Московский государственный университет имени М.В. Ломоносова
физический факультет
NeuroFuzzy Lab.

Классические методы распознавания образов. Теория вероятностей как модель регистрации данных

Зубюк Андрей Владимирович
zubjuk@physics.msu.ru
http://NeuroFuzzy.Phys.MSU.ru

Разделяющие гиперплоскости

Гиперплоскость в признаковом пространстве (в 2-мерном случае — прямая, в 3-мерном — плоскость) задаётся уравнением

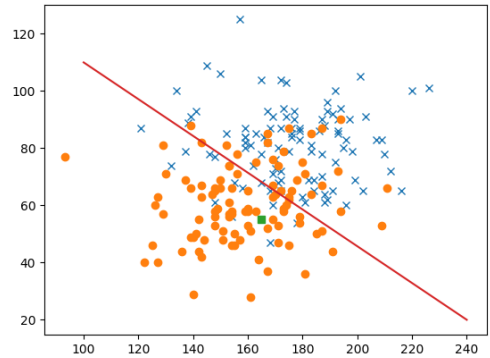
$$(w, x) + b = 0 \quad \Leftrightarrow \quad w_1 x_1 + w_2 x_2 + \ldots + b = 0.$$

Все векторы признаков x, лежащие в пространстве с одной стороны от этой гиперплоскости, удовлетворяют условию

$$(w, x) + b \leqslant 0,$$

с другой стороны — условию

$$(w, x) + b > 0.$$



$$w = \begin{pmatrix} 9 & 14 \end{pmatrix} \quad b = -2440$$

Классические и нейросетевые методы

Разделение гиперплоскостью относится к **классическим** методам распознавания образов. Наряду с такими методами как деревья решений, леса решений, корреляционный анализ, математическая морфология Серра (анализ изображений) и др.

В настоящее время **классическими** принято называть все методы, которые не связаны с искусственными нейронными сетями.

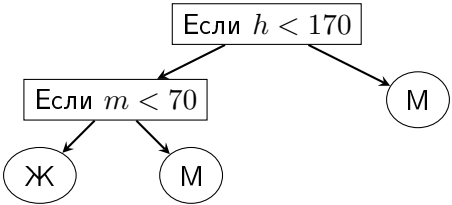
Классические и нейросетевые методы

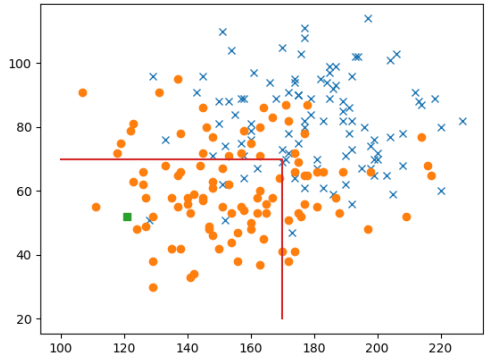
Достоинства и недостатки классических и нейросетевых методов:

- Интерпретируемость (прозрачность) классических методов.
- Чёткая очерченность области применимости классических методов.
- Возможность внедрения в классические алгоритмы априорной информации о решении, предоставленной специалистом в предметной области,.
- Возможность «отладки» классических алгоритмов.

- Высокое качество нейросетевых методов в ряде задач при наличии большого объёма обучающих примеров.
- Высокая степень автоматизации построения нейросетевых алгоритмов по обучающим примерам.

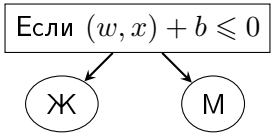
Деревья решений

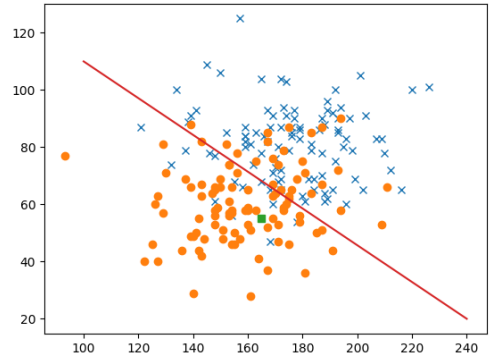




Помимо числовых признаков **деревья решений** могут использовать **категориальные** признаки: цвет волос, профессию и др. Категориальный признак принимает конечное количество значений, каждое из которых означает принадлежность объекта к некоторой категории: блондин, брюнет (по цвету волос), учитель, инженер (по профессии) и т. п.

Разделение гиперплоскостью как дерево решений



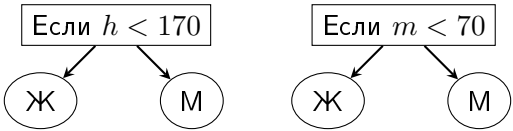


$$w = \begin{pmatrix} 9 & 14 \end{pmatrix} \quad b = -2440$$

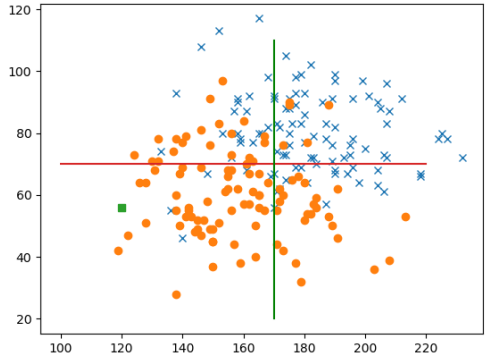
Разделение гиперплоскостью — это дерево решений, в котором в качестве «нового» признака используется $y = (w, x) + b$. Обычно при построении деревьев решений используются исходные признаки, образующие матрицу (вектор признаков) x.

Леса решений

Для решения одной и той же задачи можно построить много разных деревьев решений.



Лес решений — принятие решения в пользу того класса, «за который проголосовало» больше всего деревьев.

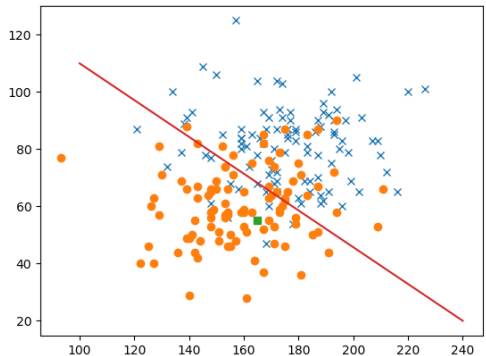


Какой метод лучше?

Фундаментальный вопрос: **чем измерить качество алгоритма**?

Стандартный ответ «в духе» машинного обучения: сосчитать количество примеров, неверно распознанных алгоритмом, отнесённое к общему количеству примеров:

1. доля женщин, принятых за мужчин,
2. доля мужчин, принятых за женщин,
3. общая доля неверно распознанных мужчин и женщин.



Лучше тот метод, для которого доля ошибок меньше.

Применительно к тестам ошибки 1 и 2 часто называют ложно-положительными и ложно-отрицательными срабатываниями алгоритма. Обычно при уменьшении одного из этих показателей растёт другой.

Какой метод лучше? Проблема с подсчётом примеров

Фундаментальная проблема: при предложенном выше подходе к измерению качества алгоритма **качество будет разным для разных наборов обучающих примеров**!

Решение проблемы: **разные наборы примеров — это выборки из одного и того же (возможно, неизвестного) распределения вероятностей**. Тогда вероятности ошибок существуют «сами по себе» — без обучающих примеров — и неизменны. А доли ошибок, сосчитанные на основе обучающих примеров, — это случайные величины, примерно равные вероятностям при большом количестве примеров.

Аналогия с бросанием монеты. Вероятности выпадения орла и решки равны 1/2 (50%). Но если 100 раз бросить монету, то орёл может выпасть 50 раз ($50/100 = 1/2$), а может 49, 51, 48 и т. д. (строго говоря — любое количество раз от 0 до 100 включительно).
Т. е. вероятность выпадения орла — «фундаментальная» константа, а доля случаев, когда выпал орёл, — случайная величина, примерно равная 1/2 при большом количестве бросаний.

Тест

При прохождении теста укажите e-mail и фамилию, имя, отчество, которые вы указали при регистрации на курс. По этим данным будут суммироваться результаты всех ваших тестов в семестре. После ответа на тест вам на почту должно прийти письмо с вашими ответами.

https://docs.google.com/forms/d/e/1FAIpQLSeJwueu5HLvTrVhazuc8AEPJW7D4ZrVzjQhAZvXDZW_eji2EQ/viewform