Метод опорных векторов

Лекция N 7 курса
“Алгоритмы для Интернета”

Юрий Лифшиц

ПОМИ РАН - СПбГУ ИТМО

Осень 2006

Рекорд

Книга 1995 года “The Nature of Statistical Learning Theory” Владимира Вапника цитируется более **6400** раз

Статистика Google Scholar

План лекции

1. Оптимальная разделяющая гиперплоскость
   - Случай линейной разделимости
   - Шумы и штрафы
2. Обучение машины опорных векторов
   - Постановка задачи квадратичной оптимизации
   - Метод последовательных мини-оптимизаций
3. Расширение пространства признаков

Часть I

Формальная постановка задачи классификации

Как выбрать лучший линейный классификатор?

Как определить уровень ошибки линейного классификатора?

Абстрактная задача классификации

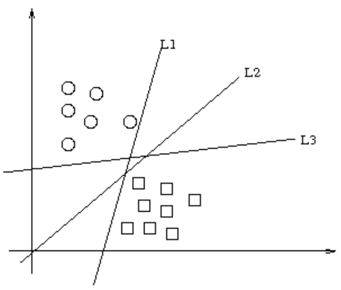
Объекты: вектора n-мерного пространства
Каждая **координата** соответствует признаку
Категории: сегодня рассматриваем только бинарную классификацию (две категории)

Учебная коллекция: даны документы (вектора) $x_1, \dots, x_n$ и их принадлежность к двум классам $y_1, \dots, y_n \in \{1, -1\}$

Цель ученых: построить по учебной коллекции **оптимальное** классифицирующее правило **заданного** вида.

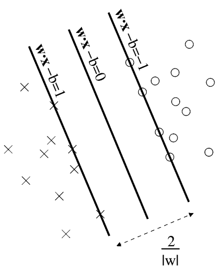
Линейный классификатор (1/3)

$w \cdot x_i > b \Rightarrow y_i = 1$
$w \cdot x_i < b \Rightarrow y_i = -1$



Линейный классификатор (2/3)

Разделение полосой:
$w \cdot x_i \geq b + \varepsilon \Rightarrow y_i = 1$
$w \cdot x_i \leq b - \varepsilon \Rightarrow y_i = -1$



Линейный классификатор (3/3)

Домножив w на $\frac{1}{\varepsilon}$, можем считать $\varepsilon = 1$:
$w \cdot x_i - b \geq 1 \Rightarrow y_i = 1$
$w \cdot x_i - b \leq -1 \Rightarrow y_i = -1$

Как выбрать наилучшую разделяющую полосу?

Выберем самую **широкую**
Факт: ширина полосы равна $\frac{2}{\|w\|}$

Формализация

Переформулировка поиска оптимальной разделяющей полосы в виде задачи **квадратичной оптимизации**:

При ограничениях

$$y_i(w \cdot x_i - b) \geq 1$$

минимизировать

$$\|w\|^2 = w \cdot w$$

Шумы и штрафы

Как определить оптимальную полосу, когда категории нельзя разделить строго?

Нужно определить штраф за неправильную классификацию учебных документов и минимизировать сумму $\|w\|^2$ и всех штрафов:

При ограничениях

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

минимизировать

$$\|w\|^2 + C \sum \xi_i$$

Часть II

Как решить задачу оптимизации?

Можно ли найти оптимальную гиперплоскость, не зная самих векторов, а только скалярные произведения между ними?

Множители Лагранжа (1/2)

Метод Лагранжа для нахождения минимума целевой функции при нескольких условиях-равенствах:

- Ввести дополнительные переменные (множители Лагранжа) λ_i и составить новую целевую функцию $F - \sum \lambda_i G_i$
- Взять все производные по старым переменным (координатам x) и приравнять их к нулю
- С помощью получившихся уравнений выразить x через $\lambda_1, \ldots, \lambda_k$
- Подставить получившуюся формулу для x в k ограничений-равенств и решить эти k уравнений с k неизвестными

Множители Лагранжа (2/2)

Метод Лагранжа для нахождения минимума целевой функции при нескольких условиях-**не**равенствах:

- Ввести дополнительные переменные (множители Лагранжа) $\lambda_i \geq 0$ и составить новую целевую функцию $F - \sum \lambda_i G_i$
- Взять все производные по старым переменным (координатам x) и приравнять их к нулю
- С помощью получившихся уравнений выразить x через $\lambda_1, \ldots, \lambda_k$
- Для каждого ограничения-неравенства верно одно из двух: (1) или в том условии выполняется строгое равенство, или (2) соответствующий множитель λ_i равен нулю. Получается k "или-уравнений" для k переменных

Применим метод Лагранжа (1/3)

Исходная задача: при ограничениях

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i$$

минимизировать

$$w \cdot w + C \sum \xi_i$$

По методу Лагранжа, нам нужен минимум по w, b, ξ_i и максимум по λ_i новой целевой функции

$$w \cdot w + C \sum \xi_i - \sum \lambda_i(\xi_i + y_i(w \cdot x_i - b) - 1)$$

при условиях

$$\xi_i \geq 0, \lambda_i \geq 0$$

Применим метод Лагранжа (2/3)

Вектор w выражается через множители Лагранжа:

$$w = \sum \lambda_i y_i x_i$$

Уравнение разделяющей гиперплоскости:

$$\sum \lambda_i y_i x \cdot x_i + b = 0$$

Постойте, а как найти b?

Применим метод Лагранжа (3/3)

Решению исходной задачи оптимизации соответствуют такие значения множителей, при которых достигается максимум

$$\sum \lambda_i - \sum_{i,j} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

при условиях

$$\sum \lambda_i y_i = 0, \qquad 0 \leq \lambda_i \leq C$$

Преимущества новой задачи

Новая (двойственная) задача: максимизировать

$$\sum \lambda_i - \sum_{i,j} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

при условиях

$$\sum \lambda_i y_i = 0, \qquad 0 \leq \lambda_i \leq C$$

Важные свойства:

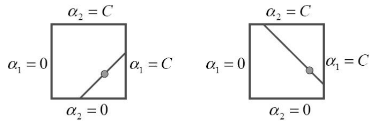
Целевая функция зависит не от самих x_i, а от скалярных произведений между ними

Только пограничные и "заграничные" учебные влияют на итоговое правило

Как решать двойственную задачу

Алгоритм Джона Платта (Microsoft Research):

1. Начать с набора λ_i удовлетворяющего ограничением
2. Выбрать (используются хитрые эвристики) пару λ_i, λ_j
3. При фиксированных значениях остальных множителей и имеющихся ограничениях $y_i\lambda_i + y_j\lambda_j = y_i\lambda_i^{old} + y_j\lambda_j^{old}$ и $0 \leq \lambda_i, \lambda_j \leq C$ выбрать оптимальную пару значений (**мини-оптимизация**)
4. Продолжать процесс до наступления стоп-условий



Часть III

Как применить разработанную технику к поиску оптимального разделяющего **эллипса**?

Обобщение метода оптимальной полосы на произвольные поверхности, заданные полиномиальными уравнениями

Идея расширенного пространства

Машина опорных векторов:

- Выбираем отображение векторов в **расширенное пространство** $x \to \phi(x)$
- Автоматически получается новая функция скалярного произведения $K(x, y) = \phi(x) \cdot \phi(y)$
- Находим линейный разделитель w, b в расширенном пространстве
- Получается классифицирующая поверхность $w \cdot \phi(x) - b$ в исходном пространстве

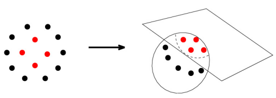
Два примера

$$\phi(x) = \phi((x_1, x_2)) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$K(x, y) = x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1x_2y_1y_2 = (x \cdot y)^2$$

$$\phi(x) = (1, c_{10\ldots0}x_1, \ldots, c_{110\ldots0}x_1x_2, \ldots, c_{0\ldots0d}x_k^d)$$

$$K(x, y) = (1 + x \cdot y)^d$$



Полиномиальные разделители

Второй пример:
$\phi(x) = (1, c_{10\ldots0}x_1, \ldots, c_{110\ldots0}x_1x_2, \ldots, c_{0\ldots0d}x_k^d)$
$K(x, y) = (1 + x \cdot y)^d$

Разделяющая гиперплоскость $w \cdot \phi(x) - b$ — это полином степени не выше d. То есть мы ищем

оптимальную полиномиальную разделяющую поверхность!

Анализ метода SVM

Преимущества:

+ На тестах превосходит другие методы
+ При различных выборах ядер можно эмулировать другие подходы
+ Теоретическое обоснование

Недостатки:

- Мало параметров для настройки
- Как выбирать ядро?
- Медленное обучение

Задача

Докажите, что гиперплоскость с максимальной шириной разделяющей полосы единственна

Главные моменты

Сегодня мы узнали:

- Метод опорных векторов сводит обучение классификатора к задаче квадратичной оптимизации
- Задача квадратичной оптимизации решается эвристическими алгоритмами путем последовательного уменьшения целевой функции
- Для построения нелинейных классификаторов используется отображение исходных объектов в расширенное пространство признаков

Вопросы?

Источники

Страница курса http://logic.pdmi.ras.ru/~yura/internet.html

Использованные материалы:

- Wikipedia
  Support Vector machine
  http://en.wikipedia.org/wiki/Support_vector_machine
- CJC Burges
  A Tutorial on Support Vector Machines for Pattern Recognition
  http://www.music.mcgill.ca/~rfergu/adamTex/references/Burges98.pdf
- Константин Воронцов
  Лекция по методу опорных векторов
  http://www.ccas.ru/voron/download/SVM.pdf
- John Platt
  Sequential Minimal Optimization
  http://research.microsoft.com/users/jplatt/smo.html