскому правилу минимальной цены):

$$x = \begin{cases} \omega_1, & \beta_0' + \boldsymbol{\beta}^T \boldsymbol{x} > 0 \\ \omega_2, & \beta_0' + \boldsymbol{\beta}^T \boldsymbol{x} < 0 \end{cases}$$

Оценка $\beta_0', \boldsymbol{\beta}$ методом максимального правдоподобия:

$$\prod_{i=1}^{N} p(c_i | x_i) \to \max_{\beta_0', \boldsymbol{\beta}}$$

где c_i - класс объекта x_i.

Многоклассовая логистическая регрессия

- Предположение:

$$\ln\left(\frac{\gamma_s p(\omega_s|x)}{\gamma_C p(\omega_C|x)}\right) = \beta_{s0} + \boldsymbol{\beta}_s^T \mathbf{x}, \quad s = 1, 2, \ldots C-1$$

- Вероятности классов (дающие эквивалентное определение):

$$\begin{aligned} p(\omega_s|x) &= \frac{exp(\beta'_{s0} + \beta_s^T x)}{1 + \sum_{s=1}^{C-1} exp(\beta'_{s0} + \beta_s^T x)}, \quad s = 1, 2, \ldots C-1 \\ p(\omega_C|x) &= \frac{1}{1 + \sum_{s=1}^{C-1} exp(\beta'_{s0} + \beta_s^T x)} \end{aligned}$$

$$\beta'_{s0} = \beta_{s0} - \ln\left(\gamma_s / \gamma_C\right)$$

- Интерпретация: soft-max от дискриминатных функций (для классов $\omega_1, \omega_2, \ldots \omega_{C-1}$) и константы (для класса ω_C).

Многоклассовая логистическая регрессия

- Решающее правило (Байесовское правило минимальной ожидаемой цены):
- $c = \arg\max_c \beta_{c0} + \beta_c^T x$, если $\beta_{c0} + \beta_c^T x > 0$ иначе сопоставить x классу C.
- Оценивание методом максимального правдоподобия:

$$\prod_{i=1}^{N} p(c_i|x_i) \to \max_{\beta_0', \boldsymbol{\beta}}$$

Функция цены

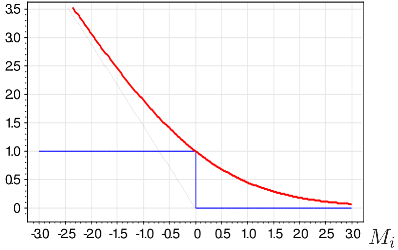
Для 2-х классов $p(y|x) = \sigma(\langle w, x\rangle y)$, где $\sigma = \frac{1}{1+e^{-z}}$, $w = [\beta_0', \beta]$, $x = [1, x_1, x_2, ...x_D]$.

Оценка методом максимального правдоподобия:

$$\prod_{i=1}^{N} \sigma(\langle w, x_i\rangle y_i) \to \max_w$$

эквивалентна

$$\sum_{i=1}^{N} \ln(1 + e^{-\langle w, x_i\rangle y_i}) \to \min_w$$



Следовательно, мажорирующая ф-ция для логистической регрессии $\mathcal{L}(M) = \ln(1 + e^{-M})$.

Метод стохастического градиента

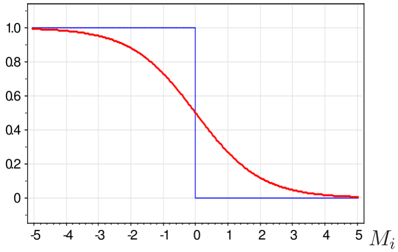
Подставляя $\mathcal{L}(M) = \ln(1 + e^{-M})$ в метод стохастического градиентного спуска, получаем:

$$w \leftarrow w + \eta\sigma(-M_i)x_i y_i$$

Правило обновления персептрона Розенблатта:

$$w \leftarrow w + \eta\mathbb{I}[M_i < 0]x_i y_i$$

- Обновление логистической регрессии - сглаженный вариант обновления персептрона Розенблатта.
- Чем существеннее ошибка - тем сильнее обновление весов.



Содержание

3 Метод опорных векторов
- Случай линейно разделимых классов
- Случай линейно неразделимых классов

Условия Куна-Таккера

Рассмотрим задачу оптимизации:

$$\begin{cases} f(x) \to \min_x \\ g_i(x) \leq 0 \qquad i = 1, 2, ...m \end{cases} \tag{1}$$

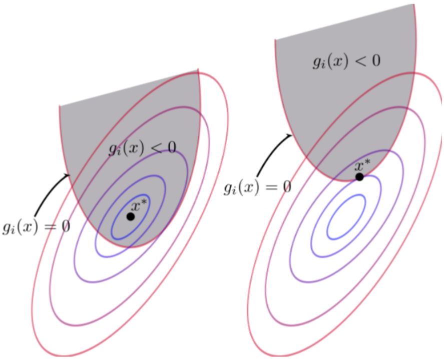
Теорема (необходимые условия оптимальности):
Пусть

- x^* - допустимое решение задачи (1),
- $f(x^*)$ и $g_i(x^*)$, $i = 1, 2, ...m$ - непрерывно-дифференцируемы в x^*.
- выполнено одно из условий регулярности

Тогда существуют такие коэффициенты $\lambda_1, \lambda_2, ...\lambda_m$, что x^* удовлетворяет условиям

$$\begin{cases} \nabla f(x^*) + \sum_{i=1}^m \lambda_i g_i(x^*) = 0 & \text{стационарность} \\ g_i(x) \leq 0 & \text{достижимость} \\ \lambda_i \geq 0 & \text{неотрицательность} \\ \lambda_i g_i(x^*) = 0 & \text{дополняющая нежесткость} \end{cases} \tag{2}$$

Иллюстрация оптимизации



Условия Куна-Таккера

Возможные условия регулярности:

- $\{\nabla g_j, j \in J\}$ - линейно независимы, где J - индексы активных ограничений $J = \{j : g_j(x^*) = 0\}$.
- Условие Слейтера: $\exists x : g_i(x) < 0 \, \forall i$ (применимо только когда $f(x)$ и $g_i(x)$, $i = 1, 2, ...m$ - выпуклы)

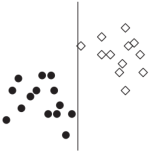
Достаточное условие оптимальности:
если $f(x)$ и $g_i(x)$, $i = 1, 2, ...m$ - выпуклы, выполнены условия (2) и условие Слейтера, то x^* - точка глобального минимума.

Выпуклые задачи оптимизации

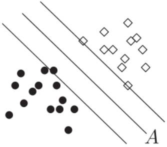
Чем удобно предположение о выпуклости $f(x)$ и $g_i(x)$, $i = 1, 2, ...m$:

- Все локальные минимумы являются глобальными минимумами
- Множество минимумов выпукло
- Если $f(x)$ - строго выпукла и минимум существует, то он единственен.

Метод опорных векторов



(a)



(b)

Метод опорных векторов

(a) (b)

Зазор

Зазор - это сумма расстояний от разделяющей гиперплоскости до множества объектов класса ω_1 и множества объектов класса ω_2 в обучающей выборке.

Основная идея

Определить линейную границу таким образом, что зазор между классами обучающей выборки был максимален.

Метод опорных векторов

Объекты x_i для $i = 1, 2, \dots n$ лежат на расстоянии $b/|w|$ от разделяющей гиперплоскости

$$\begin{cases} x_i^T w + w_0 \geq b, & y_i = +1 \\ x_i^T w + w_0 \leq -b & y_i = -1 \end{cases} \quad i = 1, 2, \dots N.$$

Это можно переписать как

$$y_i(x_i^T w + w_0) \geq b, \quad i = 1, 2, \dots N.$$

Зазор равен $2b/|w|$. Поскольку неизвестные параметры w, w_0 и b определены с точностью до мультипликативной константы, положим $b = 1$.

Постановка задачи

Постановка задачи:

$$\begin{cases} w^T w \to \min\limits_{w, w_0} \\ y_i(x_i^T w + w_0) \geq 1, \quad i = 1, 2, \dots N. \end{cases} \tag{3}$$

По теореме Куна-Таккера, оптимальное решение удовлетворяет условию:

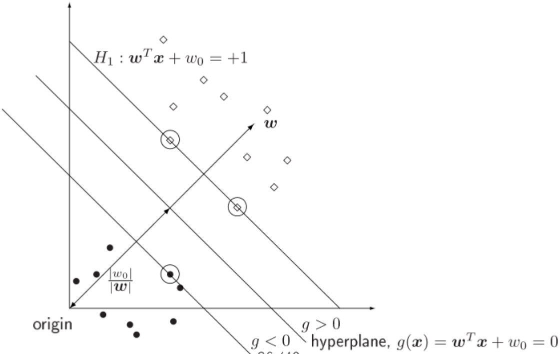
$$L_P = \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i (y_i(w^T x + w_0) - 1) \to \min_{w, w_0} \max_{\alpha}$$

с ограничениями:

$$\begin{cases} \alpha_i \geq 0, \\ y_i(x_i^T w + w_0) - 1 \geq 0, \\ \alpha_i (y_i(x_i^T w + w_0) - 1) = 0. \end{cases}$$

Опорные вектора

Условие $\alpha_i(y_i(x_i^T w + w_0) - 1) = 0$ выполнено либо когда $\alpha_i = 0$, либо когда $y_i(x_i^T w + w_0) - 1 = 0$. Случай $\alpha_i > 0$ описывает «опорные» вектора, которые лежат на расстоянии $1/|w|$ к разделяющей гиперплоскости и влияют на оптимальные веса. Другие веса не влияют на решение.



Двойственная задача

$$\frac{\partial L}{\partial w_0} = 0 : \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial w} = 0 : w = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (4)$$

Подставляя эти условия в Лагранжиан L_P, получаем:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \to \max_{\alpha}$$

α_i может быть найдено из следующей двойственной задачи:

$$\begin{cases} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \to \max_{\alpha} \\ \alpha_i \geq 0,\ i = 1, 2, ...n;\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$$

Решение

Определим $\mathcal{SV}$ как множество индексов опорных векторов. Оптимальные α_i определяют оптимальные веса w:

$$w = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i$$

w_0 может быть найдено из условия пограничности любого опорного вектора:

$$y_i(x_i^T w + w_0) = 1, \; i \in \mathcal{SV}$$

w_0, найденное из суммы пограничных условий 28 по n_{SV} опопрным векторам, будет более устойчивым:

$$n_{\mathcal{SV}} w_0 + \sum_{i \in \mathcal{SV}} x_i^T w = \sum_{i \in \mathcal{SV}} y_i$$

3 Метод опорных векторов

- Случай линейно разделимых классов
- Случай линейно неразделимых классов

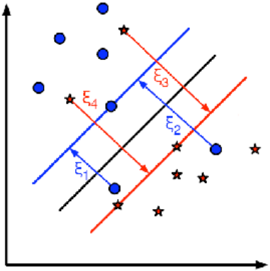
Ослабление условий оптимизации

- Пусть объекты обучающей выборки не могут быть линейно разделены на разные классы
- Оптимизационная задача модифицируется:
  - неравенства в (28) могут нарушаться на величины ξ_i
  - величины нарушений ξ_i, $i = 1, 2, ...N$ штрафуются в оптимизируемом критерии:

$$\begin{cases} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \to \min_{w,\xi} \\ y_i(w^T x_i + w_0) \geq 1 - \xi_i,\ i = 1, 2, ...N \\ \xi_i \geq 0,\ i = 1, 2, ...N \end{cases}$$

Линейно-неразделимый случай

- Новый параметр C
  - определяет цену неправильного разделения классов.
  - контролирует противоречие между простотой и точностью модели
  - выбирается на валидационном множестве
- Другие виды штрафа возможны, например $C\sum_i \xi_i^2$.



Решение для линейно-неразделимого случая

По теореме Куна-Таккера оптимальное решение также удовлетворяет: $L_P \to \min_{w,w_0,\xi} \max_{\alpha,r}$, где

$$L_P = \frac{1}{2}w^T w + C\sum_i \xi_i - \sum_{i=1}^N \alpha_i(y_i(w^T x_i + w_0) - 1 + \xi_i) - \sum_{i=1}^N r_i \xi_i$$

при ограничениях:

$$\begin{cases} \xi_i \geq 0,\ \alpha_i \geq 0,\ r_i \geq 0 \\ y_i(x_i^T w + w_0) \geq 1 - \xi_i, \\ \alpha_i(y_i(w^T x_i + w_0) - 1 + \xi_i) = 0 \\ r_i \xi_i = 0 \end{cases}$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 : \ C - \alpha_i - r_i = 0 \quad \Rightarrow \quad \alpha_i \in [0, C].$$

Типы обучающих объектов

- **Неинформативные объекты:**
  - для них $\alpha_i = 0$ ($\Leftrightarrow r_i = C \Leftrightarrow \xi_i = 0 \Leftrightarrow y_i(w^T x_i + w_0) \geq 1$)
- **Опорные вектора:**
  - для них $\alpha_i > 0$ ($\Leftrightarrow y_i(w^T x_i + w_0) = 1 - \xi_i$)
  - **граничные опорные вектора:**
    - имеют $\xi_i = 0$ ($\Leftrightarrow r_i > 0 \Leftrightarrow \alpha_i \in (0, C) \Leftrightarrow y(w^T x_i + w_0) = 1$), тогда опорный вектор лежит на расстоянии $1/|w|$ от разделяющей гиперплоскости.
  - **опорные вектора - «нарушители»:**
    - для них $\xi_i > 0$ ($\Leftrightarrow r_i = 0 \Leftrightarrow \alpha_i = C$), поэтому расстояние со знаком (зазор, margin) от них до разделяющей гиперплоскости меньше, чем $1/|w|$.
    - Если $\xi_i \in (0, 1)$, то опорный вектор все еще корректно классифицируется.
    - Если $\xi_i > 1$, то опорный вектор классифицируется неправильно.

Двойственная задача

$$\begin{aligned} \frac{\partial L_P}{\partial w_0} &= 0: \ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \frac{\partial L_P}{\partial w} &= 0: \ w = \sum_{i=1}^{N} \alpha_i y_i x_i \\ \frac{\partial L_P}{\partial \xi_i} &= 0: \ C - \alpha_i - r_i = 0 \end{aligned} \tag{5}$$

Подставляя эти условия в L_P, получим двойственную задачу:

$$\begin{cases} L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \to \max_\alpha \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases}$$

Решение

Обозначим через $\mathcal{SV}$ - множество индексов опорных векторов (для которых $\alpha_i > 0 \Leftrightarrow y(w^T x_i + w_0) = 1 - \xi_i$) и $\widetilde{\mathcal{SV}}$ - множество граничных опорных векторов
($\alpha_i \in (0, C) \Leftrightarrow \xi_i = 0,\ y(w^T x_i + w_0) = 1$)
Оптимальные α_i определяют веса w:

$$w = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i$$

w_0 может быть найдено из граничного условия на любой граничный опорный вектор $\xi_i = 0$:

$$y_i(x_i^T w + w_0) = 1,\ i \in \widetilde{\mathcal{SV}} \tag{6}$$

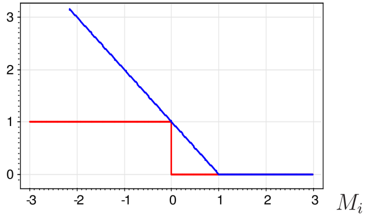
w_0, найденное из суммы (6) по всем граничным опорным векторам $i \in \widetilde{SV}$ будет более устойчиво:

$$n_{\widetilde{\mathcal{SV}}} w_0 + \sum_{i \in \widetilde{\mathcal{SV}}} x_i^T w = \sum_{i \in \widetilde{\mathcal{SV}}} y_i$$

Функция цены, соответствующая методу опорных векторов

Оптимизационная задача:

$$\begin{cases} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \to \min_{w,\xi} \\ y_i(w^T x_i + w_0) = M_i(w, w_0) \geq 1 - \xi_i, \\ \xi_i \geq 0,\ i = 1, 2, \dots N \end{cases}$$



может быть переписана как

$$\frac{1}{2C}|w|^2 + \sum_{i=1}^{N} [1 - M_i(w, w_0)]_+ \to \min_{w,\xi}$$

Таким образом, метод опорных векторов - это линейный классификатор с функцией цены $\mathcal{L}(M) = [1 - M]_+$ и L_2-регуляризацией.

Вероятностная интерпретация

Целевой критерий метода опорных векторов может быть получен, используя

$$p(y_i|x_i, w, w_0) \propto e^{-[1-M_i(w,w_0)]_+}$$

и априорное распределение на веса

$$p(w|C) \propto e^{-|w|^2/(2C)}$$

Свойства

Из (4) и (5) следует, что решение имеет вид:

$$y = \operatorname{sign}\left\{\sum_{i \in \mathcal{SV}} \alpha_i y_i < x_i, x > + w_0\right\}$$

Разреженность метода опорных векторов: решение зависит только от опорных векторов:

- поэтому объекты-выбросы сильнее влияют на решение
- возможный подход к фильтрации выбросов:
    1. решить, используя все объекты
    2. удалить объекты с минимальным отступом
    3. решить задачу заново на отфильтрованной выборке
- если только небольшая часть объектов неправильно классифицирована, то они могут быть удалены из выборки, чтобы она стала линейно разделимой
    - пропадает необходимость подбирать C.

Многоклассовая классификация

Необходимо выбрать класс среди $\omega_1, \omega_2, ...\omega_C$.

- Схема «один против всех с весами»:
    - нужно построить C бинарных классификаторов
    - нет неоднозначности
- Схема «один против одного»:
    - нужно построить $C(C-1)/2$ бинарных классификаторов
    - нужно как-то разрешать ситуации неоднозначности
- Применение изначально многоклассового алгоритма

Многоклассовый алгоритм

C линейных дискриминантных функций оцениваются одновременно:

$$g_k(x) = (w^k)^T x + w_0^k$$

Линейно разделимый случай:

$$\begin{cases} \sum_{k=1}^{C} (w^k)^T w^k \to \min_w \\ (w^{c(i)})^T x + w_0^{c(i)} - (w^k)^T x - w_0^k \geq 1 \, \forall k \neq c(i), \; i = 1, 2, \ldots N \end{cases}$$

Линейно неразделимый случай:

$$\begin{cases} \sum_{k=1}^{C} (w^k)^T w^k + C \sum_{i=1}^{N} \xi_i \to \min_w \\ (w^{c(i)})^T x + w_0^{c(i)} - (w^k)^T x - w_0^k \geq 1 - \xi_i \, \forall k \neq c(i), \; i = 1, 2, \ldots N \\ \xi_i \geq 0 \end{cases}$$