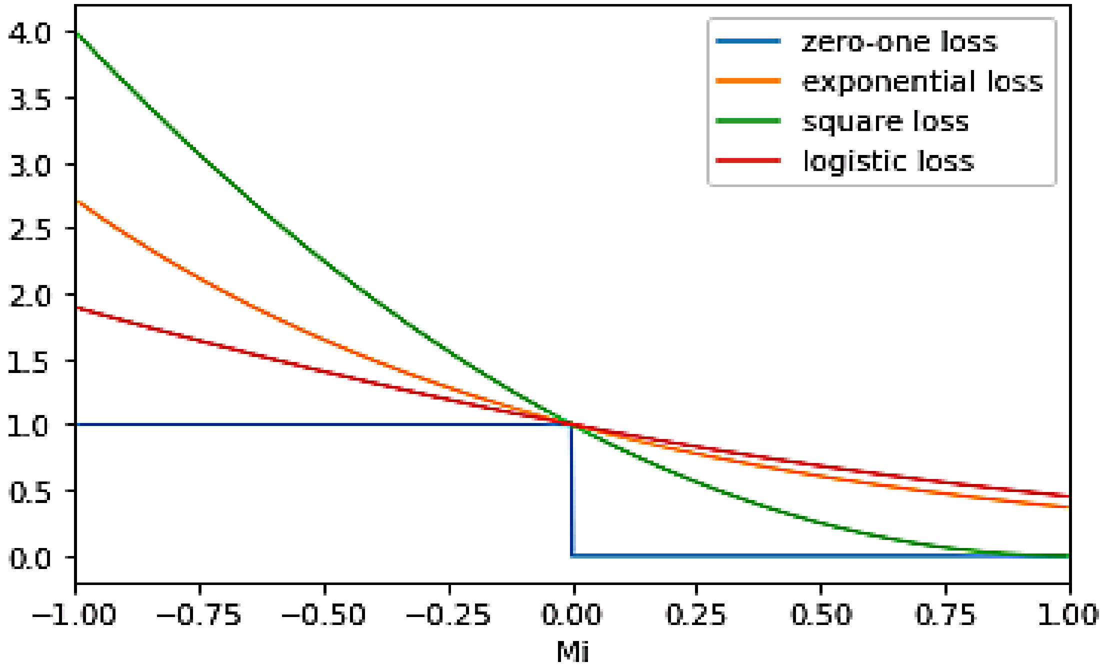
Машины опорных векторов

Корлякова М.О.

2022

Рабочие Потери



$$[M_i < 0] \leq \tilde{L}(M_i).$$

$$Q(a, X) \leq \tilde{Q}(a, X) = \frac{1}{l} \sum_{i=1}^{l} \tilde{L}(M_i) \to \min_{\omega}$$

— логарифмическая $\mathscr{L}(M) = \log(1 + e^{-M})$

— кусочно-линейная $\mathscr{L}(M) = (1 - M)_+$.

— экспоненциальная $\mathscr{L}(M) = e^{-M}$

Отделимость

1) Непустые выпуклые множества A и B называются (собственно) отделимыми, если существует аффинная гиперплоскость

H = $\{x \in \mathrm{R}^n \mid v^T x = b\}$, $v \in \mathrm{R}^n$, $v \neq 0$,

такая, что множества A, B лежат в противоположных по отношению к H замкнутых полупространствах и, по крайней мере, одно из множеств A, B не содержится в H. В этом случае гиперплоскость H отделяет множества A, B.

Отделимость

2) Непустые выпуклые множества A и B называются строго отделимыми, если существуют две различные параллельные гиперплоскости

$H1 = \{ x \in R^n | v^T x = b1\}$, $H2 = \{x \in R^n \mid v^T x = b2\}$, $v \in R^n$,

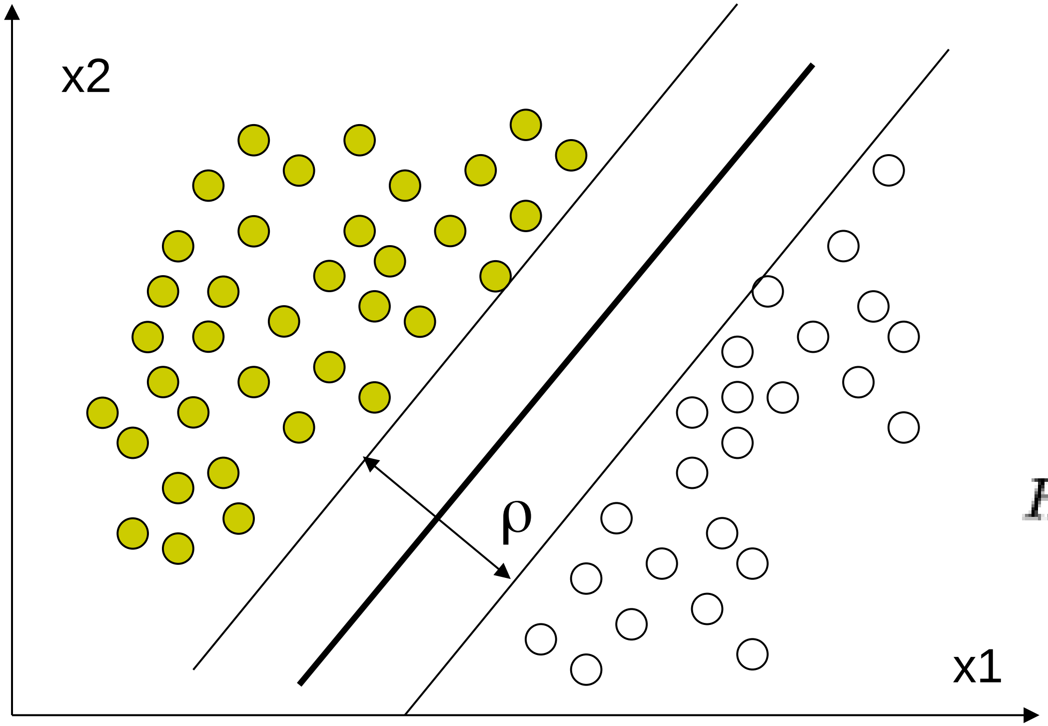
такие, что

inf $x \in A v^T x > b1 > b2 >$ sup $x \in B v^T x$.

Отделимость

Теорема 1.
Пусть A и B — непустые выпуклые множества в R^n, $A \cap B = \varnothing$. Тогда множества A и B отделимы.

Линейно-разделимые модели



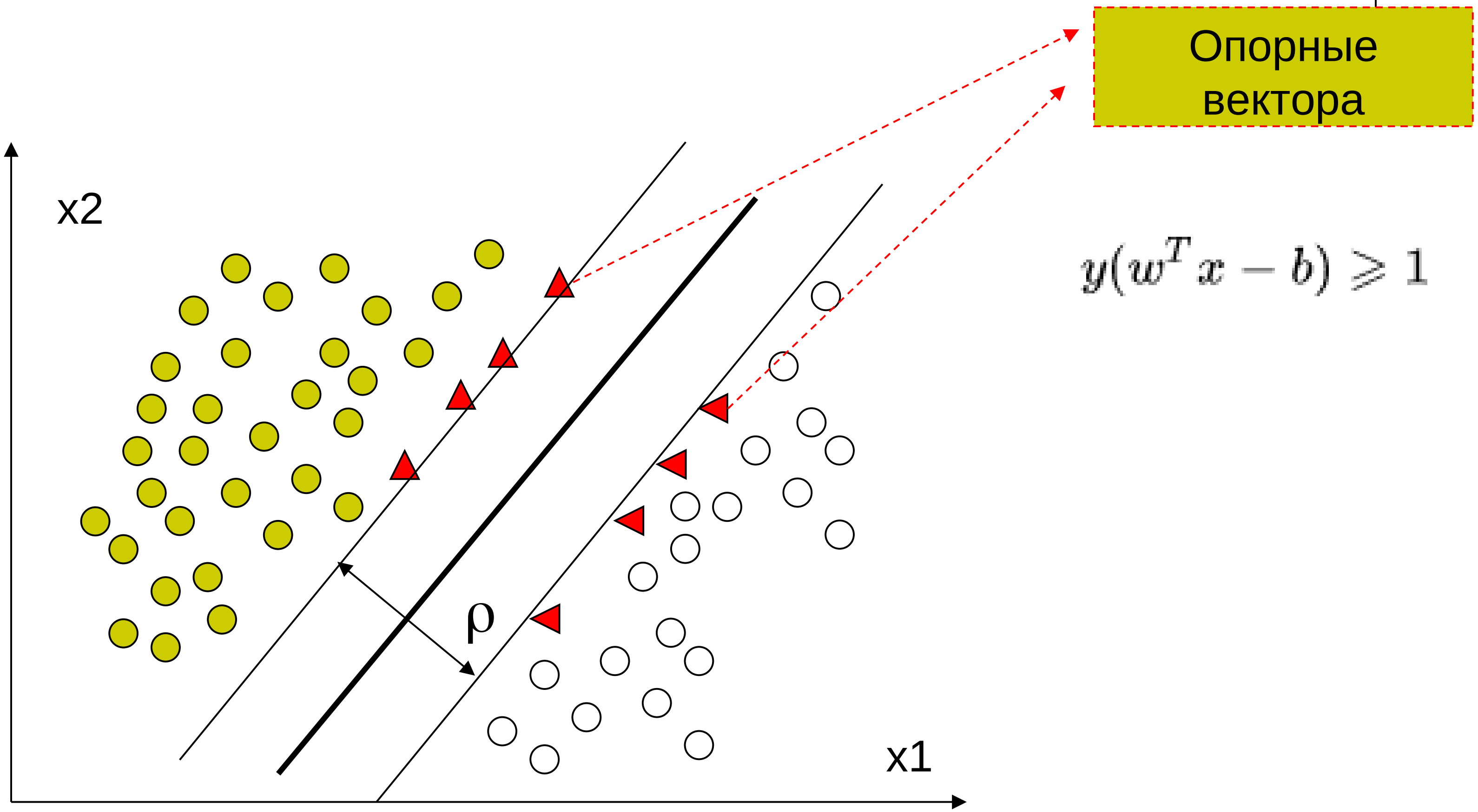
$$M = y(w^T x - b)$$

Hinge_loss

$$H = max(0, 1 - y(w^T x))$$

Вапник В.Н. – Support Vector Machine - SVM

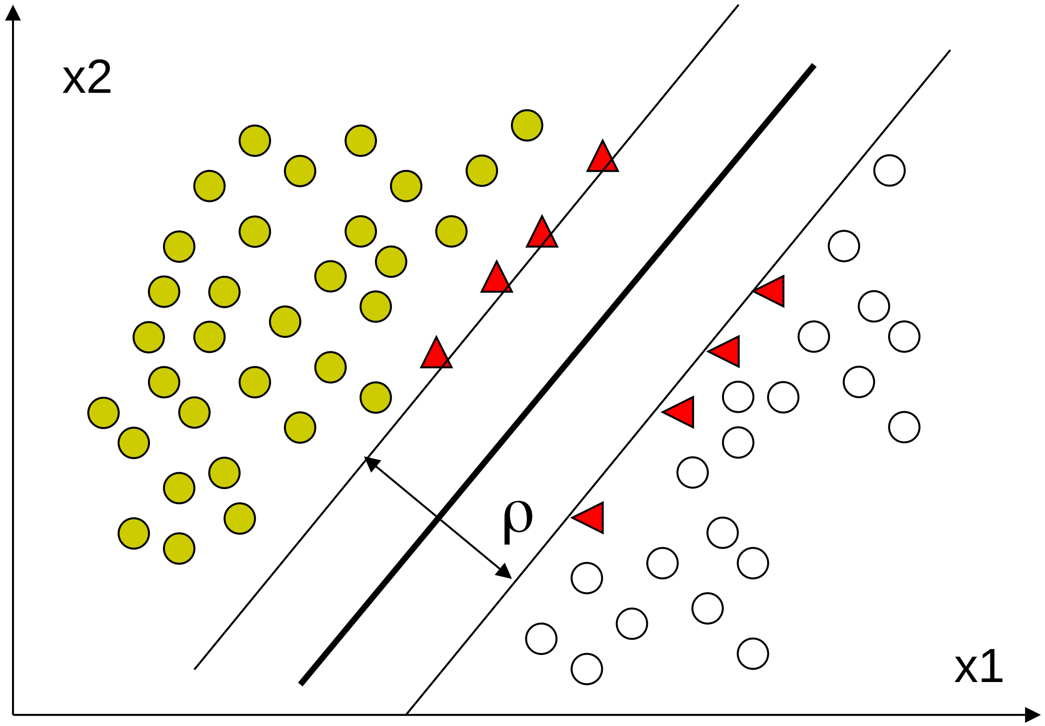
Опорные вектора



$$y(w^T x - b) \geqslant 1$$

C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998

Опорные вектора и поля



$$2/\|w\| \to max$$

$$\|w\| \to min$$

$$(w^T w)/2 \to min$$

Опорные вектора и поля

$$\langle (x_+ - x_-), w/\|w\| \rangle = (\langle x_+, w\rangle - \langle x_-, w\rangle)/\|w\| =$$

$$= ((b+1) - (b-1))/\|w\| = 2/\|w\|$$

- Для опорных векторов

  $w\, x_i + b = \pm 1$
- Расстояние от точки до гиперплоскости

  $|\, x_i w + b \,| \,/\, \|w\|$
- Отступ – поля

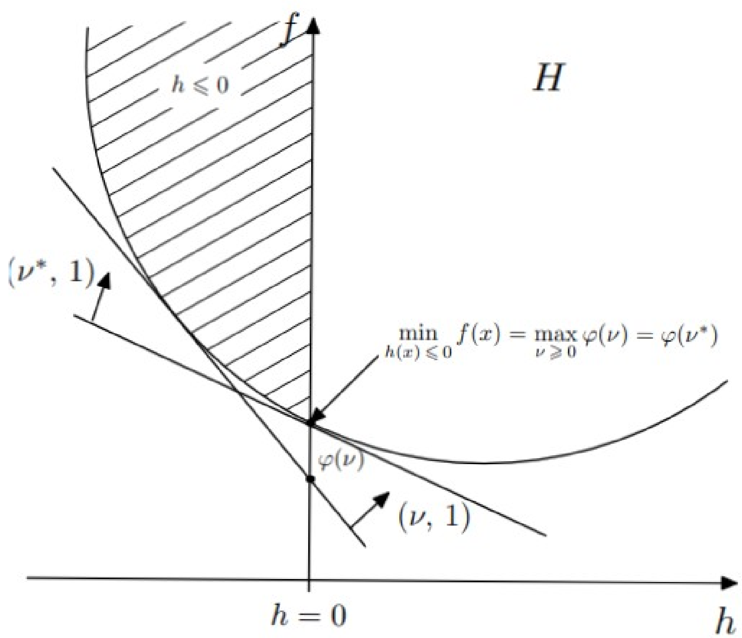
  $\rho = 2 \,/\, \|w\|$

- **Максимизация полей = минимизация $\|w\|$**

Опорные вектора и поля

Максимизация полей = минимизация ||w||

- Задачей выпуклого программирования называется задача минимизации выпуклой функции на выпуклом множестве



$f : X \to \mathrm{R}^n$ — выпуклая функция с областью определения dom $f = X$.

Выпуклой задачей без ограничений называется задача

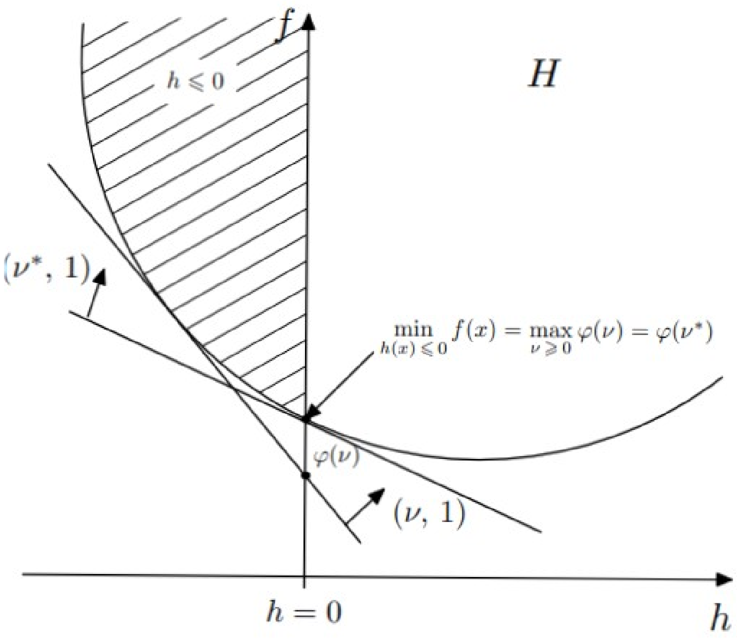
$$\min_{x \in X} f(x).$$

В выпуклой задаче локальный минимум всегда будет и глобальным

Опорные вектора и поля

Максимизация полей = минимизация $||w||$

- Задачей выпуклого программирования называется задача минимизации выпуклой функции на выпуклом множестве



$f : X \to R^n$ — выпуклая функция с областью определения dom $f = X$.

Выпуклой задачей с ограничениями называется задача

$$\min_{x \in X} f(x).$$

hi , $i = 1, \ldots, p$ — достаточно гладкие выпуклые функции

Опорные вектора и поля

Максимизация полей = минимизация $||w||$

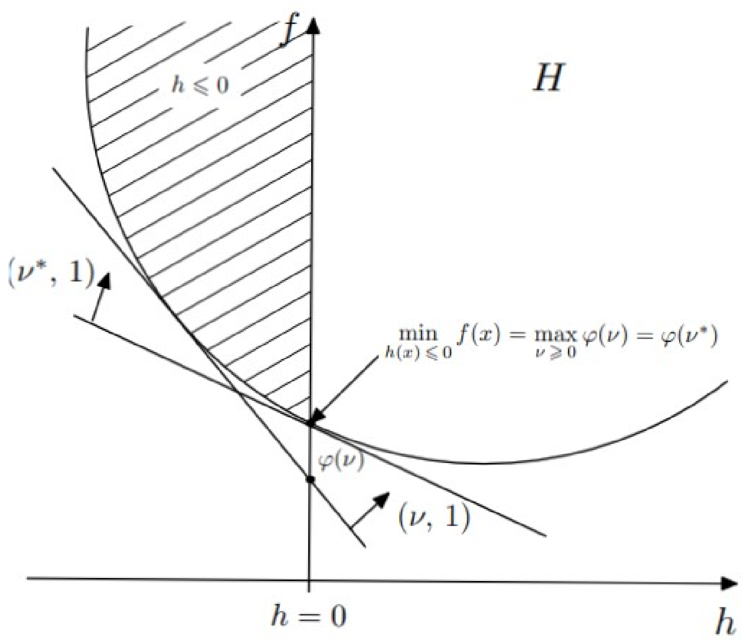
- Задачей выпуклого программирования называется задача минимизации выпуклой функции на выпуклом множестве

$f : X \to R^n$ — выпуклая функция с областью определения dom $f = X$.

Двойственная функция Лагранжа для задачи определяется как

$$\varphi(\nu) = \inf \{x \in Q,\ L(x, \nu)\}$$

$$= \inf\{x \in Q,\ (f(x) + \sum \alpha_i h_i(x)),\ i = 1,p\}$$



Опорные вектора и поля

Теорема Куно -Такера

Пусть выполняются условия задачи выпуклости.

- Тогда если x^* — решение задачи , то существует ненулевой вектор множителей Лагранжа $\alpha = (\alpha_i)$ i = 1,p, такой, что для функции Лагранжа

$$L(x, \nu) = f(x)\,\alpha_0 + \sum \alpha_i\, h_i(x) \quad i = 1,p$$

Выполняются:

1) принцип минимума для функции Лагранжа

$\min\limits_{x \in Q} L(x, \nu) = L(x^*, \alpha);$

2) условия дополняющей нежёсткости

$\alpha_i h_i(x^*) = 0,\ i = 1, \ldots, p;$

3) условия неотрицательности

$\alpha_i > 0,\ i = 0, \ldots,$

Если $\alpha_0 \neq 0$, то условия 1) – 3) предыдущего утверждения достаточны для того, чтобы допустимая точка x^* была решением задачи

Опорные вектора и поля

Линейное программирование :

Задача минимизации линейной формы при линейных ограничениях

Джордж Стиглер, возникла в американской армии в 30–40-е годы XX века.
Задача о диете:
выбрать суточный набор продуктов питания минимальной стоимости, который будут обеспечивать необходимую норму питательных веществ.

В 1940 г. Л. В. Канторович опубликовал статью «Об одном эффективном методе решения некоторых классов экстремальных проблем»

Д. Данциг (1914–2005) в 1947 г. разработал первый вычислительный алгоритм для решения задач ЛП — симплекс-метод

Опорные вектора и поля

Задача квадратичного программирования:

Задача минимизации квадратичной формы при линейных ограничениях

$$x^T C x + d^T x \to \min$$
$$Ax >= b$$

Функция Лагранжа:

$$L(x, v) = x^T Cx + d^T x + \alpha^T (Ax - b).$$

Двойственная функция:

$$\varphi(v) = \inf_x L(x, v).$$

Опорные вектора и поля

Задача квадратичного программирования:

Необходимым и достаточным условием экстремума при $C > 0$ в точке x^*, $Ax^* <= b$ является существование

$$\alpha^* \in R^m,\ \alpha^* > 0 \text{ таких, что}$$

$$2Cx^* + d + A^T\alpha^* = 0,\ <h\alpha^*, Ax^* - b_i> = 0.$$

Кроме того, при $C > 0$ двойственная задача для задачи квадратичного программирования :

$$\min 1/4(C^{-1}(-d - A^T\alpha))^T(-d - A^T) + b^T\alpha$$

$$\alpha > 0$$

В случае $C > 0$ для прямой и двойственной задач выполняется свойство сильной двойственности, т.е. оптимальные значения прямой и двойственной задач квадратичного программирования совпадают

Прямая задача оптимизации

- min Φ(w)= 1/2 $w^T w$ по w

при $d_i(w^T x_i+b)\geq 1$

J(w,b,α)= 1/2 $w^T w$ - $\Sigma\alpha_i$ (d_i ($w^T x_i$ +b)-1)

Условие оптимальности:

$$\frac{\partial J(w,b,\alpha)}{\partial w}=0 \quad \Rightarrow \frac{\partial J(w,b,\alpha)}{\partial w}=w-\sum_{i=1}^{N}\alpha_i d_i x_i=0$$

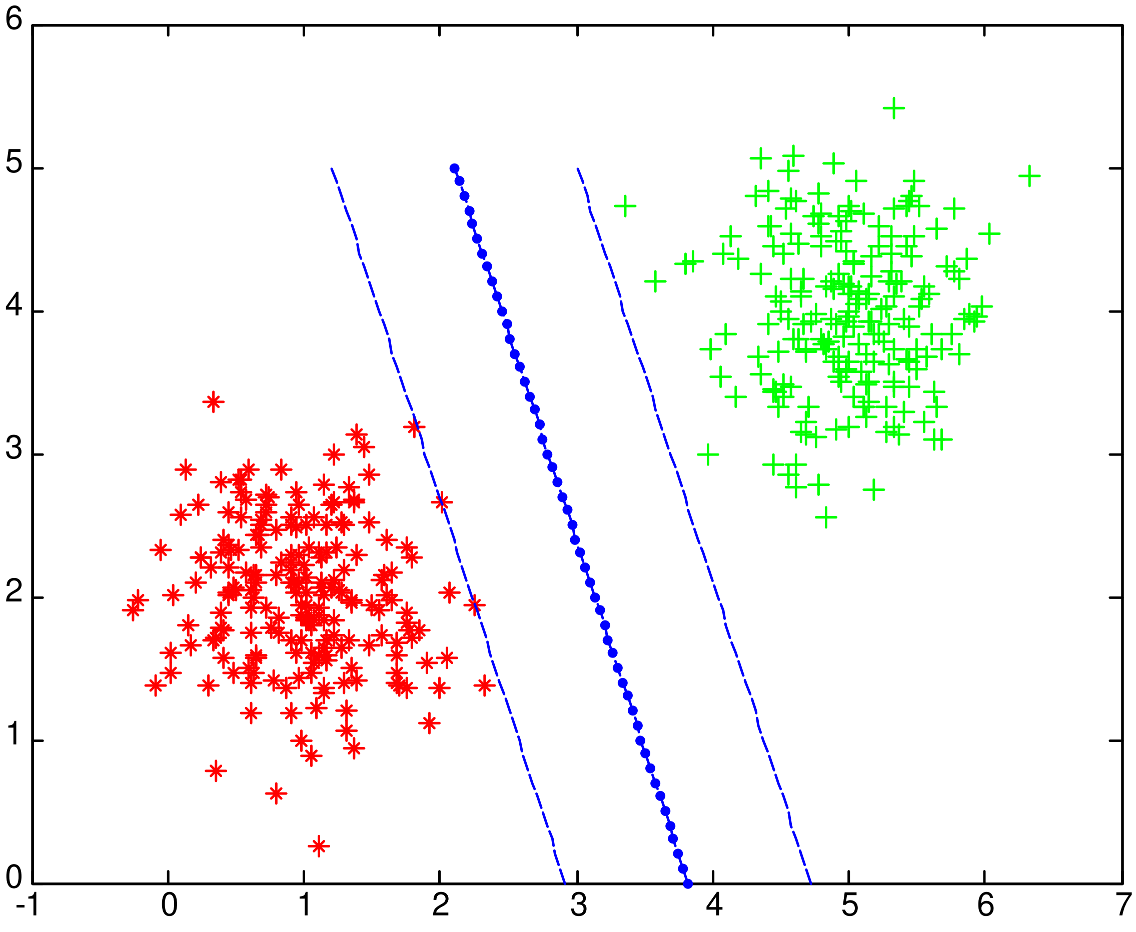
$$\frac{\partial J(w,b,\alpha)}{\partial b}=0 \quad \Rightarrow \sum_{i=1}^{N}\alpha_i d_i=0$$

Решение

- $w = \sum \alpha_i \, y_i \, x_i$
  - x_i – опорные вектора
  - α_i – обученные веса
- Для большей части векторов вес = 0!
- Все вектора, для которых вес >0 называются опорными
- Определяется только опорными векторами

Решение задачи оптимизации для линейно-разделимых образов

Линейно-разделимые образы



Расстояние от плоскости

Двойственная задача оптимизации

$$\max Q(\alpha)=\sum_{i=1}^{N} \alpha_i-1/2 \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i \alpha_j d_i d_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$ при условии:

- $\sum \alpha_i d_i=0$
- $0\leq \alpha_i$

$\boldsymbol{w}_0=\sum \alpha_{0i} d_i \boldsymbol{x}_i$,

$b_0=1- \boldsymbol{w}_0^T x_s$

Статистические свойства

$$h \le \min\left\{\left[\frac{D^2}{\rho^2}\right], m_0\right\} + 1$$

D – диаметр наименьшего шара

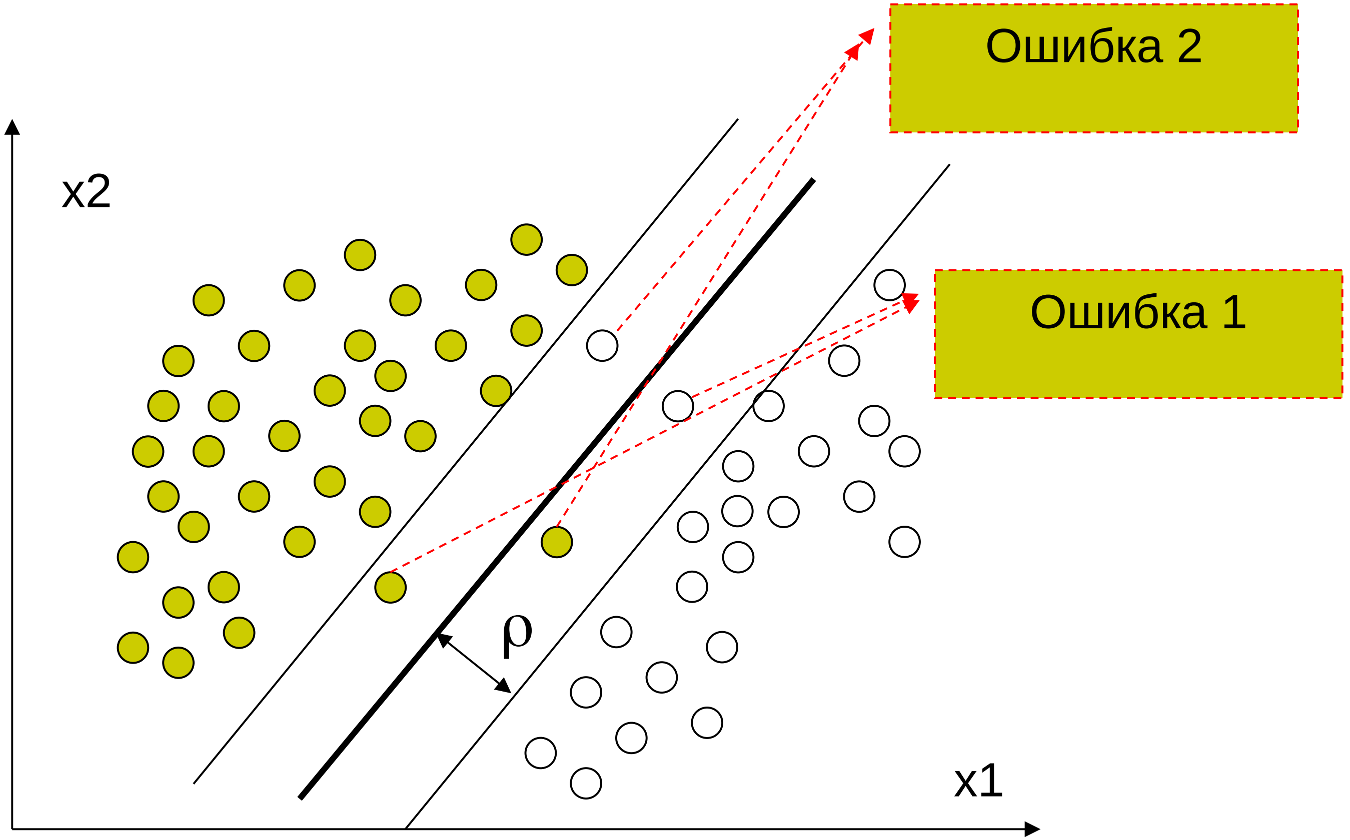
ρ - граница

h –VCdim оптимальной гиперплоскости

m0 – размерность входного пространства

Задача оптимизации

- Нелинейное пространство примеров
- $\{(x_i, d_i)\}$



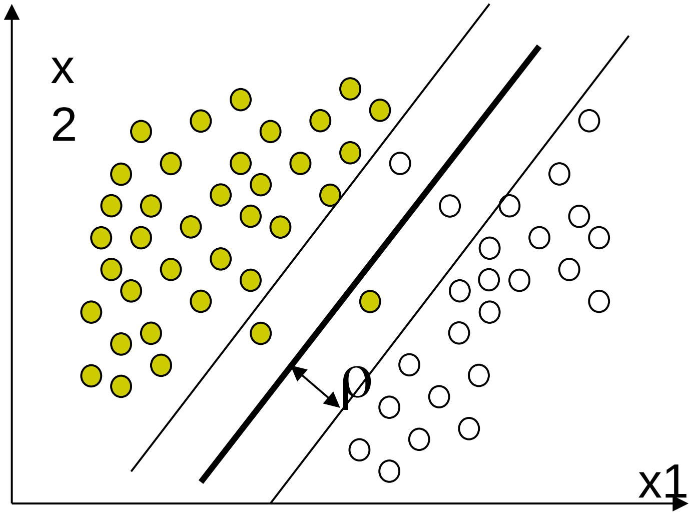
- $d_i(w^T x_i + b)$ - ξ_i, i=1,N

- $0 \le \xi_i \le 1$ – верная классификация
- $\xi_i > 1$ – ошибки

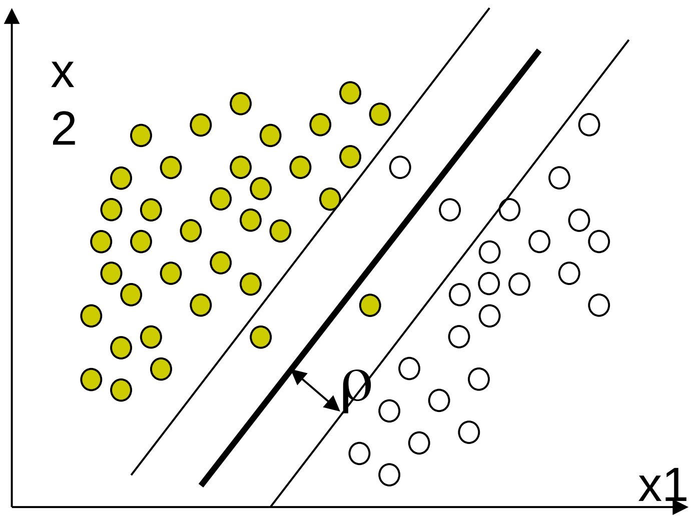
- Опорные вектора

Для $\xi_i > 0$

$d_i(w^T x_i + b) = 1 - \xi_i$



$$\begin{cases} (w^T w)/2 + \alpha \sum \xi_i \to min \\ y(w^T x_i - b) \geqslant 1 - \xi_i \\ \xi_i \geqslant 0 \end{cases}$$



$$Penalty = \sum [M_i < 0]$$

$$[M_i < 0] = \begin{cases} 1 & \text{, если } M_i < 0 \\ 0 & \text{, если } M_i \geqslant 0 \end{cases}$$

$$Penalty = \sum [M_i < 0] \leqslant \sum (1 - M_i)_+ = \sum max(0, 1 - M_i)$$

Задача оптимизации

- min $\Phi(\xi)=\Sigma I(\xi_i - 1)$ по w при $d_i(w^T x_i+b)\geq 1-\xi_i$

- $I(\xi_i-1)$ – индикатор
  - $I(\xi_i - 1)=0$ при $\xi_i \leq 1$
  - $I(\xi_i-1)=1$ при $\xi_i > 1$

$\Phi(\xi)=\Sigma I(\xi_i - 1)$ аппроксимируем

$\Phi(w, \xi)= 1/2\, w^T w + C\Sigma\xi_i$, где i=1,N

Прямая задача оптимизации

- $\{(x_i, d_i)\}$, где i=1,N

min $\Phi(\xi, w)=1/2\ w^T w + C\Sigma\xi_i$,

При ограничениях :

  - $\xi_i \geq 0$
  - $d_i(w^T x_i + b) \geq 1-\xi_i$
  - C>0

Двойственная задача оптимизации

- $\max Q(\alpha)=\sum_{i=1}^{N} \alpha_i - 1/2 \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j$ при условии:
  - $\sum \alpha_i d_i = 0$
  - $0 \le \alpha_i \le C$
  - $C > 0$

$\mathbf{w}_0 = \sum \alpha_{0i} d_i x_i$, где вектора x_i опорные

Отбор опорных векторов

$(\alpha_{0i}(d_i(w^T x_i + b) - 1 + \xi_i) = 0)$

$0 < \alpha_i < C$

Разработка SVM систем

- Перейти к линейной модели большей размерности.
- Построить гиперплоскость.

Теорема Ковера

- Условия
  - Преобразование нелинейно
  - Размерность нового пространства большая

Поиск преобразования

- Вычисление скалярных произведений в многомерном пространстве вычислительно сложно
- вместо прямого вычисления преобразования φ(x), определим ядро скалярного произведения $\mathbf{K}(\mathbf{x},\mathbf{x}_i)=\varphi(\mathbf{x})^T\varphi(\mathbf{x}_i)$
- ядро должно удовлетворять условию Мерсера (*Mercer's condition)*
  - Матрица K(xi,xj) должна быть неотрицательно определенной

Ядро скалярного произведения

$$\left\{\phi_j(x)\right\}_{j=1}^{m1}$$

$$\sum_{j=1}^{m1} w_j \phi_j(x) + b = 0$$

$$\phi_0(x) = 1, w_0 = b \;\Rightarrow\; \sum_{j=0}^{m1} w_j \phi_j(x) = 0$$

$$\bar{\phi}(x) = \left[\phi_0(x) \ldots \phi_{m1}(x)\right]^T$$

$$W^T \bar{\phi}(x) = 0$$

$$w = \sum_{i=1}^{N} \alpha_i d_i \bar{\phi}(x_i)$$

$$\sum_{i=1}^{N} \alpha_i d_i \bar{\phi}^T(x_i) \bar{\phi}(x_i) = 0$$

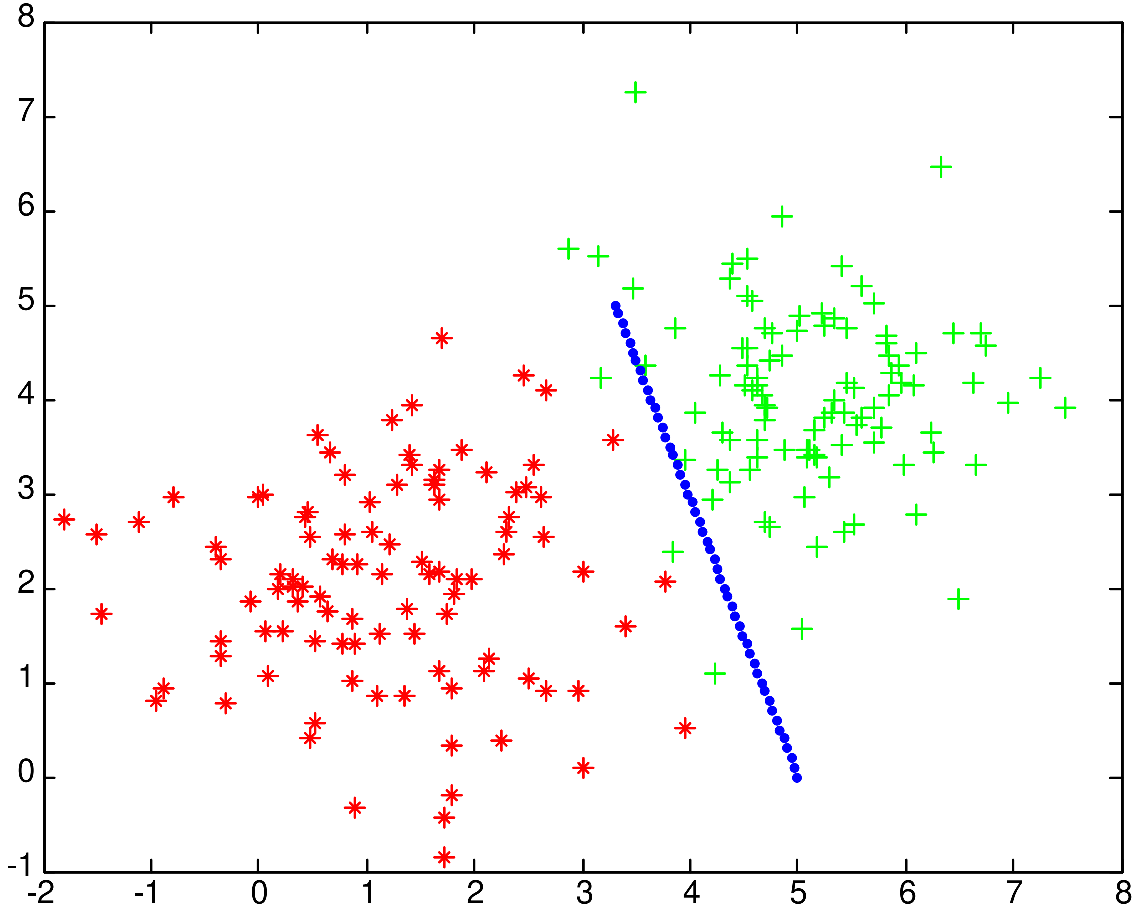
Ядро скалярного произведения

- $\mathbf{K}(x,x_i)=\varphi(x)^T\varphi(x_i) = \Sigma\varphi_j(x)\,\varphi_j(x_i)$ j=0.m1, i=1.N

$\Sigma\ \alpha_i\, d_i\, \mathbf{K}(x,x_i) = 0$

Использование ядра

- Выбрать признаки
- Выбрать ядро для этого вектор-признака
- Вычислить матрицу значений ядра для каждой пары примеров из обучающей выборки
- применить SVM для получения весов и опорных векторов

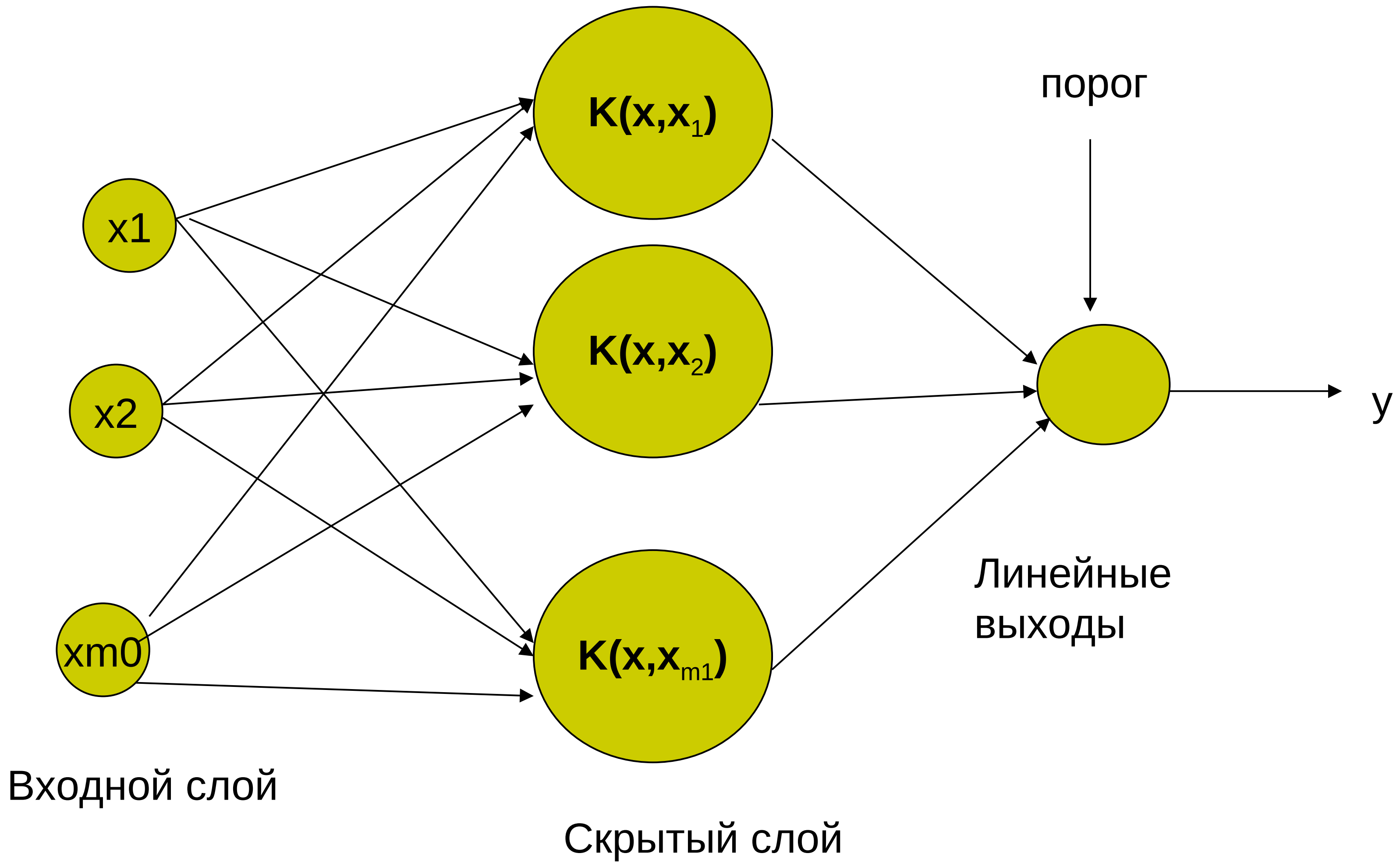


8
7
6
5
4
3
2
1
0
-1
-2
-1
0
1
2
3
4
5
6
7
8

Примеры SVM

Тип SVM	ядро	
полиномиальные	$(x^T x_i + 1)^p$	
Сети RBF	$\exp(-1/(2\sigma^2) \|\|x-x_i\|\|^2)$	
2-й перцептрон	$\mathrm{Th}(\beta 0\ x^T x_i + \beta 1)$	

Модель SVM



порог
x1
x2
xm0
K(x,x1)
K(x,x2)
K(x,xm1)
y
Линейные
выходы
Входной слой
Скрытый слой

XOR

- x1 x2 d
- $\mathbf{x}_1$ -1 -1 -1
- $\mathbf{x}_2$ -1 +1 +1
- $\mathbf{x}_3$ +1 -1 +1
- $\mathbf{x}_4$ +1 +1 -1

- $K(x,x_i)=|1+x^T x_i|^2$
- $x=[x_1,x_2]^T$
- $X_i=[x_{i1},x_{i2}]^T$

- $K(x,x_i)=1+x_1^2x_{i1}^2+2x_1x_{i1}x_2x_{i2}+x_2^2x_{i2}^2+2x_1x_{i1}+2x_2x_{i2}$

- $\varphi(\mathbf{x})=[1, x_1^2, \sqrt{2}\, x_1 x_2, x_2^2, \sqrt{2}\, x_1 , \sqrt{2}\, x_2]$
- $\varphi(\mathbf{x}_\mathbf{i})=[1, x_{1\mathbf{i}}^2, \sqrt{2}\, x_{1\mathbf{i}} x_{2\mathbf{i}}, x2_\mathbf{i}^2, \sqrt{2}\, x_{1\mathbf{i}} , \sqrt{2}\, x_{2\mathbf{i}}]$

ядро

$$K = \begin{matrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{matrix}$$

Целевая функция

- $Q(a) = a1 + a2 + a3 + a4 – ½ (9\ a1^2 - 2\ a1\ a2 – 2\ a1\ a3 + 2\ a1\ a4 + 9\ a2^2 + 2\ a2\ a3 – 2\ a2\ a4$ _ $9\ a3^2 – 2\ a3\ a4 + 9\ a4^2)$
- система

9 a1 – a2 –a3 =a4 =1

- a1 =2 a2 =a3 – a4 =1

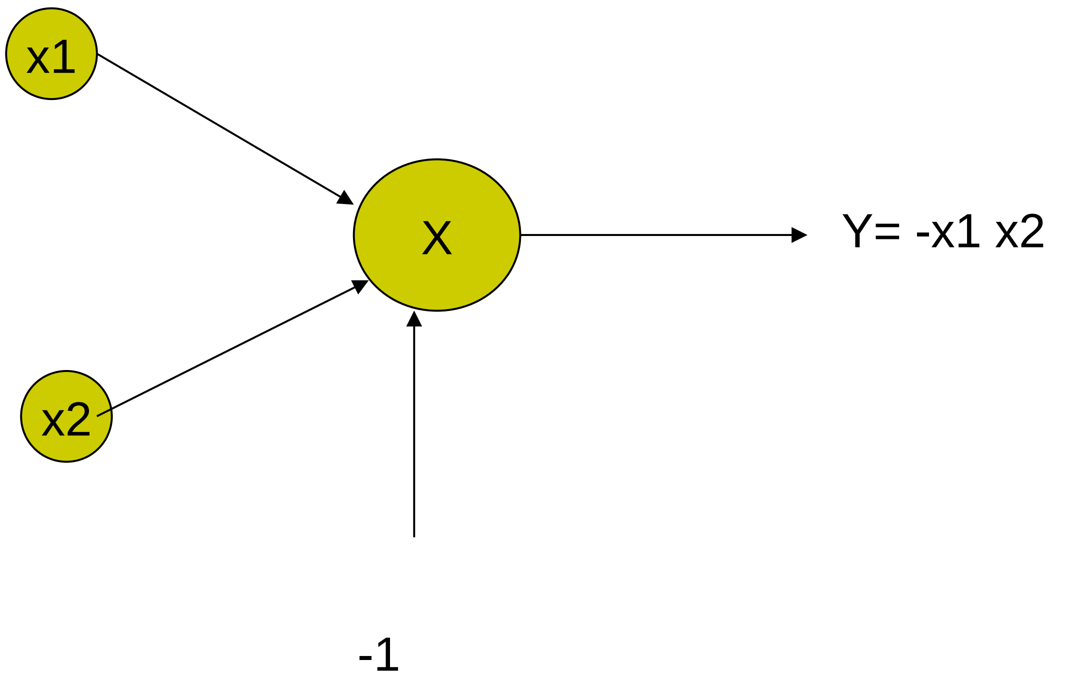
- a1 = a2 = 9 a3 – a4 =1

a1 – a2 – a3 = 9 a4 = 1

Множители Лагранжа

- $a1 = a2 = a3 = a4 = 1/8$
- $ai > 0$ - все примеры опорные

сеть



Нелинейная регрессия

- $L\varepsilon(d,y)=\{|d-y|-\varepsilon$ при $|d-y|>\varepsilon, 0\}$

- $d_i - \boldsymbol{w}^T \varphi(\mathbf{x_i}) <= \varepsilon + \xi i$
- $\boldsymbol{w}^T \varphi(\mathbf{x_i})$ **-** $d_i <= \varepsilon + \xi i'$

SVM

- Плюсы
  - Много доступных библиотек: http://www.kernel-machines.org/software
  - Мощный и гибкий подход на основе ядер
  - Работает хорошо, даже для маленьких обучающих выборок
- Минусы
  - Нет прямых многоклассовых метод, нужно объединять двухклассовые
  - Вычисления, память
  - При обучении нужно строить полную матрицу ядра для всех примеров
  - Обучение занимает много времени для больших задач

литература

- Математические методы распознавания образов. Курс лекций. МГУ, ВМиК, кафедра «Математические методы прогнозирования» Местецкий Л.М., 2002–2004
- Хайкин С. Нейрокомпьютеры: полный курс. – М.:Вильямс – 2006 (стр 418-435)