Классические модели машинного обучения для классификации

О переобучении

Всегда модели обучаются, минимизируя функцию потерь, Loss. Чтобы справиться с переобучением к ней добавляют слагаемое:

$$\hat{L} = L + \alpha(w_1^2 + w_2^2 + \cdots + w_n^2)$$

$$\hat{L} = L + \alpha(|w_1| + |w_2| + \cdots + |w_n|)$$

Модели

- Деревья решений
- Случайный лес
- Метод опорных векторов (SVM)
- Метод k ближайших соседей
- Логистическая регрессия

1 Деревянные модели

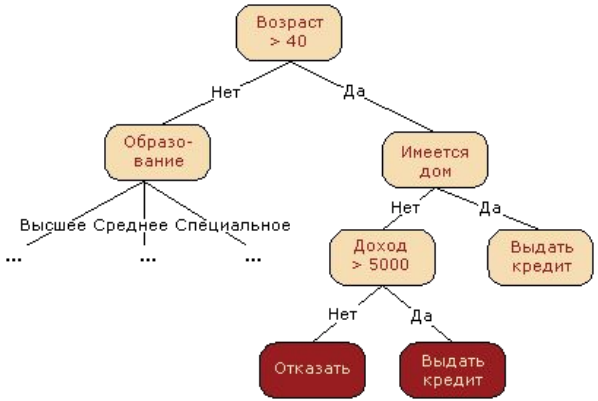
Дерево решений

Строит правила в виде дерева.

Гиперпараметры:

Max_depth – максимальная глубина дерева

min_samples_leaf – минимальное число объектов в листе

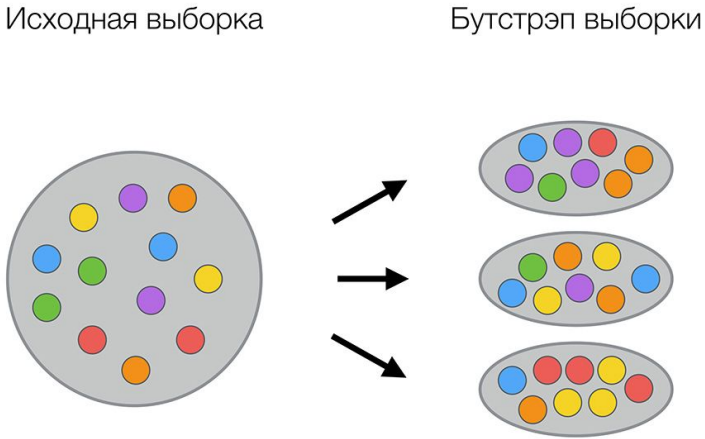


Дерево решений

- Плюсы:
  - Легко интерпретируется
  - Дает оценку важности признаков
  - Не требует масштабирования признаков
- Минусы:
  - Легко переобучается
  - Неустойчиво к шумам в данных

Случайный лес: бутстреп

Для обучения каждого дерева из исходных данных берется случайная подвыборка с возвращением



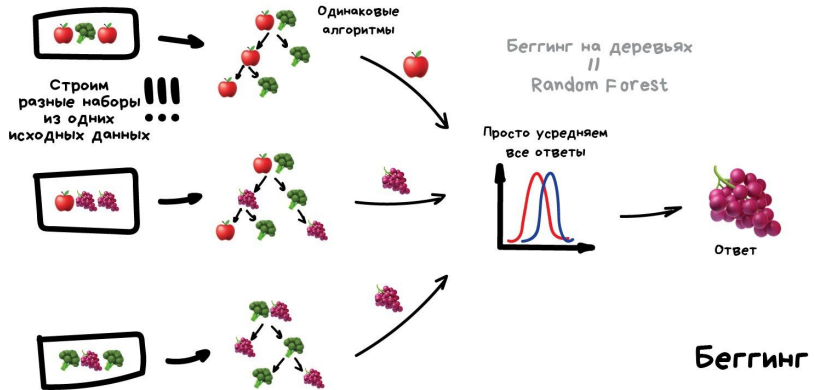
Случайный лес

Гиперпараметры:

max_depth – ограничение глубины деревьев

n_estimators – число деревьев

min_samples_leaf – минимальное число объектов в листе



Случайный лес: оценка

- Плюсы:
  - Устойчив к переобучению
  - Дает оценку важности признаков
  - Хорошо параллелится
- Минусы:
  - Сложно интерпретировать результаты
  - Плохо работает при большом числе признаков

2

Другие модели

Метод опорных векторов

Гиперпараметры:

kernel – ядро

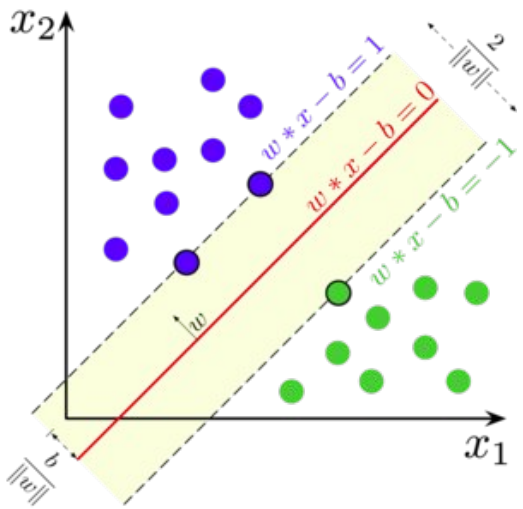
linear – простое линейное

poly – степенное

rbf – радиальное

p – степень (для степенного)

C – параметр регуляризации



Метод опорных векторов

- Плюсы:
  - Устойчив к переобучению, особенно если признаков много
  - Много ядер на выбор, есть попроще, есть посложнее
  - Третий плюс
- Минусы:
  - Неинтуитивные гиперпараметры
  - Нужно масштабировать признаки
  - Долго обучается, если выборка большая

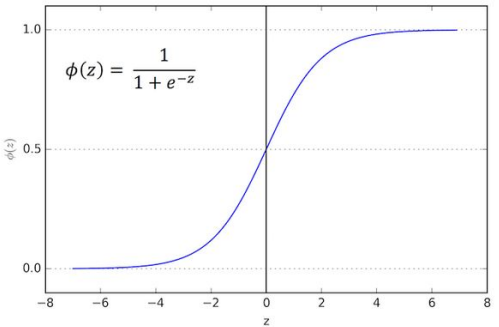
Логистическая регрессия

Ищет функцию в виде: $\sigma(w_1x_2 + w_2x_2 + \cdots + w_nx_n)$

Гиперпараметры:

penalty – регуляризация

C – параметр регуляризации



Логистическая регрессия

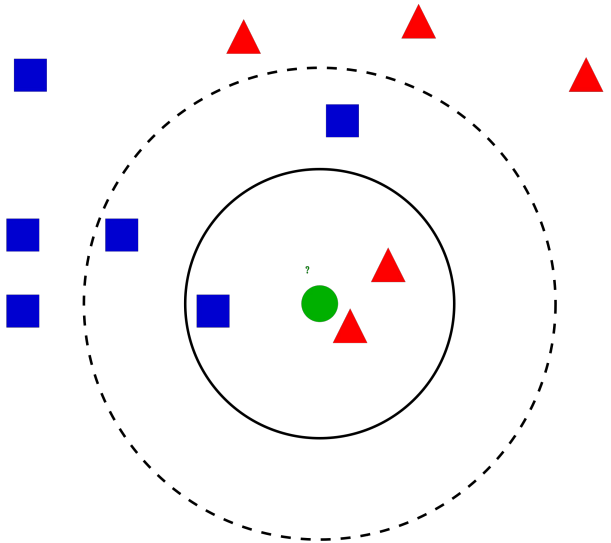
- Плюсы:
  - Выдает вероятность
  - Мало гиперпараметров
  - С переобучением легко бороться регуляризацией
- Минусы:
  - Может требовать масштабирования признаков
  - Линейная

KNN

Гиперпараметры:

n_neighbours – число соседей

metric – функция расстояния



KNN: оценка

- Плюсы:
  - Не требует обучения в привычном смысле, оно мгновенное
  - Можно интерпретировать результат
  - Может решать сложные зависимости
- Минусы:
  - Неинтуитивно выбирать метрику
  - Плохо работает если признаков много
  - Долго работает, если обучающая выборка большая
  - Требует масштабирования признаков