*Вопросы занятия*

1. Задача классификации: постановка и примеры 2. Дерево решений: как его построить?

3. Достоинства и недостатки деревьев решений.

4. Визуализируем принятие решений и предсказания алгоритма.

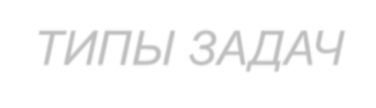
*В конце занятия научимся:*

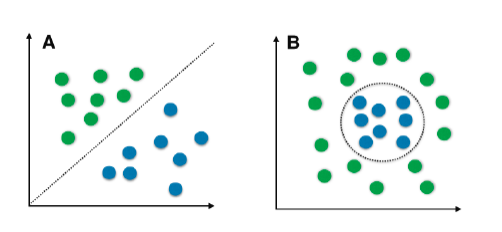
- применять *алгоритм классификации*, принятие решений которого *можно проинтерпретировать;*

*- измерять качество* решений в задачах классификации;

- *оценивать важность* фичей.

*1. ЗАДАЧА КЛАССИФИКАЦИИ*

*ТИПЫ ЗАДАЧ*

• классификация • ранжирование • регрессия

• кластеризация

*1. ЗАДАЧА КЛАССИФИКАЦИИ*



*ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ*

***Скоринг.*** Оценка риска выдачи клиенту кредита? (banking, insurance)

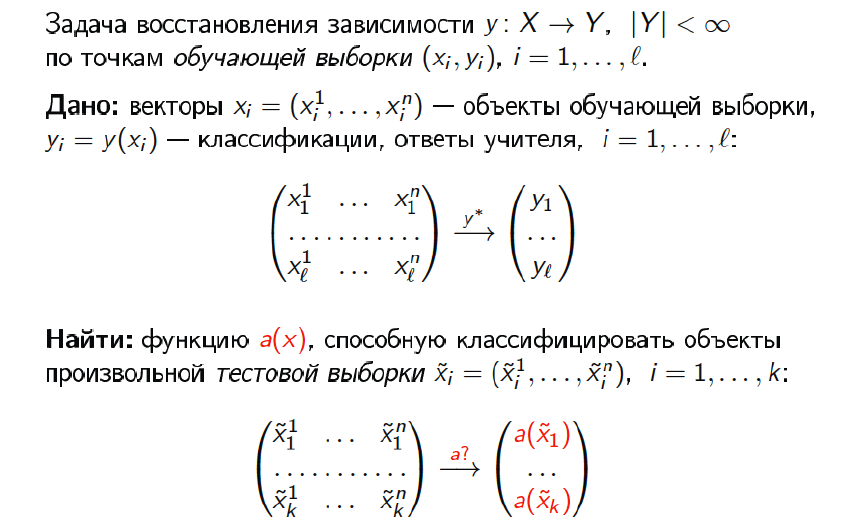


***Отток.*** Перестанет ли пользоваться клиент услугами компании? Перестанет ли, если дать ему бонус? (marketing)

***Intent recognition.***О чём говорит пользователь в своём обращении? (может быть несколько intent’ов, может быть древовидная структура)

***Image recognition.***Распознавание образов

*1. ЗАДАЧА КЛАССИФИКАЦИИ*



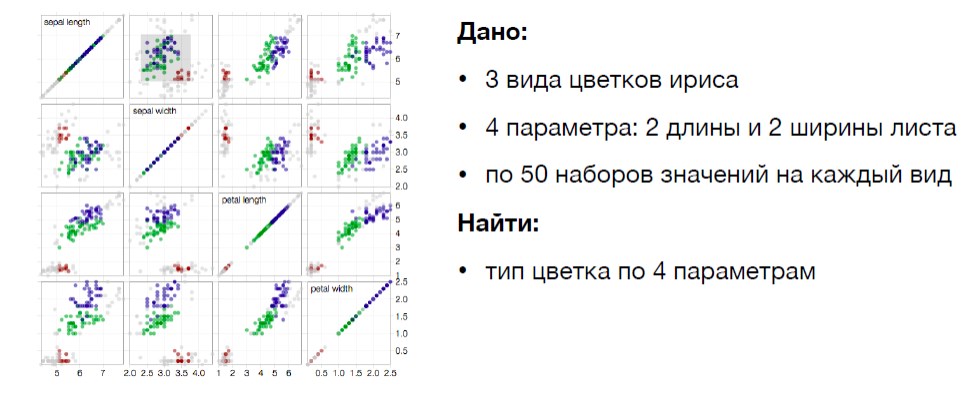
*ПОСТАВНОВКАЗАДАЧИ*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



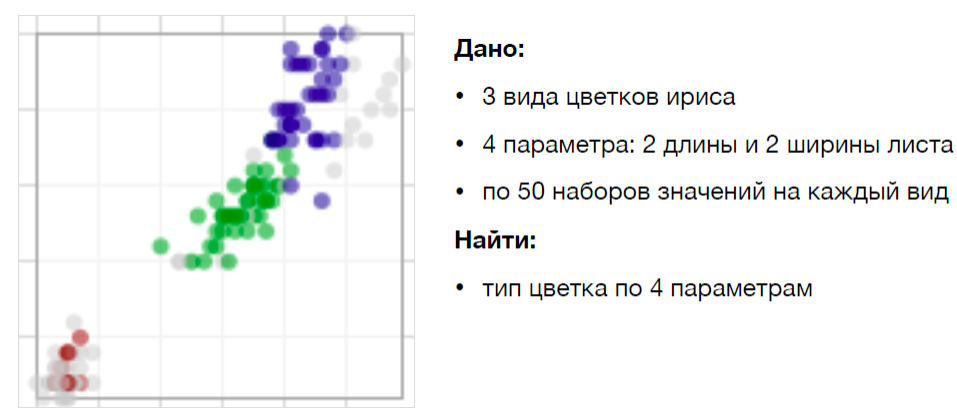
*ЦВЕТКИ ИРИСА: ЗАДАЧА*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



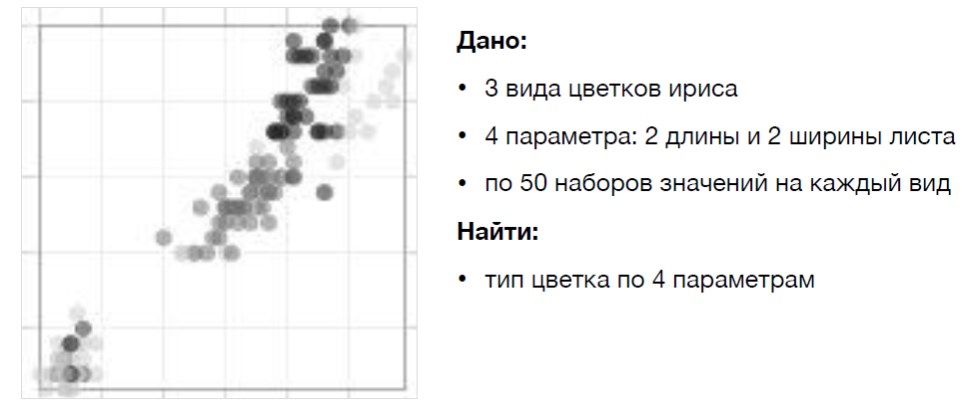
*ЦВЕТКИ ИРИСА: ДАННЫЕ*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



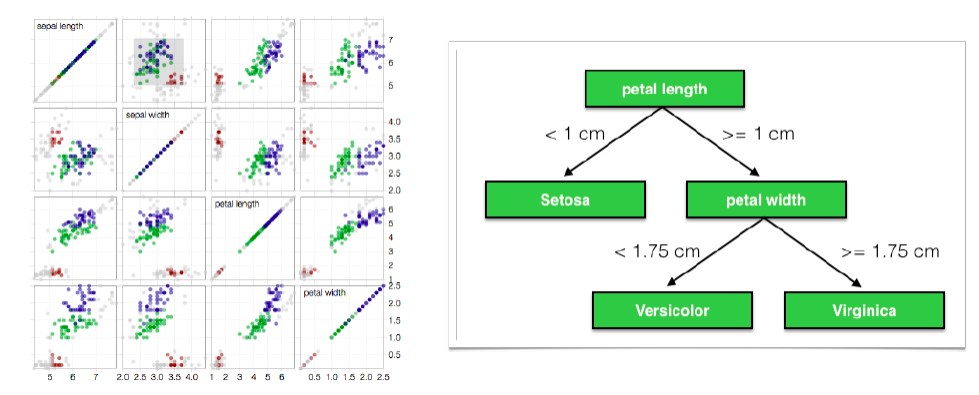
*ЦВЕТКИ ИРИСА: ДАННЫЕ*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



*ЦВЕТКИ ИРИСА: ДАННЫЕ*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*

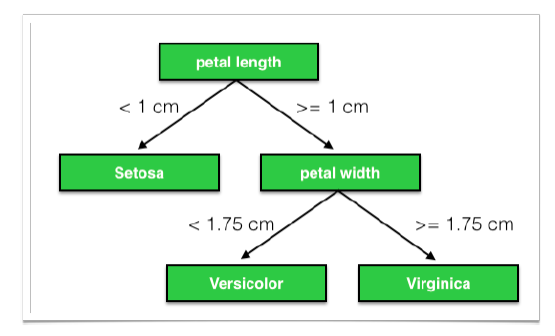


*ЦВЕТКИ ИРИСА: РЕШАЮЩЕЕДЕРЕВО*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



*ПОСТРОЕНИЕДЕРЕВА*

**Определить:**

-вид правила разбиения

-критерий информативности разбиения

-критерий останова

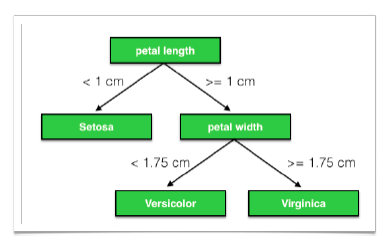
-метод стрижки

- обработка пропусков

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



*ВИД ПРАВИЛАРАЗБИЕНИЯ*

- **одномерное**: сравниваетсязначение одной фичи вектора x

- **линейное**: сравниваетсялинейнаякомбинация фичей x

- **метрическое**:

расстояние до точки признакового пространства

*здесь используется одномерный предикат: сравнение идёт лишь по одной фиче из вектора признаков*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ.ФУНКЦИОНАЛКАЧЕСТВА РАЗБИЕНИЯ*



*ФУНКЦИОНАЛ КАЧЕСТВА РАЗБИЕНИЯ*

*Идея:*

- взять признак

- отсортировать его по возрастанию

- в зависимости от целевой переменной установить порог разделения выборки на две, максимально снижая численно выражаемый разброс внутри каждой из 2 групп

- подобрать лучшее с точки зрения улучшения разбиение

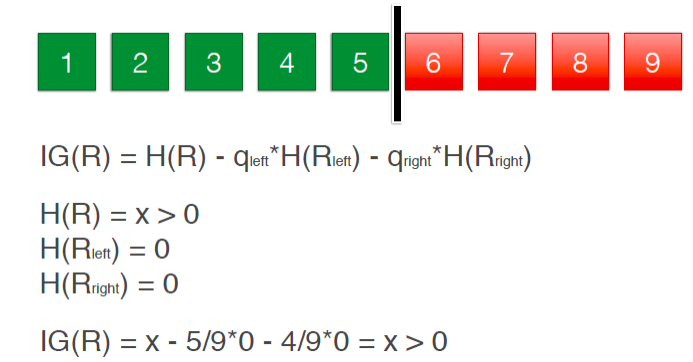
*Вопрос:* как измерить улучшение?

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ.ФУНКЦИОНАЛКАЧЕСТВА РАЗБИЕНИЯ*



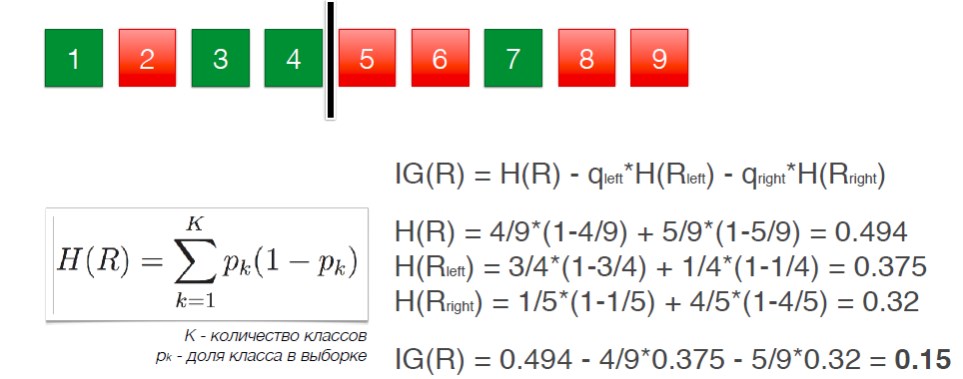
*ИЗМЕРЕНИЕПОЭТАПНОГО УЛУЧШЕНИЯ*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ.ФУНКЦИОНАЛКАЧЕСТВА РАЗБИЕНИЯ*



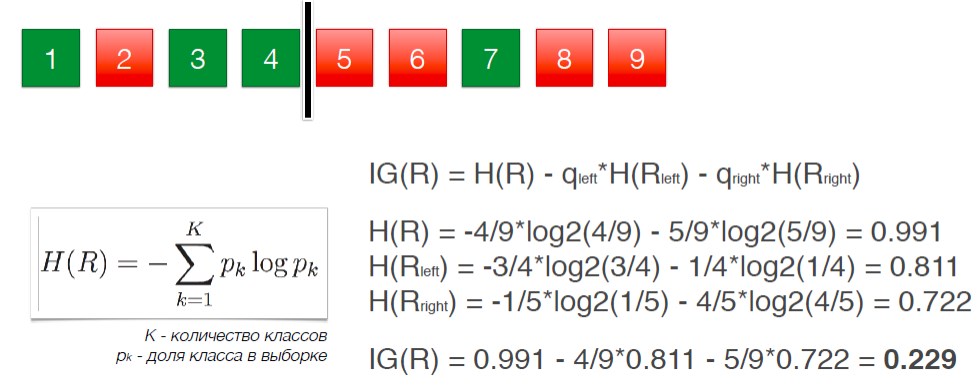
*ИЗМЕРЕНИЕПОЭТАПНОГО УЛУЧШЕНИЯ*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ.ФУНКЦИОНАЛКАЧЕСТВА РАЗБИЕНИЯ*



*КРИТЕРИЙДЖИНИ*

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ.ФУНКЦИОНАЛКАЧЕСТВА РАЗБИЕНИЯ*



*ЭНТРОПИЙНЫЙКРИТЕРИЙ*

*ПРАКТИЧЕСКИЕ ЗАДАНИЯ*

1. Построитькритерии информативности: джини и энтропийный

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



*КРИТЕРИИОСТАНОВА*

-останов,когда в каждом листе объекты только одного класса

-ограничение max глубины дерева

-ограничение min число объектовв листьях

- требованиеулучшения функционала качества при дроблении не менее, чем x или на x%

*2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ*



*ПРОБЛЕМА ПРОПУСКОВ*

- удалить объекты с пропусками из обучающей;

- замена на значения вне средние, медианные;

- заменить на значения вне области значений фич;

- модифицировать алгоритм построения и работы дерева: включать элементы с пропусками в обе ветки дерева, но взвешивать качество разбиения по объёму пропусков

*3. ДОСТОИНСТВА ИНЕДОСТАТКИДЕРЕВЬЕВРЕШЕНИЙ*



*ДОСТОИНСТВА*

- легко интерпретировать, визуализировать, «белый ящик»;

- простота подготовки данных: не требуется нормализация, dummy переменные

- скорость работы.

*3. ДОСТОИНСТВА ИНЕДОСТАТКИДЕРЕВЬЕВРЕШЕНИЙ*



*НЕДОСТАТКИ* - острая проблема переобучения;

- неустойчивость;

- не учитывает нелинейные зависимости или даже простые линейные, которые идут не по осям координат;

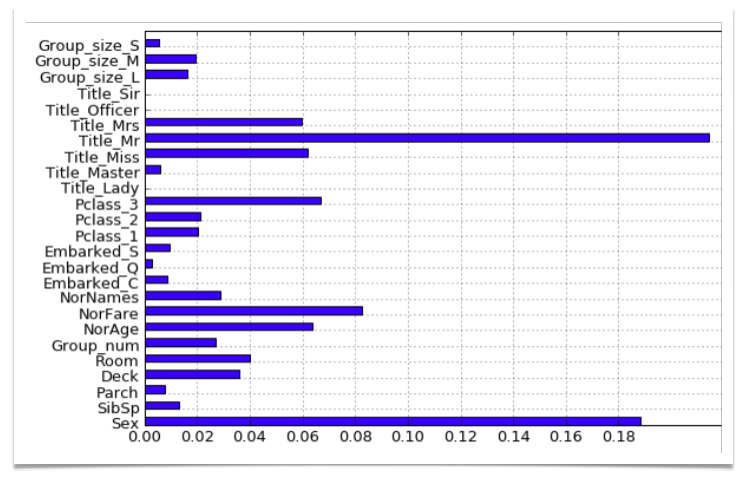
- чувствителен к несбалансированным классам;

- хорошо интерполирует, плохо экстраполирует.

*ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ РЕАЛИЗАЦИЯВ SKLEARN*



Деревья могут оценивать важность фичей. Метод: feature\_importances\_

Например,судяпо решению,на выживаемостьна Титанике сильнее всего влияли:

\* наличие в обращении «Mr.» \* пол

\* уровень дохода

\* проживаниев 3 классе \* возраст

\* наличие в обращении «Mrs» / «Miss»

*ПРАКТИЧЕСКИЕ ЗАДАНИЯ*

2. Обучить дерево решений на цветкахириса

3. Нарисоватьдерево принятий решений

4. Оценить важностьфичей

*ЧТО МЫ СЕГОДНЯ УЗНАЛИ*

1. Деревья решений, объединённые в «лес», составляют одни из наиболее сильных алгоритмов. По одиночке же они являются слабыми, зато очень легко интерпретируемыми и визуализируемыми алгоритмами.

2. Деревья позволяют оценивать важность признаков.