Введение в анализ данных

Лекция 15

Метод опорных векторов

Евгений Соколов

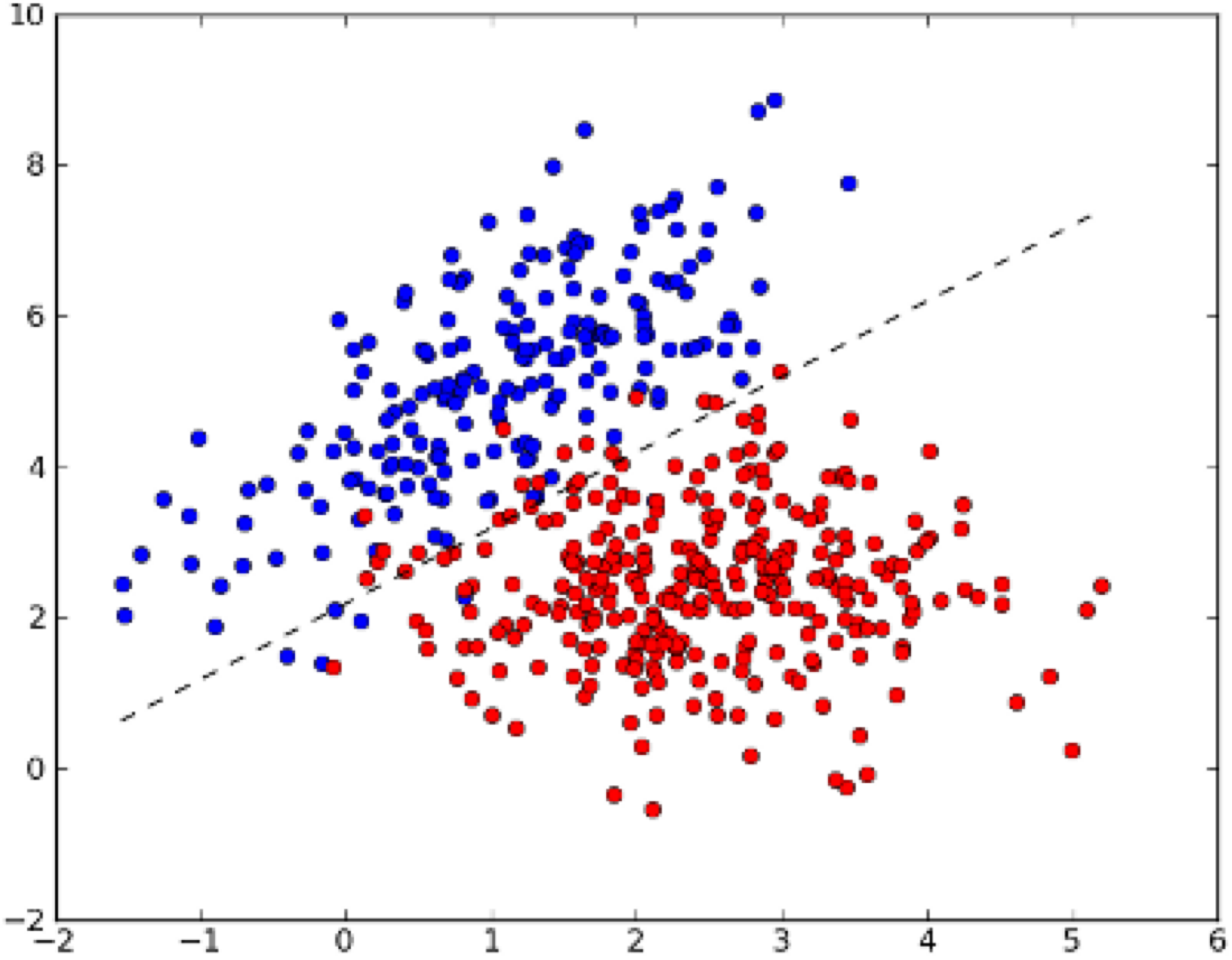
esokolov@hse.ru

НИУ ВШЭ, 2019

Логистическая регрессия

- Линейная модель классификации: $a(x) = \operatorname{sign} \langle w, x \rangle$
- Позволяет оценивать вероятности: $\pi(x) = \sigma(\langle w, x \rangle)$

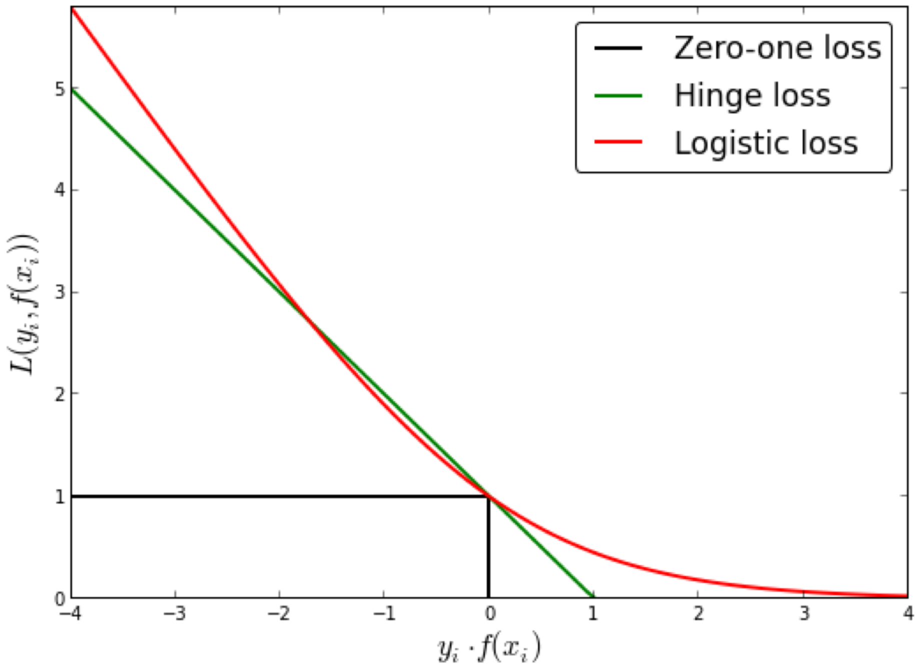
$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log_2(1 + \exp(-y_i \langle w, x_i \rangle)) \to \min_w$$



10
8
6
4
2
0
−2
−2
−1
0
1
2
3
4
5
6

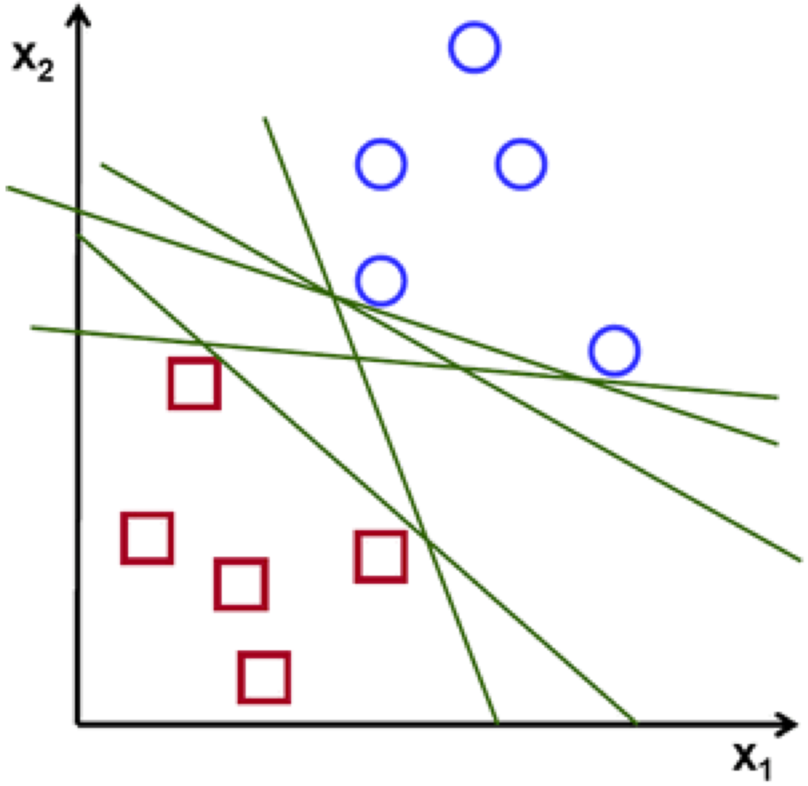
Смысл штрафа

- Отступ должен быть как можно больше
- Логично, поскольку вероятность должна быть как можно ближе к нулю или единице
- А если забыть про вероятности и требовать только корректной классификации?



Идея максимизации отступа

Объекты должны быть как можно дальше от разделяющей прямой



Идея максимизации отступа

Предположение:

выборку можно идеально разделить линейным классификатором

Т.е. существует такое w, что для всех объектов обучающей выборки

$$y_i(\langle w, x_i \rangle + w_0) > 0$$

Идея максимизации отступа

$$a(x) = \text{sign}(\langle w, x \rangle + w_0)$$

Подготовка: если поделить w и w_0 на положительное число, то ответы классификатора не поменяются

Пример:

$$\text{sign}(10 * x_1 + 4 * x_2 + 2) = \text{sign}(5 * x_1 + 2 * x_2 + 1)$$

Идея максимизации отступа

$$a(x) = \text{sign}(\langle w, x \rangle + w_0)$$

Подготовка: если поделить w и w_0 на положительное число, то ответы классификатора не поменяются

Поделим так, что выполнено **условие нормировки**

$$\min_{x \in X} |\langle w, x \rangle + w_0| = 1$$

(на обучающей выборке минимальный модуль прогноза равен 1)

Геометрия линейного классификатора

- Расстояние от точки до гиперплоскости $\langle w, x \rangle + w_0 = 0$:

$$\frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

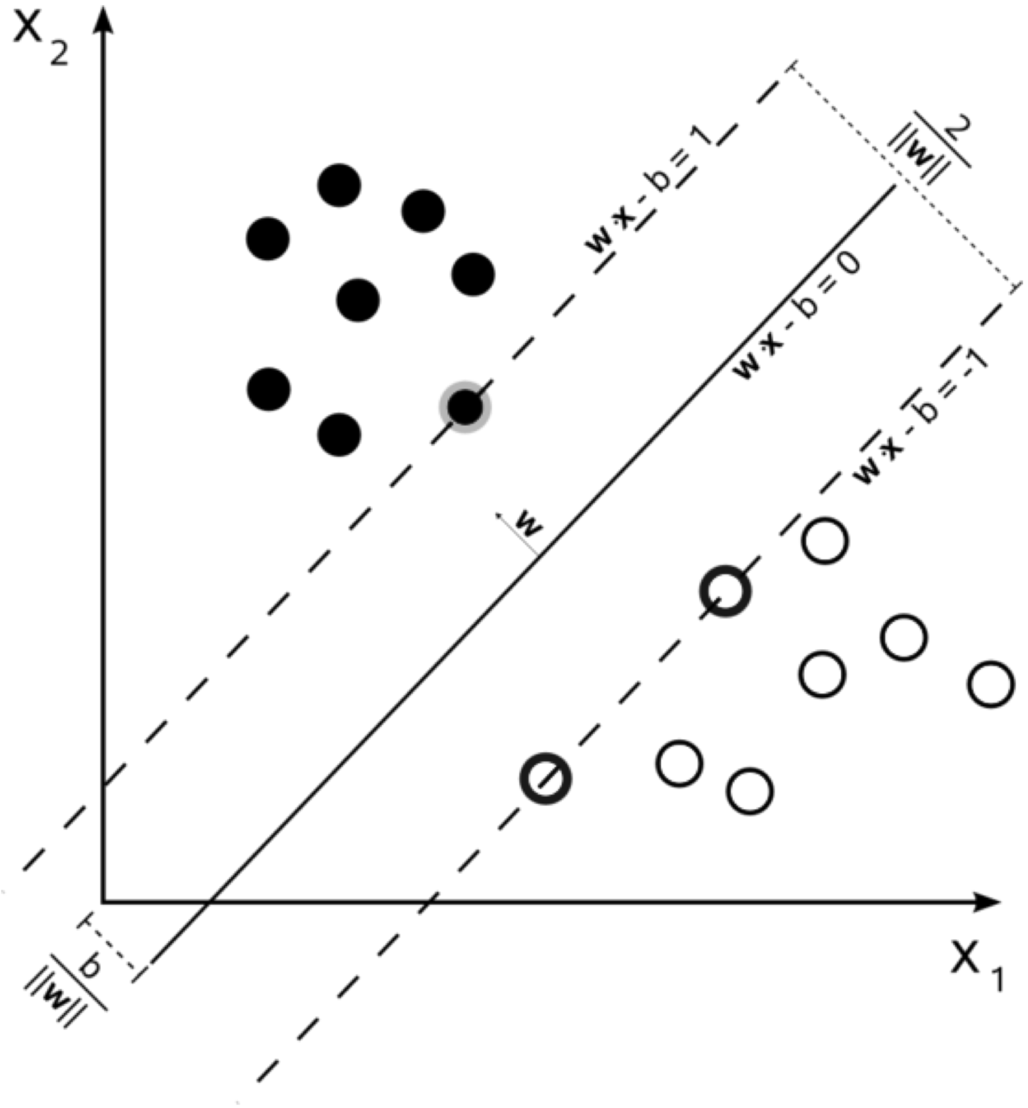
- Чем больше $\langle w, x \rangle$, тем дальше объект от разделяющей гиперплоскости

Идея максимизации отступа

Минимальное расстояние от объекта обучающей выборки до разделяющей гиперплоскости:

$$\min_{x \in X} \frac{|\langle w, x \rangle + w_0|}{\|w\|} = \frac{1}{\|w\|} \min_{x \in X} |\langle w, x \rangle + w_0| = \frac{1}{\|w\|}$$

То есть ширина разделяющей полосы равна $\frac{2}{\|w\|}$



X_2
X_1
$\mathbf{w} \cdot \mathbf{x} - b = 1$
$\mathbf{w} \cdot \mathbf{x} - b = 0$
$\mathbf{w} \cdot \mathbf{x} - b = -1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$

Задача максимизации ширины полосы

$$\begin{cases} \|w\|^2 \to \min\limits_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

Неравенство равносильно двум другим неравенствам:

- $y_i(\langle w, x_i \rangle + w_0) > 0$ — все объекты должны правильно классифицироваться
- $|\langle w, x \rangle + w_0| \geq 1$ — условие нормировки

Задача максимизации ширины полосы

$$\begin{cases} \|w\|^2 \to \min\limits_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

Для решения существуют специальные методы оптимизации

Задача максимизации ширины полосы

- Всё это время мы изучали задачу, которая никогда не встречается!
- Вряд ли настоящие данные будут линейно разделимыми

Задача максимизации ширины полосы

$$\begin{cases} \|w\|^2 \to \min\limits_{w, w_0} \\ \boxed{y_i(\langle w, x_i \rangle + w_0) \geq 1} \end{cases}$$

- Условие невозможно выполнить для всех объектов на линейно неразделимой выборке
- Сделаем условие более мягким

Задача максимизации ширины полосы

$$\begin{cases} \|w\|^2 \to \min\limits_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

- Теперь мы разрешаем отступу на некоторых объектах быть меньше единицы
- Что не так с этой задачей?

Задача максимизации ширины полосы

$$\begin{cases} \|w\|^2 \to \min\limits_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

- Теперь мы разрешаем отступу на некоторых объектах быть меньше единицы
- Что не так с этой задачей?
- Можно взять $\xi_i = +\infty$, и тогда подойдёт решение $w = 0$

Метод опорных векторов (SVM)

$$\begin{cases} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \to \min\limits_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Параметр C отвечает за баланс между шириной полосы и качеством классификации:

- При больших C стараемся не допускать ошибок
- При малых C разрешаем не обращать внимание на много объектов

Метод опорных векторов (SVM)

$$\begin{cases} \|w\|^2 + C\sum_{i=1}^{\ell} \xi_i \to \min\limits_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Последние два условия равносильны одному равенству:

$$\xi_i = \max\bigl(0, 1 - y_i(\langle w, x_i \rangle + w_0)\bigr)$$

Метод опорных векторов (SVM)

Последние два условия равносильны одному равенству:

$$\xi_i = \max\big(0, 1 - y_i(\langle w, x_i \rangle + w_0)\big)$$

Эквивалентная задача:

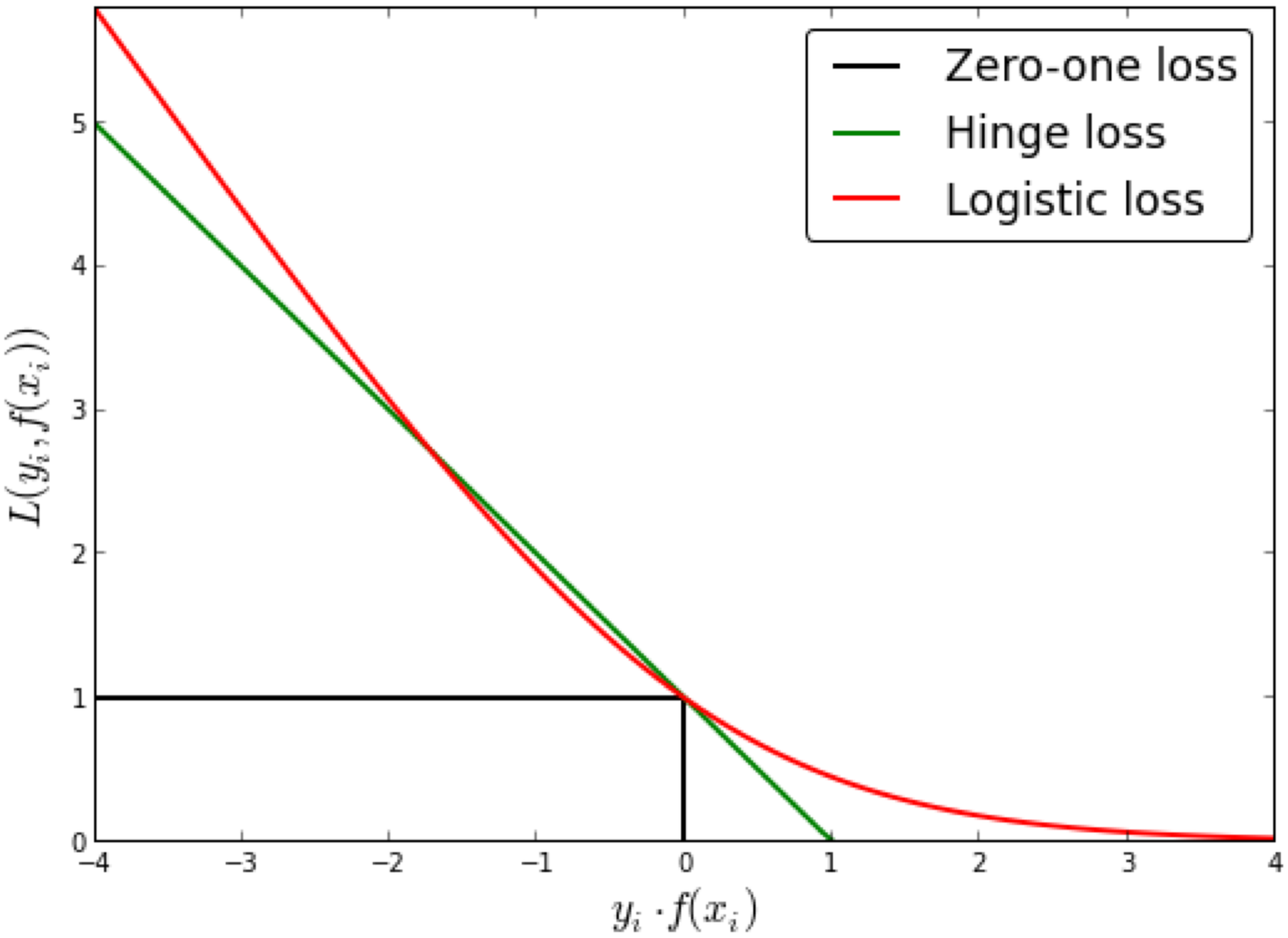
$$\|w\|^2 + C \sum_{i=1}^{\ell} \max\big(0, 1 - y_i(\langle w, x_i \rangle + w_0)\big) \to \min_{w, w_0}$$

Метод опорных векторов (SVM)

$$\|w\|^2 + C \sum_{i=1}^{\ell} \max\big(0, 1 - y_i(\langle w, x_i \rangle + w_0)\big) \to \min_{w, w_0}$$

Регуляризатор

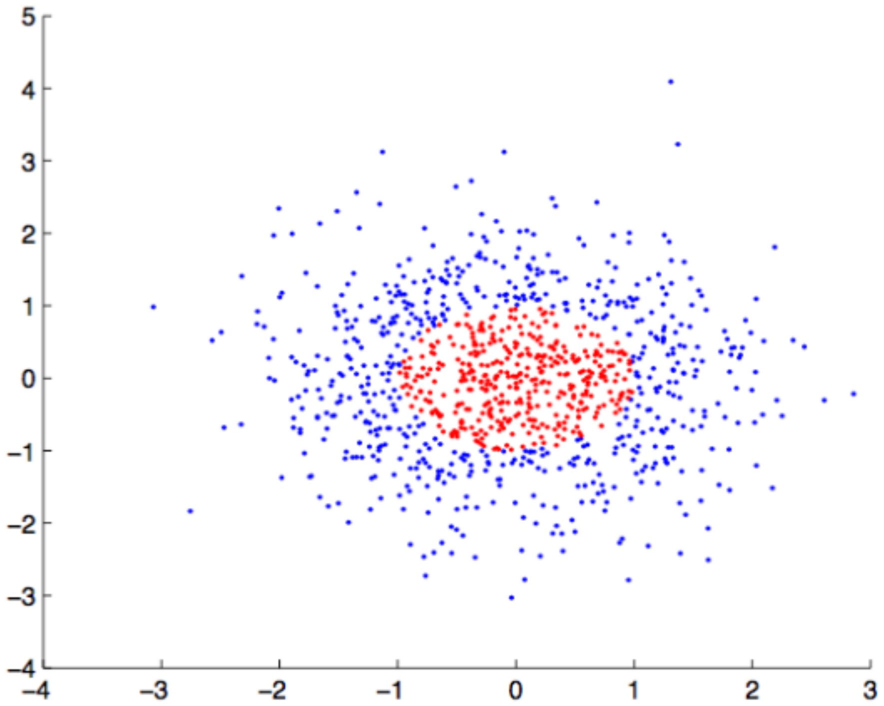
Ошибка с функцией потерь
$L(M) = \max(0, 1 - M)$

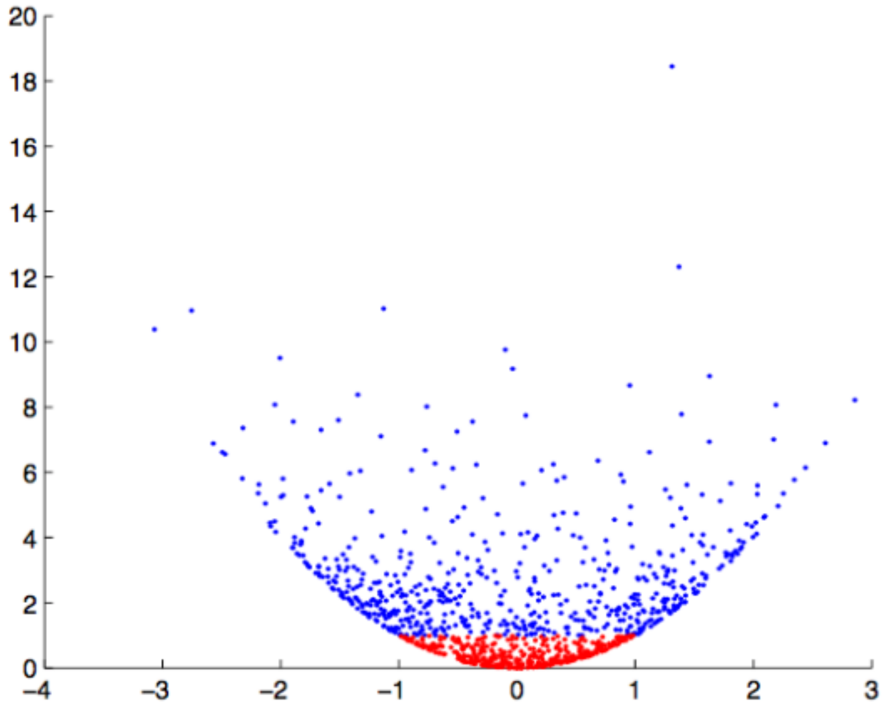


Zero-one loss
Hinge loss
Logistic loss
$L(y_i, f(x_i))$
$y_i \cdot f(x_i)$
0
1
2
3
4
5
−4
−3
−2
−1
0
1
2
3
4

Как сделать линейную модель нелинейной?

Нелинейные признаки





Нелинейные признаки

- Можно добавлять к исходным признакам их нелинейные преобразования
- Пример: $x_i^2,\ x_i x_j$
- Если исходных признаков 100, то парных — несколько тысяч
- Если исходных признаков 1000, то парных — сотни тысяч
- Проблемы с памятью и производительностью

Нелинейные признаки

- Допустим, мы полностью переходим к парным признакам:

$$\phi(x) = \left(x_i x_j\right)_{i,j=1}^{\ell}$$

- Скалярное произведение:

$$\langle \phi(x), \phi(z) \rangle = \sum_{i=1}^{\ell} x_i x_j z_i z_j = \sum_{i=1}^{\ell} x_i z_i \sum_{j=1}^{\ell} x_j z_j = \langle x, z \rangle^2$$

- При некоторых нелинейных признаках скалярные произведения считаются легко и без дополнительной памяти

Ядровой переход

Двойственная задача

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

Что общего у этой задачи и SVM?

Двойственная задача

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

Что общего у этой задачи и SVM?

$$w = \sum_{i=1}^{\ell} \lambda_i y_i x_i$$

Двойственная задача

$$w = \sum_{i=1}^{\ell} \lambda_i y_i x_i$$

Модель:

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle + w_0\right)$$

Двойственная задача

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

- Важно: задача зависит от объектов только через их скалярные произведения!
- Можно заменить на скалярные произведения, соответствующие нелинейным признакам!

Двойственная задача

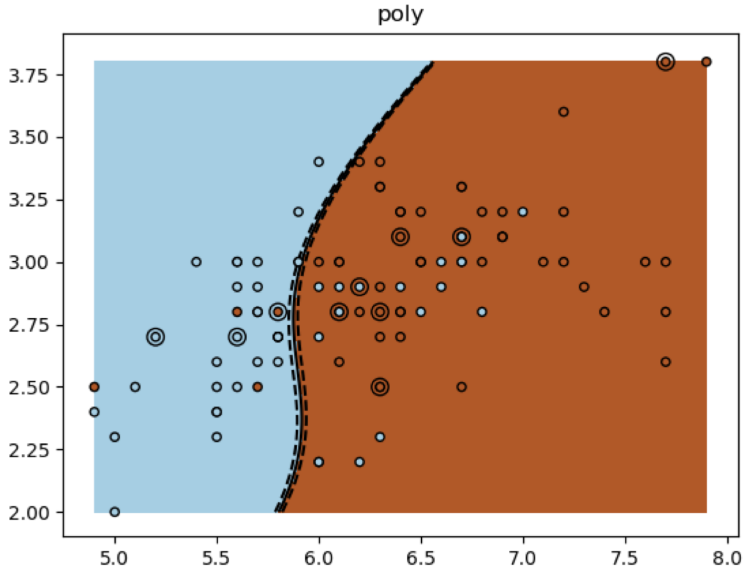
$$\begin{cases} \sum_{i=1}^{\ell} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \to \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, \ell, \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases}$$

Здесь $K(x, z)$ — **ядро**, новое скалярное произведение

Примеры ядер

$$K(x, z) = (\langle x, z \rangle + R)^m$$

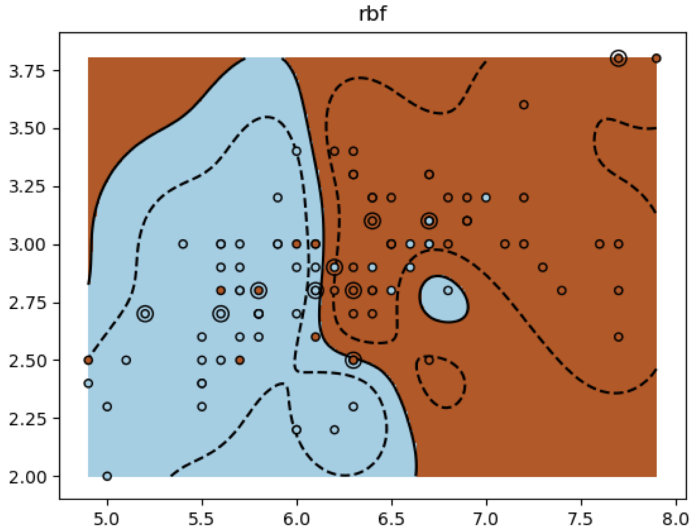
- Полиномиальное ядро
- Соответствует добавлению всех мономов степени не выше m



Примеры ядер

$$K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$$

- Гауссово ядро
- Соответствует добавлению **всех** мономов



Метод опорных векторов

- Требует правильной классификации объектов и наличия небольшого отступа
- Может быть преобразован так, что использует только скалярные произведения
- Ядра позволяют без дополнительных затрат подменить исходные признаковые описания на нелинейные