Лекция 1
Метод опорных векторов. Ядра.

Метод опорных векторов (SVM)

Повторение: задача классификации

$D = \{(x_i, y_i)\}_{i=1}^{N}$ — набор данных

$x_i \in \mathbb{R}^n$ — признаки

$y_i \in \{+1, -1\}$ — ответы

$a(x; \mu)$ — алгоритм классификации

μ — параметры алгоритма

$L(D, \mu) = \frac{1}{N} \sum\limits_{i=1}^{N} [a(x_i, \mu) \neq y_i]$ — ошибка алгоритма

$\hat{\mu} = \arg\min\limits_{\mu} L(D, \mu)$ — обучение алгоритма

Повторение: задача классификации

$D = \{(x_i, y_i)\}_{i=1}^{N}$ — набор данных

$x_i \in \mathbb{R}^n$ — признаки

$y_i \in \{+1, -1\}$ — ответы

$a(x; \mu)$ — алгоритм классификации

μ — параметры алгоритма

$L(D, \mu) = \frac{1}{N} \sum\limits_{i=1}^{N} [a(x_i, \mu) \neq y_i]$ — ошибка алгоритма

$\hat{\mu} = \arg\min\limits_{\mu} L(D, \mu)$ — обучение алгоритма

Однозначно ли выбирается $\hat{\mu}$?

Повторение: линейный классификатор

$h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b$

$h(x) = 0$ — гиперплоскость

$$a(x; \mu) = \begin{cases} +1, & h(x) > 0; \\ -1, & h(x) < 0. \end{cases}$$

$\mu = (w, b)$

w — вектор весов

b — смещение (bias)

Нормаль гиперплоскости

a_1, a_2 — две произвольные точки на гиперплоскости $h(x) = w^T x + b$.

$$h(a_1) = w^T a_1 + b = 0$$
$$h(a_2) = w^T a_2 + b = 0$$
$$w^T (a_1 - a_2) = 0$$

$(a_1 - a_2)$ — вектор в гиперплоскости
w — нормаль к гиперплоскости

Расстояние от гиперплоскости

Пусть $x \in \mathbb{R}^n$ — произвольная точка.

Обозначим x_P — ее проекция на гиперплоскость h,
r — расстояние от x до гиперплоскости (со знаком)

$$x = x_P + r\frac{w}{\|w\|}$$

Расстояние от гиперплоскости

Пусть $x \in \mathbb{R}^n$ — произвольная точка.

Обозначим x_P — ее проекция на гиперплоскость h,
r — расстояние от x до гиперплоскости (со знаком)

$$x = x_P + r\frac{w}{\|w\|}$$

Как выразить расстояние через параметры h?

Расстояние от гиперплоскости

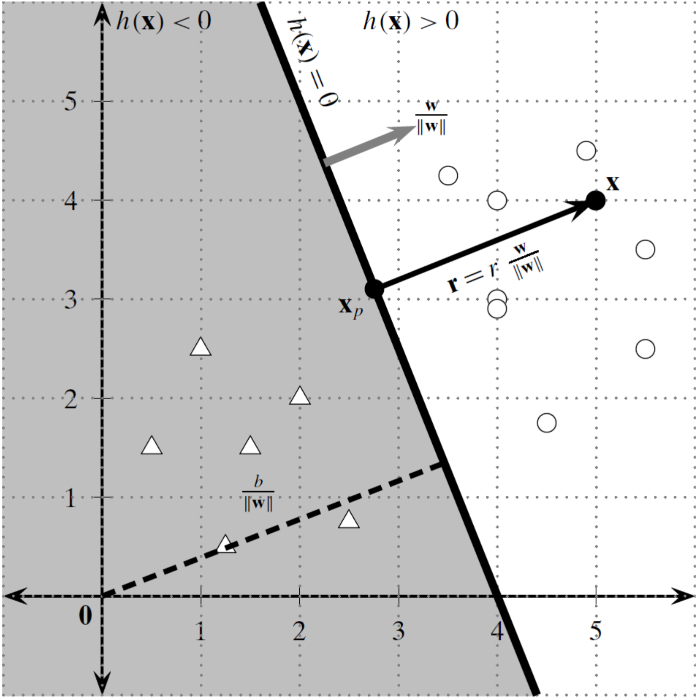
$$x = x_P + r\frac{w}{\|w\|}$$

$$h(x) = w^T\left(x_P + r\frac{w}{\|w\|}\right) + b = w^T x_P + b + r\frac{w^T w}{\|w\|} =$$

$$= h(x_P) + r\|w\| = r\|w\|$$

$$r = \frac{h(x)}{\|w\|}$$

$$|r| = yr = \frac{yh(x)}{\|w\|}$$



$h(\mathbf{x}) < 0$
$h(\mathbf{x}) > 0$
$h(\mathbf{x}) = 0$
$\frac{\mathbf{w}}{\|\mathbf{w}\|}$
$\mathbf{x}$
$\mathbf{x}_p$
$\mathbf{r} = r\ \frac{\mathbf{w}}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
0
1
2
3
4
5

Отступ классификатора

$D = \{(x_i, y_i)\}_{i=1}^N$ — набор данных

$$\delta^* = \min_{x_i} |r_i| = \min_{x_i} \frac{y_i h(x_i)}{\|w\|}$$

δ^* — отступ (margin) классификатора

Вектора, на которых достигается минимальное расстояние, называются опорными.

Отступ классификатора

$D = \{(x_i, y_i)\}_{i=1}^{N}$ — набор данных

$$\delta^* = \min_{x_i} |r_i| = \min_{x_i} \frac{y_i h(x_i)}{\|w\|}$$

δ^* — отступ (margin) классификатора

Вектора, на которых достигается минимальное расстояние, называются опорными.

$$\delta^* \to \max_{w}$$

Каноническая гиперплоскость

$$h(x) = w^T x + b \qquad h'(x) = cw^T x + cb$$

Уравнения задают одну и ту же гиперплоскость

Каноническая гиперплоскость

$$h(x) = w^T x + b \qquad h'(x) = cw^T x + cb$$

Уравнения задают одну и ту же гиперплоскость

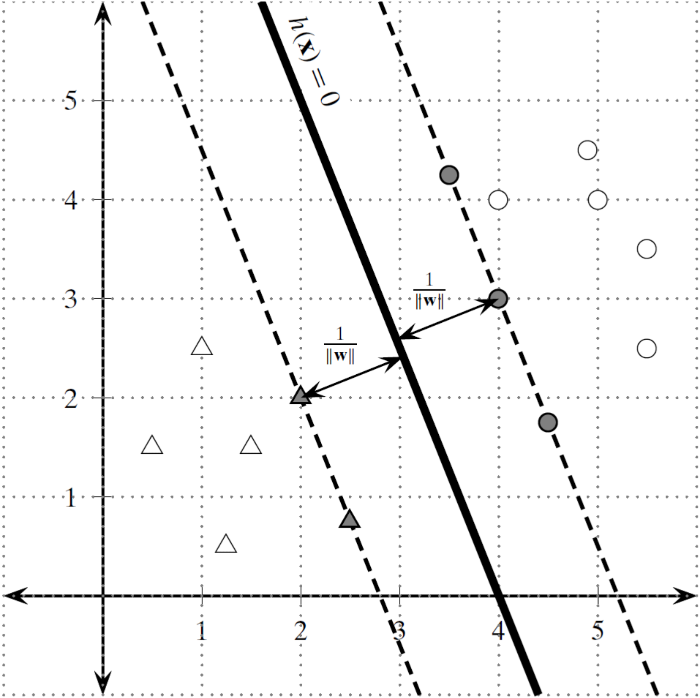
Пусть x_k — опорный вектор. Выберем w, b так, чтобы

$$y_k h(x_k) = 1$$

Тогда

$$r_k = \frac{y_k h(x_k)}{\|w\|} = \frac{1}{\|w\|}$$

(для всех опорных векторов)



$h(\mathbf{x}) = 0$
$\frac{1}{\|\mathbf{w}\|}$
$\frac{1}{\|\mathbf{w}\|}$
5
4
3
2
1
1
2
3
4
5

SVM (линейная разделимость)

$D = \{(x_i, y_i)\}_{i=1}^N$ — обучающая выборка (линейно разделимая)

Классификатор:

$$a(x; w, b) = \begin{cases} +1, & w^T x + b \geqslant 0; \\ -1, & w^T x + b < 0. \end{cases}$$

Задача

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{при условии} \quad y_i(w^T x_i + b) \geqslant 1.$$

$$i = 1, \ldots, N.$$

SVM (линейная разделимость)

$D = \{(x_i, y_i)\}_{i=1}^N$ — обучающая выборка (линейно разделимая)

Классификатор:

$$a(x; w, b) = \begin{cases} +1, & w^T x + b \geqslant 0; \\ -1, & w^T x + b < 0. \end{cases}$$

Задача

$$\min_{w,b} \|w\|^2 \quad \text{при условии} \quad y_i(w^T x_i + b) \geqslant 1.$$

$$i = 1, \dots, N.$$

Отсутствие линейной разделимости

Не существует решений для

$$y_i(w^T x_i + b) \geqslant 1, \quad i = 1, \ldots, N.$$

Отсутствие линейной разделимости

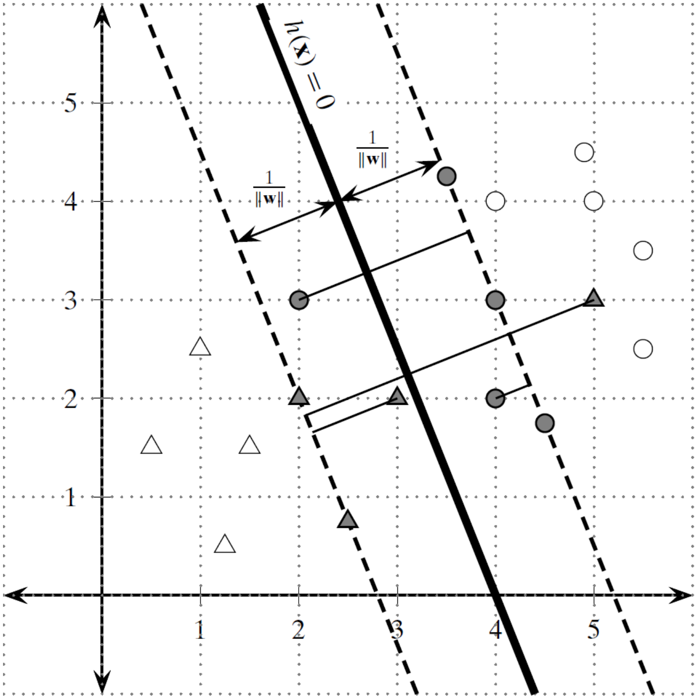
Не существует решений для

$$y_i(w^T x_i + b) \geqslant 1, \quad i = 1, \ldots, N.$$

Разрешим некоторым объектам нарушать условие

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \qquad \xi_i \geqslant 0,$$
$$i = 1, \ldots, N.$$

- $\xi_i = 0$ — обычный объект
- $0 < \xi_i \leqslant 1$ — объект попадает в отступ, но классифицируется верно
- $\xi_i > 1$ — объект классифицируется неверно



$h(\mathbf{x}) = 0$
$\frac{1}{\|\mathbf{w}\|}$
$\frac{1}{\|\mathbf{w}\|}$
5
4
3
2
1
1
2
3
4
5

SVM (Общий случай)

Первая попытка:

$$\min_{w,b} \|w\|^2 \quad \text{при условии}$$

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0.$$

$$i = 1, \dots, N.$$

SVM (Общий случай)

Первая попытка:

$$\min_{w,b} \|w\|^2 \quad \text{при условии}$$

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0.$$

$$i = 1, \ldots, N.$$

Проблема: будет большая ошибка классификации.

SVM (Общий случай)

$$\min_{w,b} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \quad \text{при условии}$$

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0.$$

$$i = 1, \ldots, N.$$

C — параметр регуляризации

SVM (Общий случай)

$$\min_{w,b} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \quad \text{при условии}$$

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0.$$

$$i = 1, \ldots, N.$$

C — параметр регуляризации

- $C \to 0$ — сильная регуляризация, слабо учитываются данные
- $C \to \infty$ — слабая регуляризация, настройка на данные

Преимущества и недостатки SVM

Преимущества и недостатки SVM

Преимущества

- Достаточно эффективное решение
- Высокая обобщающая способность

Недостатки

- Не очень высокая устойчивость к шуму
  Опорные вектора могут быть шумовыми
- Непонятно, как выбирать C
  Обычно выбирается кросс-валидацией

Ядра в SVM

Расширение признаков

Как добиться линейной разделимости?
Ответ: нужно добавить нелинейные признаки.

x — исходные признаки
$\varphi(x)$ — расширенные признаки.

Примеры для $x = (x_1, x_2)$:

- $\varphi(x) = (x_1, x_2, x_1^2, x_2^2, x_1 x_2)$
- $\varphi(x) = (x_1, x_2, \ln x_1, \ln x_2)$

Расширение признаков

Как добиться линейной разделимости?
Ответ: нужно добавить нелинейные признаки.

x — исходные признаки
$\varphi(x)$ — расширенные признаки.

Примеры для $x = (x_1, x_2)$:

- $\varphi(x) = (x_1, x_2, x_1^2, x_2^2, x_1 x_2)$
- $\varphi(x) = (x_1, x_2, \ln x_1, \ln x_2)$

Проблемы при большом числе признаков:

- Вычислительная сложность
- Проклятие размерности

Преобразование SVM

Исходная задача

$$\min_{w,b} \|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad i = 1, \dots, N$$

$$\xi_i \geqslant 0, \quad i = 1, \dots, N$$

Исходная решаюшая функция

$$h(z) = w^T z + b$$

Преобразование SVM

Преобразованная задача

$$\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \right)$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Преобразованная решающая функция

$$h(z) = \sum_{\alpha_i > 0} \alpha_i y_i x_i^T z + \operatorname*{average}_{i, 0 < \alpha_i < C} \left(y_i - \sum_{\alpha_j > 0} \alpha_j y_j x_j^T x_i \right)$$

Преобразование SVM

Преобразованная задача

$$\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \right)$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Преобразованная решающая функция

$$h(z) = \sum_{\alpha_i > 0} \alpha_i y_i \langle x_i, z \rangle + \operatorname*{average}_{i, 0 < \alpha_i < C} \left(y_i - \sum_{\alpha_j > 0} \alpha_j y_j \langle x_j, x_i \rangle \right)$$

Преобразование SVM

Преобразованная задача

$$\max_{\alpha}\left(\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j)\right)$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

Преобразованная решающая функция

$$h(z) = \sum_{\alpha_i > 0}\alpha_i y_i K(x_i, z) + \operatorname*{average}_{i, 0 < \alpha_i < C}\left(y_i - \sum_{\alpha_j > 0}\alpha_j y_j K(x_j, x_i)\right)$$

Расширение признаков и ядра

x — исходные признаки
$\varphi(x)$ — расширенные признаки.

$K(x, y) = \langle \varphi(x), \varphi(y) \rangle$ — ядро.

Расширение признаков и ядра

x — исходные признаки
$\varphi(x)$ — расширенные признаки.

$K(x,y) = \langle \varphi(x), \varphi(y) \rangle$ — ядро.

Норма

$$\|\varphi(x)\|^2 = \langle \varphi(x), \varphi(x) \rangle = K(x,x)$$

Расстояние

$$\begin{aligned}\rho(\varphi(x), \varphi(y))^2 = \|\varphi(x) - \varphi(y)\| = \\ = \langle \varphi(x) - \varphi(y), \varphi(x) - \varphi(y) \rangle = \\ = \langle \varphi(x), \varphi(x) \rangle + \langle \varphi(y), \varphi(y) \rangle - 2\langle \varphi(x), \varphi(y) \rangle = \\ = K(x,x) + K(y,y) - 2K(x,y).\end{aligned}$$

Ядровые методы

Идея: работать только с ядрами.

Пример

$x = (x_1, x_2)$

$\varphi(x) = (x_1^2, x_2^2, \sqrt{2}\, x_1 x_2)$

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle =$$
$$= x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 y_1 x_2 y_2 = (x_1 y_1 + x_2 y_2)^2 = \langle x, y \rangle^2$$

Ядровые методы

Идея: работать только с ядрами.

Пример

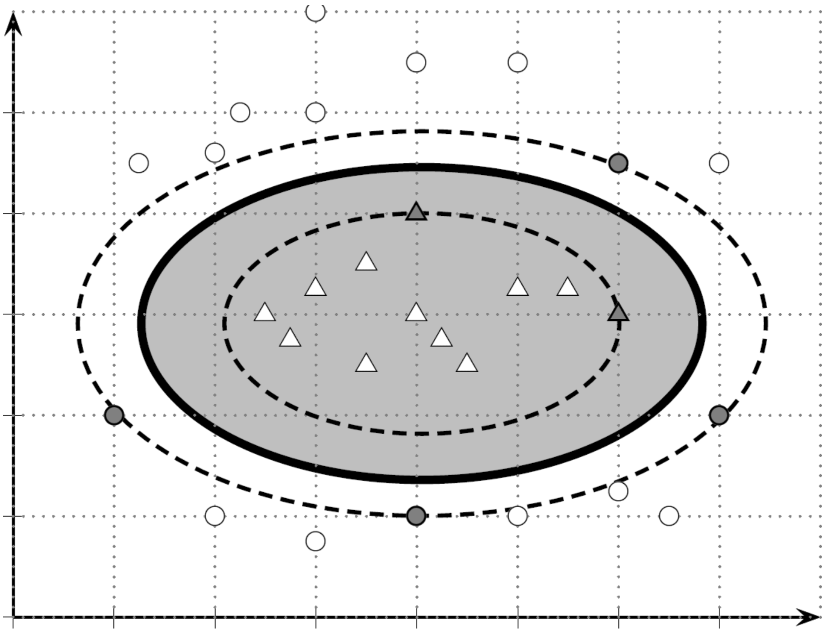
$x = (x_1, x_2)$

$\varphi(x) = (x_1^2, x_2^2, \sqrt{2}\, x_1 x_2)$

$$K(x,y) = \langle \varphi(x), \varphi(y) \rangle =$$
$$= x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 y_1 x_2 y_2 = (x_1 y_1 + x_2 y_2)^2 = \langle x, y \rangle^2$$

Другие ядра

- $K(x,y) = \langle x, y \rangle^d$
- $K(x,y) = (\langle x, y \rangle + 1)^d$
- $K(x,y) = e^{-\|x-y\|^2}$



Где еще применяются ядра

Как составлять ядра

Стандартные ядра

- $K(x, y) = 1$
- $K(x, y) = \langle x, y \rangle$
- $K(x, y) = e^{-\|x-y\|^2}$
- $K(x, y) = e^{-\|x-y\|}$

Преобразование ядер

- $K(x, y) = K_1(x, y) K_2(x, y)$
- $K(x, y) = C_1 K_1(x, y) + C_2 K_2(x, y)$

Напоминание: метод главных компонент

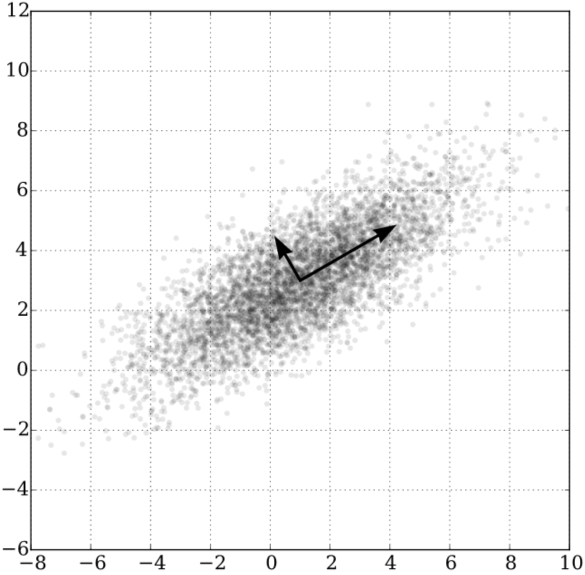
X — матрица объекты-признаки
Строим новые признаки:

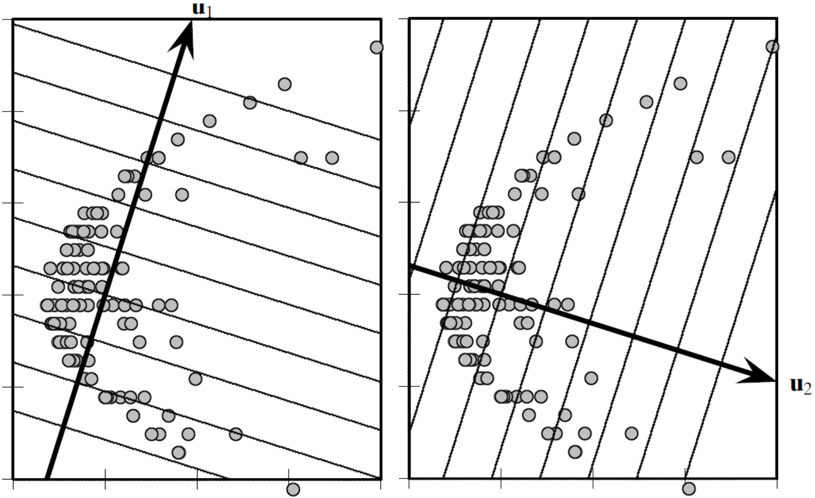
$$Z = XW^T$$

Хотим выбрать D самых «интересных» признаков, т.е. с наибольшей дисперсией:

$$\sum_{k=1}^{D} w_k^T X^T X w_k \to \max_W$$

$$W^T W = E$$


12
10
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6
8
10



$\mathbf{u}_1$
$\mathbf{u}_2$

Ядровой метод главных компонент

$$\sum_{k=1}^{D} w_k^T X^T X w_k \to \max_W$$

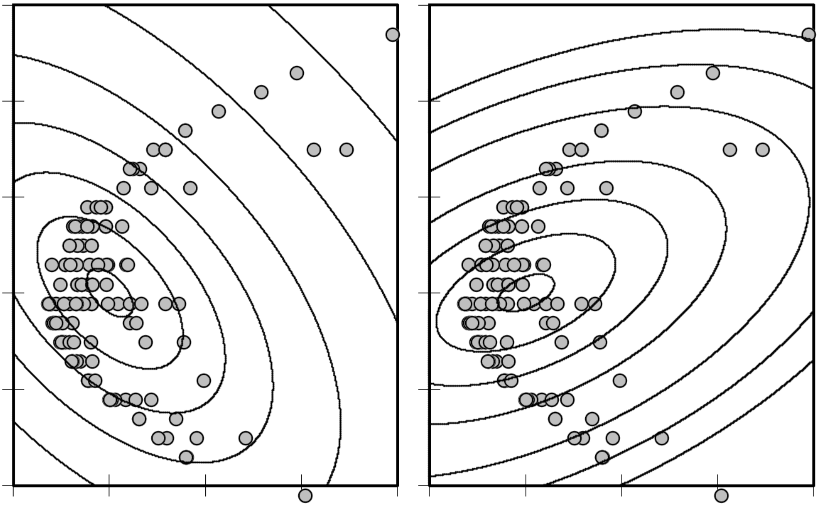
$$W^T W = E$$

Матрица коварияции (x_i — столбцы)

$$X^T X = \sum_i \sum_j x_i^T x_j$$

Матрица коварияции с ядром:

$$X^T X = \sum_i \sum_j K(x_i, x_j)$$



Ядра для строк

S — некоторая последовательность «слов»
(например, текст или последовательность ДНК)

Ядра для строк

S — некоторая последовательность «слов»
(например, текст или последовательность ДНК)

Возможные признаки
(например, $S =$ «dog and cat and cow»)

- Сколько раз встретилось каждое слово
  (dog: 1, and: 2, cat: 1, cow: 1)

Ядра для строк

S — некоторая последовательность «слов»
(например, текст или последовательность ДНК)

Возможные признаки
(например, S = «dog and cat and cow»)

- Сколько раз встретилось каждое слово
(dog: 1, and: 2, cat: 1, cow: 1)
- Сколько раз встретились последовательности слов длины d
(«dog and»: 1, «and cat»: 1, ..., «cow and»: 0, ...)

Ядра для строк

S — некоторая последовательность «слов»
(например, текст или последовательность ДНК)

Возможные признаки
(например, S = «dog and cat and cow»)

- Сколько раз встретилось каждое слово
(dog: 1, and: 2, cat: 1, cow: 1)
- Сколько раз встретились последовательности слов длины d
(«dog and»: 1, «and cat»: 1, ..., «cow and»: 0, ...)
- Сколько раз встретилась каждая последовательность
(..., and: 2, ..., «dog and»: 1, ... «dog and cat»: 1, ...)

Ядра для строк

- Очень много признаков
- Много нулевых значений

Ядра для строк

- Очень много признаков
- Много нулевых значений

Ядро: число совпадений последовательностей

A = «dog and cat and cow»
B = «cat and cat and cat»

$$
\begin{aligned}
K(A, B) &= \\
&= 1 \times 0 + 2 \times 2 + 1 \times 3 + 1 \times 0 + \\
&+ 1 \times 0 + 1 \times 2 + 1 \times 2 + 1 \times 0 + \\
&+ 1 \times 0 + 1 \times 1 + 1 \times 0 + \\
&+ 1 \times 0 + 1 \times 0 + \\
&+ 1 \times 0 = \\
&= 12.
\end{aligned}
$$

Материалы

Метод опорных векторов (SVM)

- Zaki, Meira. Data Mining and Analysis, гл. 21.

Ядра

- Zaki, Meira. Data Mining and Analysis, гл. 5.