Л е к ц и я № 10

Метод случайного леса

Разработал профессор П. М. Васильев
Кафедра фармакологии и биоинформатики

Для студентов, обучающихся по направлению 06.03.01 «Биология»
профили Биохимия, Генетика
при изучении дисциплины «Цифровые технологии в биологии»

План лекции

- **Что такое метод случайного леса**
- **Дерево решений**
- **Бэггинг и бутстрэп**
- **Алгоритм метода случайного леса**
- **Параметризация метода случайного леса**
- **Достоинства и недостатки метода случайного леса**

Метод случайного леса

Random Forest, RF

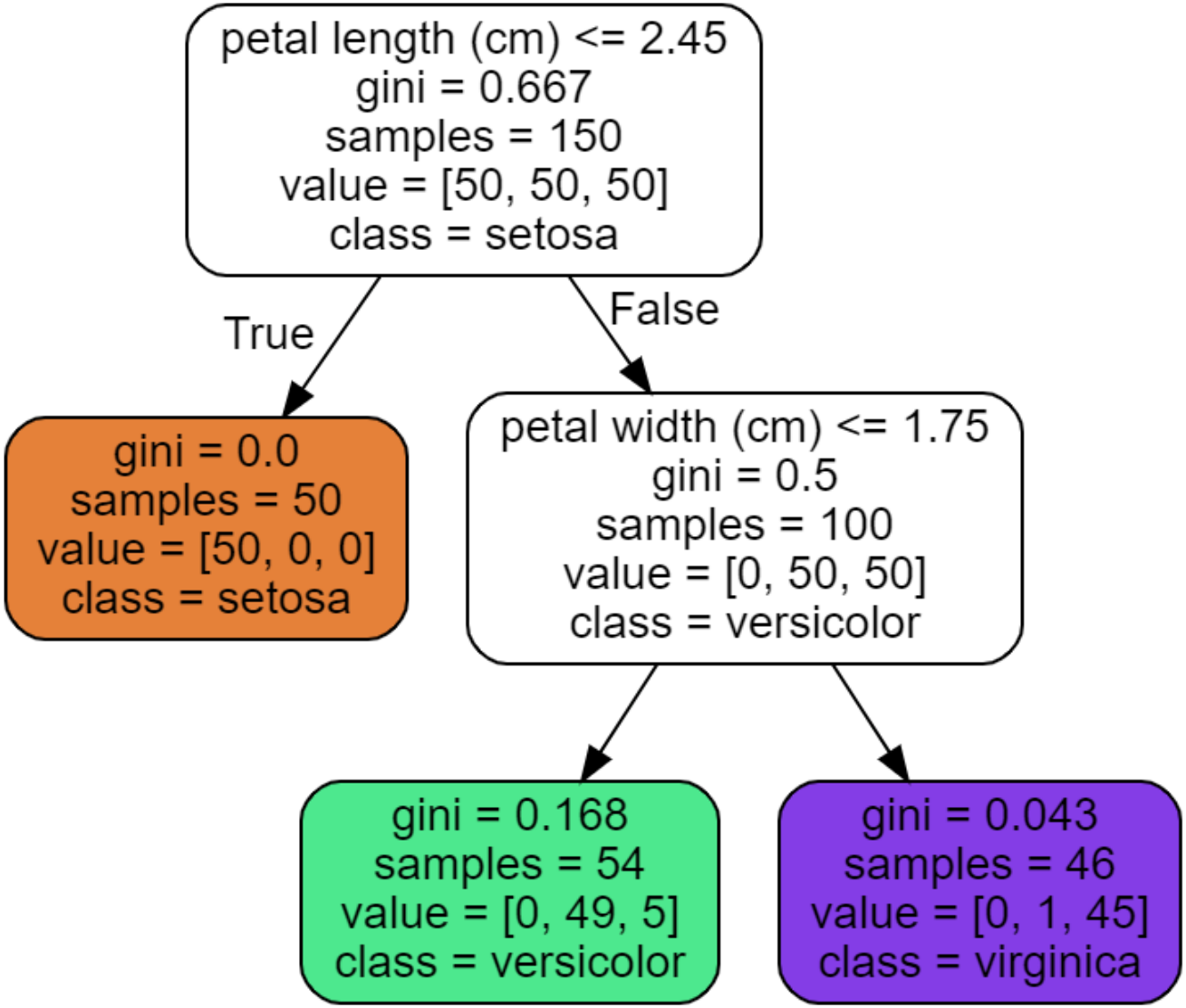
Метод бинарной классификации, основанный на использовании большого ансамбля решающих деревьев, построенных на случайных подпространствах, ошибки каждого из которых взаимно компенсируют друг друга, обеспечивая высокую точность классификации.

Дерево решений

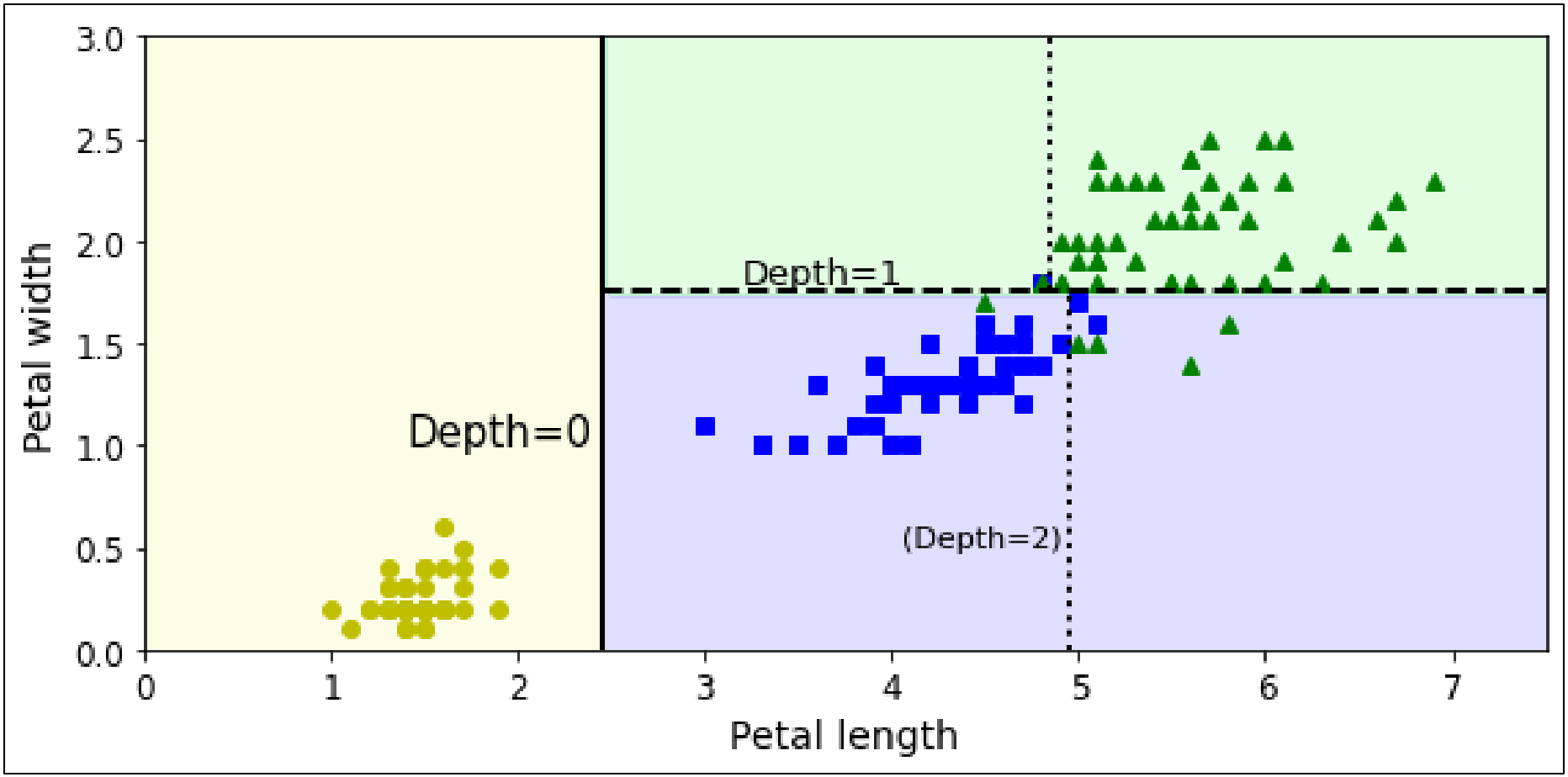
Decision Tree

Объединение логических условий в структуру дерева, состоящего из листьев (вершин графа) и веток (ребер графа).

Дерево решений



Дерево решений

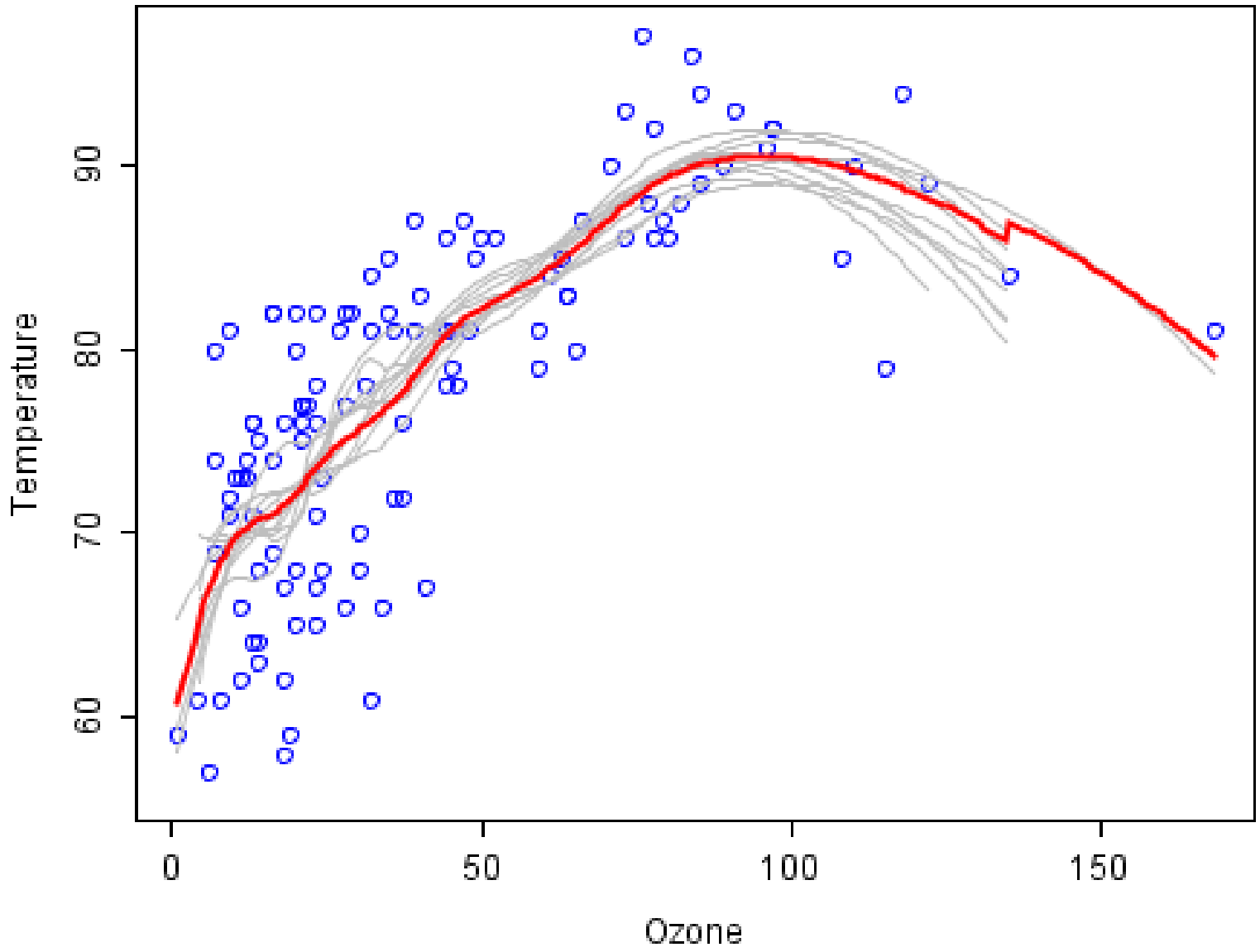


Бэггинг

Bagging, Bootstrap aggregating

Обобщение ансамбля моделей, обученных на множестве подвыборок, сформированных случайным образом из исходной обучающей выборки. Представляет собой один из способов построения усредненной модели.

Бэггинг



Бутстрэп

Bootstrapping

Метод исследования распределения статистик любых вероятностных распределений, основанный на многократной генерации выборок методом Монте-Карло на базе имеющейся исходной выборки. Является методом непараметрической статистики.

Алгоритм метода случайного леса

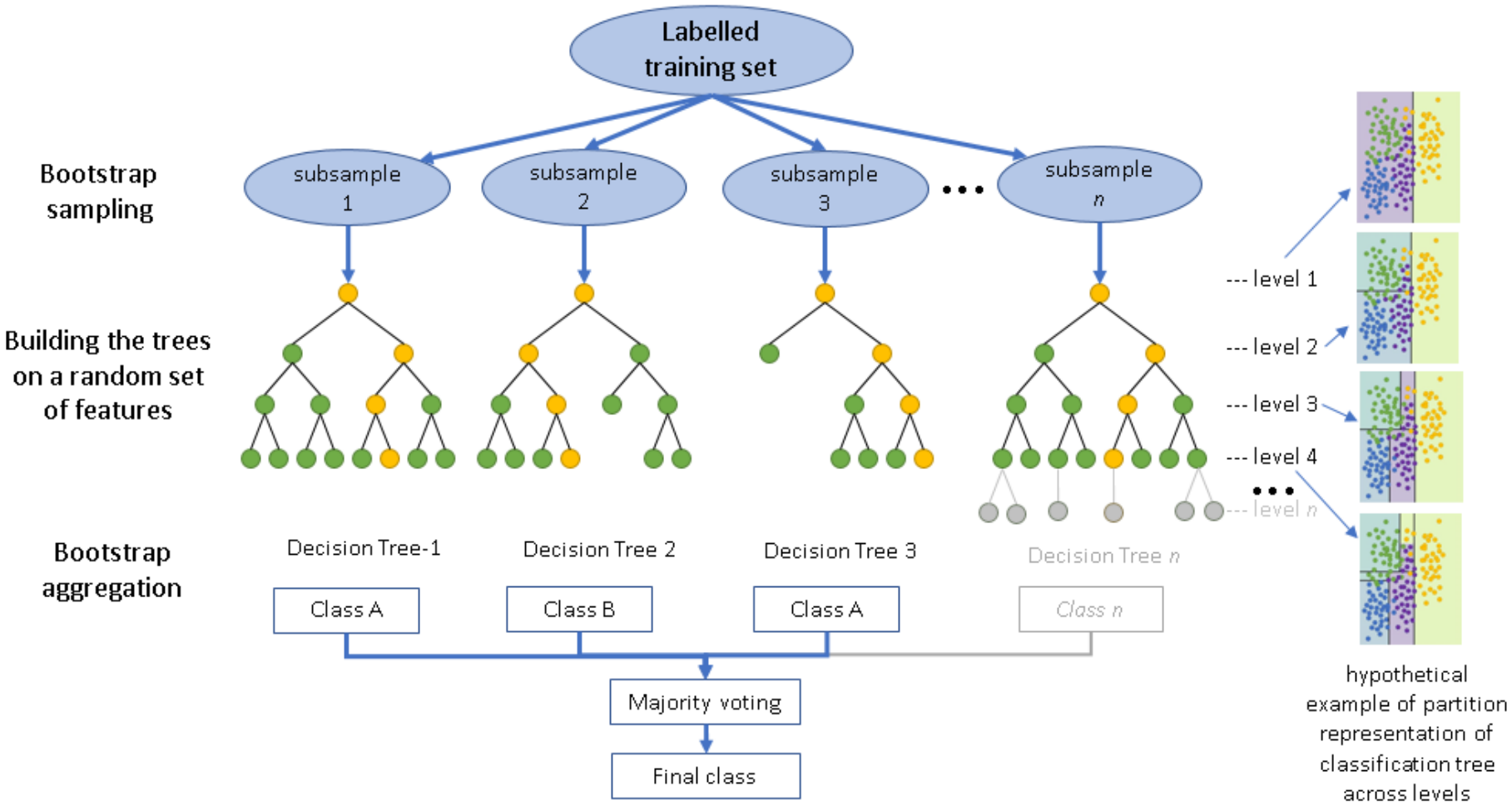
N – число наблюдений в обучающей выборке;
M – число признаков в обучающей выборке;
m – число признаков для обучения, обычно ~squart(M).

1. Создаем из обучающей выборки случайную повторную подвыборку размером *N*.
2. Построим в пространстве *m* признаков решающее дерево, классифицирующее наблюдения данной подвыборки.
3. Решающее дерево строим до полного исчерпания подвыборки.
4. Создаем ансамбль решающих деревьев, многократно повторяя пп. 1 - 3.

Алгоритм метода случайного леса

5. Проводим путем голосования ансамблевую классификацию наблюдений тестовой выборки.
6. Каждое дерево ансамбля относит классифицируемый объект к одному из двух классов; побеждает класс, за который проголосовало наибольшее число деревьев.
7. Подбираем оптимальное число деревьев, соответствующее минимуму ошибки классификации тестовой выборки.

Схема метода случайного леса



Итоговый классификатор

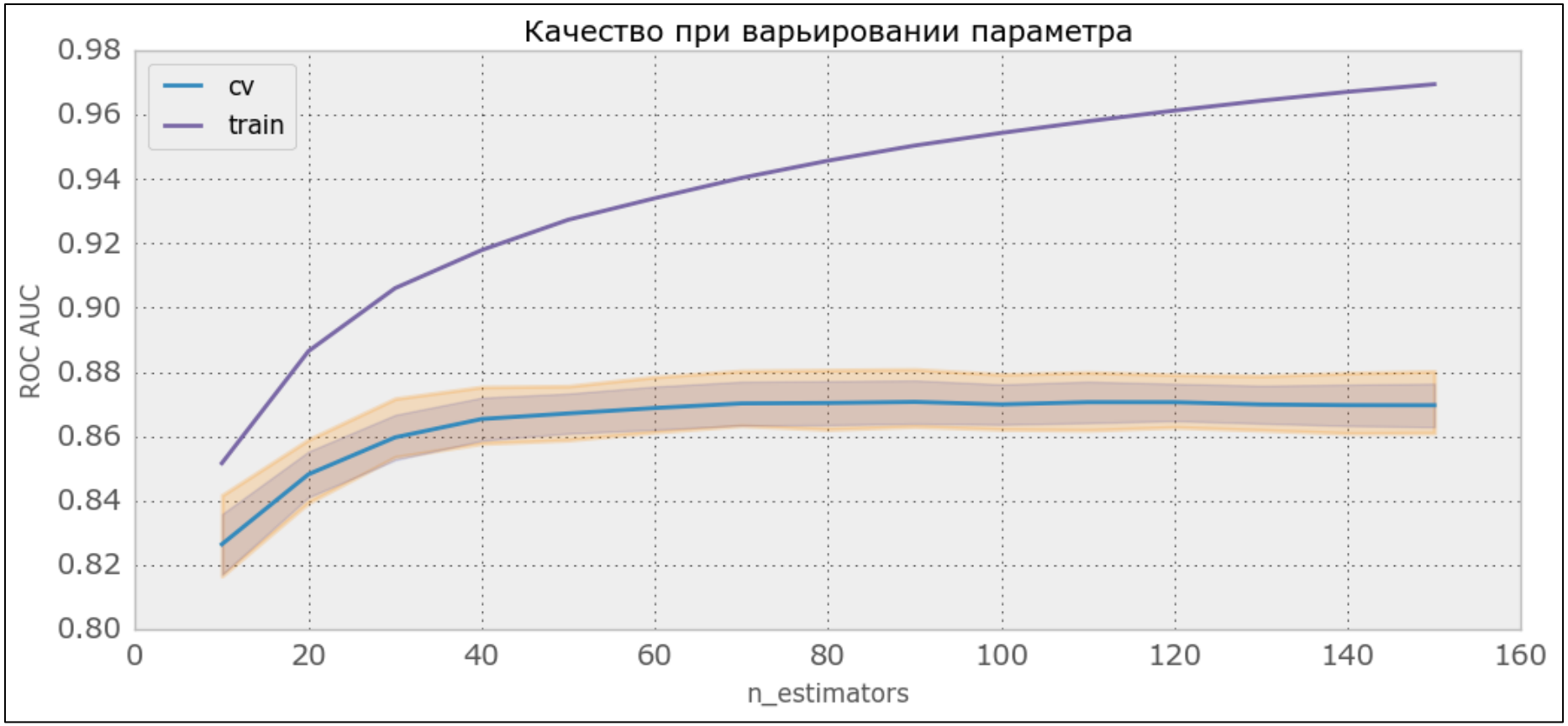
$$\mathbf{a}(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^{K} \mathbf{b}_i(\mathbf{x})$$

K – число деревьев;

$\boldsymbol{b}_i$ – i-е решающее дерево;

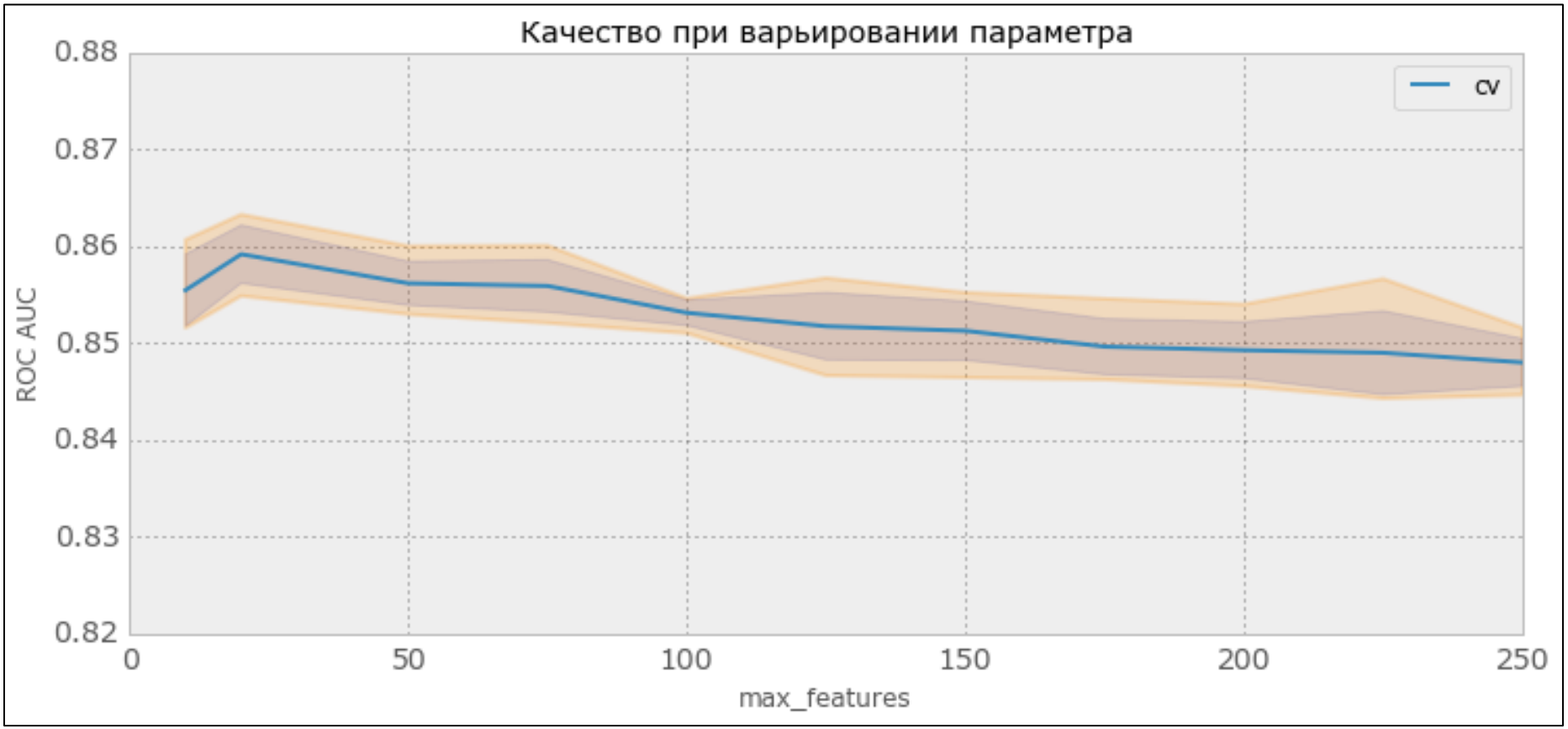
$\boldsymbol{x}$ – обучающая (тестовая) подвыборка.

Влияние числа деревьев



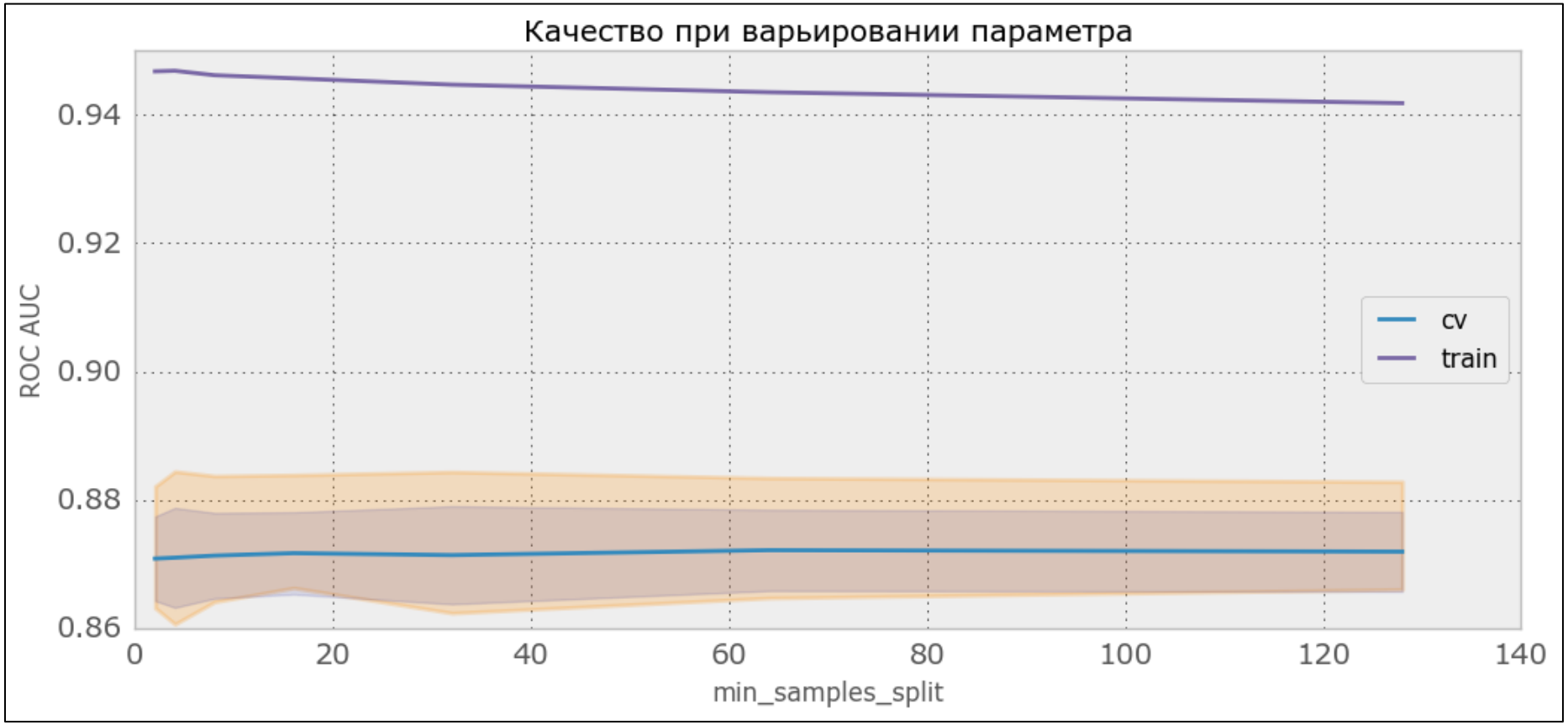
30 по умолчанию

Влияние числа признаков



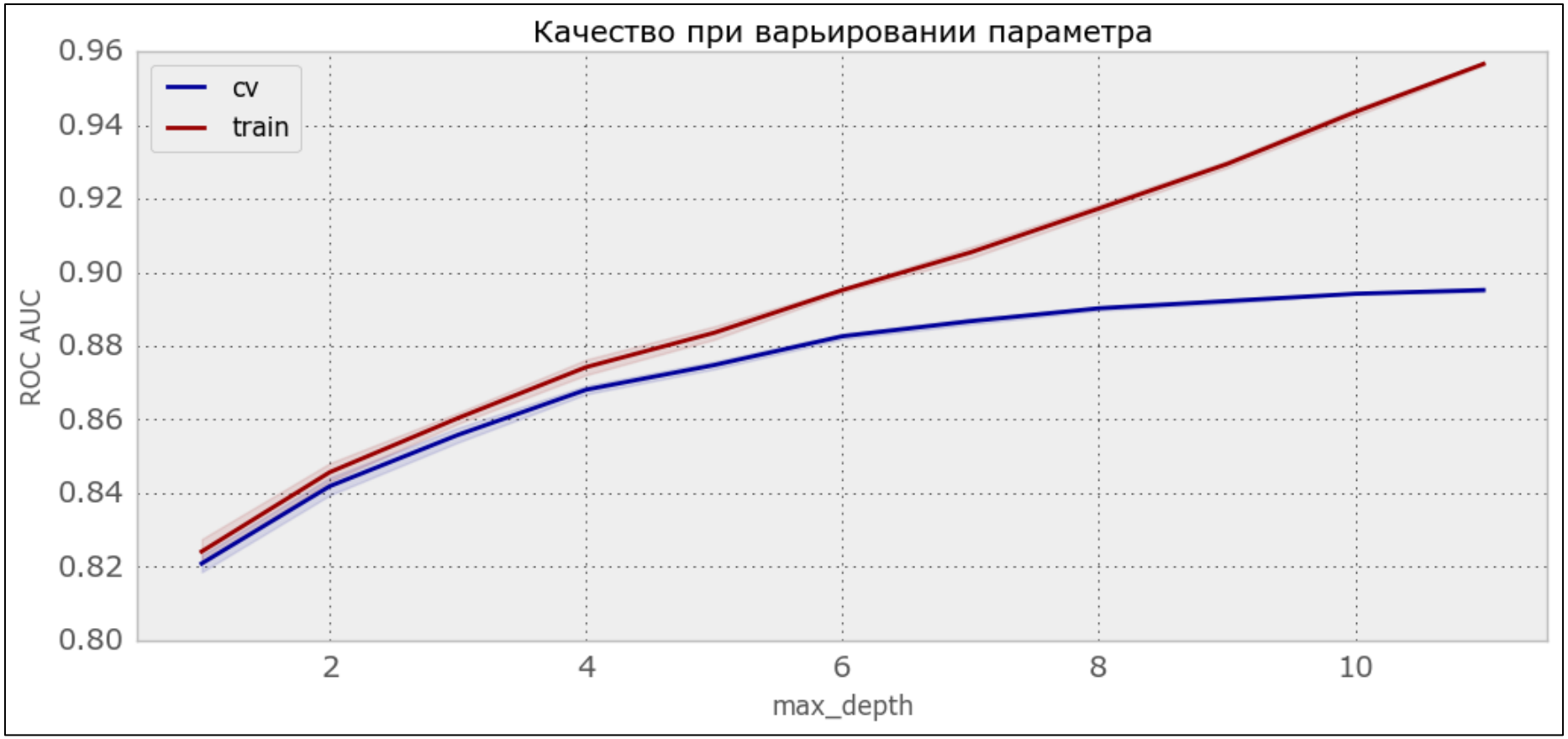
4 по умолчанию

Влияние объема подвыборки



≥5 по умолчанию

Влияние глубины деревьев



≤10 по умолчанию

Достоинства метода случайного леса

1. Способен обрабатывать данные большой размерности.
2. Применим для *k*-нарной классификации.
3. Нечувствителен к любым монотонным преобразованиям исходных переменных.
4. Эффективно обрабатывает как непрерывные, так и дискретные признаки.
5. Мало чувствителен к выбросам в данных из-за случайного семплирования.
6. Хорошо работает с пропущенными данными.
7. Хорошо распараллеливается и масштабируется.

Недостатки метода случайного леса

1. Требует значительных вычислительных ресурсов.
2. Большой размер модели.
3. Склонен к переобучению на зашумленных данных.
4. Результаты сложно интерпретировать.
5. Плохо работает на разреженных признаках.
6. Не обладает возможностью экстраполяции.
7. На коррелированных данных склонен к недообучению.


To be continued ...