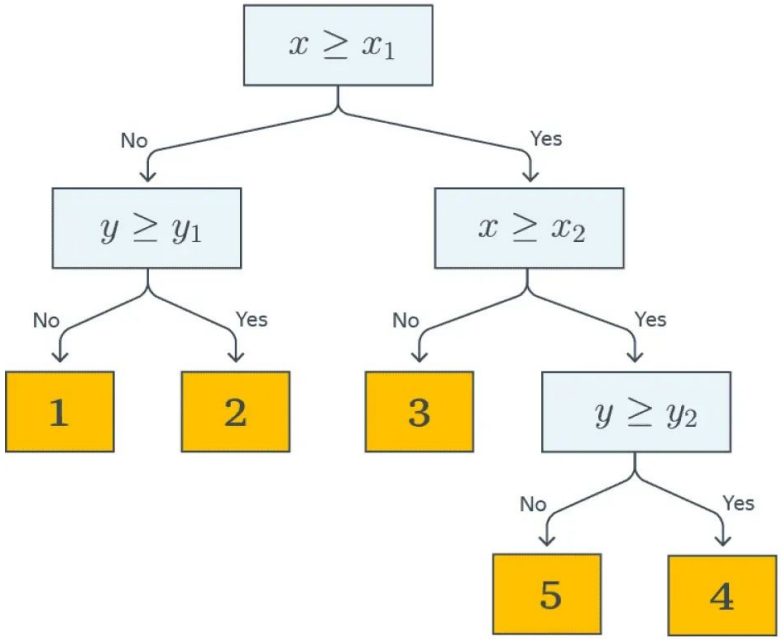
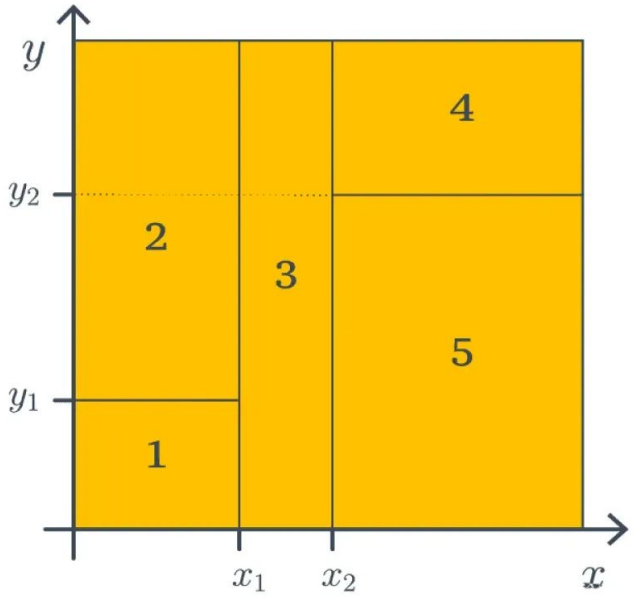
Классификационное дерево и Random Forest

Подготовила:

Кривоногова Татьяна, 25222

Классификационное решающее дерево





Параметры решающего дерева

- Критерий разделения
  - criterion ('entropy' или 'log_loss' — энтропия или информационный выигрыш, 'gini' — индекс Джини)

Параметры решающего дерева

- Критерий разделения
  - criterion ('entropy' или 'log_loss' — энтропия или информационный выигрыш, 'gini' — индекс Джини)
- Параметры остановки
  - max_depth: Максимальная глубина дерева
  - min_samples_split: Минимальное количество объектов выборки, необходимое для разделения внутреннего узла
  - min_samples_leaf: Минимальное количество объектов выборки, которое должно находиться в листовом узле
  - max_leaf_nodes: Максимальное количество листьев в дереве
  - min_impurity_decrease: Минимальное уменьшение неоднородности, необходимое для разделения

Параметры решающего дерева

- Критерий разделения
  - criterion ('entropy' или 'log_loss' — энтропия или информационный выигрыш, 'gini' — индекс Джини)
- Параметры остановки
  - max_depth: Максимальная глубина дерева
  - min_samples_split: Минимальное количество образцов, необходимое для разделения внутреннего узла
  - min_samples_leaf: Минимальное количество образцов, которое должно находиться в листовом узле
  - max_leaf_nodes: Максимальное количество листьев в дереве
  - min_impurity_decrease: Минимальное уменьшение неоднородности, необходимое для выполнения разделения
- Параметры стратегии поиска разделения
  - splitter: Стратегия выбора разделения в узле ('best' — перебирает все возможные разделения и выбирает лучшее по критерию разделения, 'random' — выбирает случайное разделение из лучших max_features вариантов)
  - max_features: Количество признаков, рассматриваемых при поиске лучшего разделения (None или 'auto' — рассматриваются все признаки, 'sqrt' или 'log2' — рассматривается квадратный корень или логарифм от общего числа признаков, целое число — точное количество признаков, дробное число от 0.0 до 1.0 — доля от общего числа признаков)

Свойства

1. Выученная функция — кусочно-постоянная, из-за чего производная равна нулю везде, где задана. Градиентные методы при поиске оптимального решения не подходят.
2. Дерево решений не сможет экстраполировать зависимости за границы области значений обучающей выборки.
3. Дерево решений способно идеально приблизить обучающую выборку и ничего не выучить (то есть такой классификатор будет обладать низкой обобщающей способностью): для этого достаточно построить такое дерево, в каждый лист которого будет попадать только один объект.

Решающее дерево

Плюсы

- Интерпретируемость
- Не требует подготовки данных (нормализации, заполнения пропусков)
- Работает с разнотипными данными (числовыми и категориальными)
- Хорошо обнаруживает нелинейные зависимости

Минусы

- Проблема получения оптимального дерева решений является NP-полной задачей
- Склонность к переобучению
- Чувствительность к малым изменениям данных
- Не умеет экстраполировать за пределы обучающих данных
- Плохо обнаруживает линейные зависимости

Случайный лес

Ансамбль из решающих деревьев

Бэггинг - уменьшает разброс модели

1. Для построения i-го дерева:
   - сначала, как в обычном бэггинге, из обучающей выборки X выбирается с возвращением случайная подвыборка X_i того же размера, что и X;
   - в процессе обучения каждого дерева в каждой вершине случайно выбираются n < N признаков, где N — полное число признаков, и среди них ищется оптимальный сплит; такой приём позволяет управлять степенью скоррелированности базовых алгоритмов.
2. Чтобы получить предсказание ансамбля на тестовом объекте, усредняем отдельные ответы деревьев (для регрессии) или берём самый популярный класс (для классификации).

- Используем глубокие и переобученные деревья
- Чем больше признаков, тем больше корреляция между деревьями и тем меньше чувствуется эффект от ансамблирования. Чем меньше признаков, тем слабее сами деревья.
- Имеет смысл построить график ошибки от числа деревьев и ограничить размер леса в тот момент, когда ошибка перестанет значимо уменьшаться

Случайный лес

Плюсы

- Высокая точность предсказания
- Устойчивость к переобучению
- Устойчивость к выбросам и шуму
- Высокая параллелизуемость и масштабируемость
- Оценка ошибки без отдельной валидационной выборки

Минусы

- Низкая интерпретируемость
- Большой объем памяти и медленное предсказание
- Плохо экстраполирует за пределы обучающих данных

Источники

1. https://mipt-stats.gitlab.io/courses/ad_fivt/trees.html
2. https://education.yandex.ru/handbook/ml/article/reshayushchiye-derevya
3. https://education.yandex.ru/handbook/ml/article/ansambli-v-mashinnom-obuchenii
4. https://neerc.ifmo.ru/wiki/index.php?title=Дерево_решений_и_случайный_лес
5. https://ru.wikipedia.org/wiki/Дерево_решений
6. https://ru.wikipedia.org/wiki/Метод_случайного_леса
7. https://www.youtube.com/playlist?list=PL4_hYwCyhAvZyW6qS58x4uElZgAkMVUvj