

Разбор летучки

Лекция 9

Метод опорных векторов.

Екатерина Тузова

Постановка задачи

$$X = \mathbb{R}^n, Y = \{-1, +1\}$$

$X^l = (x_i, y_i)_{i=1}^l$ – обучающая выборка

Линейный классификатор:

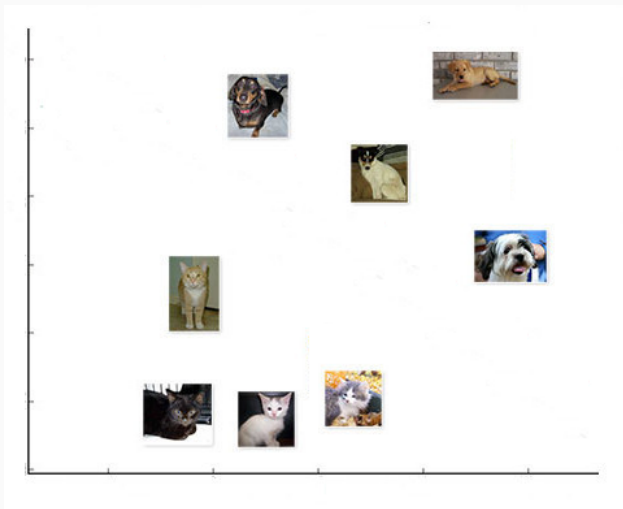
$$a(\mathbf{x}, \mathbf{w}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - w_0)$$

Найти:

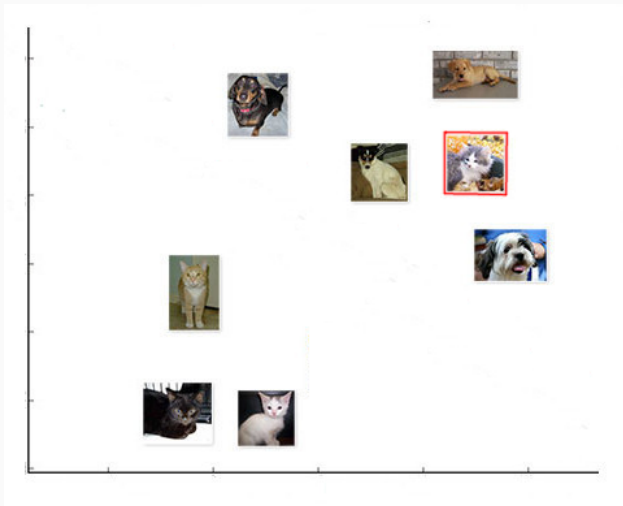
$(n - 1)$ -мерную гиперплоскость, которая разделяет данные как можно лучше.

Линейно разделимая выборка

Линейно разделимая выборка



Линейно неразделимая выборка



Линейно разделимая выборка

Выборка линейно разделима, если отступ на каждом объекте положителен.

$$\exists \mathbf{w}, w_0 : M_i(\mathbf{w}, w_0) = y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - w_0) > 0, \quad i = 1, \dots, l$$

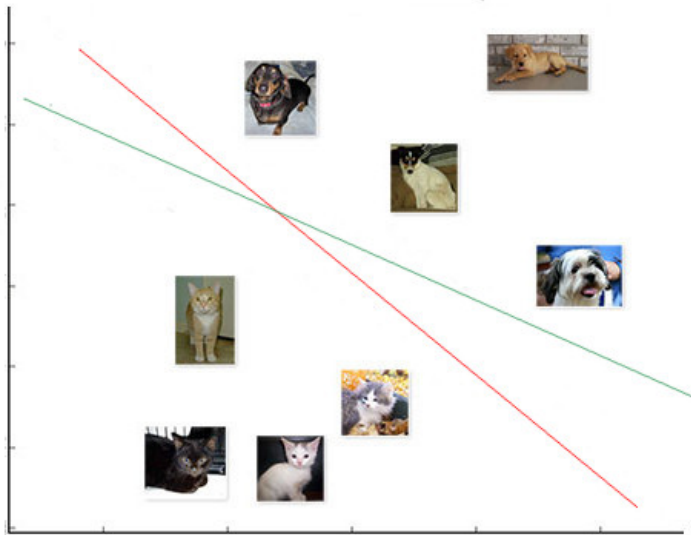
Линейно разделимая выборка

Выборка линейно разделима, если отступ на каждом объекте положителен.

$$\exists \mathbf{w}, w_0 : M_i(\mathbf{w}, w_0) = y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - w_0) > 0, \quad i = 1, \dots, l$$

Нормировка: $\min_{i=1, \dots, l} M_i(\mathbf{w}, w_0) = 1$

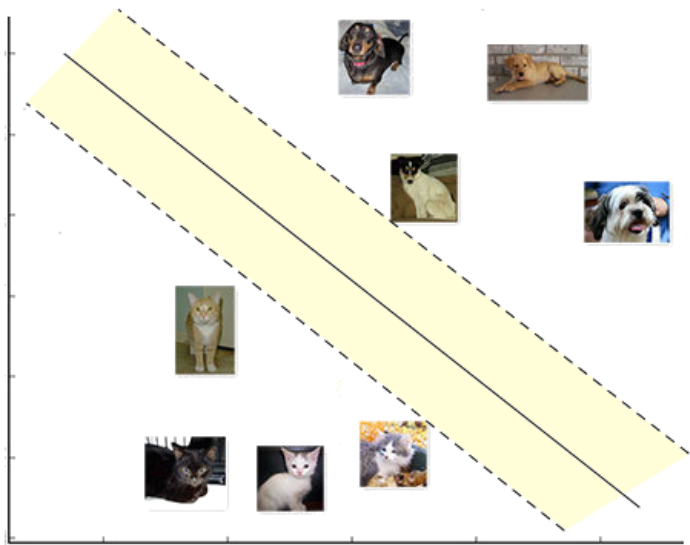
Разделяющая гиперплоскость



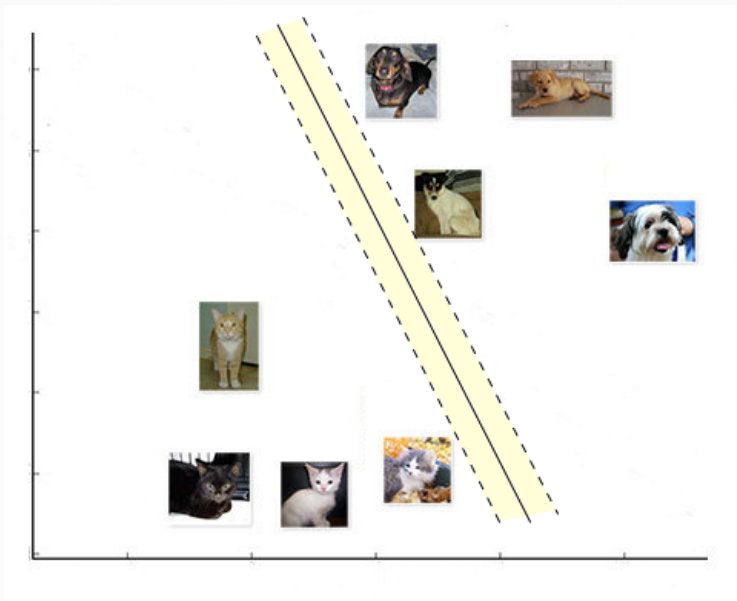
Максимизация отступа

Идея: Будем искать такую разделяющую поверхность, которая обеспечивает разделяющую полосу максимальной ширины.

Пример



Пример



Опорная гиперплоскость

Опорная гиперплоскость

Гиперплоскость называется опорной для множества точек X , если все точки из X лежат по одну сторону от этой гиперплоскости.

$$a(\mathbf{x}, \mathbf{w}, w_0) = \langle \mathbf{x}, \mathbf{w} \rangle - w_0 = 0$$

Как посчитать расстояние от точки до гиперплоскости?

Опорная гиперплоскость

Гиперплоскость называется опорной для множества точек X , если все точки из X лежат по одну сторону от этой гиперплоскости.

$$a(\mathbf{x}, \mathbf{w}, w_0) = \langle \mathbf{x}, \mathbf{w} \rangle - w_0 = 0$$

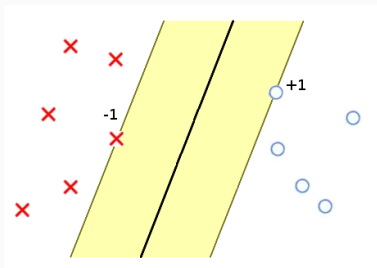
Расстояние от точки до гиперплоскости = $\frac{|a(\mathbf{x}, \mathbf{w}, w_0)|}{\|\mathbf{w}\|}$

Максимизация отступа

Идея: Максимизировать отступ между двумя параллельными опорными плоскостями, а затем провести параллельную им плоскость на равных расстояниях.

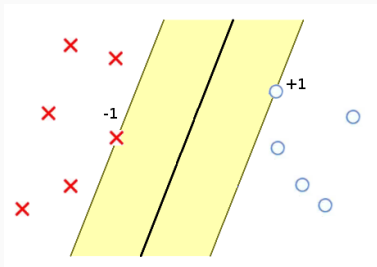
Оптимальная разделяющая гиперплоскость

Как выглядит разделяющая полоса?



Оптимальная разделяющая гиперплоскость

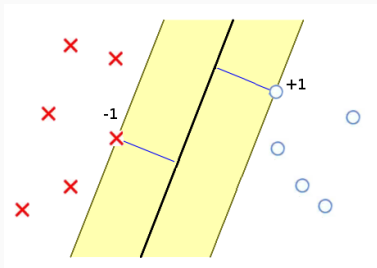
Разделяющая полоса: $\{\mathbf{x} : -1 \leq \langle \mathbf{w}, \mathbf{x} \rangle - w_0 \leq 1\}$



Оптимальная разделяющая гиперплоскость

Разделяющая полоса: $\{ \mathbf{x} : -1 \leq \langle \mathbf{w}, \mathbf{x} \rangle - w_0 \leq 1 \}$

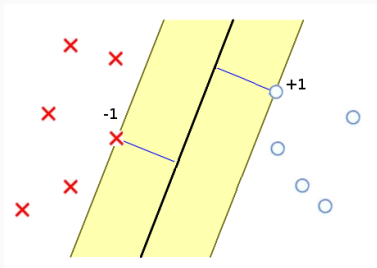
Ширина разделяющей полосы?



Оптимальная разделяющая гиперплоскость

Разделяющая полоса: $\{ \mathbf{x} : -1 \leq \langle \mathbf{w}, \mathbf{x} \rangle - w_0 \leq 1 \}$

Ширина разделяющей полосы: $\frac{\langle \mathbf{x}_+, \mathbf{w} \rangle + \langle \mathbf{x}_-, \mathbf{w} \rangle}{\| \mathbf{w} \|} = \frac{2}{\| \mathbf{w} \|} \rightarrow \max$



Оптимальная разделяющая гиперплоскость

Линейно разделяемая выборка:

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}} \\ M_i(\mathbf{w}, w_0) \geq 1 \quad i = 1, \dots, l \end{cases}$$

Оптимальная разделяющая гиперплоскость

Линейно разделяемая выборка:

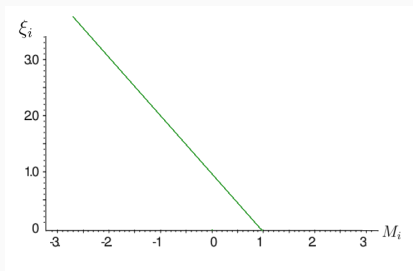
$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}} \\ M_i(\mathbf{w}, w_0) \geq 1 \quad i = 1, \dots, l \end{cases}$$

Линейно неразделимая выборка – надо ослабить имеющиеся условия.

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i \rightarrow \min_{\mathbf{w}, \xi} \\ M_i(\mathbf{w}, w_0) \geq 1 - \xi_i \quad i = 1, \dots, l \\ \xi_i \geq 0 \quad i = 1, \dots, l \end{cases}$$

Оптимальная разделяющая гиперплоскость

$$\begin{cases} \xi_i \geq 1 - M_i(\mathbf{w}, w_0) \\ \xi_i \geq 0 \end{cases} \Rightarrow \xi_i = 1 - M_i(\mathbf{w}, w_0)$$



Задача безусловной минимизации

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i \rightarrow \min_{\mathbf{w}, \xi} \\ \xi_i = 1 - M_i(\mathbf{w}, w_0) \end{cases}$$

Задача безусловной минимизации

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i \rightarrow \min_{\mathbf{w}, \xi} \\ \xi_i = 1 - M_i(\mathbf{w}, w_0) \end{cases}$$

Задача безусловной минимизации:

$$C \sum_{i=1}^l (1 - M_i(\mathbf{w}, w_0)) + \frac{1}{2} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$

Минимизация эмпирического риска

$$Q(\mathbf{w}) = \sum_{i=1}^l [M_i(\mathbf{w}, w_0) < 0] \leq \sum_{i=1}^l \mathcal{L}(M_i(\mathbf{w}, w_0)) \rightarrow \min_{\mathbf{w}}$$

Минимизация эмпирического риска

$$Q(\mathbf{w}) = \sum_{i=1}^l [M_i(\mathbf{w}, w_0) < 0] \leq \sum_{i=1}^l \mathcal{L}(M_i(\mathbf{w}, w_0)) \rightarrow \min_{\mathbf{w}}$$

Штраф за увеличение нормы вектора весов:

$$Q_{\tau} = Q + \frac{\tau}{2} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$

Минимизация эмпирического риска

$$Q(\mathbf{w}) = \sum_{i=1}^l [M_i(\mathbf{w}, w_0) < 0] \leq \sum_{i=1}^l \mathcal{L}(M_i(\mathbf{w}, w_0)) \rightarrow \min_{\mathbf{w}}$$

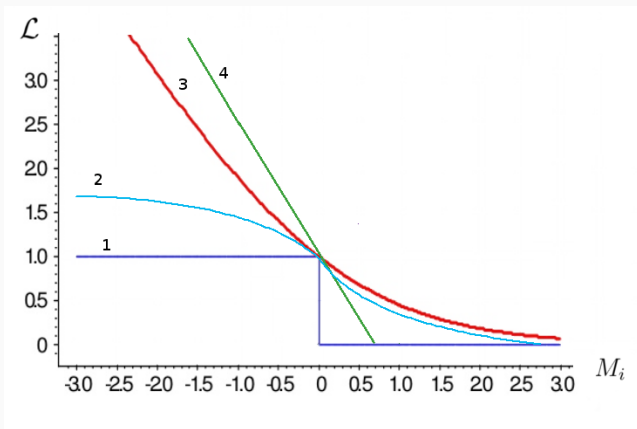
Штраф за увеличение нормы вектора весов:

$$Q_{\tau} = Q + \frac{\tau}{2} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$

Метод опорных векторов:

$$C \sum_{i=1}^l (1 - M_i(\mathbf{w}, w_0)) + \frac{1}{2} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$

Примеры \mathcal{L}

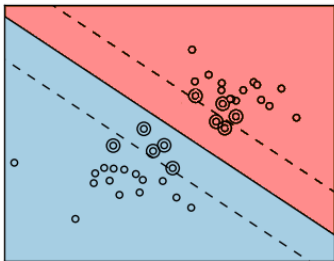


1. $[M_i(\mathbf{w}, w_0) < 0]$
2. $L(M) = \log_2(1 + e^{-M})$ – логарифмическая
3. $S(M) = 2(1 + e^M)^{-1}$ – сигмоидная

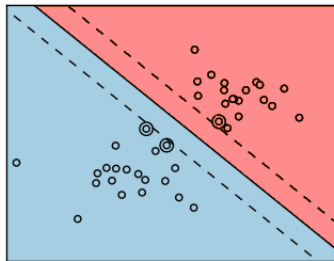
$$\sum_{i=1}^l (1 - M_i(\mathbf{w}, w_0)) + \frac{1}{2C} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$

На что влияет параметр C ?

Выбор параметра C



Маленький C
Сильная регуляризация



Большой C
Слабая регуляризация

Условие Каруша-Куна-Такера

$$\begin{cases} f(x) \rightarrow \min \\ g_i(x) \leq 0, i = 1, \dots, m \\ h_j(x) = 0, j = 1, \dots, k \end{cases}$$

Условие Каруша-Куна-Такера

$$\begin{cases} f(x) \rightarrow \min \\ g_i(x) \leq 0, i = 1, \dots, m \\ h_j(x) = 0, j = 1, \dots, k \end{cases}$$

Двойственная задача:

$$\begin{cases} \mathcal{L}(x; \mu, \alpha) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \alpha_j h_j(x) \\ \frac{\partial \mathcal{L}}{\partial x} = 0 \\ g_i(x) \leq 0, h_j(x) = 0 \\ \mu_i \geq 0 \\ \mu_i g_i(x) = 0 \end{cases}$$

Двойственная задача SVM

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i \rightarrow \min_{\mathbf{w}, \xi} \\ M_i(\mathbf{w}, w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Двойственная задача SVM

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i \rightarrow \min_{\mathbf{w}, \xi} \\ M_i(\mathbf{w}, w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Двойственная задача:

$$\begin{cases} \mathcal{L} = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^l \alpha_i (M_i(\mathbf{w}, w_0) - 1) - \sum_{i=1}^l \xi_i (\alpha_i + \mu_i - C) \\ \xi_i \geq 0, \quad \alpha_i \geq 0, \quad \mu_i \geq 0 \\ \alpha_i = 0 \text{ либо } M_i(\mathbf{w}, w_0) = 1 - \xi_i, \quad i = 1, \dots, l \\ \mu_i = 0 \text{ либо } \xi_i = 0, \quad i = 1, \dots, l \end{cases}$$

Двойственная задача SVM

$$\mathcal{L}(\mathbf{w}, w_0, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^l \alpha_i (M_i(\mathbf{w}, w_0) - 1) - \sum_{i=1}^l \xi_i (\alpha_i + \mu_i - C)$$

Двойственная задача SVM

$$\mathcal{L}(\mathbf{w}, w_0, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^l \alpha_i (M_i(\mathbf{w}, w_0) - 1) - \sum_{i=1}^l \xi_i (\alpha_i + \mu_i - C)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = 0$$

Двойственная задача SVM

$$\mathcal{L}(\mathbf{w}, w_0, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^l \alpha_i (M_i(\mathbf{w}, w_0) - 1) - \sum_{i=1}^l \xi_i (\alpha_i + \mu_i - C)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = - \sum_{i=1}^l \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^l \alpha_i y_i = 0$$

Двойственная задача SVM

$$\mathcal{L}(\mathbf{w}, w_0, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^l \alpha_i (M_i(\mathbf{w}, w_0) - 1) - \sum_{i=1}^l \xi_i (\alpha_i + \mu_i - C)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = - \sum_{i=1}^l \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^l \alpha_i y_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = -\alpha_i - \mu_i + C = 0 \quad \Rightarrow \quad \mu_i + \alpha_i = C$$

Двойственная задача SVM

$$\begin{cases} -\mathcal{L}(\alpha) = -\sum_{i=1}^l \alpha_i + \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \rightarrow \min_{\alpha} \\ 0 \leq \alpha_i \leq C \\ \sum_{i=1}^l \alpha_i y_i = 0 \end{cases}$$

Двойственная задача SVM

Решение исходной задачи выражается через решение двойственной:

$$\begin{cases} \mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i \\ w_0 = \langle \mathbf{w}, \mathbf{x}_i \rangle - y_i \end{cases}$$

Двойственная задача SVM

Решение исходной задачи выражается через решение двойственной:

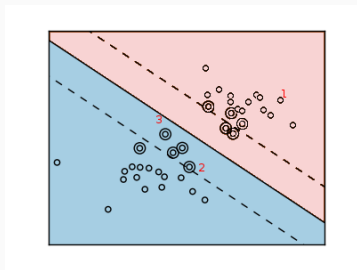
$$\begin{cases} \mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i \\ w_0 = \langle \mathbf{w}, \mathbf{x}_i \rangle - y_i \end{cases}$$

Линейный классификатор:

$$a(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^l \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle - w_0\right)$$

Понятие опорного вектора

- $\alpha_i = 0, M_i \geq 1$ – неинформативные объекты
- $0 < \alpha_i < C, M_i = 1$ – опорные объекты
- $\alpha_i = C, M_i < 1$ – опорные объекты-нарушители



Kernel trick

Kernel trick

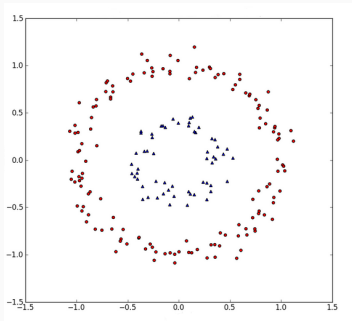
$$\langle \mathbf{x}_i, \mathbf{x} \rangle \rightarrow K(\mathbf{x}_i, \mathbf{x})$$

$\psi : X \rightarrow H$, H - Гильбертово пространство

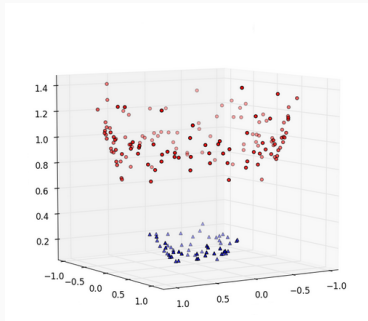
$$K(\mathbf{x}_i, \mathbf{x}) = \langle \psi(\mathbf{x}_i), \psi(\mathbf{x}) \rangle_H$$

- $K(\mathbf{x}_i, \mathbf{x}) = K(\mathbf{x}, \mathbf{x}_i)$
- неотрицательно определена

Переход к более высокой размерности

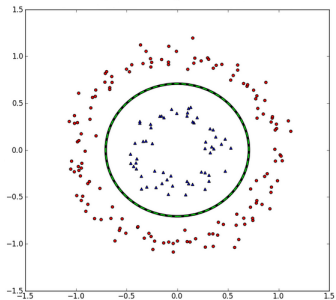


Данные в \mathbb{R}^2

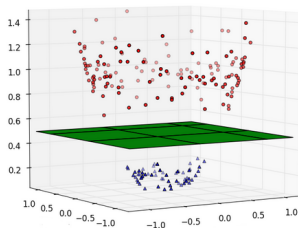


Данные в \mathbb{R}^3

Переход к более высокой размерности

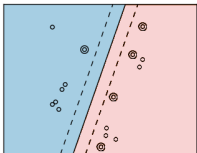


Данные в \mathbb{R}^2

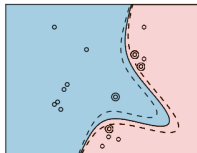


Данные в \mathbb{R}^3

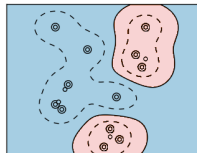
Примеры ядер



Линейное
 $\langle x, x' \rangle$



Полиномиальное
 $(\langle x, x' \rangle + 1)^3$

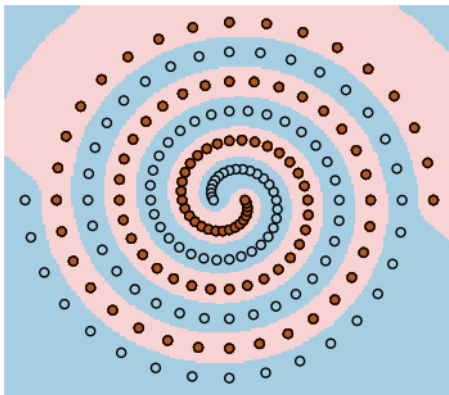


Гауссовское
 $\exp(-\beta \|x - x'\|^2)$

Примеры ядер

Гауссовское

$$\exp(-\beta \|x - x'\|^2)$$



Конструктивные методы получения ядер

- $K(\mathbf{x}_i, \mathbf{x}) = \langle \mathbf{x}_i, \mathbf{x} \rangle$
- $K(\mathbf{x}_i, \mathbf{x}) = \text{const}$
- $K(\mathbf{x}_i, \mathbf{x}) = K_1(\mathbf{x}_i, \mathbf{x})K_2(\mathbf{x}_i, \mathbf{x})$
- $K(\mathbf{x}_i, \mathbf{x}) = \alpha_1 K_1(\mathbf{x}_i, \mathbf{x}) + \alpha_2 K_2(\mathbf{x}_i, \mathbf{x})$ при $\alpha_1, \alpha_2 > 0$
- $\forall \psi : X \rightarrow \mathbb{R} \quad K(\mathbf{x}_i, \mathbf{x}) = \psi(\mathbf{x}_i)\psi(\mathbf{x})$
- $\forall \phi : X \rightarrow X \quad K(\mathbf{x}_i, \mathbf{x}) = K_0(\phi(\mathbf{x}_i), \phi(\mathbf{x}))$

- $K(\mathbf{x}_i, \mathbf{x}) = \langle \mathbf{x}_i, \mathbf{x} \rangle^d$
- $K(\mathbf{x}_i, \mathbf{x}) = (\langle \mathbf{x}_i, \mathbf{x} \rangle + 1)^d$
- $K(\mathbf{x}_i, \mathbf{x}) = \sigma(\langle \mathbf{x}_i, \mathbf{x} \rangle)$

Двуслойная нейросеть с функцией активации σ

- $K(\mathbf{x}_i, \mathbf{x}) = th(k_0 + k_1 \langle \mathbf{x}_i, \mathbf{x} \rangle), \quad k_0, k_1 \geq 0$
- $K(\mathbf{x}_i, \mathbf{x}) = \exp(-\beta \|\mathbf{x} - \mathbf{x}_i\|^2)$

+ Задача имеет единственное решение

Достоинства и недостатки

- + Задача имеет единственное решение
- + Число опорных векторов определяется автоматически

Достоинства и недостатки

- + Задача имеет единственное решение
- + Число опорных векторов определяется автоматически
- Неустойчивость к шуму

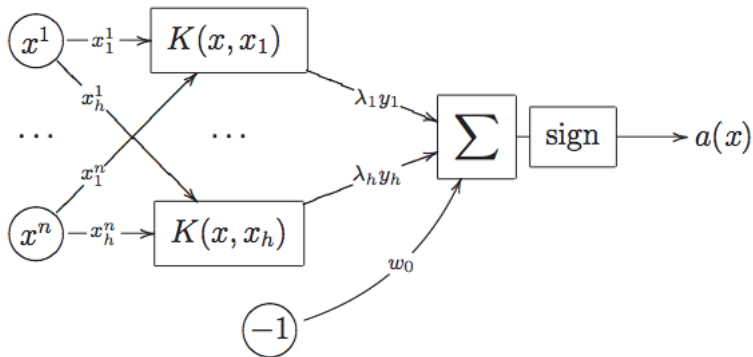
Достоинства и недостатки

- + Задача имеет единственное решение
- + Число опорных векторов определяется автоматически
- Неустойчивость к шуму
- Нет общих подходов к оптимизации ядра под задачу

Достоинства и недостатки

- + Задача имеет единственное решение
- + Число опорных векторов определяется автоматически
- Неустойчивость к шуму
- Нет общих подходов к оптимизации ядра под задачу
- Подбор константы C

SVM как двуслойная нейросеть



h – количество опорных объектов

Вопросы?

Что почитать по этой лекции

- T. Hastie, R. Tibshirani "The Elements of Statistical Learning" Chapter 12
- Andrew Ng CS229 Lecture notes
- M. Morhi, A. Rostamizadeh, A. Talwalkar "Foundations of Machine Learning"

На следующей лекции

- Линейная регрессия
- Многомерная регрессия
- Проблема мультиколлинеарности
- Сингулярное разложение
- Гребневая регрессия
- Lasso