Логические методы классификации

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».
Лекция 4.

Общая характеристика

Логические методы — широко используемый класс методов.

Основные варианты:

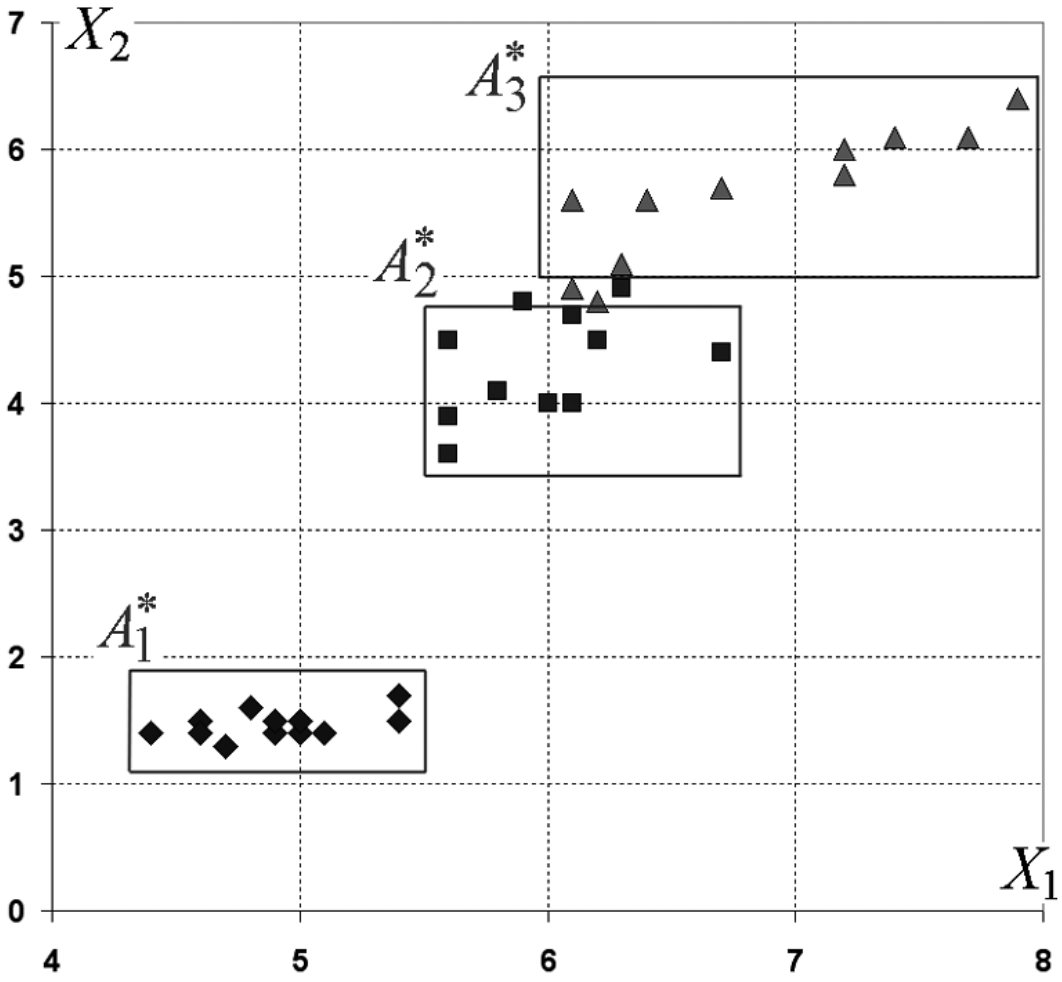
- поиск закономерностей, решающие списки,
- решающие деревья.

Свойства:

- работа в разнотипном пространстве,
- работа с пропусками,
- интерпретируемость решений.

Бустинг на деревьях — один из лучших методов.

Логические закономерности для задачи Iris



Понятие закономерности

X – пространство значений прогнозирующих переменных,
$Y = \{0, 1, ...\}$ – прогнозируемая переменная, $\varphi : X \to \{0, 1\}$ – предикат.

$V = \left((x^i, y^i) \,|\, i = \overline{1, N}\right)$ – выборка объектов,
M – из них 1-го класса,
n – число точек, на которых предикат истинный,
m – из них 1-го класса.

«Хороший» предикат — закономерность:

$$a = \frac{m}{M} \to \max, \quad b = \frac{n - m}{n} \to \min .$$

Статистический критерий

Вероятность при отборе n объектов получить m из них 1-го класса:

$$\mathsf{P}(m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} = \frac{C_n^m C_{N-n}^{M-m}}{C_N^M}.$$

Критерий «неслучайности» отбора:

$$\mathsf{P}(m \geqslant m_0) = \sum_{m=m_0}^{M} \mathsf{P}(m) < \alpha.$$

Информационный критерий

Формула Стирлинга:

$$\ln(k!) \approx k \ln k - k + \frac{1}{2} \ln(2k\pi) + \frac{1}{12k} - \frac{1}{360k^2}.$$

Количество информации:

$$G = H\left(\frac{M}{N}\right) - \frac{n}{N} \cdot H\left(\frac{m}{n}\right) - \frac{N-n}{N} \cdot H\left(\frac{M-m}{N-n}\right).$$

$$H(p) = -p \ln p - (1-p) \ln(1-p), \quad G \approx -\frac{1}{N} \ln \mathsf{P}(m).$$

Принцип равномерной сходимости

Статистический критерий $\mathsf{P}(m \geqslant m_0) < \alpha$ характеризует «случайность» только для априорно выбранной закономерности.

Поскольку возможных предикатов много, вероятность того, что хотя бы на одном значение критерия будет «хорошим», больше.

Вероятность зависит от сложности предиката.

Поиск закономерностей

Схема направленного поиска закономерностей («жадный» алгоритм):

- дискретизация признаков (границы между проекциями точек выборки);
- интервальные предикаты;
- конъюнкции элементарных предикатов.

Известные алгоритмы:

- KORA,
- ТЭМП.

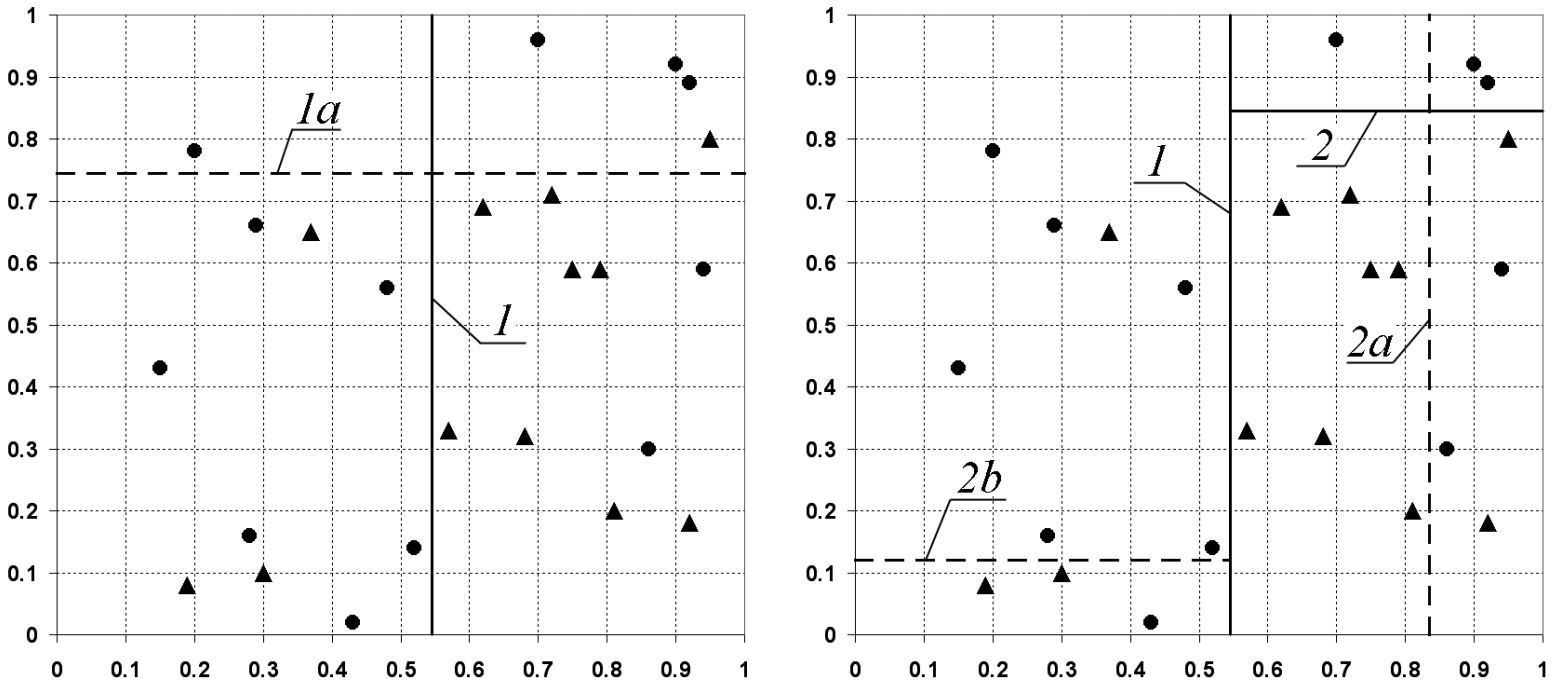
Решающие списки

Метод классификации на основе списка закономерностей.

- Формируем упорядоченный по информативности список закономерностей.
- Решение принимаем по первой закономерности, которой удовлетворяет объект.

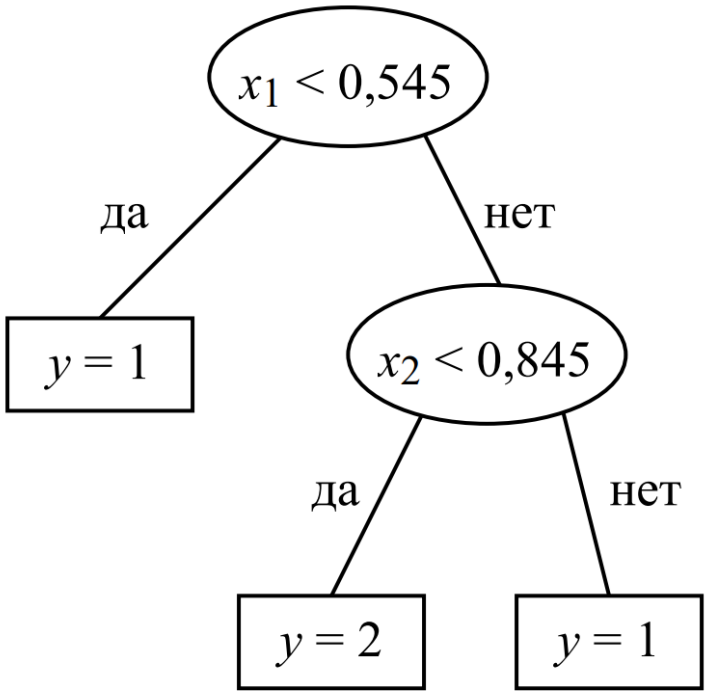
Можно применять голосование.

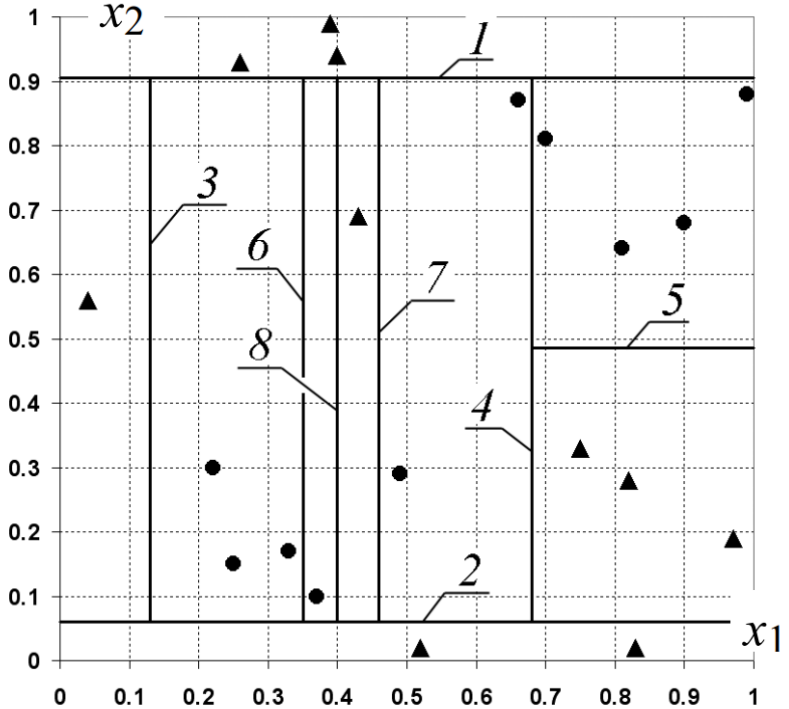
Процесс построения дерева для задачи Iris



Дерево представляет собой последовательное разбиение пространства значений на области.

Представление дерева





Можно получить не лучшее дерево (пример справа для задачи XOR).

Критерии ветвления

Критерий ветвления основывается на критерии качества дерева (строится путём сравнения вариантов дерева).

Критерии оценки дерева (для классификации):

- число ошибок классификации (для детерминированного прогноза),
- критерий Джини (Gini impurity) — это по сути число ошибок для вероятностного прогноза,
- информационный (log loss, на основе функции правдоподобия для вероятностного прогноза).

Первый критерий работает плохо, остальные сопоставимо.

Алгоритмы

Варианты алгоритма построения дерева:

- жадный (без учёта будущих ветвлений),
- рекурсивный (предикат в узле выбирается с учётом ветвления на нижнем уровне, решает задачу XOR),
- неограниченный (строит дерево, затем его оптимизирует всё доступное время).

Сейчас на практике используется «жадный» алгоритм.

Выводы

Недостатки логических методов:

- невозможность «гладких» решений,
- много эвристик (настраиваемых параметров),
- вычислительная трудоёмкость нахождения точных решений,
- неизвестен лучший критерий ветвления.

Актуальность:

- возможность получить интерпретируемое решение,
- используется в составе бустинга.