Отбор информативных признаков в методе опорных векторов

Андрющенко Анастасия Михайловна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: Коробейников А.И.
Рецензент: к.ф.-м.н., д. Алексеева Н.П.

Санкт-Петербург
2010г.

Введение

Задача отбора информативных признаков в задачах классификации способствует:

- уменьшению ошибки предсказания;
- определению значимых признаков;
- уменьшению размерности данных.

В работе используется модификация метода опорных векторов (support vector machines, SVM (Vapnik, 1995)).

SVM. Задача классификации

Данные:

	X_1	$\cdots$	X_m
$\boldsymbol{x}_1$	$x_1^{(1)}$	$\cdots$	$x_1^{(m)}$
$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_n$	$x_n^{(1)}$	$\cdots$	$x_n^{(m)}$

$\boldsymbol{x}_i \in \mathbb{R}^m$ — наблюдения

$(X_1, \ldots, X_m)$ — признаки

Метки: $\boldsymbol{x}_i \to y_i \in \{\pm 1\}$

Задача классификации:

построить f : $f(\boldsymbol{x}_i) = y_i \quad \forall i = 1 \ldots m.$

Задача отбора признаков:

$(X_1, \ldots, X_m) \to (X_{i_1}, \ldots, X_{i_l})$

SVM. Задача классификации

Данные:

	X_1	$\cdots$	X_m
$\boldsymbol{x}_1$	$x_1^{(1)}$	$\cdots$	$x_1^{(m)}$
$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_n$	$x_n^{(1)}$	$\cdots$	$x_n^{(m)}$

$\boldsymbol{x}_i \in \mathbb{R}^m$ — наблюдения

$(X_1, \ldots, X_m)$ — признаки

Метки: $\boldsymbol{x}_i \to y_i \in \{\pm 1\}$

Задача классификации:

$$\text{построить } f: \quad f(\boldsymbol{x}_i) = y_i \quad \forall i = 1 \ldots m.$$

Задача отбора признаков:

$$(X_1, \ldots, X_m) \to (X_{i_1}, \ldots, X_{i_l})$$

SVM. Задача классификации

Данные:

	X_1	$\cdots$	X_m
$\boldsymbol{x}_1$	$x_1^{(1)}$	$\cdots$	$x_1^{(m)}$
$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_n$	$x_n^{(1)}$	$\cdots$	$x_n^{(m)}$

$\boldsymbol{x}_i \in \mathbb{R}^m$ — наблюдения

$(X_1, \ldots, X_m)$ — признаки

Метки: $\boldsymbol{x}_i \to y_i \in \{\pm 1\}$

Задача классификации:

$$\text{построить } f: \quad f(\boldsymbol{x}_i) = y_i \quad \forall i = 1 \ldots m.$$

Задача отбора признаков:

$$(X_1, \ldots, X_m) \to (X_{i_1}, \ldots, X_{i_l})$$

Линейные SVM

Целевая функция SVM:

$$f_{\boldsymbol{\sigma}}(\boldsymbol{x}) = \operatorname{sign}\left(\langle \boldsymbol{w}, \boldsymbol{x}\rangle - b\right).$$

Параметры $(\boldsymbol{w}, b)$ получены решением:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \longrightarrow \min_{\boldsymbol{w},b,\boldsymbol{\xi}}$$

$$y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - b\right) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0, \quad i = 1, \ldots, n.$$

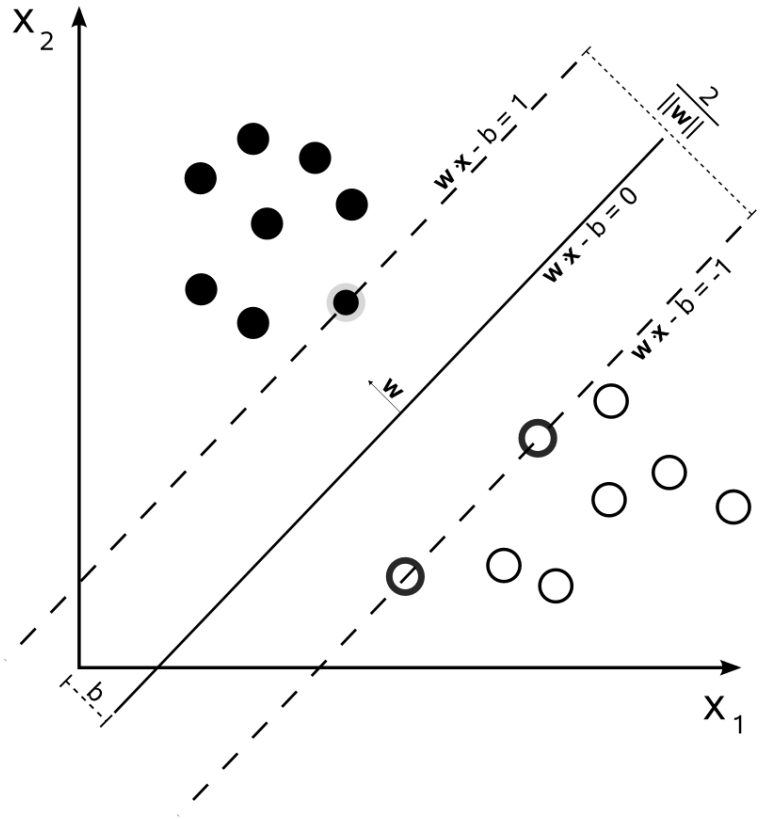
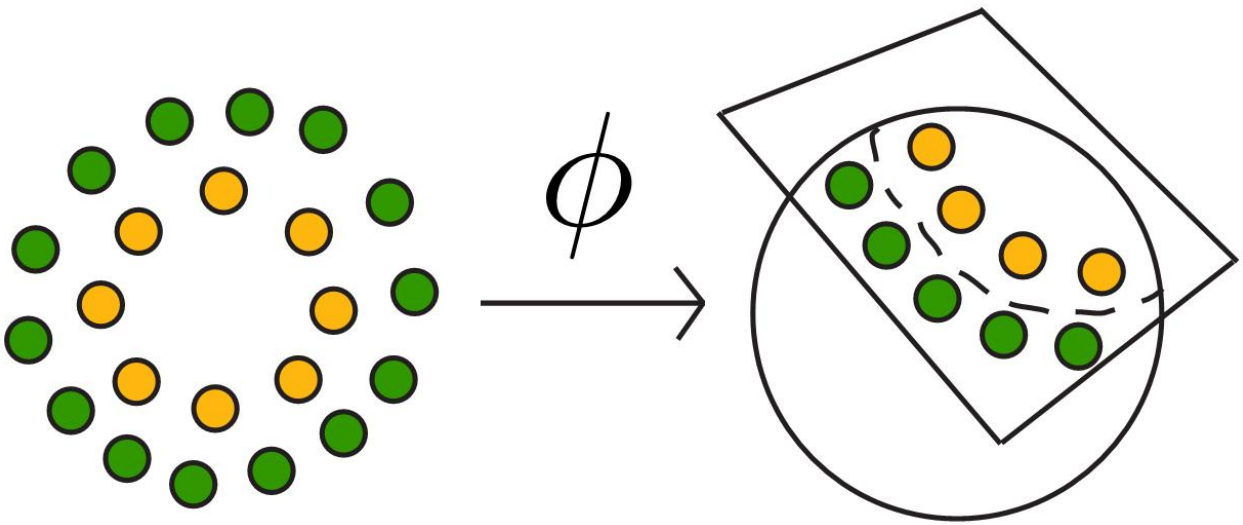


Рис.: Гиперплоскость.

Нелинейные SVM

Данные не всегда отделимы в исходном пространстве
$\Rightarrow$ они рассматриваются в спрямляющем:

$$\phi : \mathbb{R}^m \to \mathcal{H}$$



Нелинейные SVM. Kernel trick

Введем ядро:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}} .$$

Целевая функция SVM:

$$f_{\boldsymbol{\sigma}}(\boldsymbol{x}) = \operatorname{sign}\left(\langle \boldsymbol{w}, \phi_{\boldsymbol{\sigma}}(\boldsymbol{x}) \rangle_{\mathcal{H}} - b \right) = \operatorname{sign}\left(\sum_{i=1}^{n_s} \alpha_i y_i K_{\boldsymbol{\sigma}}(\boldsymbol{s}_i, \boldsymbol{x}) - b \right).$$

Примеры ядер:

- Линейное: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$.
- Полиномиальное: $K_{\gamma,r,d}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + r)^d$, $\gamma > 0$.
- Экспоненциально-радиальное (RBF — radial basis function): $K_\gamma(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}$.

SVM. Отбор признаков

Рассматриваются ядра:

$$K(\boldsymbol{x}, \boldsymbol{z}) = \ell\left(\|\boldsymbol{\sigma}(\boldsymbol{x} - \boldsymbol{z})\|^2\right) = \ell\left(\sum_{k=1}^{m} \sigma_k^2 (x_k - z_k)^2\right),$$

где ℓ — монотонная функция. Масштабирующие коэффициенты $\{\sigma_k\}_{k=1}^m$ отражают степень влияния признака на результат классификации.

Задача:

1. Реализация алгоритма пересчета масштабирующих коэффициентов (Grandvalet, Canu, 2003) на R.
2. Проверка и исследование на модельных данных.
3. Анализ реальных данных из кардиологии.

SVM. Отбор признаков

Рассматриваются ядра:

$$K(\boldsymbol{x}, \boldsymbol{z}) = \ell\left(\|\boldsymbol{\sigma}(\boldsymbol{x} - \boldsymbol{z})\|^2\right) = \ell\left(\sum_{k=1}^{m} \sigma_k^2 (x_k - z_k)^2\right),$$

где ℓ — монотонная функция. Масштабирующие коэффициенты $\{\sigma_k\}_{k=1}^m$ отражают степень влияния признака на результат классификации.

Задача:

1. Реализация алгоритма пересчета масштабирующих коэффициентов (Grandvalet, Canu, 2003) на R.
2. Проверка и исследование на модельных данных.
3. Анализ реальных данных из кардиологии.

SVM. Отбор признаков

Параметры классификатора: C, σ_0.

Обучающий критерий:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \longrightarrow \min_{\boldsymbol{\sigma},\boldsymbol{w},b,\boldsymbol{\xi}},$$

$$y_i(\langle \boldsymbol{w}, \phi_{\boldsymbol{\sigma}}(\boldsymbol{x}_i)\rangle_{\mathcal{H}} - b) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0, \quad i = 1, \ldots, n,$$

$$\frac{1}{m}\sum_{k=1}^{m}\sigma_k^2 = \sigma_0^2.$$

Подход к решению

Вместо сложной задачи итерационно решается несколько более простых:

1. Зафиксировать $\boldsymbol{\sigma}$ и построить SVM.
2. Метод позволяет получить $\boldsymbol{w}$, $\sum\limits_{i=1}^{n} \xi_i$ и $\partial \sum\limits_{i=1}^{n} \xi_i / \partial \boldsymbol{\sigma}$ в виде функций от $\boldsymbol{\sigma}$:

$$g(\boldsymbol{\sigma}) = \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i.$$

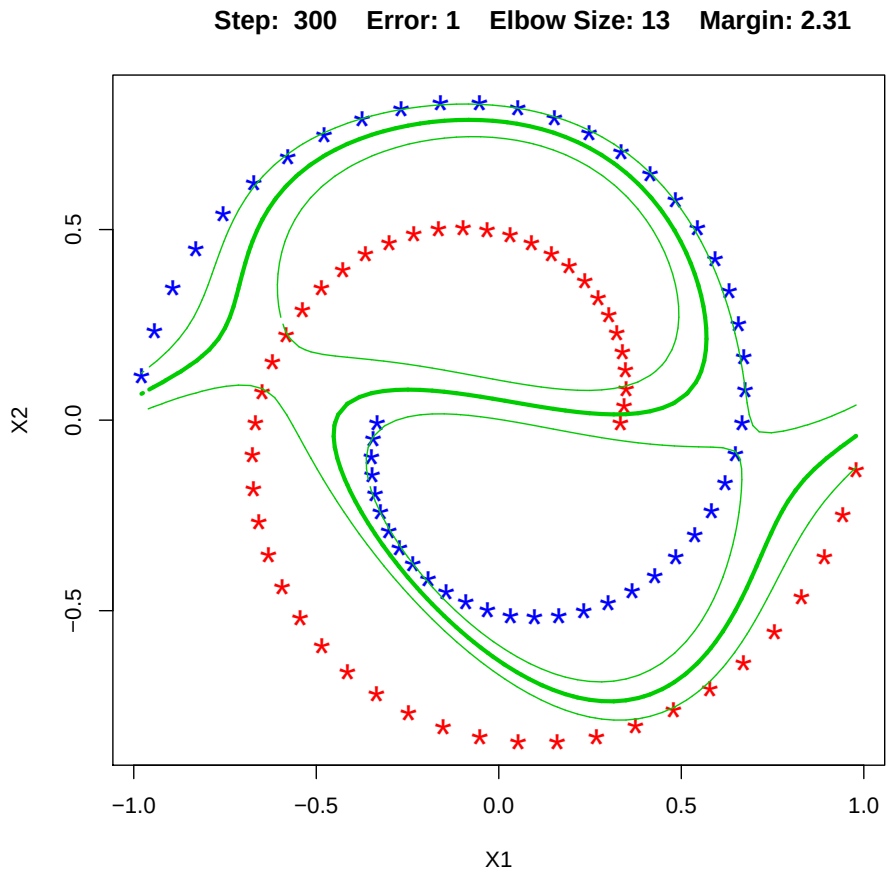
$$g(\boldsymbol{\sigma}) \to \min_{\boldsymbol{\sigma}}.$$

3. На шаге l, начиная с $\boldsymbol{\sigma}^{(l)}$, вычислить оптимальные $(\hat{\boldsymbol{w}}(\boldsymbol{\sigma}^{(l)}), \hat{b}(\boldsymbol{\sigma}^{(l)}))$. $\boldsymbol{\sigma}^{(l+1)}$ определяется с помощью метода сопряженных градиентов.

Модельные данные

Данные:

1. Координаты двух спиралей. Значимость признаков одинакова.



Модельные данные

Данные:

1. Координаты двух спиралей. Значимость признаков одинакова.
2. Данные из (Chapelle, Vapnik, 2002).
   - Линейно отделимые данные. 6 признаков из 10 значимы, остальные шум.
   - Линейно не отделимые данные. 2 признака из 10 значимы, остальные шум.

Полученные масштабирующие коэффициенты σ соответствуют характеру данных.

Реальные данные

Данные о 422 пациентах, перенесших операцию на открытом сердце.

Значимые признаки определялись для следующих задач:

- бинарная классификация: пациенты с наличием или отсутствием ПКТС;
- тернарная классификация: ПКТС отсутствует/ранний/поздний.

50 признаков: 22 количественных, 28 категориальных.

	Возр.	ИМТ	Кардиопл.	ФВ	Эозин.	Л1	Л7
1	67	23	13.50	68	0	8.70	7.10
2	57	20	6.30	68	1	9.90	6.70
3	54	37	9.50	64	1	14.70	7.00
4	57	28	13.50	69	0	17.20	10.40
				...			

Результаты

Бинарная задача

Всего 34 значимых признака:

- 14 количественных;
- 17 категориальных;
- 3 градации у двух категориальных признаков.

Адреналин	0.04*	Дренир.1	0.21	Тип опер.2	0.11
Возраст	0.22	Дренир.2	0.13	Тип опер.3	0.07*
Аллергия	0.16	Дренир.3	0.09*	Тип опер.4	0.02*
Анемия	0.01*	Дренир.пот.	0.19	Тип опер.5	0.01*
Антиког.тер.	0.08*	Дренир.вр.	0.05*	Тип опер.6	0.06*
Переж.аорты	0.14	Фр.выброса	0.17	Веноз.застой	0.13
Арт.шунт	0.05*	Эозинофилы	0.24	Темп.прайм.	0.19
Аутоимунные	0.01*	СОЭ	0.19	МНО	0.06*
ИМТ	0.24	EuroSCORE	0.19	Апп.иск.кр.	0.14
Число шунтов	0.09*	EuroSCORE2	0.09*	Реперфузия	0.07*
Кардиоплегия	0.08*	ССН	0.14	Дыхат.недост.	0.13
Коронарогр.	0.28	Гиперглик.	0.16	Наруш.ритма	0.11
		...			

Значимые признаки, * не значимые.

Результаты

Тернарная задача

Рассматривались 3 бинарные задачи. Для каждой из них получены масштабирующие коэффициенты.

Признак	n-el	l-ne	e-nl	Признак	n-el	l-ne	e-nl
Адреналин	0.01*	0.68	0.78	СОЭ	0.01*	0.01*	0.51
Возраст	0.01*	0.99	0.98	EuroSCORE	1.77	0.34	0.37
Переж.аорты	0.01*	0.25	0.01*	EuroSCORE2	1.15	0.45	0.61
ИМТ	3.16	1.02	1.18	Ин.поддержка	0.01*	0.01*	0.07*
Число шунтов	0.01*	0.31	0.19	Лейк.д.1	1.05	0.65	0.68
Кардиоплегия	0.01*	0.16	0.25	Лейк.д.7	2.35	0.63	0.18
Хрон.серд.нед.	0.01*	0.04	0.21	Мезатон	0.01*	0.01*	0.01*
Дрен.потери	0.01*	0.94	0.97	Темп.прайм.	0.01*	0.49	0.32
Дренир.вр.	3.31	0.41	0.69	МНО	1.29	0.18	0.43
Фр.выброса	4.64	0.97	0.71	Апп.иск.кр.	3.41	0.74	1.01
Эозинофилы	0.01*	0.34	0.49	Реперфузия	0.01*	0.27	0.46

Значимые признаки, * не значимые.

Заключение

Итоги:

- Реализован алгоритм отбора информативных признаков на языке R.
- Его работа проверена и исследована на модельных данных.
- Осуществлен анализ реальных данных.

Перспективы:

- Создание приложения.
- Комбинирование алгоритма с другими методами.