Анализ данных

Семинар 1. Введение в анализ данных.

12 января 2016

Что такое анализ данных?

Анализ данных, или **машинное обучение** — это наука, изучающая способы извлечения закономерностей из ограниченного количества примеров.

Что такое анализ данных?

Примеры задач:

- для нового письма определить, является ли оно спамом;
- оценить стоимость квартиры по имеющимся характеристикам;
- разбить пользователей социальной сети на несколько «плотных» групп (*кластеров*);
- выбрать товары/фильмы, которые могли бы заинтересовать данного пользователя интернет-магазина/онлайн-кинотеатра;
- определить, реклама какого продукта с наибольшей вероятностью заинтересует данного пользователя соц. сети;
- автоматическая постановка диагноза по набору симптомов за последнюю неделю;
- управление предметами реального мира при помощи ментальных команд;
- и т.д.

Данные

- Как правило, входные данные представлены в виде набора объектов, для которых известен ответ на поставленный вопрос, — обучающей выборки (training set).
- При этом каждый объект в рамках поставленной задачи описывается набором наперед заданных характеристик (признаков).
- В этом случае удобнее всего представлять данные в виде таблицы/матрицы;
- Ответ на поставленный в рамках задачи вопрос называется целевым признаком.
- Поскольку фиксированный признак для всех объектов имеет одинаковый смысл, то для всех объектов он определен на одном и том же множестве.

Данные

- Таким образом, обучающая выборка представлена в виде набора пар $X^L = (X_1, y_1), \ldots, (X_L, y_L)$, где
  X_i — описание i-ого объекта, вектор значений признаков,
  y_i — ответ на поставленный в рамках решаемой задачи вопрос для i-ого объекта.

Данные

В зависимости от множества значений, которые может принимать признак, можно выделить несколько типов:

- вещественные;
- категориальные;
- номинальные;
- бинарные;
- прочие.

Данные

В зависимости от постановки вопроса и множества значений ответов в обучающей выборке задачи можно разделить на несколько типов.

- Задачи классификации — значения целевого признака принадлежат некоторому ограниченному множеству, необходимо ответить на вопрос «какой из групп принадлежит объект?».
- Задачи регрессии — значения целевого признака являются вещественными, необходимо ответить на вопрос «какое значение примет данная величина для объекта?».

Постановка задачи

Цель: с учётом имеющихся данных научиться как можно «лучше» давать ответ на поставленный вопрос для любого нового объекта.

- Что значит «лучше»?

Постановка задачи

Цель: с учётом имеющихся данных научиться как можно «лучше» давать ответ на поставленный вопрос для любого нового объекта.

- Что значит «лучше»?
- Оптимизация некоторой функции — степени того, насколько найденное решение адекватно (*оптимизационного критерия*)

Постановка задачи

Цель: с учётом имеющихся данных научиться как можно «лучше» давать ответ на поставленный вопрос для любого нового объекта.

- Что значит «лучше»?
- Оптимизация некоторой функции — степени того, насколько найденное решение адекватно (*оптимизационного критерия*)
- «Лучшее» решение — то, при котором значение критерия достигает оптимума

Постановка задачи

Примеры оптимизационных критериев:

- классификация: $Q(a, X^L) = \frac{1}{L} \sum_{i=1}^{L} [a(X_i) \neq y_i]$;

Постановка задачи

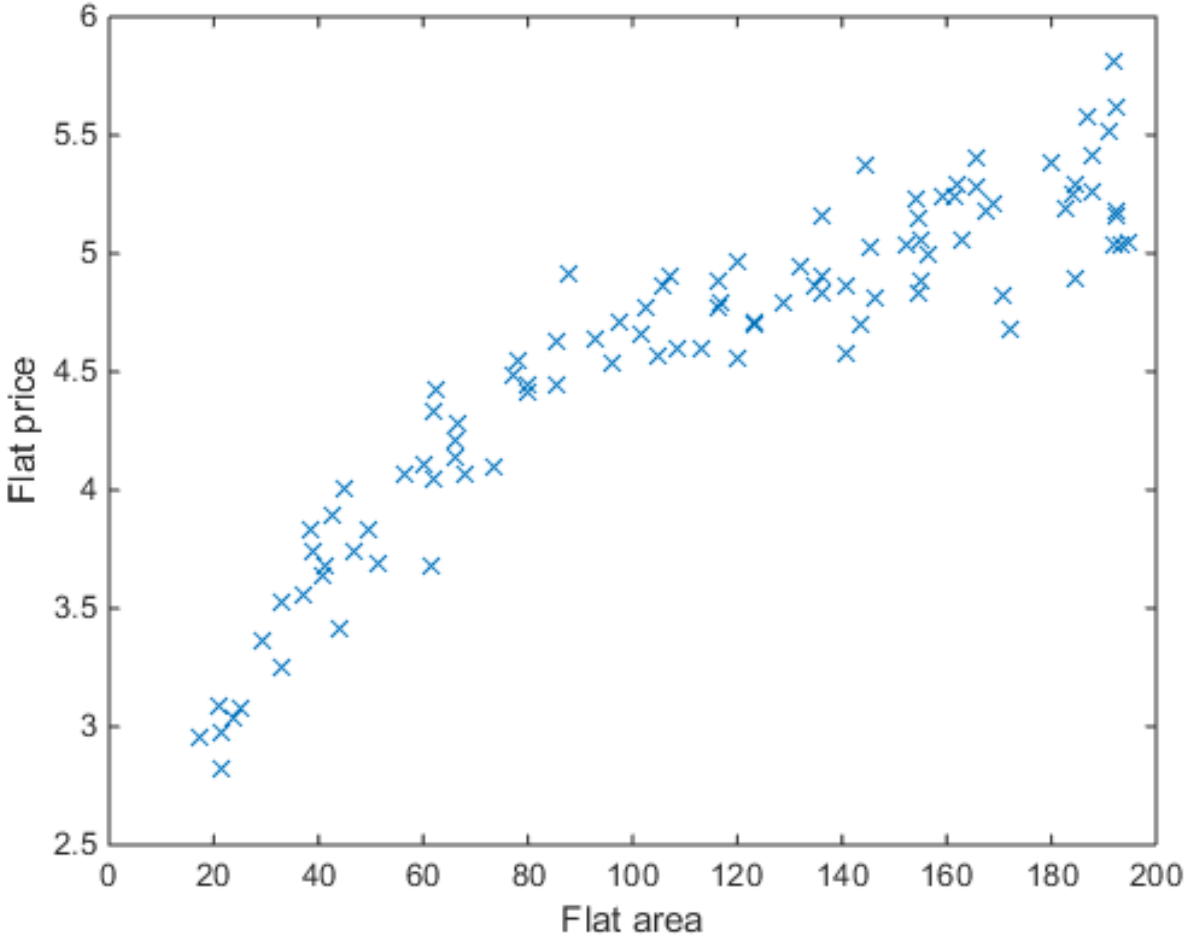
Примеры оптимизационных критериев:

- классификация: $Q(a, X^L) = \frac{1}{L} \sum_{i=1}^{L} [a(X_i) \neq y_i]$;
- регрессия: $Q(a, X^L) = \frac{1}{L} \sum_{i=1}^{L} (a(X_i) - y_i)^2$.

Модель

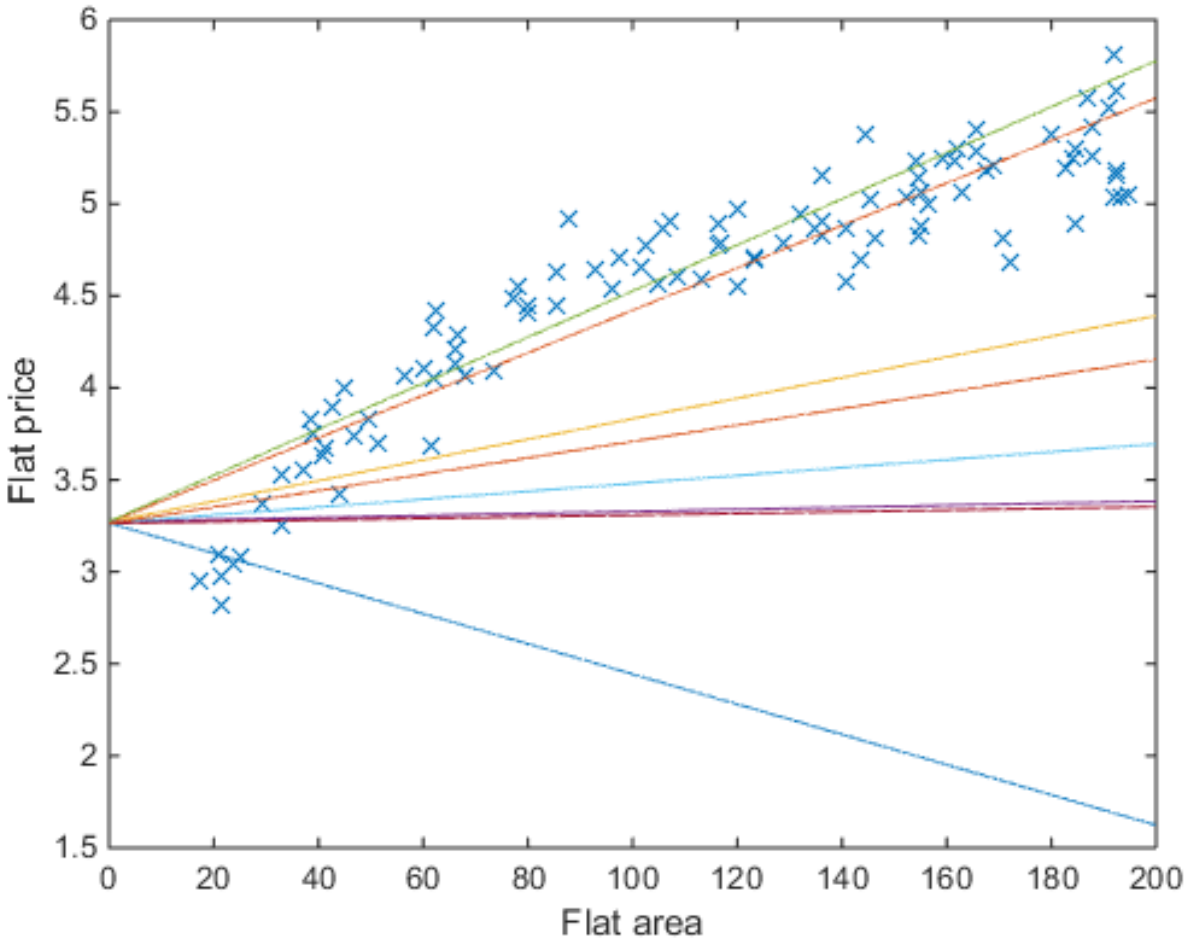
- Прежде чем решать задачу, необходимо выбрать модель, в соответствии с которой будут порождаться ответы на новых объектах.
- Чаще всего это некоторое семейство однотипных функций, отличающихся лишь набором параметров.

Модель


Flat price
Flat area
6
5.5
5
4.5
4
3.5
3
2.5
0
20
40
60
80
100
120
140
160
180
200

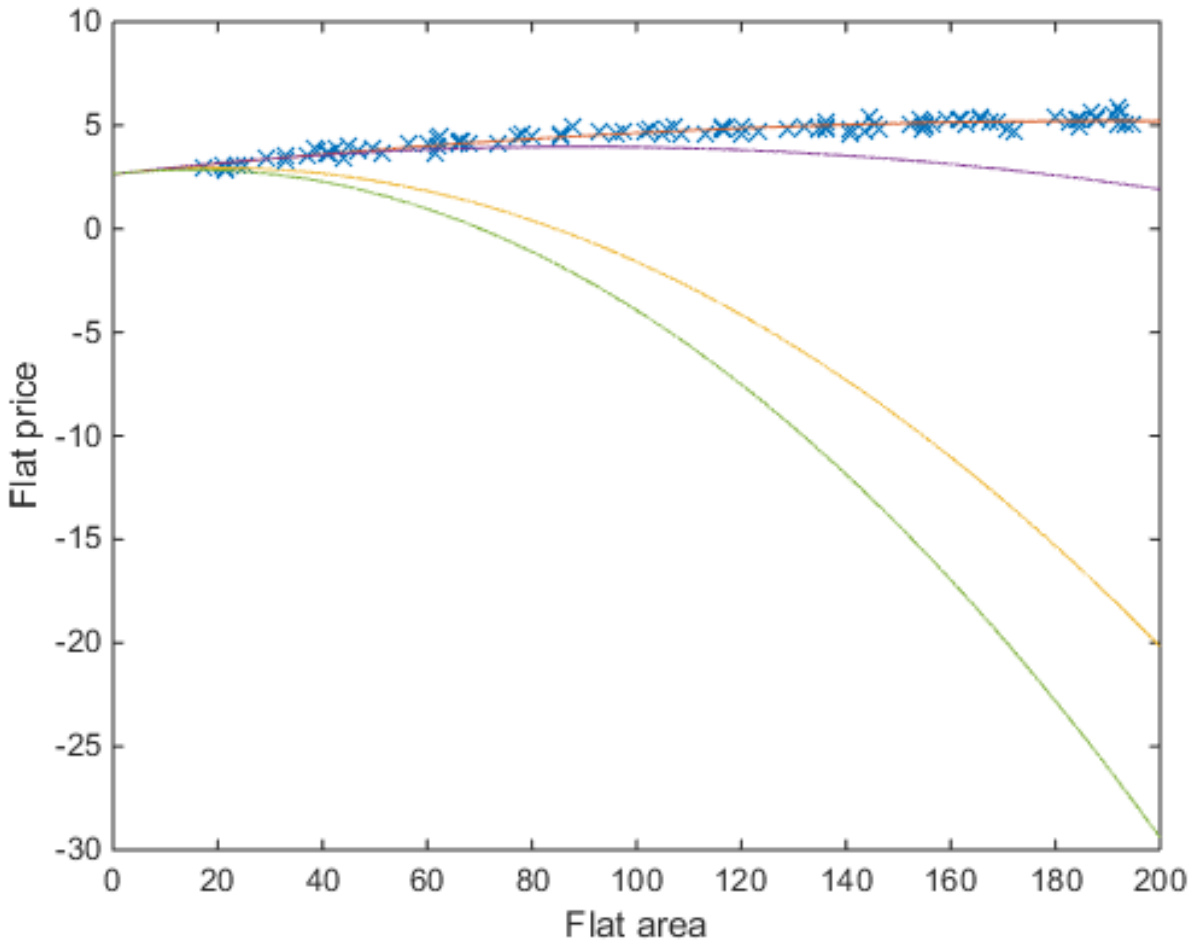
Модель

$$f(x) = a * x + b$$

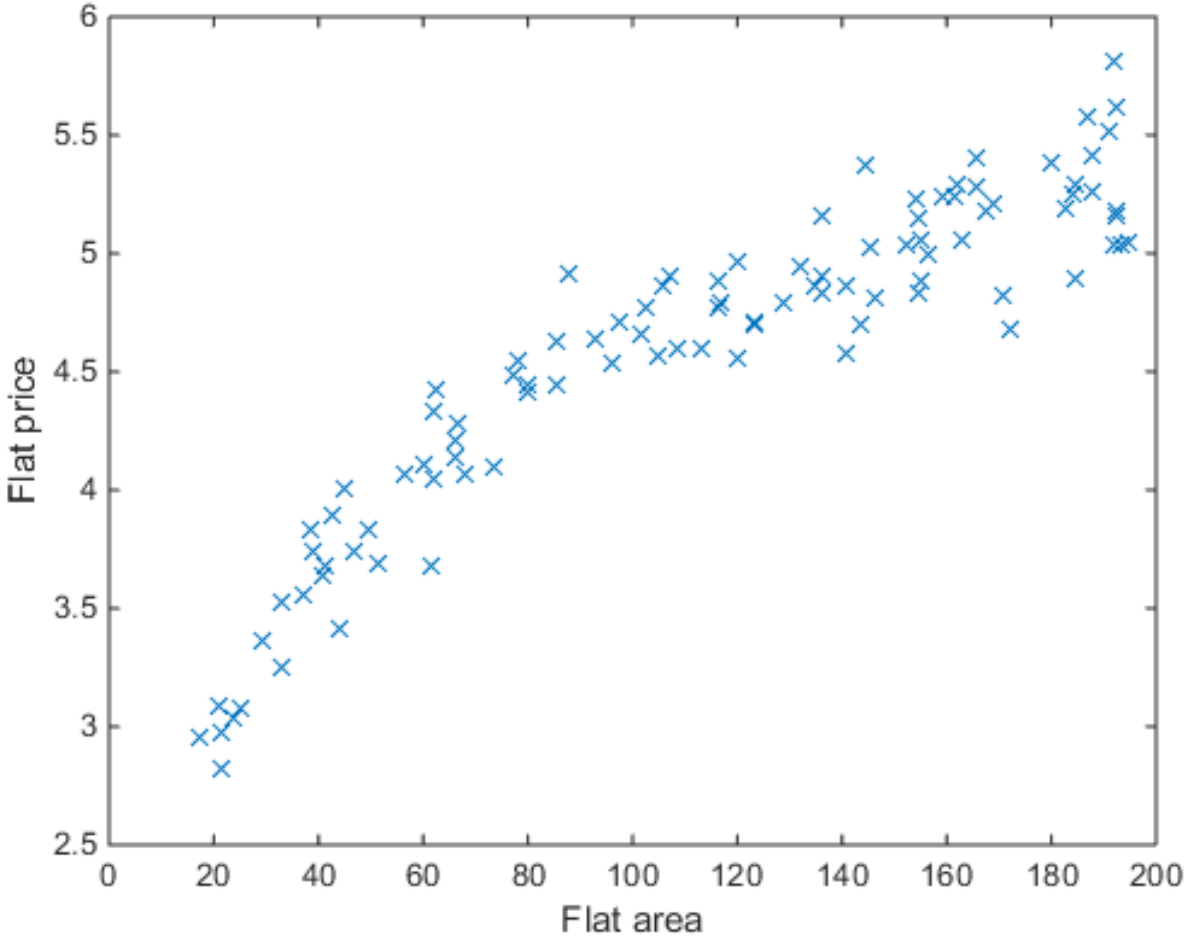


Модель

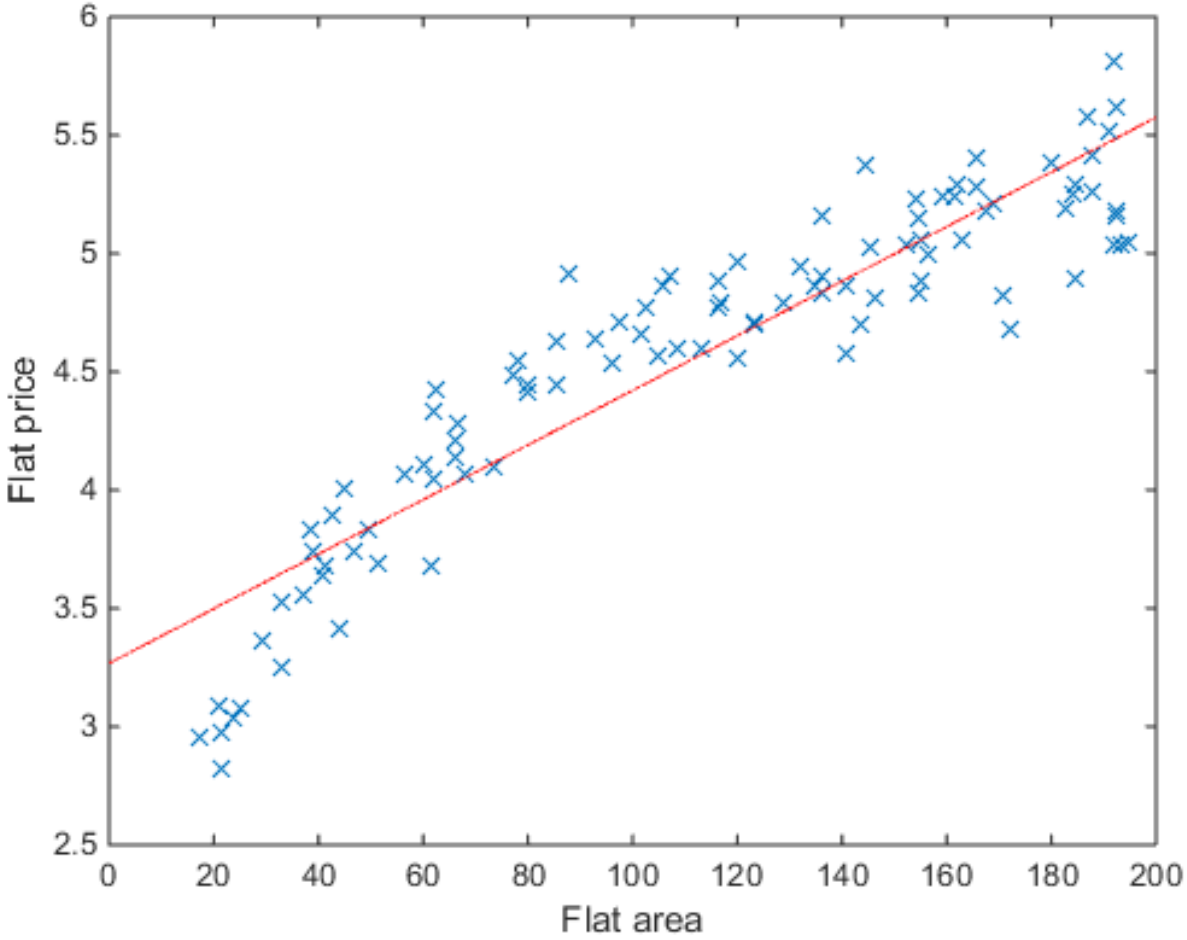
$$f(x) = a * x^2 + b * x + c$$



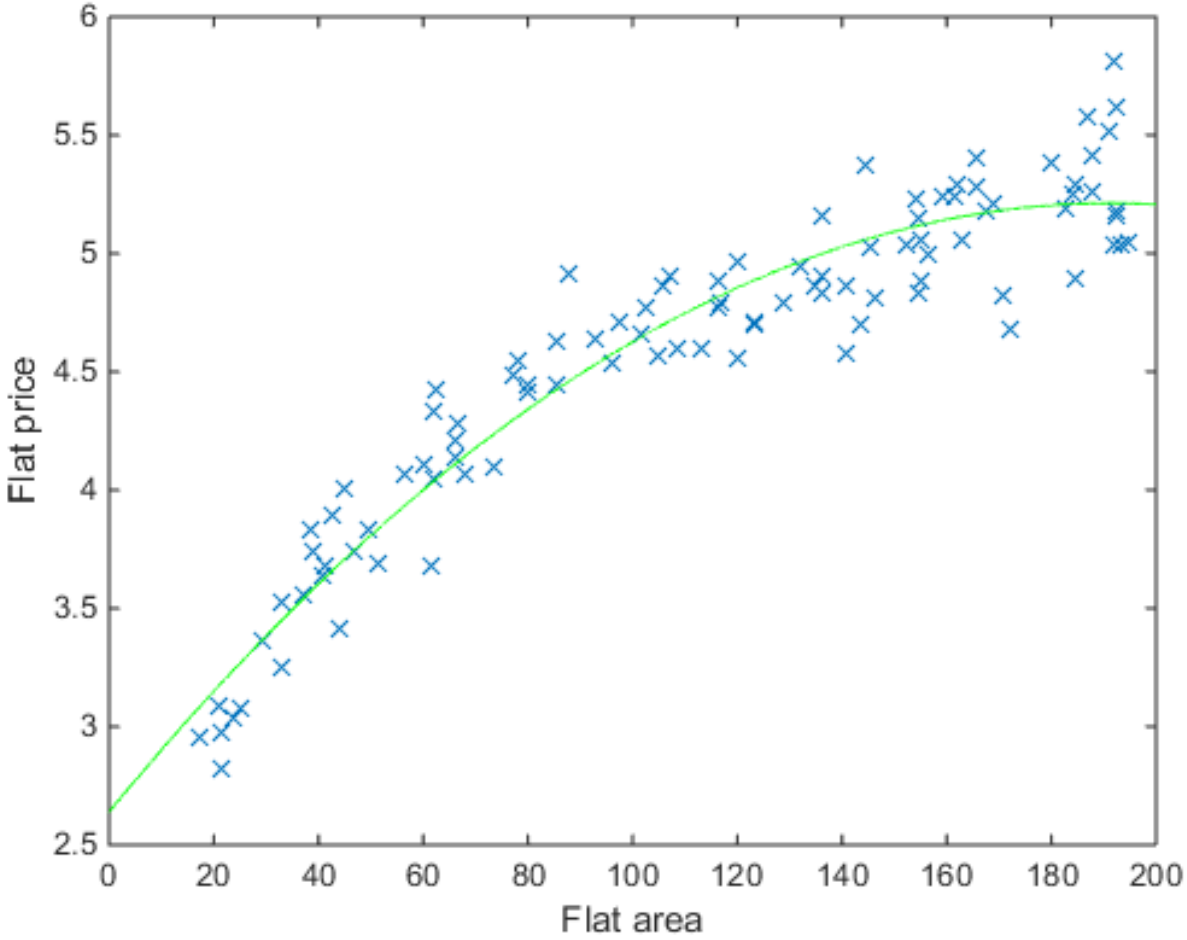
Модель


Flat price
Flat area
6
5.5
5
4.5
4
3.5
3
2.5
0
20
40
60
80
100
120
140
160
180
200

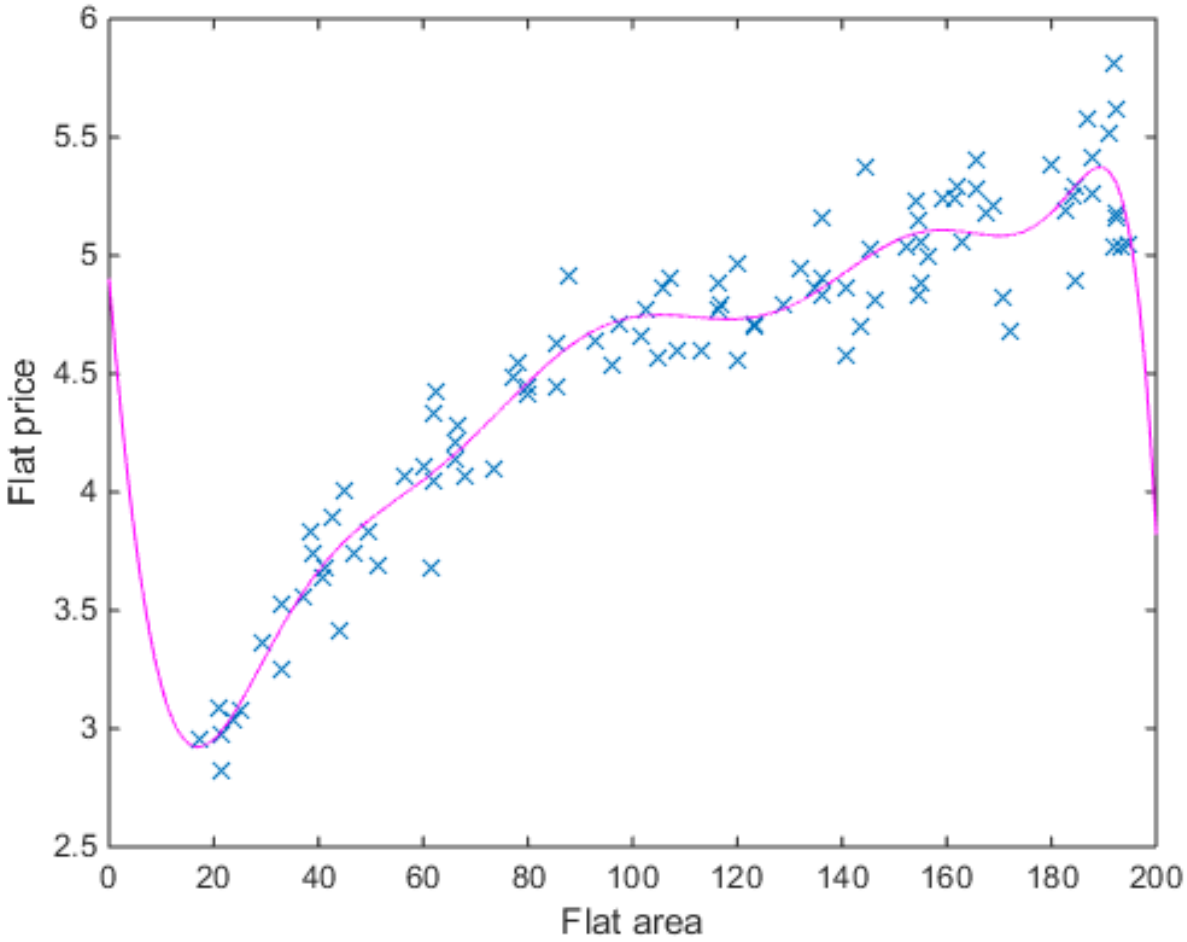
Модель


Flat price
Flat area
6
5.5
5
4.5
4
3.5
3
2.5
0
20
40
60
80
100
120
140
160
180
200

Модель


Flat price
Flat area
6
5.5
5
4.5
4
3.5
3
2.5
0
20
40
60
80
100
120
140
160
180
200

Модель


Flat price
Flat area
6
5.5
5
4.5
4
3.5
3
2.5
0
20
40
60
80
100
120
140
160
180
200

Переобучение

- Переобучение — ситуация, при которой алгоритм хорошо работает для объектов из обучающей выборки, но при этом сильно ошибается на новых объектах.
- Другими словами, алгоритм учится, но теряет способность обобщать.

Процесс решения задачи

Таким образом, процесс решения задачи может быть разбит на несколько составляющих:

- получение данных;
- выбор модели, определение набора её параметров;
- выбор оптимизационного критерия;
- нахождение оптимальных значений параметров модели (с учётом предыдущих пунктов);
- получение ответов на новых объектах.