Отбор информативных признаков в методе опорных векторов

Тикка Анна Алексеевна, гр. 522

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

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

2014г.

Постановка задачи классификации

- $\mathbf{X} \subset \mathbb{R}^d$ — множество наблюдений
- $\mathbf{Y} = \{-1, 1\}$ — бинарная классификация
- $(x_1, y_1), \ldots, (x_n, y_n)$ — обучающая выборка, $x_i \in \mathbf{X}, y_i \in \mathbf{Y}$
- $y_i = g(x_i),\ i = 1 \ldots n,\ g : \mathbf{X} \to \mathbf{Y}$
- Задача классификации: аппроксимировать зависимость g на всем пространстве $\mathbf{X}$

SVM. Линейный классификатор.

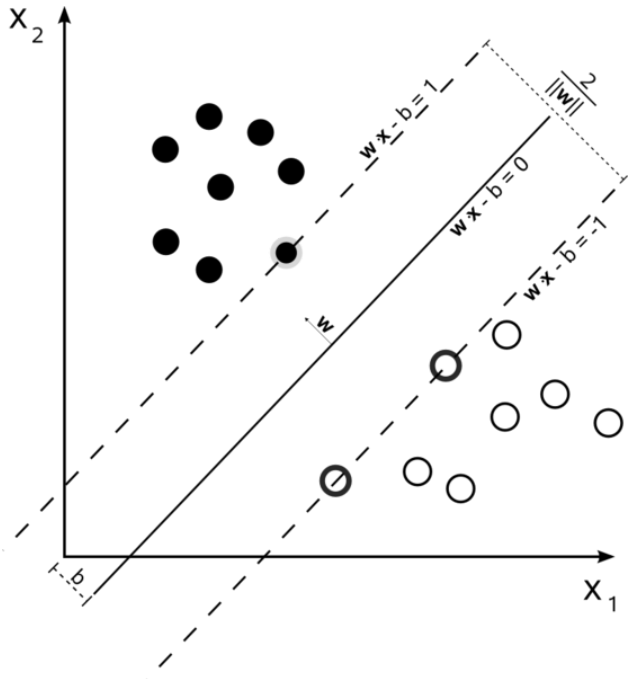
- Метод опорных векторов (SVM) (Vapnik, 1995)
- Разделяющая гиперплоскость:

$$f(x) = \langle w, x \rangle + b = 0$$

- $y = \text{sign}(f(x))$
- Задача оптимизации:

$$\frac{1}{2}\|w\|^2 \to \min_{w,b}$$

$$y_i(\langle w, x_i \rangle + b) \geq 1, \quad i = 1 \ldots n$$



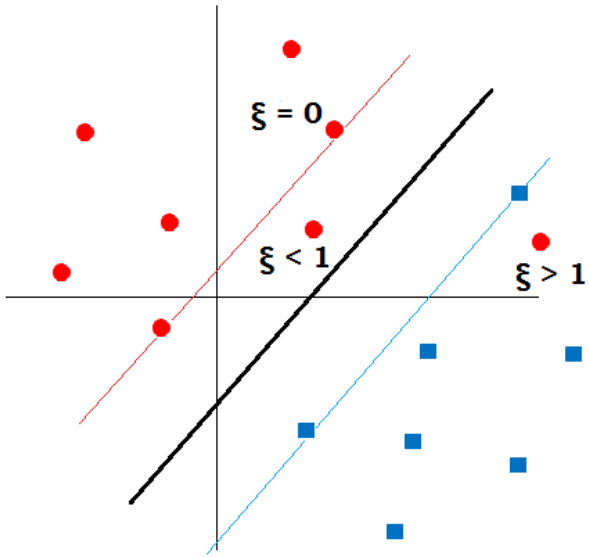
SVM. Нарушение ограничений.

- $\xi_i \geq 0$ — степень нарушения ограничений
  C — штраф за нарушение ограничений
- Задача оптимизации:

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

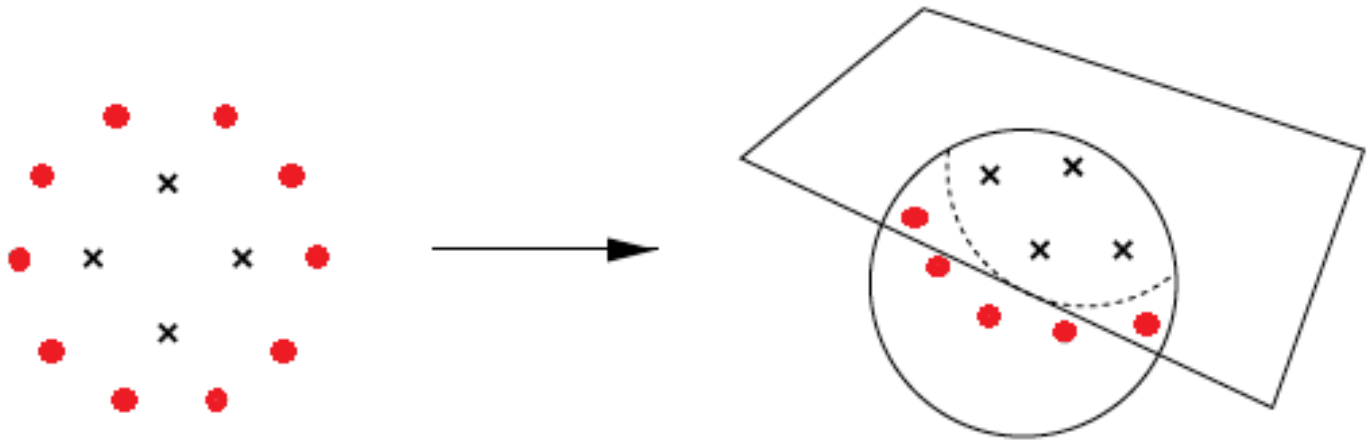
$$y_i(\langle w, x_i\rangle + b) \geq 1 + \xi_i$$

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



Kernel trick. Спрямляющее пространство.

- Спрямляющее отображение: $\varphi : \mathbf{X} \to \mathbf{H}$, $\mathbf{H}$ — гильбертово
- Решение ищем в $\mathbf{H}$, используется только $K(x, x')$
- $K(x, x') = \langle \varphi(x), \varphi(x') \rangle_H$ — функция ядра

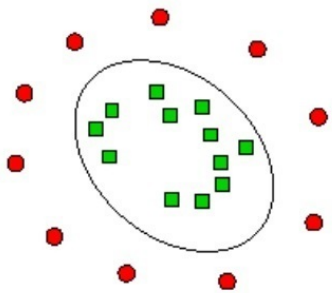


- Примеры ядер:
  - полиномиальное ядро
    $K(x, x') = (\langle x, x' \rangle + 1)^d$, d — степень полинома
  - радиальная базисная функция
    $K(x, x') = e^{-\gamma \|x - x'\|^2}$, $\gamma > 0$

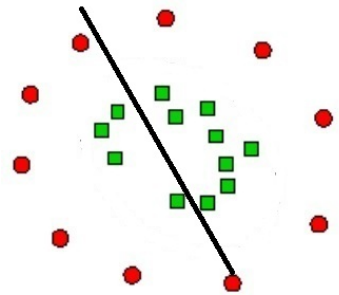
Подбор параметров. Проблема переобучения.

- Параметр C, параметры ядра
- Цель подбора: уменьшить ошибку классификации:

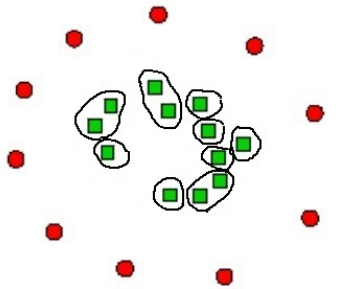
$$\sum_{i=1}^{n} [y_i \neq f(x_i)]$$



а)
Оптимальное разделение



б)
Недообучение



в)
Переобучение

Рис.: Примеры классификатора для двух признаков.

Задачи

- Применить SVM к реальным кардиологическим данным

  Проблема: Возникает переобучение: все наблюдения являются опорными векторами, ошибка классификации: 0

- Рассмотреть робастые модификации стандартного SVM
- Отобрать информативные признаки

Отбор информативных признаков

- Исходные данные могут содержать неинформативные признаки, которые ухудшают точность предсказания или не влияют на результат классификации
- Это может привести к увеличению ошибки классификации, увеличению времени вычислений, переобучению
- Подходы к решению задачи
  1. Пошаговый метод
     Критерий отбора — процент ошибок классификации.
  2. Одновременный отбор признаков
     SCAD SVM (Zhang, 2006), Elastic SCAD SVM (Becker, 2011)
     Критерий отбора — веса признаков

SVM как модель со штрафом

- Задача SVM:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{w,b,\xi}$$
$$y_i(\langle w, x_i\rangle + b) \geq 1 + \xi_i$$
$$\xi_i \geq 0, \quad i = 1\ldots n$$

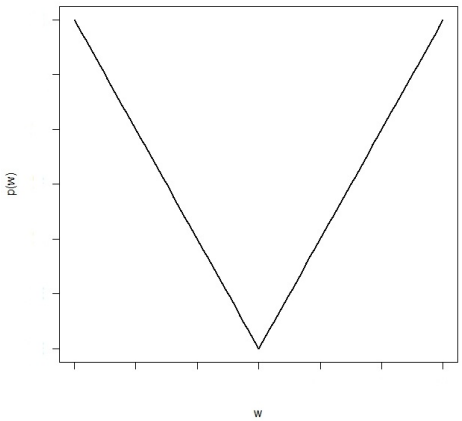
- $f(x) = \langle w, x\rangle + b$
  $p(w) = \lambda\|w\|^2$ — штрафная функция
- Эквивалентная формулировка:

$$p(w) + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(x_i)]_+ \to \min_{w,b}$$

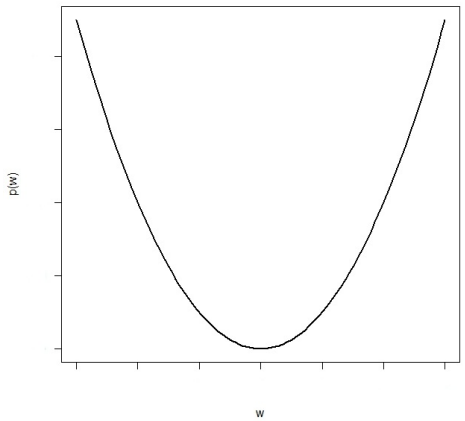
Штрафные функции

$$p(w)_{L^1} = \lambda \|w\|_1$$

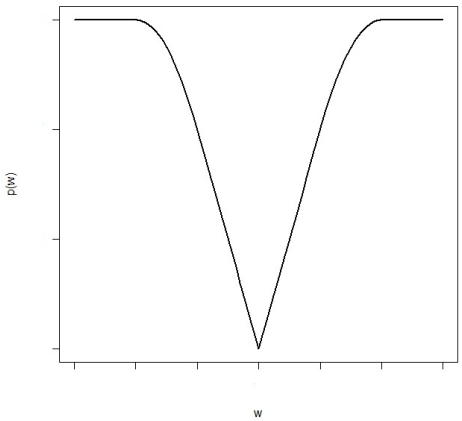
$$p(w)_{SCAD} = \begin{cases} \lambda|w|, & |w| \leq \lambda \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)}, & \lambda < |w| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2}, & |w| > a\lambda \end{cases}$$



а) L^1



б) L^2



в) SCAD

Сравнение штрафных функций

- L^2 не может быть использована для отбора признаков (Bradley, 1998)
- Штрафная функция с L^1 нормой позволяет отбирать информативные признаки
- Преимущества SCAD
  - При малых значениях w функция SCAD соответствует L^1
  - При больших значениях w SCAD использует в качестве штрафа константу, что позволяет получить устойчивость к выделяющимся наблюдениям

SCAD SVM. Elastic SCAD SVM.

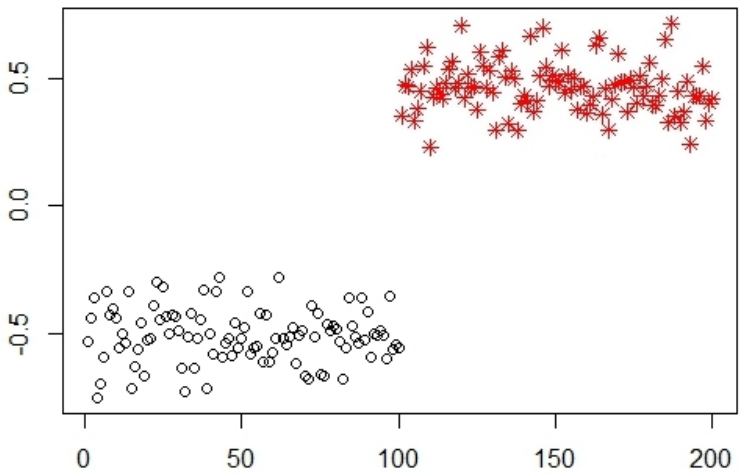
- SCAD SVM

$$\sum_{j=1}^{q} p_{\lambda}\left(w_j\right)_{SCAD} + \frac{1}{n}\sum_{i=1}^{n}\left[1 - y_i\left(b + wx_i\right)\right]_+ \to \min_{w,b}$$

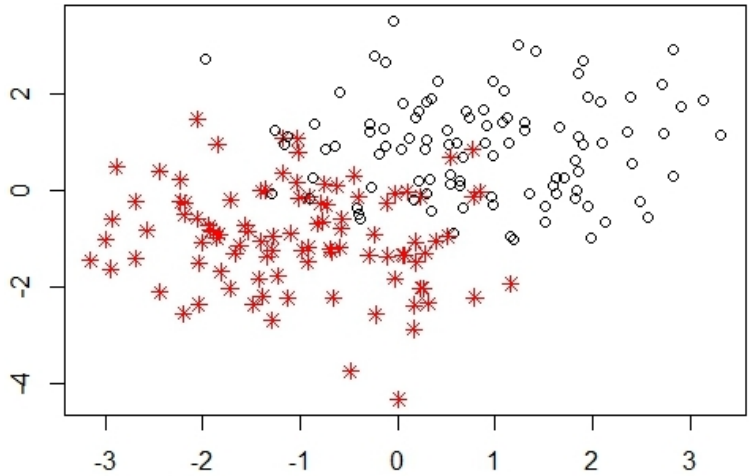
- Elastic SCAD SVM

$$\sum_{j=1}^{q} p_{\lambda_1}\left(w_j\right)_{SCAD} + \lambda_2\|w\|_1 + \frac{1}{n}\sum_{i=1}^{n}\left[1 - y_i\left(b + wx_i\right)\right]_+ \to \min_{w,b}$$

Применение методов на модельных данных



а)



б)

а) 100 наблюдений, 10 дополнительных признаков, ошибка классификации: 0

б) 100 наблюдений, 10 дополнительных признаков, ошибка классификации SCAD SVM: 0.07, Elastic SCAD SVM: 0.06

Информативными выбраны изначальные признаки.

Постперикардиотомный синдром (ПКТС)

- 428 пациента, перенесшие операцию на открытом сердце
- 50 признаков описывают состояние пациентов до операции и во время нее
- Две группы:
  1. В послеоперационном периоде развился ПКТС — 257 человек
  2. Не выявлено клинических проявлений ПКТС — 171 человек
- Отдельно рассматриваются: подгруппа мужчин, подгруппа мужчин с типом операции — коронарное шунтирование

Результаты. SCAD SVM.

Подгруппа мужчин с типом операции — коронарное шунтирование, ошибка классификации: 0.27

- Лейкоциты на 7 сутки
- Коронография
- Эозинофилы
- ИМТ
- Шунты
- Фракция выброса
- ЭКК
- Температура
- Время реперфузии
- Время пережатия аорты
- Длительность дренирования раны
- Гипертоническая болезнь
- Гипергликемия п/о
- Длительность кардиоплегии
- Трасилол
- Инфекционный процесс в п/о периоде

50 признаков $\rightarrow$ 16 признаков

Результаты. Elastic SCAD SVM.

Подгруппа мужчин с типом операции — коронарное шунтирование, ошибка классификации: 0.25

- Лейкоциты на 7 сутки
- Коронография
- Эозинофилы
- ИМТ
- Шунты
- Фракция выброса
- ЭКК
- Температура
- Время реперфузии
- Время пережатия аорты
- Длительность дренирования раны
- Гипертоническая болезнь
- Гипергликемия п/о
- Длительность кардиоплегии
- Лейкоциты в 1 сутки
- Возраст
- СОЭ

50 признаков $\rightarrow$ 17 признаков