Машинное обучение

Лекция 6. Метод опорных векторов.

Катя Тузова

Разбор летучки

Почему линейная зависимость признаков приводит к переобучению?

Разбор летучки

$a(\mathbf{x}, \mathbf{w}) = sign(\langle \mathbf{w}, \mathbf{x} \rangle)$
Линейная зависимость признаков:
$\forall \mathbf{x} \exists \mathbf{u} : \langle \mathbf{u}, \mathbf{x} \rangle = 0$
$\Rightarrow \forall \gamma : a(\mathbf{x}, \mathbf{w}) = sign(\langle \mathbf{w} + \gamma \mathbf{u}, \mathbf{x} \rangle)$

Алгоритм a' работает точно также как исходный a.
А значит мы можем получить любое решение из семейства $\mathbf{w} + \gamma \mathbf{u}$

Разбор летучки

Каким способом оценивается функция потерь при стохастическом градиенте?

Разбор летучки

Input: X^l, α, η
Output: $w_0, w_1, \ldots, w_n$

Перемешать данные в X^l
Инициализировать: w_j, $j = 0, \ldots, n$

$$Q(\mathbf{w}) = \sum_{i=1}^{l} \mathcal{L}(\langle \mathbf{w}, \mathbf{x_i} \rangle y_i)$$

Повторить пока Q и/или w не стабилизируются:

- Взять x_i из X^l
- Потеря: $\varepsilon_i = \mathcal{L}(\langle \mathbf{w}, \mathbf{x_i} \rangle y_i)$
- Градиентный шаг: $w = w - \alpha \mathcal{L}'(\langle \mathbf{w}, \mathbf{x_i} \rangle y_i)\mathbf{x_i} y_i$
- Оценить $Q = (1 - \eta)Q + \eta \varepsilon_i$

Разбор летучки

Каким образом сокращаются веса при градиентном спуске? И для чего?

Разбор летучки

Штраф за увеличение нормы вектора весов:
$Q_\tau = Q + \frac{\tau}{2}\|\mathbf{w}\|^2 \to \min\limits_{\mathbf{w}}$

Градиент:
$\triangledown Q_\tau = \triangledown Q + \tau \mathbf{w}$

Градиентный шаг:
$\mathbf{w} = \mathbf{w}(1 - \alpha\tau) - \alpha \triangledown Q(\mathbf{w})$
τ – параметр регуляризации

Разбор летучки

Для чего можно делать пробные случайные шаги?

Разбор летучки

Выбивание из локальных минимумов.

Почему возникает необходимость в изобретении метода стохастического градиента?

Разбор летучки

$$\mathbf{w} = \mathbf{w} - \alpha \sum_{i=1}^{l} \mathcal{L}'(\langle \mathbf{w}, \mathbf{x_i} \rangle y_i)\mathbf{x_i} y_i$$

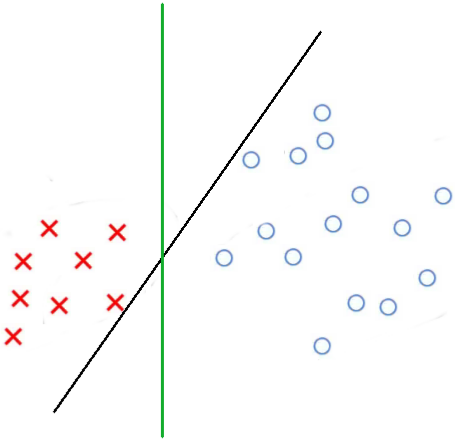
Постановка задачи

$X = \mathbb{R}^n$, $Y = \{-1, +1\}$
$X^l = (x_i, y_i)_{i=1}^l$ – обучающая выборка

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

Как можно лучше – это как?

Пример



Постановка задачи

Как можно лучше:
Два разделенных класса должны лежать как можно дальше от разделяющей гиперплоскости.

Опорная гиперплоскость

Опорная гиперплоскость

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

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

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.

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

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

Максимизиция отступа

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

Постановка задачи

$X^l = (x_i, y_i)_{i=1}^l$
$Y = \{-1, +1\}$

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

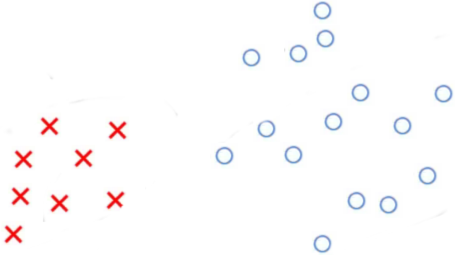
Линейно разделимая выборка

Линейно разделимая выборка

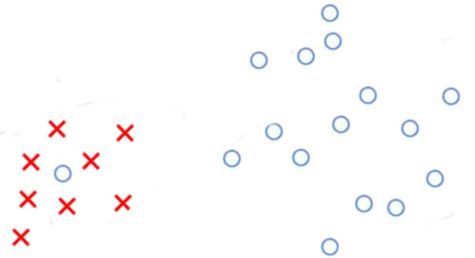
Выборка линейно разделима, если отступ на каждом объекте положителен.

$\exists \mathbf{w}, w_0 : M_i(\mathbf{w}, w_0) = y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle - w_0) > 0$, i=1, . . . , l

Линейно разделимая выборка



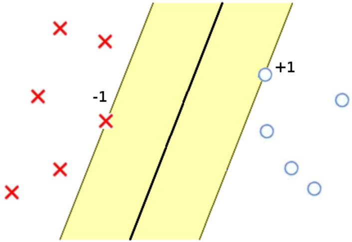
Линейно неразделимая выборка



Оптимальная разделяющая гиперплоскость

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

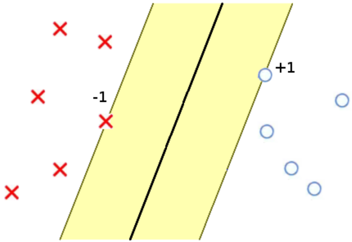
Как выглядит разделяющая полоса?



Оптимальная разделяющая гиперплоскость

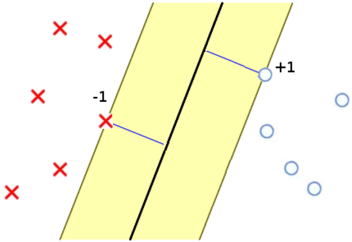
Нормировка: $\min\limits_{i=1,\dots,l} M_i(\mathbf{w}, w_0) = 1$

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



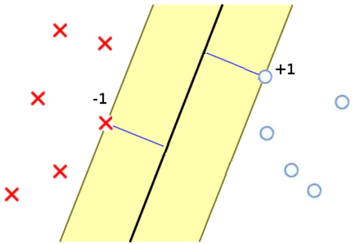
Оптимальная разделяющая гиперплоскость

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: $\frac{\langle \mathbf{x}_+, \mathbf{w} \rangle + \langle \mathbf{x}_-, \mathbf{w} \rangle}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \to \max$



Оптимальная разделяющая гиперплоскость

Линейно разделимая выборка:

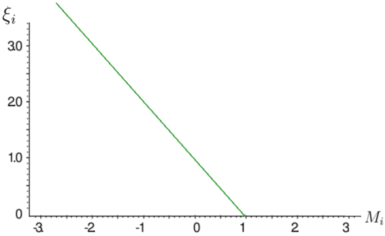
$$\begin{cases} \|\mathbf{w}\|^2 \to \min\limits_{\mathbf{w}} \\ M_i(\mathbf{w}, w_0) \geq 1 \end{cases}$$

Линейно неразделимая выборка – надо ослабить имеющиеся условия.

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

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

$$\begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{l}\xi_i \to \min\limits_{\mathbf{w},\xi} \\ \xi_i = 1 - M_i(\mathbf{w}, w_0) \end{cases}$$

Задача безусловной минимизации:

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

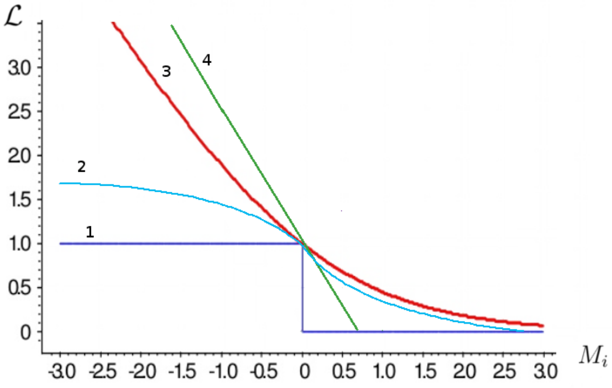
Минимизация эмпирического риска

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

Штраф за увеличение нормы вектора весов:
$Q_\tau = Q + \frac{\tau}{2}\|\mathbf{w}\|^2 \to \min\limits_{\mathbf{w}}$

Метод опорных векторов:
$C\sum\limits_{i=1}^{l}(1 - M_i(\mathbf{w}, w_0)) + \frac{1}{2}\|\mathbf{w}\|^2 \to \min\limits_{\mathbf{w}}$

Примеры $\mathcal{L}$



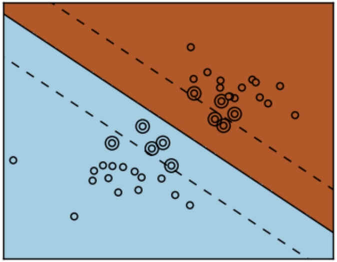
1. $[M_i(\mathbf{w}, w_0) < 0]$
2. $L(M) = \log_2(1 + e^{-M})$ – логарифмическая
3. $S(M) = 2(1 + e^{M})^{-1}$ – сигмоидная
4. $V(M) = (1 - M)_+$ – кусочно-линейная

Вопрос

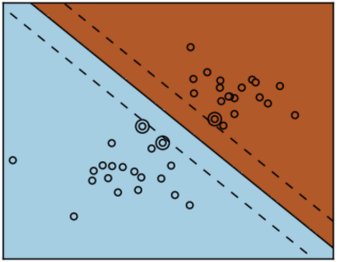
$$\sum_{i=1}^{l}(1 - M_i(\mathbf{w}, w_0)) + \frac{1}{2C}\|\mathbf{w}\|^2 \to \min_{\mathbf{w}}$$

На что влияет параметр C?

Выбор параметра C



Маленький C
Сильная регуляризация



Большой C
Слабая регуляризация

Пример из Python scikit-learn: http://scikit-learn.org/dev

Условие Каруша-Куна-Такера

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

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

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

Двойственная задача:

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

Двойственная задача SVM

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

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x_i} = 0 \qquad \Rightarrow \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x_i} = 0$$
$$\frac{\partial \mathcal{L}}{\partial w_0} = -\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0$$
$$\frac{\partial \mathcal{L}}{\partial \xi_i} = -\alpha_i - \mu_i + C = 0 \qquad \Rightarrow \mu_i + \alpha_i = C$$

Двойственная задача SVM

$$\begin{cases} -\mathcal{L}(\alpha) = -\sum\limits_{i=1}^{l} \alpha_i + \frac{1}{2} \sum\limits_{i=1}^{l} \sum\limits_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \mathbf{x_i}, \mathbf{x_j} \rangle \to \min\limits_{\alpha} \\ 0 \leq \alpha_i \leq C \\ \sum\limits_{i=1}^{l} \alpha_i y_i = 0 \end{cases}$$

Двойственная задача SVM

Решение исходной задачи выражается через решение двойственной:

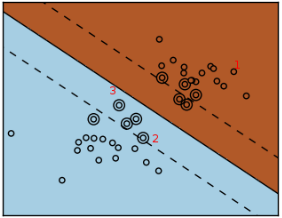
$$\begin{cases} \mathbf{w} = \sum\limits_{i=1}^{l} \alpha_i y_i \mathbf{x_i} \\ w_0 = \langle \mathbf{w}, \mathbf{x_i} \rangle - y_i \end{cases}$$

Линейный классификатор:

$$a(\mathbf{x}) = \mathrm{sign}(\sum_{i=1}^{l} \alpha_i y_i \langle \mathbf{x_i}, \mathbf{x} \rangle - w_0)$$

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



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:

$$a(\mathbf{x}) = \mathrm{sign}(\sum_{i=1}^{l} \alpha_i y_i \langle \mathbf{x_i}, \mathbf{x} \rangle - w_0)$$

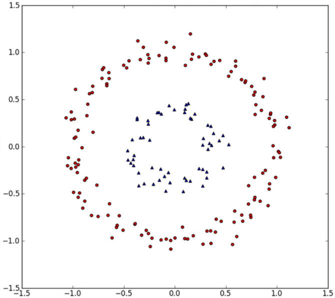
Kernel trick

$\langle \mathbf{x_i}, \mathbf{x} \rangle \to K(\mathbf{x_i}, \mathbf{x})$

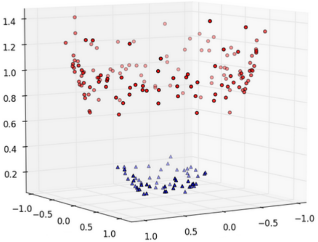
$\psi : X \to H$, H - Гильбертово пространство
$K(\mathbf{x_i}, \mathbf{x}) = \langle \psi(\mathbf{x_i}), \psi(\mathbf{x}) \rangle_H$

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

Переход к более высокой размерности

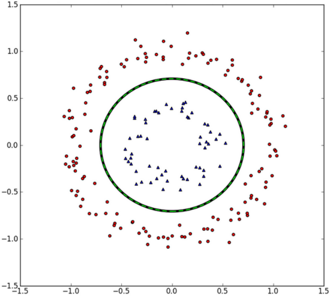


Данные в $\mathbb{R}^2$

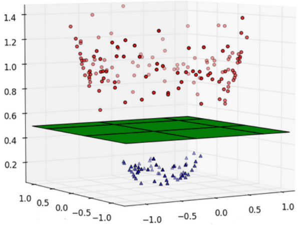


Данные в $\mathbb{R}^3$

Переход к более высокой размерности

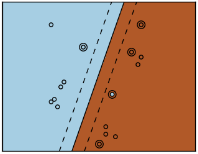


Данные в $\mathbb{R}^2$

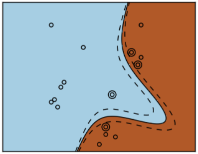


Данные в $\mathbb{R}^3$

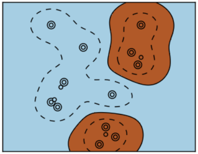
Примеры ядер



Линейное
$\langle x, x' \rangle$



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



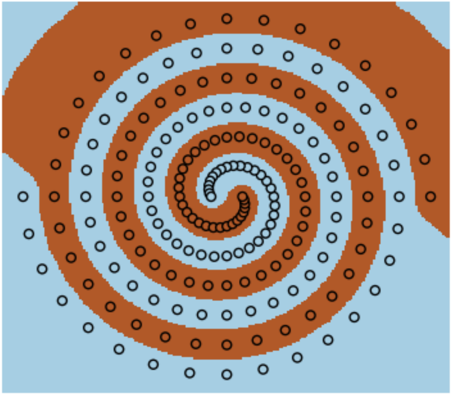
Гауссовское
$exp(-\beta \|x - x'\|^2)$

Пример из Python scikit-learn: http://scikit-learn.org/dev

Примеры ядер

Гауссовское

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



Конструктивные методы получения ядер

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

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

На следующей лекции

– Ликбез по использованию Python, Numpy, ...