



Лекция 9: линейные методы

Евгений Борисов

четверг, 15 ноября 2018 г.

Линейные методы

методы ML

- статистические -
восстановить плотность, определить вероятность
- метрические -
померять расстояния, определить ближайших
- логические -
построить правило (комбинацию предикатов)
- линейные -
построить разделяющую поверхность

Линейные методы: о задаче классификации

метки классов

$$Y = \{-1, 1\}$$

размеченные данные

$$X = (x, y)$$

Линейные методы: о задаче классификации

метки классов

$$Y = \{-1, 1\}$$

размеченные данные

$$X = (x, y)$$

алгоритм классификации

$$a(x, w) = \text{sign}(f(x, w))$$

Линейные методы: о задаче классификации

метки классов

$$Y = \{-1, 1\}$$

размеченные данные

$$X = (x, y)$$

алгоритм классификации

$$a(x, w) = \text{sign}(f(x, w))$$

дискриминантная функция

$$f(x, w)$$

вектор параметров

$$w$$

Линейные методы: о задаче классификации

метки классов

$$Y = \{-1, 1\}$$

размеченные данные

$$X = (x, y)$$

алгоритм классификации

$$a(x, w) = \text{sign}(f(x, w))$$

дискриминантная функция

$$f(x, w)$$

вектор параметров

$$w$$

разделяющая поверхность

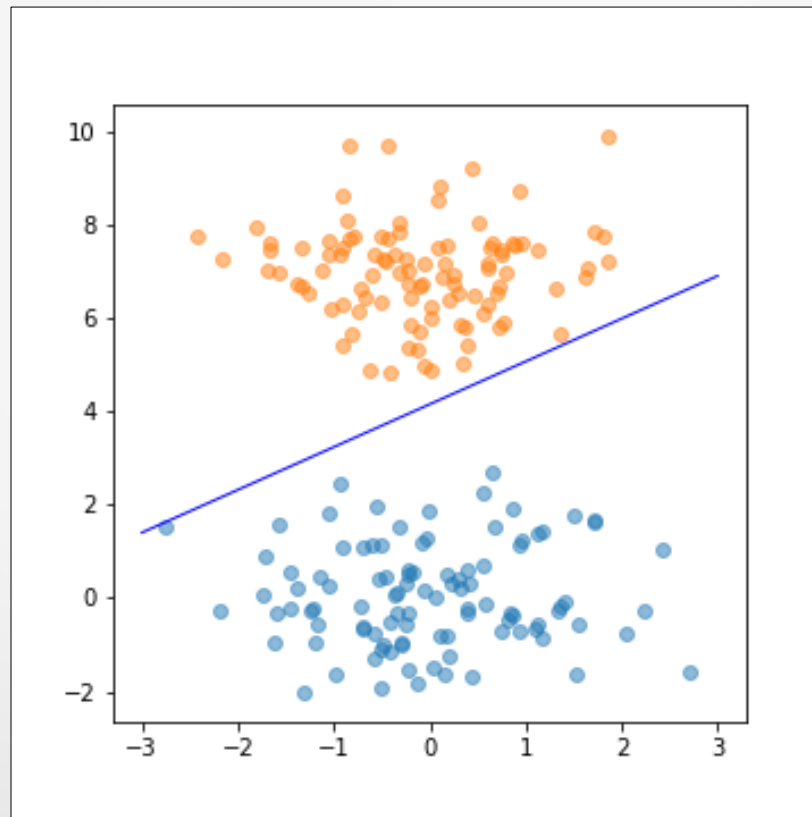
$$f(x, w) = 0$$

Линейные методы: разделяющая поверхность

пример: линейно разделимые данные

разделяющая поверхность - прямая

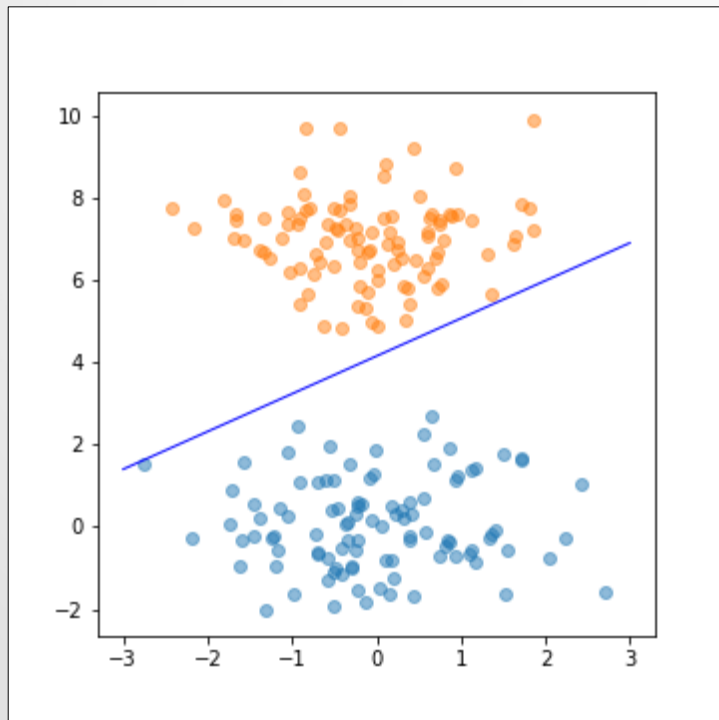
$$w_1 \cdot x + w_0 = 0$$



Линейные методы: отступы

отступ - насколько далеко объект x от разделяющей поверхности

$$M(x, w) = y \cdot f(x, w)$$



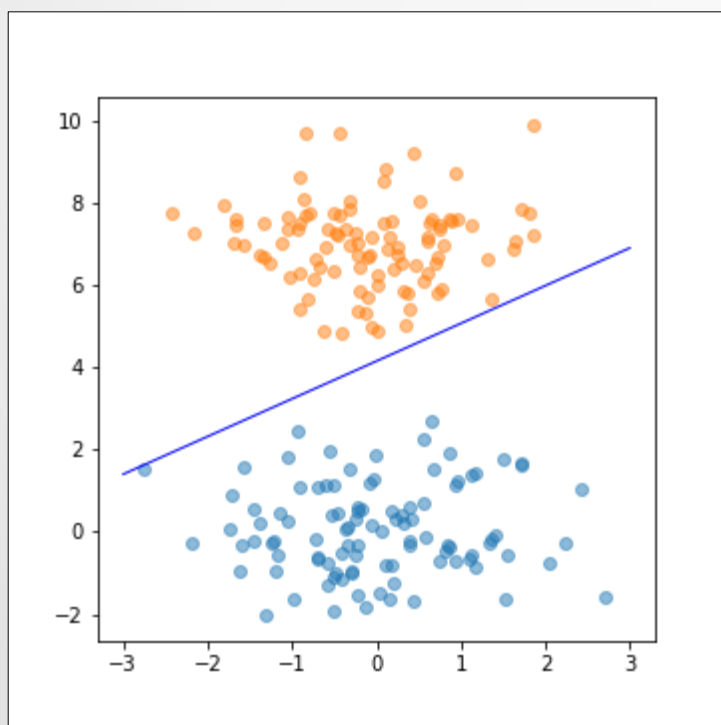
$y \in \{-1, 1\}$ - метка класса

$f(x, w)$ - дискриминантная функция

Линейные методы: отступы

отступ - насколько далеко объект x от разделяющей поверхности

$$M(x, w) = y \cdot f(x, w)$$



$y \in \{-1, 1\}$ - метка класса

$f(x, w)$ - дискриминантная функция

$M(x, w) < 0$ - алгоритм ошибается на x

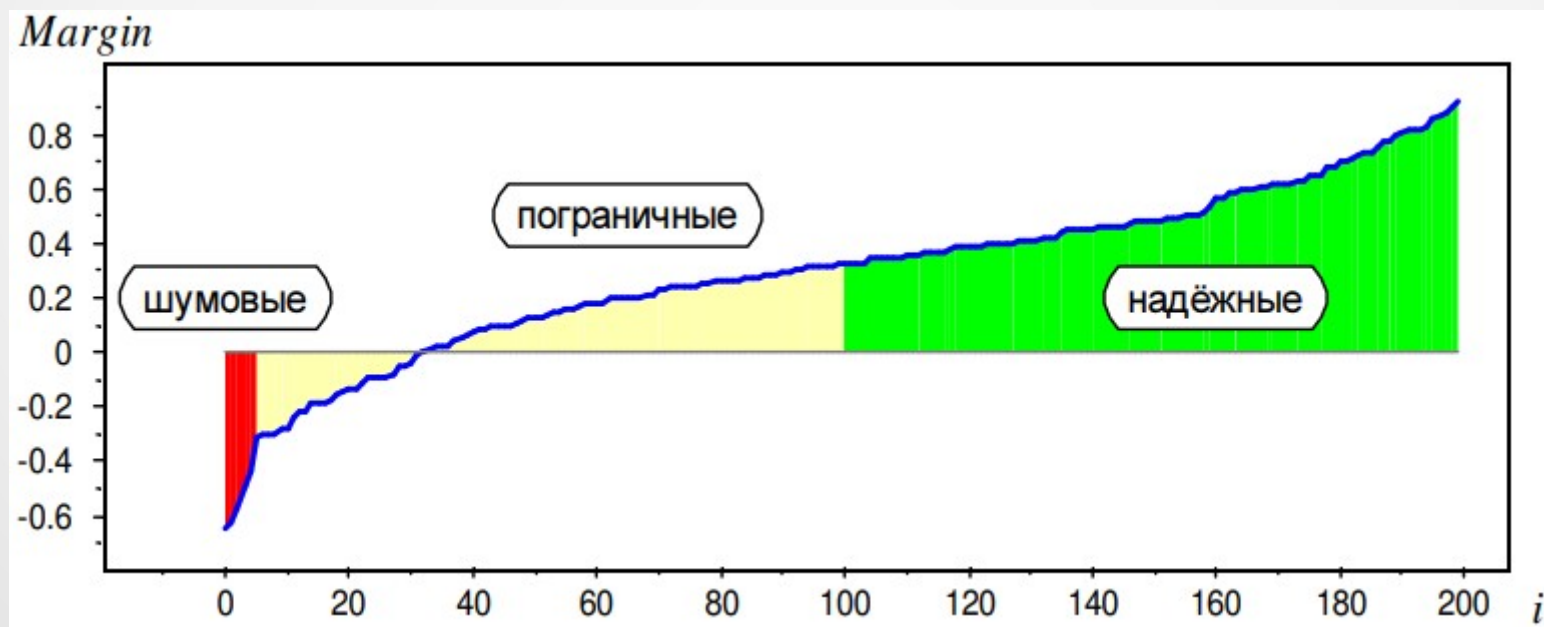
Линейные методы: отступы

отступ - насколько далеко объект от разделяющей поверхности

$$M(x, w) = y \cdot f(x, w)$$

$y \in \{-1, 1\}$ - метка класса

$f(x, w)$ - дискриминантная функция



$M(x, w) < 0$ - алгоритм ошибается на x

Линейные методы: эмпирический риск

функционал эмпирического риска, (число ошибок)

$$Q(x, w) = \sum_x [M(x, w) < 0]$$

$M(x, w) = y \cdot f(x, w)$ - отступ объекта x

$y \in \{-1, 1\}$ - метка класса

$f(x, w)$ - дискриминантная функция

$M(x, w) < 0$ - алгоритм ошибается на x

Линейные методы: функция потери

функционал эмпирического риска

$$Q(x, w) = \sum_x [M(x, w) < 0]$$

Линейные методы: функция потерь

функционал эмпирического риска

$$Q(x, w) = \sum_x [M(x, w) < 0]$$

[$M < 0$] это пороговая функция,
не учитываем значение отступа M ,
оптимизировать не удобно,
заменим её...

Линейные методы: функция потерь

функционал эмпирического риска

$$Q(x, w) = \sum_x [M(x, w) < 0]$$

[$M < 0$] это пороговая функция,
не учитываем значение отступа M ,
оптимизировать не удобно,
заменим её...

построим аппроксимацию Q

введём **функцию потерь $L(M)$**
(невозрастающая, неотрицательная)

$$\tilde{Q}(x, w) = \sum_x L(M(x, w)) \rightarrow \min$$

$$Q(x, w) \leq \tilde{Q}(x, w)$$

Линейные методы

функционал эмпирического риска

$$Q(x, w) = \sum_x [M(x, w) < 0]$$

$[M < 0]$ это пороговая функция, оптимизировать не удобно, заменим её...

варианты для замены $[M < 0]$

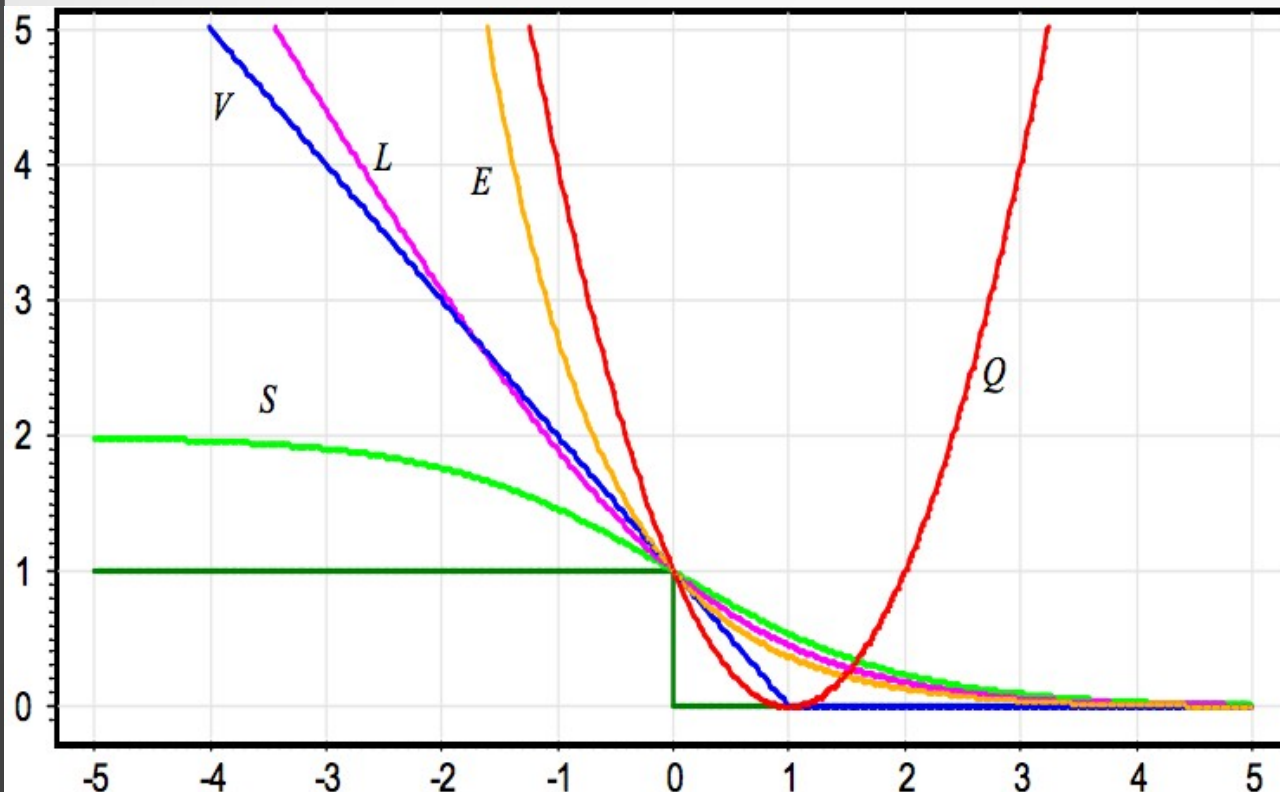
$$L(M) = \log_2 \left(1 + \frac{1}{\exp(M)} \right) \quad \text{логарифмическая}$$

$$V(M) = (1 - M)_+ \quad \text{кусочно-линейная}$$

$$Q(M) = (1 - M)^2 \quad \text{квадратичная}$$

$$E(M) = \frac{1}{\exp(M)} \quad \text{экспоненциальная}$$

$$S(M) = \frac{1}{2 \cdot (1 + \exp(M))} \quad \text{сигмоид}$$



Линейные методы: линейный классификатор

$$a(x, w) = \text{sign} \left(\sum_{i=1}^n x_i \cdot w_i - w_0 \right) = \text{sign}(\langle x, w \rangle)$$

$M(x, w) = \langle x, w \rangle \cdot y$ - отступ на объекте x класса y

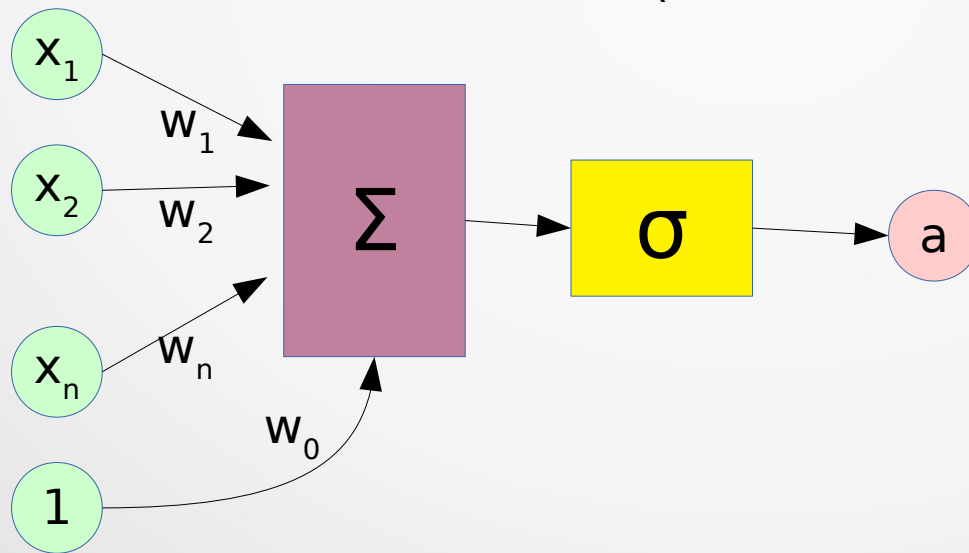
$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

Линейные методы: линейный классификатор

линейная модель МакКаллока-Питтса (1943) (формальный нейрон)

$$a(x, w) = \sigma \left(\sum_{i=1}^n x_i \cdot w_i - w_0 \right) = \sigma(\langle x, w \rangle)$$

σ - функция активации нейрона
(можно использовать **sign**)



Линейные методы

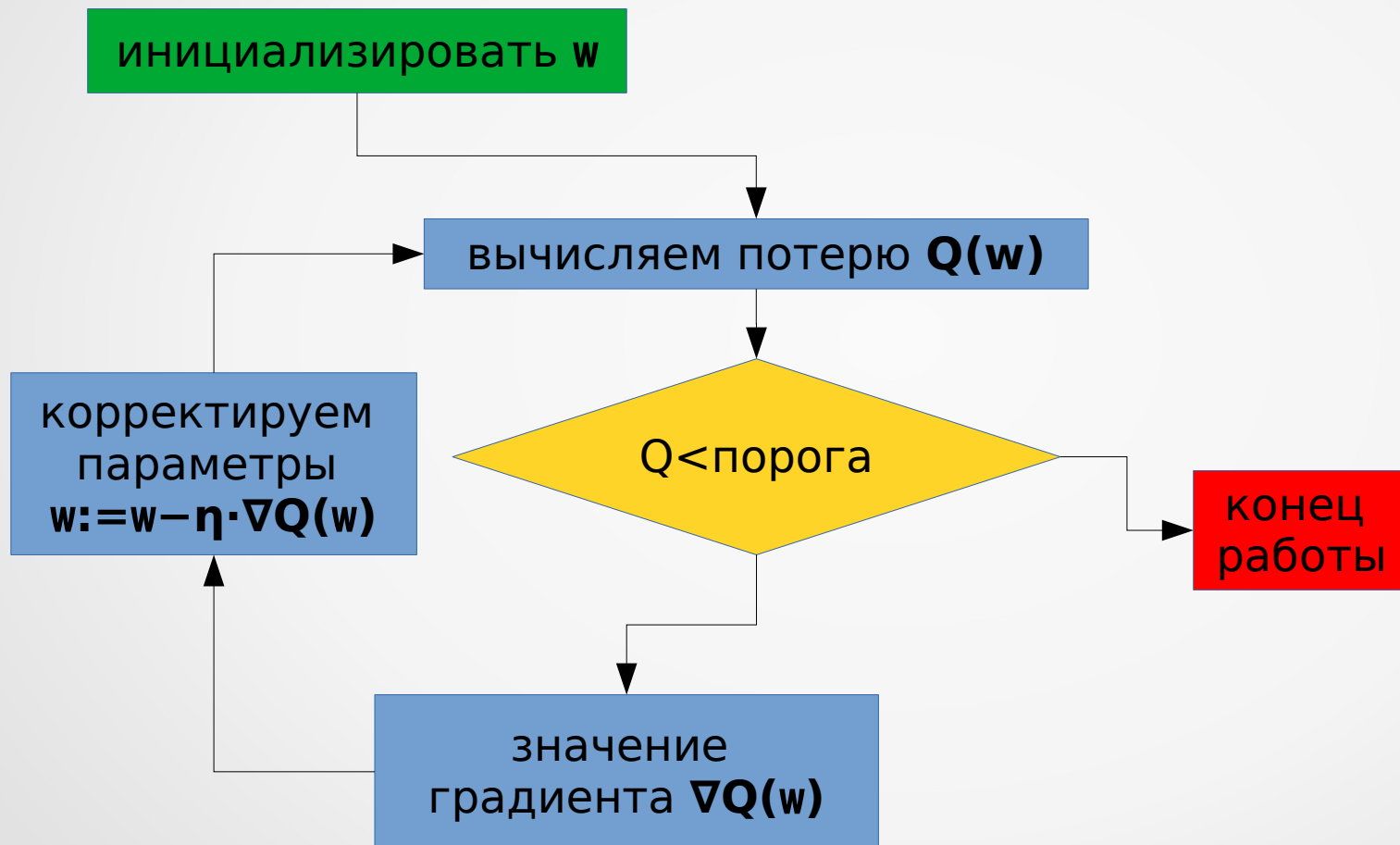
обучение классификатора как задача оптимизации

$$Q(w; X) = \sum_{x \in X} L(\langle x, w \rangle \cdot y) \rightarrow \min_w$$

можно использовать градиентные методы

$$\nabla Q(w) = \left(\frac{\partial Q(w)}{\partial w_j} \right)_{j=0}^n \text{ - вектор градиента ф-ции } Q$$

Линейные методы: градиентный спуск (GD)



Линейные методы: стохастический градиентный спуск (SGD)

инициализировать w

вычисляем суммарную
потерю $Q(w)$ на X

$Q < \text{порога}$

конец
работы

корректируем
суммарную потерю
 $Q := \lambda Q_j + (1 - \lambda)Q$

выбираем
случайный x_j

вычисляем значение
градиента $\nabla Q(w, x_j)$

вычисляем
потерю для объекта x_j
 $Q_j = Q(w, x_j)$

корректируем
параметры
 $w := w - \eta \cdot \nabla Q(w, x_j)$

Линейные методы

частный случай 1:

адаптивный линейный элемент ADALINE, Видроу, Хофф (1960)

задача регрессии

$$X = \mathbb{R}^n \quad y \in \mathbb{R} \quad a(x, w) = \langle x, w \rangle$$

Линейные методы

частный случай 1:

адаптивный линейный элемент ADALINE, Видроу, Хофф (1960)

задача регрессии

$$X = \mathbb{R}^n \quad y \in \mathbb{R}$$

$$a(x, w) = \langle x, w \rangle$$

$$L(a, y) = \frac{1}{2} \cdot (a(x, w) - y)^2$$

Линейные методы

частный случай 1:

адаптивный линейный элемент ADALINE, Видроу, Хофф (1960)

задача регрессии

$$X = \mathbb{R}^n \quad y \in \mathbb{R} \quad a(x, w) = \langle x, w \rangle \quad L(a, y) = \frac{1}{2} \cdot (a(x, w) - y)^2$$

градиентный шаг - **дельта правило**

$$w := w - \eta (\langle x, w \rangle - y) \cdot x$$

$(\langle x, w \rangle - y)$ - ошибка на объекте x

Линейные методы

частный случай 2:
правило обучения Хебба (1949)

задача классификации

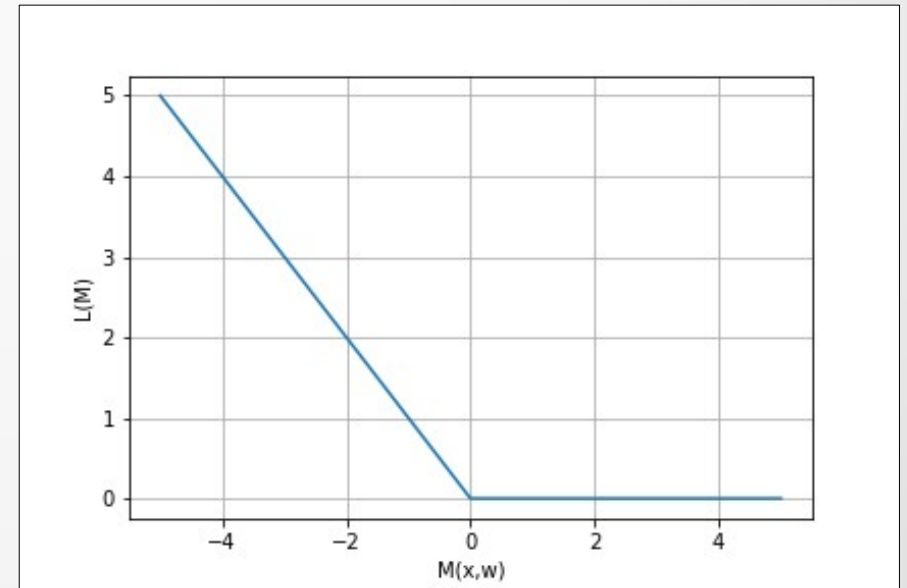
$$x \in \mathbb{R}^n \quad y \in \{-1, 1\} \quad a(x, w) = \text{sign}(\langle x, w \rangle)$$

Линейные методы

частный случай 2:
правило обучения Хебба (1949)

задача классификации

$$x \in \mathbb{R}^n \quad y \in \{-1, 1\} \quad a(x, w) = \text{sign}(\langle x, w \rangle) \quad L(a, y) = (-\langle x, w \rangle \cdot y)_+$$



Линейные методы

частный случай 2:
правило обучения Хебба (1949)

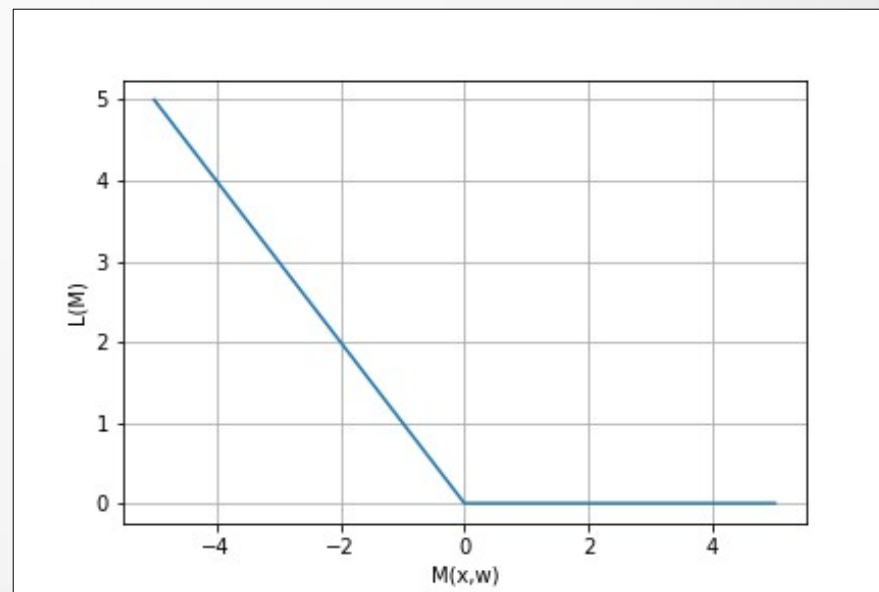
задача классификации

$$x \in \mathbb{R}^n \quad y \in \{-1, 1\} \quad a(x, w) = \text{sign}(\langle x, w \rangle) \quad L(a, y) = (-\langle x, w \rangle \cdot y)_+$$

градиентный шаг

$$[\langle x, w \rangle \cdot y < 0] \Rightarrow w := w + \eta \cdot y \cdot x$$

параметры корректируем
только в случае ошибки



Линейные методы

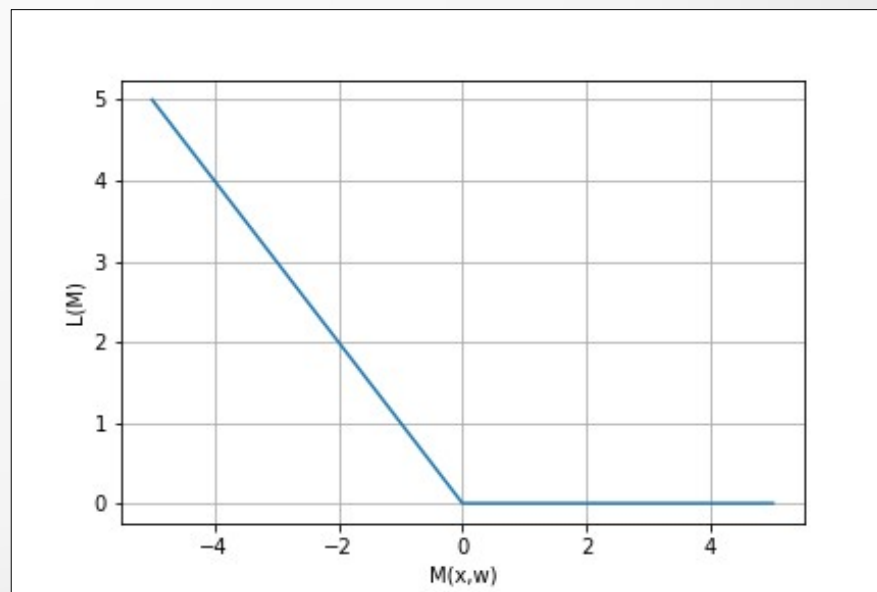
частный случай 3:
правило обучения Розенблатта (1957)

задача классификации

$$x \in \{0,1\}^n \quad y \in \{0,1\} \quad a(x, w) = \text{sign}(\langle x, w \rangle) \quad L(a, y) = (-\langle x, w \rangle \cdot y)_+$$

градиентный шаг

$$w := w - \eta \cdot (a(x, w) - y) \cdot x$$



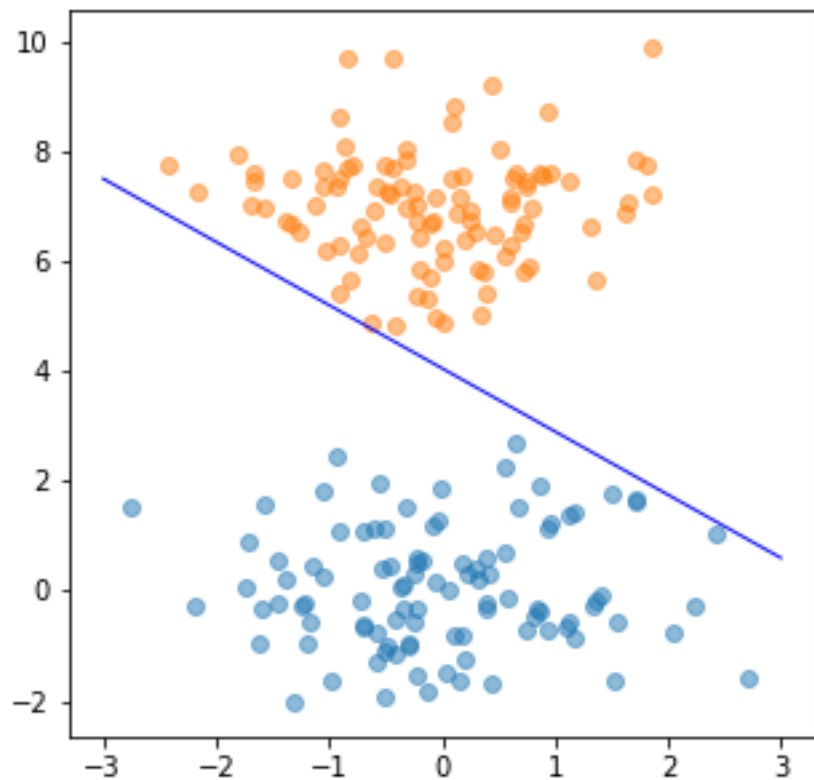
Линейные методы

«зоопарк» методов

- вид разделяющей поверхности $f(\mathbf{x}, \mathbf{w})$
(линейная, нелинейная)
- вид функции потерь $L(\mathbf{M})$
- вид метода оптимизации $Q(\mathbf{w}) \rightarrow \min$

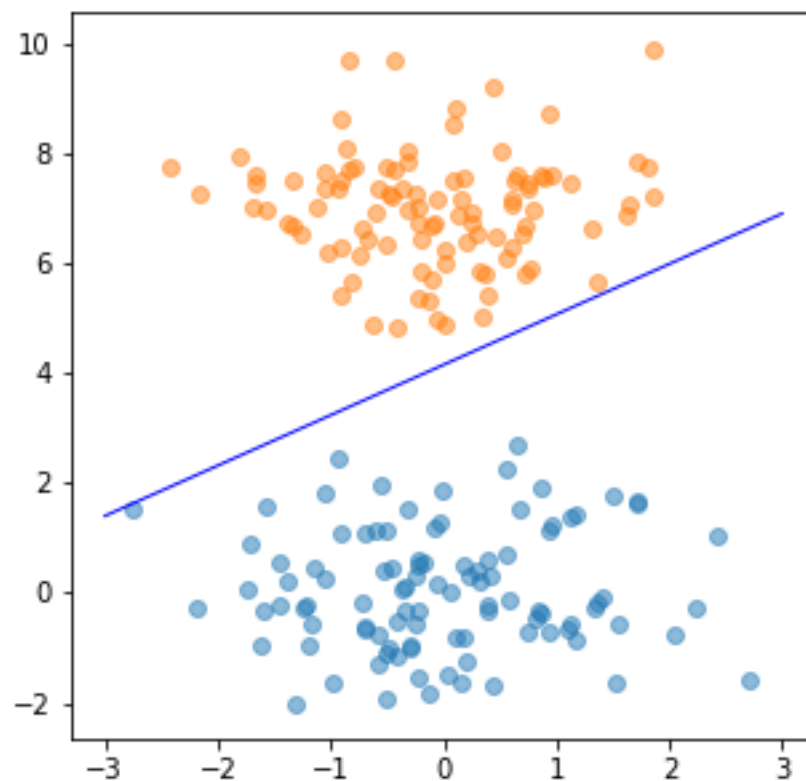
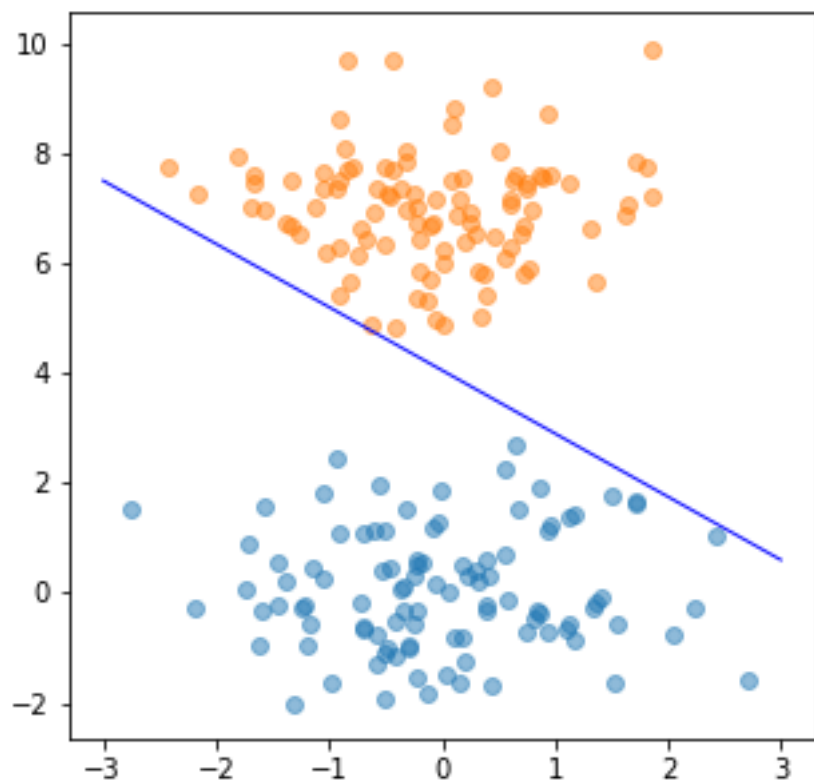
Линейные методы: SVM

рассмотрим линейно разделимый набор



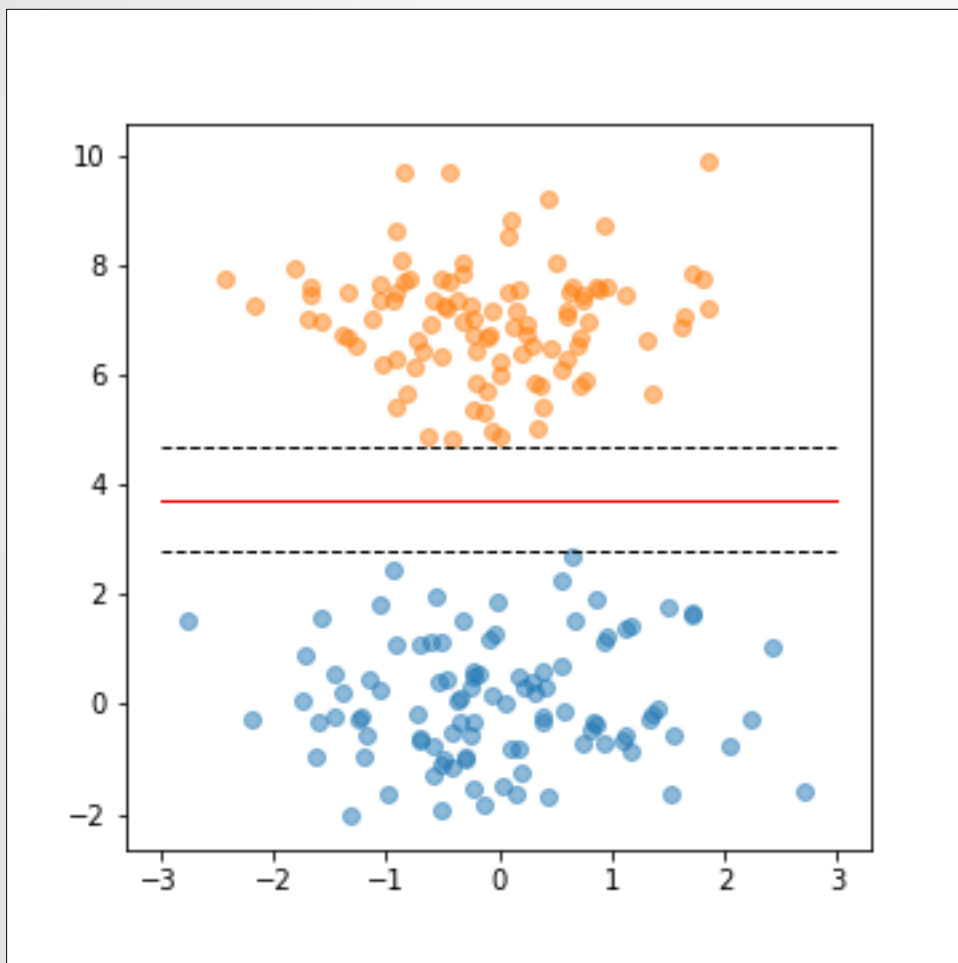
Линейные методы: SVM

рассмотрим линейно разделимый набор
много разделяющих гиперплоскостей



Линейные методы: SVM

разделительная полоса



цель: увеличить отступы,
получить полосу максимальной ширины

Линейные методы: SVM

модель: машина опорных векторов (SVM)

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

задача классификации

$$x \in \mathbb{R}^n \quad y \in \{-1, +1\}$$

Линейные методы: SVM

модель: машина опорных векторов (SVM)

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

задача классификации

$$x \in \mathbb{R}^n \quad y \in \{-1, +1\}$$

обучение классификатора
это задача оптимизации
функционала эмпирического риска

$$\sum_i [a(x_i; w, w_0) \neq y_i] = \sum_x [M(x, w, w_0) < 0] \rightarrow \min_{w, w_0}$$

отступ на объекте **x** класса **y**

$$M(x, w, w_0) = (\langle x, w \rangle - w_0) \cdot y$$

Линейные методы: SVM

модель: машина опорных векторов (SVM)

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

задача классификации

$$x \in \mathbb{R}^n \quad y \in \{-1, +1\}$$

обучение классификатора
это задача оптимизации
функционала эмпирического риска

$$\sum_i [a(x_i; w, w_0) \neq y_i] = \sum_x [M(x, w, w_0) < 0] \rightarrow \min_{w, w_0}$$

отступ на объекте **x** класса **y**

$$M(x, w, w_0) = (\langle x, w \rangle - w_0) \cdot y$$

замена пороговой ф-ции потери на кусочно линейную

$$\sum_x [M(x, w, w_0) < 0] \leq \sum_x (1 - M(x, w, w_0))_+ \rightarrow \min_{w, w_0}$$

Линейные методы: SVM

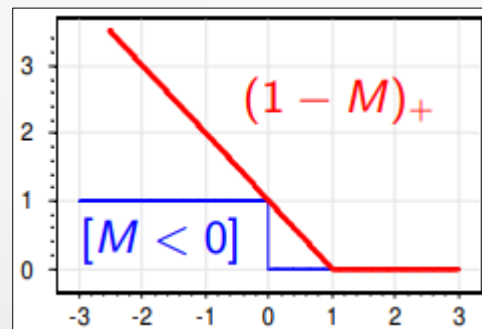
замена пороговой ф-ции потери на кусочно линейную

$$\sum_x \left(1 - M(x, w, w_0)\right)_+ + \frac{1}{2C} \|w\|^2 \rightarrow \min_{w, w_0}$$

аппроксимация штрафует за приближение к границе классов

регуляризация штрафует за неустойчивые решения

увеличиваем разделительную полосу (зазор между классами)



Линейные методы: SVM

$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) \cdot y_i$ - отступ на объекте x_i

для линейно разделимого набора все отступы $M > 0$

Линейные методы: SVM

$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) \cdot y_i$ - отступ на объекте x_i

для линейно разделимого набора все отступы $M > 0$

введём нормировку отступов

$$\min_i (M_i(w, w_0)) = 1$$

Линейные методы: SVM

$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) \cdot y_i$ - отступ на объекте x_i

для линейно разделимого набора все отступы $M > 0$

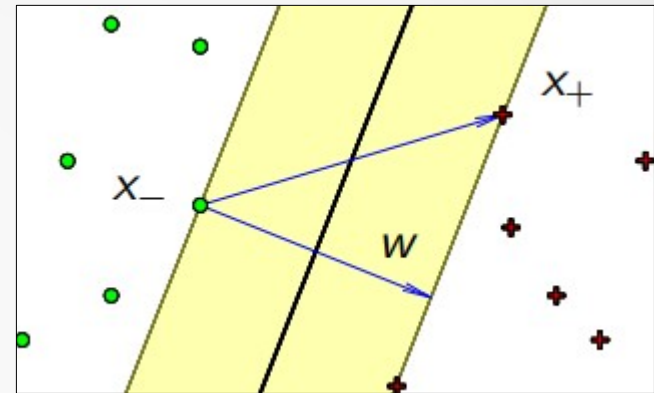
введём нормировку отступов

$$\min_i (M_i(w, w_0)) = 1$$

крайние точки классов,
ограничивающие разделяющую полосу

$$\exists x_+ : \langle w, x_+ \rangle - w_0 = +1$$

$$\exists x_- : \langle w, x_- \rangle - w_0 = -1$$



Линейные методы: SVM

$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) \cdot y_i$ - отступ на объекте x_i

для линейно разделимого набора все отступы $M > 0$

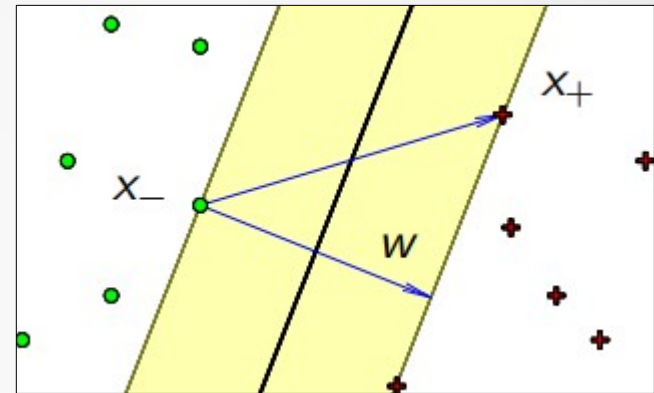
введём нормировку отступов

$$\min_i (M_i(w, w_0)) = 1$$

крайние точки классов,
ограничивающие разделяющую полосу

$$\exists x_+ : \langle w, x_+ \rangle - w_0 = +1$$

$$\exists x_- : \langle w, x_- \rangle - w_0 = -1$$



разделяющая полоса

$$\{x : -1 \leq (\langle x_i, w \rangle - w_0) \leq 1\}$$

Линейные методы: SVM

$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) \cdot y_i$ - отступ на объекте x_i

для линейно разделимого набора все отступы $M > 0$

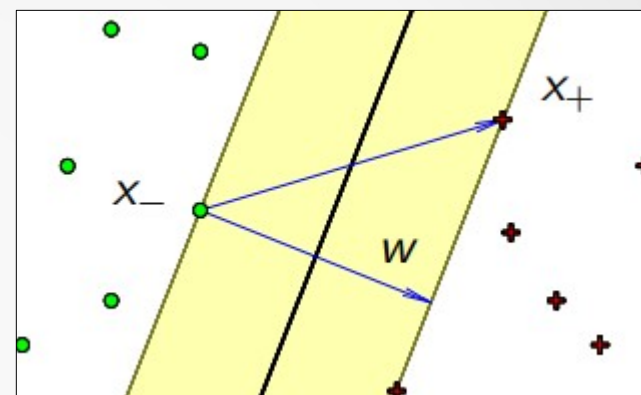
введём нормировку отступов

$$\min_i (M_i(w, w_0)) = 1$$

крайние точки классов,
ограничивающие разделяющую полосу

$$\exists x_+ : \langle w, x_+ \rangle - w_0 = +1$$

$$\exists x_- : \langle w, x_- \rangle - w_0 = -1$$



разделяющая полоса

$$\{x : -1 \leq (\langle x_i, w \rangle - w_0) \leq 1\}$$

ширина разделяющей полосы

$$\frac{\langle x_+, w \rangle - \langle x_-, w \rangle}{\|w\|} = \frac{1 + w_0 - (-1 + w_0)}{\|w\|} = \frac{2}{\|w\|} \rightarrow \max$$

добавка
регуляризации

$$\|w\| \rightarrow \min$$

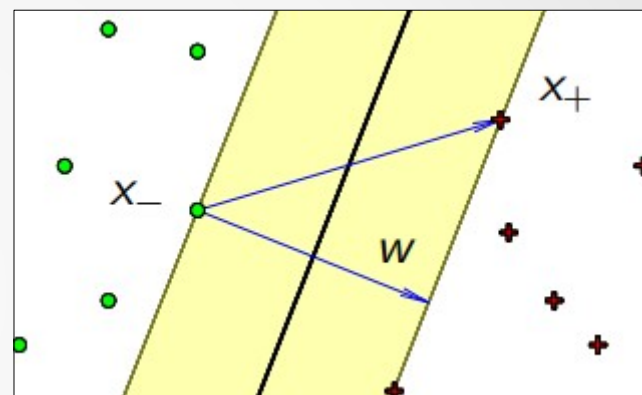
Линейные методы: SVM

постановка задачи

для линейно разделимого набора

$$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) \cdot y_i$$

$$\begin{cases} \|w\|^2 \rightarrow \min_{w, w_0} \\ M_i(w, w_0) \geq 1 \end{cases}$$



Линейные методы: SVM

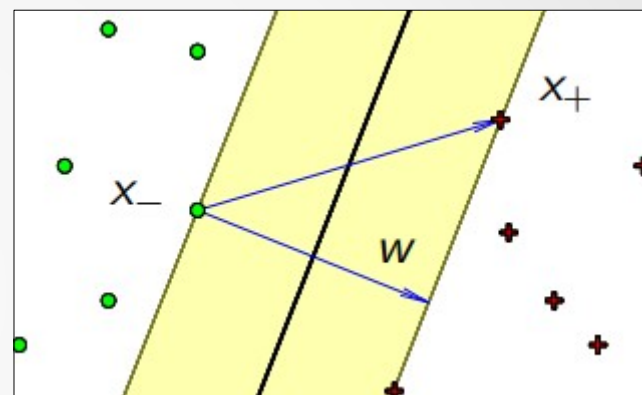
постановка задачи

для линейно разделимого набора

$$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) \cdot y_i$$

$$\begin{cases} \|w\|^2 \rightarrow \min_{w, w_0} \\ M_i(w, w_0) \geq 1 \end{cases}$$

для линейно НЕразделимого набора
система неравенств несовместна
решения нет



Линейные методы: SVM

эвристика для линейно НЕразделимого набора

ослабим ограничения

$$\left\{ \begin{array}{l} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \rightarrow \min_{w, w_0, \xi} \\ M_i(w, w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{array} \right.$$

$$\left\{ \begin{array}{l} \xi_i \geq 1 - M_i \Rightarrow \xi_i = 1 - M_i \\ \xi_i \geq 0 \end{array} \right.$$

эквивалентная задача оптимизации

$$C \cdot \sum_i \left(1 - M_i(x, x_0) \right)_+ + \frac{1}{2} \|w\|^2 \rightarrow \min_{w, w_0}$$

задача выпуклой квадратичной оптимизации

Линейные методы: SVM

решение задачи выпуклой квадратичной оптимизации

применение условий Каруша-Куна-Таккера

выписываем ф-цию Лагранжа и ищем её седловую точку
(приравниваем к нулю производную)

Функция Лагранжа: $\mathcal{L}(w, w_0, \xi; \lambda, \eta) =$

$$= \frac{1}{2} \|w\|^2 - \sum_{i=1}^{\ell} \lambda_i (M_i(w, w_0) - 1) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

Линейные методы: SVM

разбираем объекты x_i на три типа

1. $\lambda_i = 0; \eta_i = C; \xi_i = 0; M_i \geq 1$.
— периферийные (неинформативные) объекты.
2. $0 < \lambda_i < C; 0 < \eta_i < C; \xi_i = 0; M_i = 1$.
— **опорные** граничные объекты.
3. $\lambda_i = C; \eta_i = 0; \xi_i > 0; M_i < 1$.
— **опорные**-нарушители.

опорным назовём объект x_i , для которого $\lambda_i \neq 0$

$$a(x) = \text{sign} \left(\sum_i \lambda_i y_i \langle x_i, x \rangle - w_0 \right)$$

Линейные методы: SVM

для решения задачи оптимизации и нахождения опорных объектов применяется алгоритм SMO (sequential minimal optimization)

после нахождения опорных объектов

классификатор приобретает следующий вид

$$a(x) = \text{sign} \left(\sum_i \lambda_i y_i \langle x_i, x \rangle - w_0 \right) \quad \begin{cases} w = \sum_{i=1}^{\ell} \lambda_i y_i x_i; \\ w_0 = \langle w, x_i \rangle - y_i, \end{cases}$$

Линейные методы: SVM

нелинейное обобщение - kernel trick

вместо скалярного произведения

будем использовать функцию-ядро

$$a(x) = \text{sign} \left(\sum_i \lambda_i y_i K(x_i, x) - w_0 \right)$$

функция K - ядро

если для него существует отображение,
удовлетворяющее условиям скалярного произведения

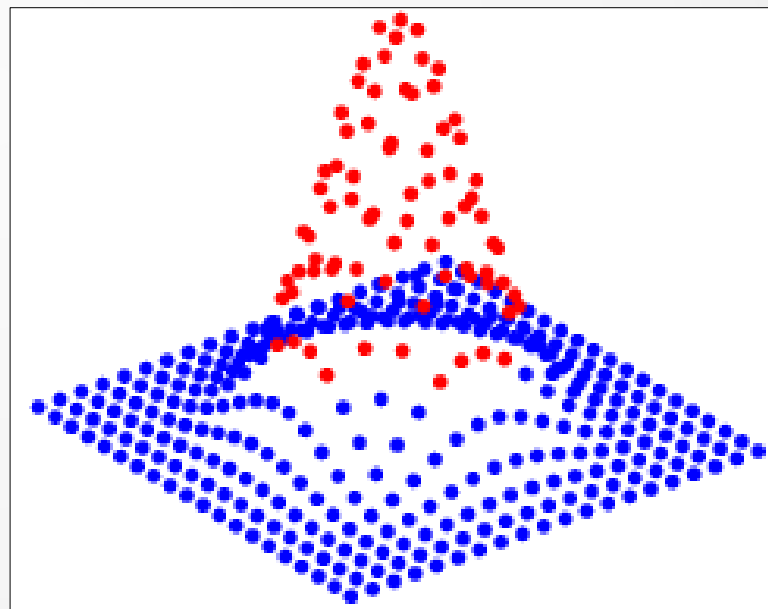
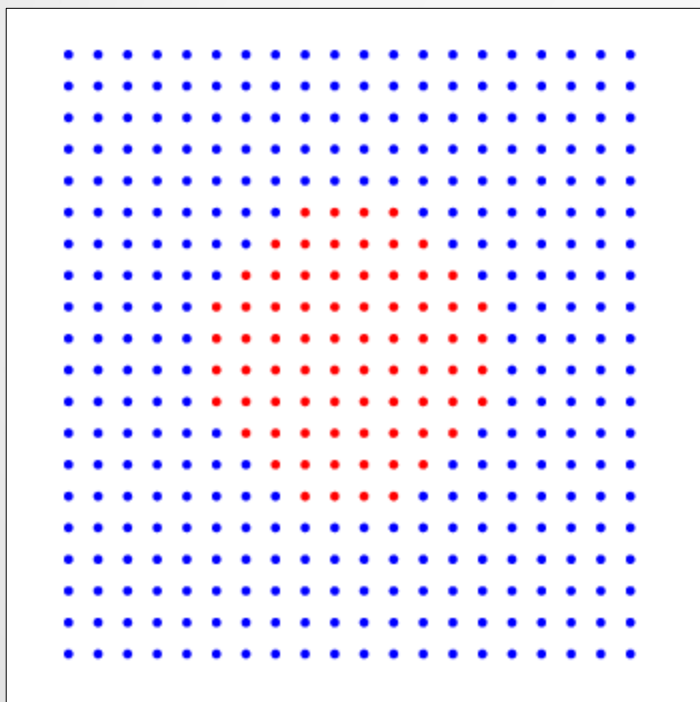
$$\exists \psi: K(x, x') = \langle \psi(x), \psi(x') \rangle$$

функция K симметрична и неотрицательно определена

Линейные методы: SVM

kernel trick

повышаем размерность пространства
линейно неразделимая задача
превращается в линейно разделимую



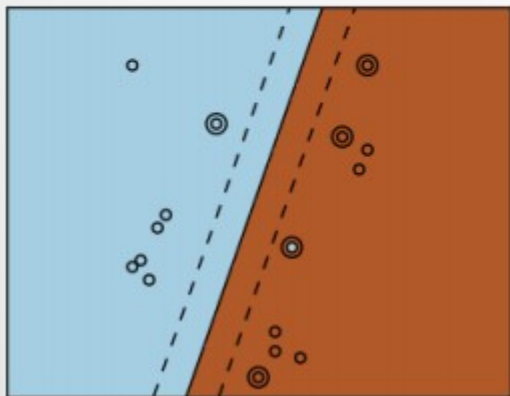
Линейные методы: SVM

kernel trick

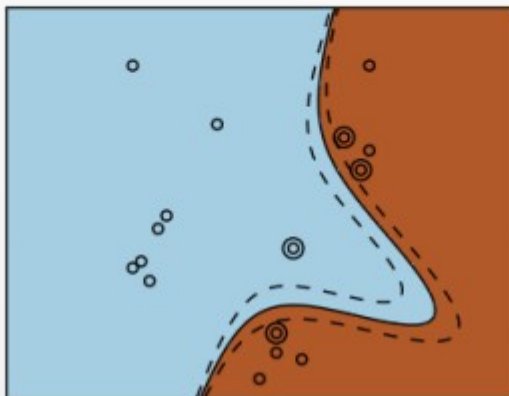
повышаем размерность пространства
линейно неразделимая задача
превращается в линейно разделимую

Примеры с различными ядрами $K(x, x')$

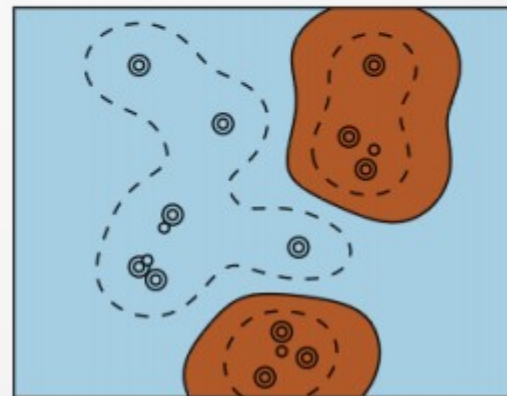
линейное
 $\langle x, x' \rangle$



полиномиальное
 $(\langle x, x' \rangle + 1)^d, \quad d=3$



гауссовское (RBF)
 $\exp(-\gamma \|x - x'\|^2)$



Линейные методы: итог

- линейные методы строят разделяющие поверхности в пространстве признаков
- использования нелинейных поверхностей позволяет разделять линейно неразделимые наборы
- аппроксимация пороговой ф-ции потерь позволяет использовать градиентные методы оптимизации
- метод стохастического градиента SGD хорошо подходит для обучения на больших данных
- метод обучения SVM как задача выпуклой квадратичной оптимизации имеет единственное решение
- для обучения SVM применяется алгоритм SMO (sequential minimal optimization)
- применение ядер позволяет SVM разделять линейно неразделимые наборы, общих подходов для выбора ядер нет

Линейные методы: литература

git clone https://github.com/mechanoid5/ml_lectorium.git

- Борисов Е.С. Классификатор на основе машины опорных векторов.
<http://mechanoid.kiev.ua/ml-svm.html>
- К.В. Воронцов Линейные методы классификации: метод стохастического градиента.
- К.В. Воронцов Линейные методы классификации: метод опорных векторов.

Линейные методы



Вопросы ?

Линейные методы: практика

источники данных для экспериментов



`sklearn.datasets`

UCI Repository

kaggle

