Деревья решений

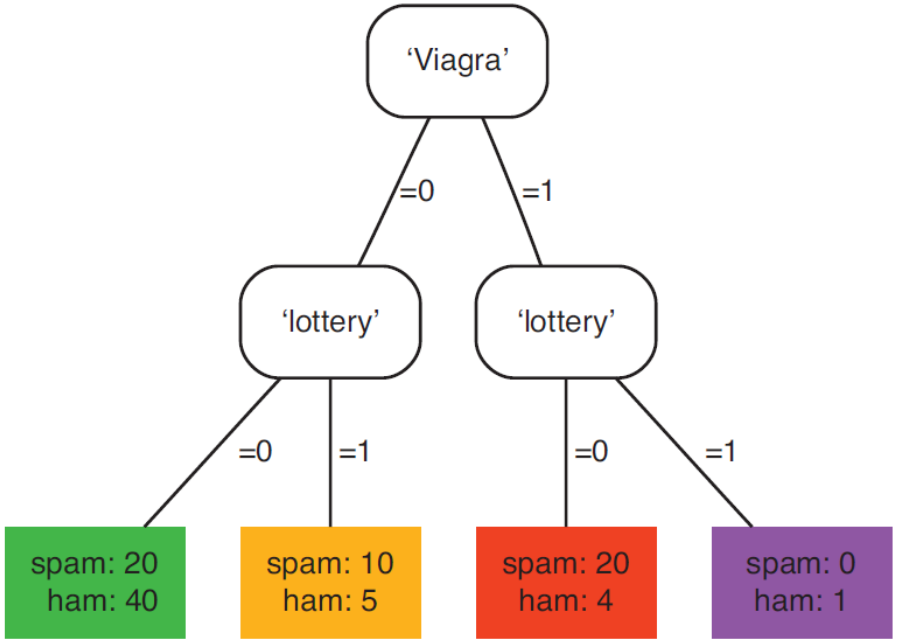
Лекция 4

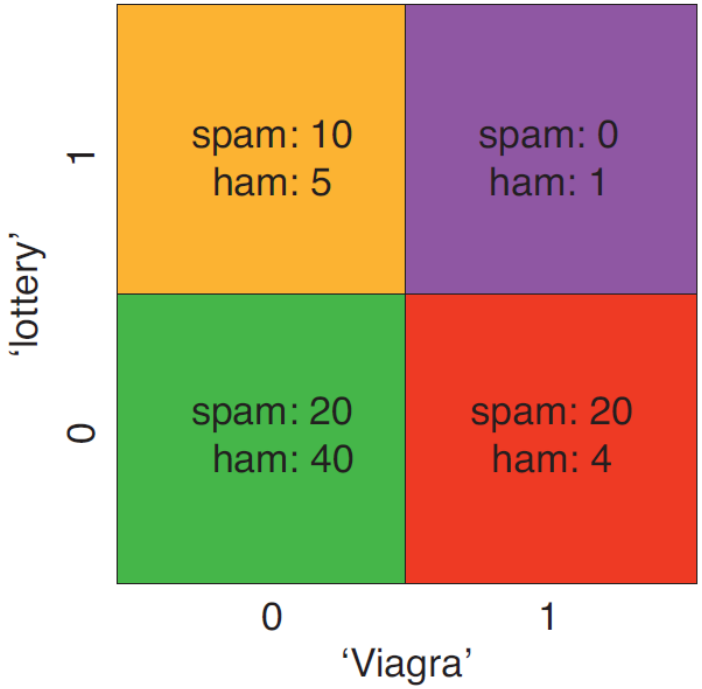
Логические методы

- Логические методы классификации:
  - деревья решений (decision trees)
  - решающие списки (decision lists)
  - взвешенное голосование правил (weighted voting)
  - поиск ассоциативных правил (association rules)
- Классификатор представляет собой набор правил вида ЕСЛИ … ТО …
- Правила индуктивно генерируются на основе обучающих данных

Деревья решений

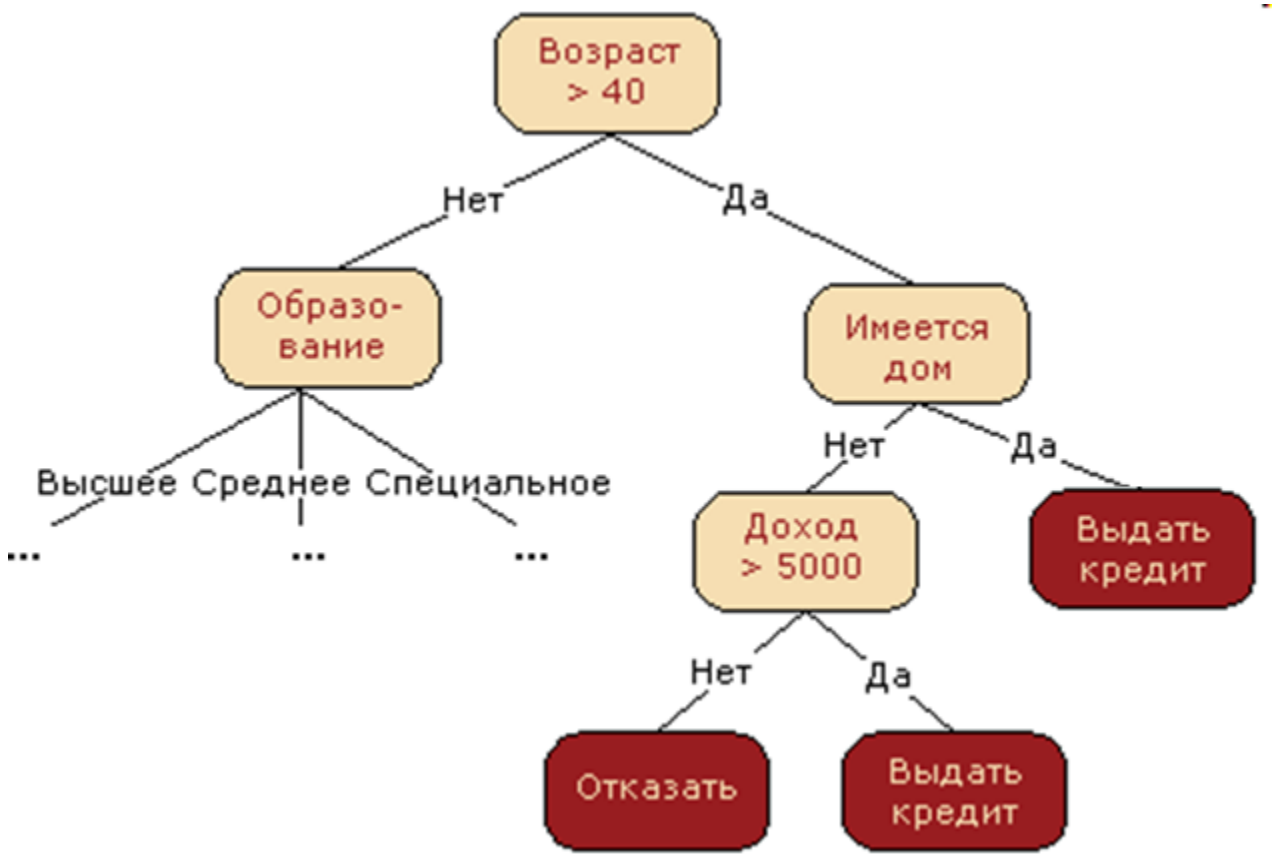
- Пример дерева решений (классификация):





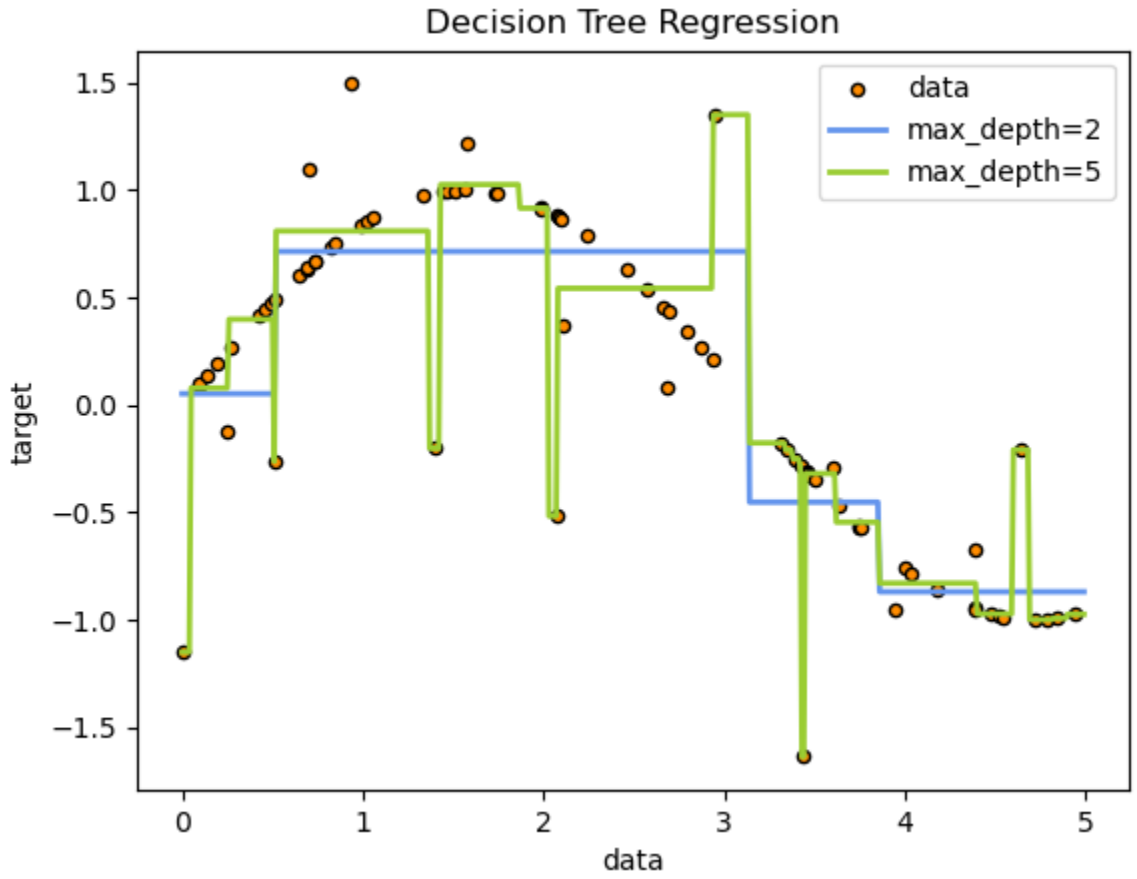
Деревья решений

- Пример дерева решений (классификация):



Деревья решений

- Пример дерева решений (регрессия):



Деревья решений

- Деревья решений позволяют решать задачи классификации и регрессии
- По дереву решений легко можно построить набор правил
  - нужно пройти от корня дерева до каждого листа
- Для повышения обобщающей способности дерева решений:
  - число листьев должно быть как можно меньше
  - листья должны покрывать подвыборки примерно одинаковой мощности
- Задача построения дерева минимальной сложности, правильно классифицирующего заданную выборку, в общем случае является NP-полной
- На практике используются различные эвристики

Алгоритмы построения

- Существует множество алгоритмов построения деревьев решений:
  - CART (Л. Брейман и др., 1984)
  - ID3, C4.5, See5/C5.0 (Дж. Р. Куинлан, 1986)
  - CHAID – Chi-square Automatic Interaction Detector (Г. Касс, 1980 г.)
  - CN2 (П. Кларк. Т. Ниблетт, 1988 г.)
  - MARS (Фридман Дж., 1991 г.)
  - OC1 (Murthy S., 1994 г.)
  - …

Алгоритмы построения

Общий принцип:

- на каждом шаге алгоритма имеется множество обучающих примеров T
- каждый пример задается множеством признаков $F = \{f_1, \dots, f_m\}$
- если множество T содержит примеры только одного класса, то формируется *лист* дерева
- если множество *T* содержит примеры разных классов, то формируется *внутренний узел* дерева:
  - выбирается один из признаков f_i
  - значения признака $f_i = \{d_1, \dots, d_k, \dots, d_n\}$
  - множество T делится на n подмножеств:

$$T_1, \dots, T_k, \dots, T_n$$

  - подмножество T_k содержит примеры только с одним значением признака $f_i = d_k$

Алгоритмы построения

В алгоритме построения дерева решений должны быть определены следующие правила:

- правило выбора признака f_i на каждом шаге
- правило остановки разбиения
- правило отсечения

Алгоритм ID3

- Алгоритм ID3 (1986 г.)
  – John Ross Quinlan
- Iterative Dichotomiser 3

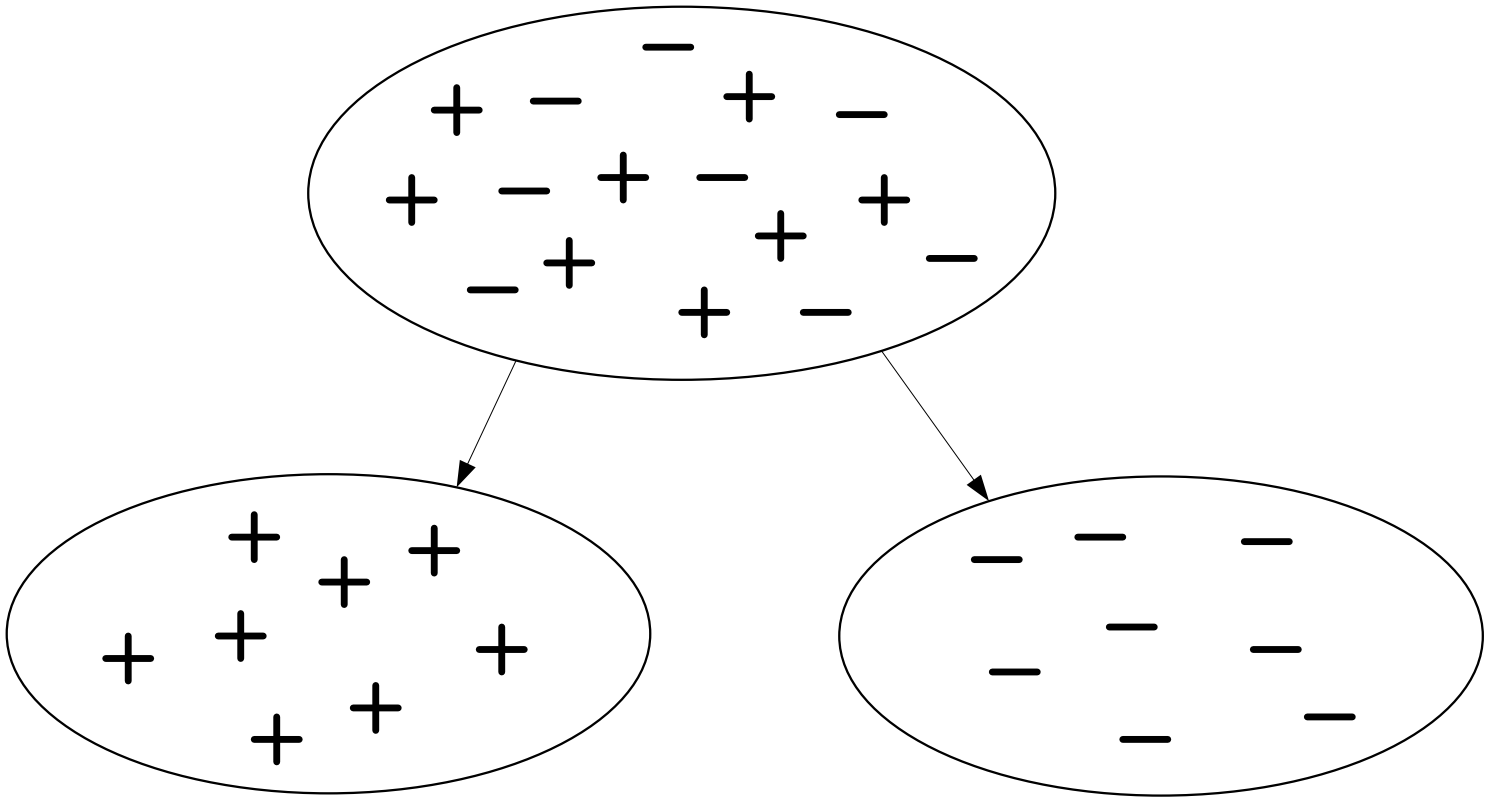
Алгоритм ID3

Правило выбора признака

- Выберем некоторый признак f_i и проведем по нему разбиение множества T
- Как узнать, хорошее или плохое разбиение получили?
- Единственная информация – распределение примеров по классам

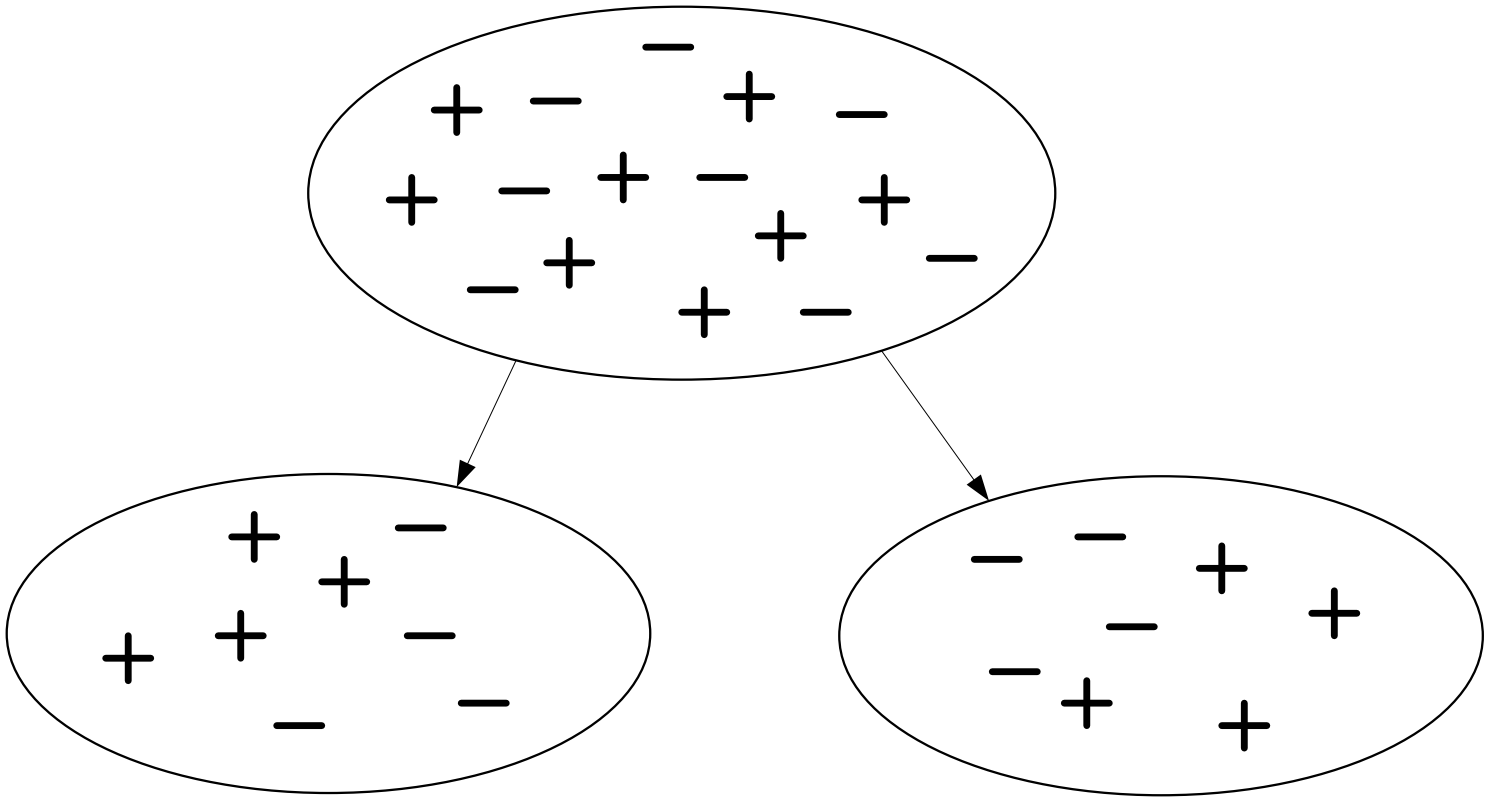
Алгоритм ID3

- Наилучший вариант:



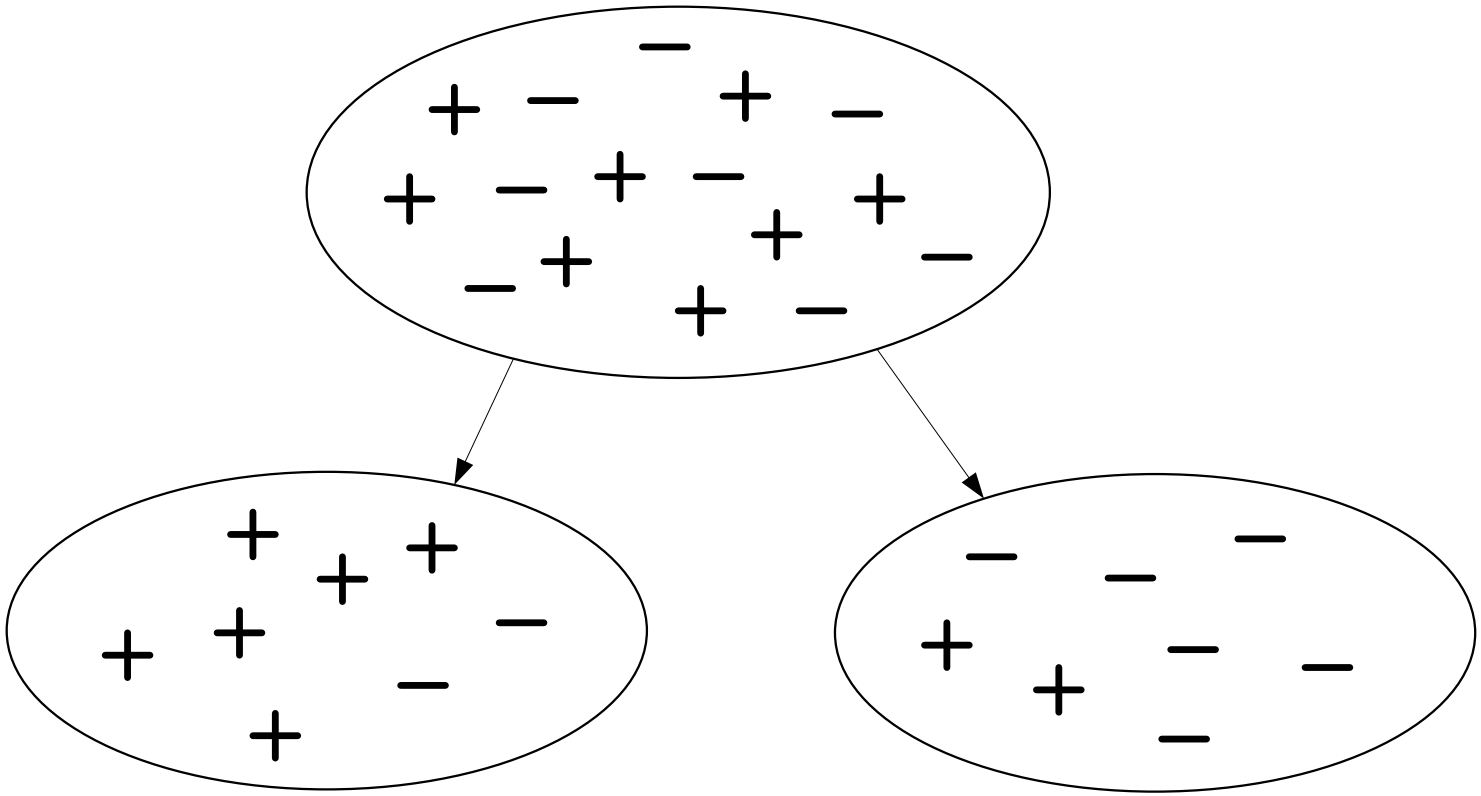
Алгоритм ID3

- Наихудший вариант:



Алгоритм ID3

- Средний вариант:



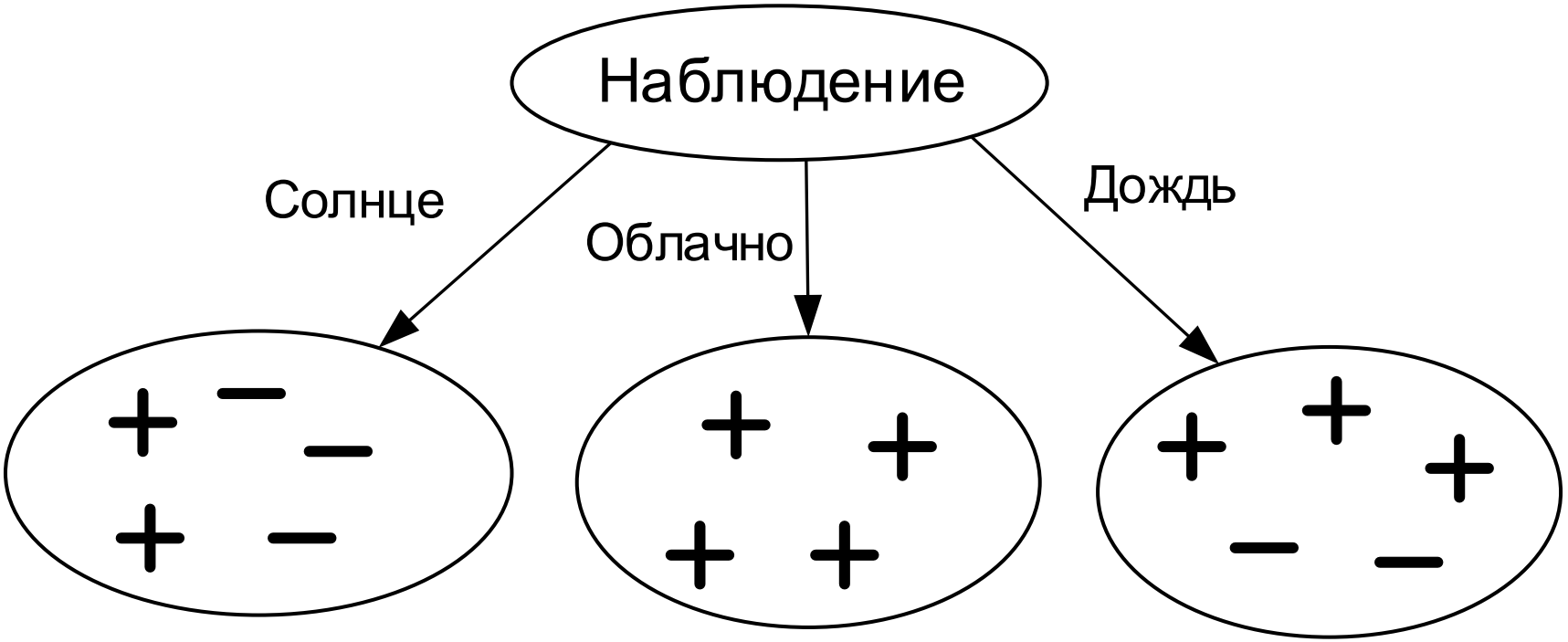
Алгоритм ID3

- Рассмотрим пример «Состоится ли игра?»
- Разобьем множество исходных примеров поочередно по каждому из четырех признаков

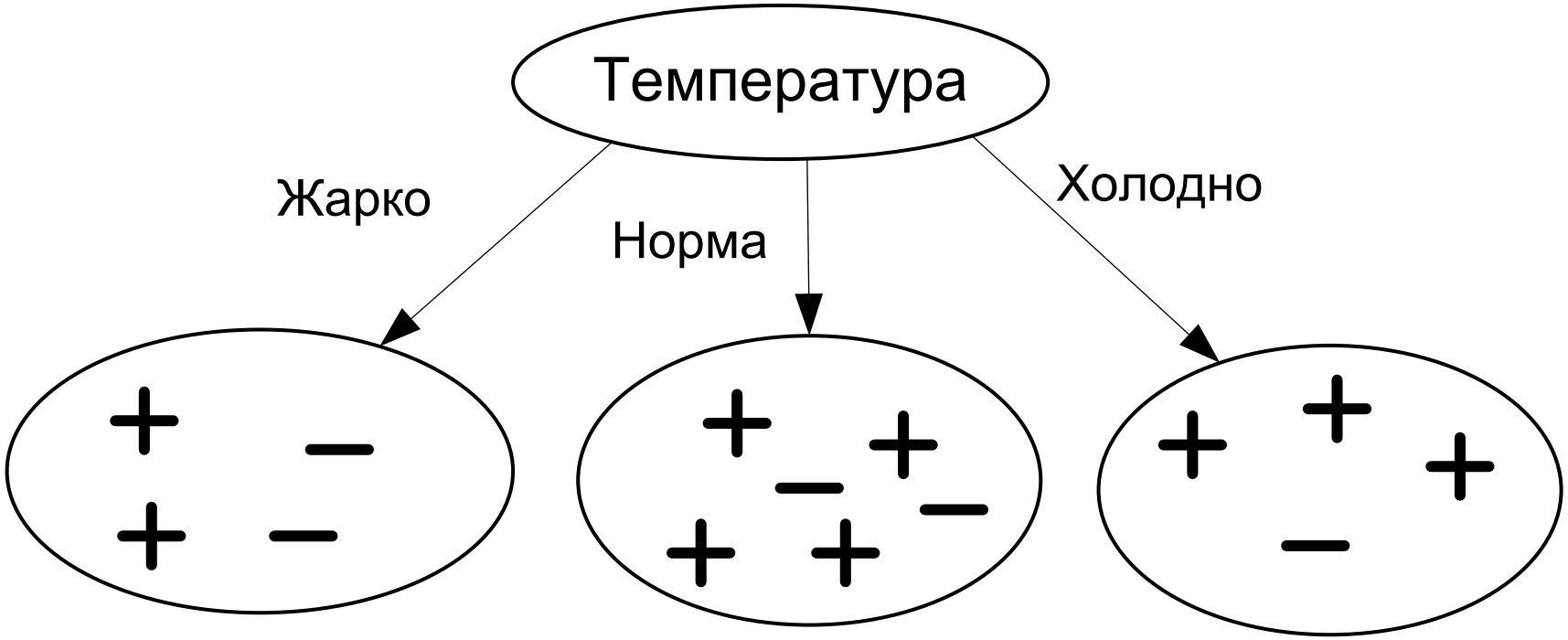
Состоится ли игра?

Наблюдение	Температура	Влажность	Ветер	Игра
Солнце	Жарко	Высокая	Нет	Нет
Солнце	Жарко	Высокая	Есть	Нет
Облачность	Жарко	Высокая	Нет	Да
Дождь	Норма	Высокая	Нет	Да
Дождь	Холодно	Норма	Нет	Да
Дождь	Холодно	Норма	Есть	Нет
Облачность	Холодно	Норма	Есть	Да
Солнце	Норма	Высокая	Нет	Нет
Солнце	Холодно	Норма	Нет	Да
Дождь	Норма	Норма	Нет	Да
Солнце	Норма	Норма	Есть	Да
Облачность	Норма	Высокая	Есть	Да
Облачность	Жарко	Норма	Нет	Да
Дождь	Норма	Высокая	Есть	Нет

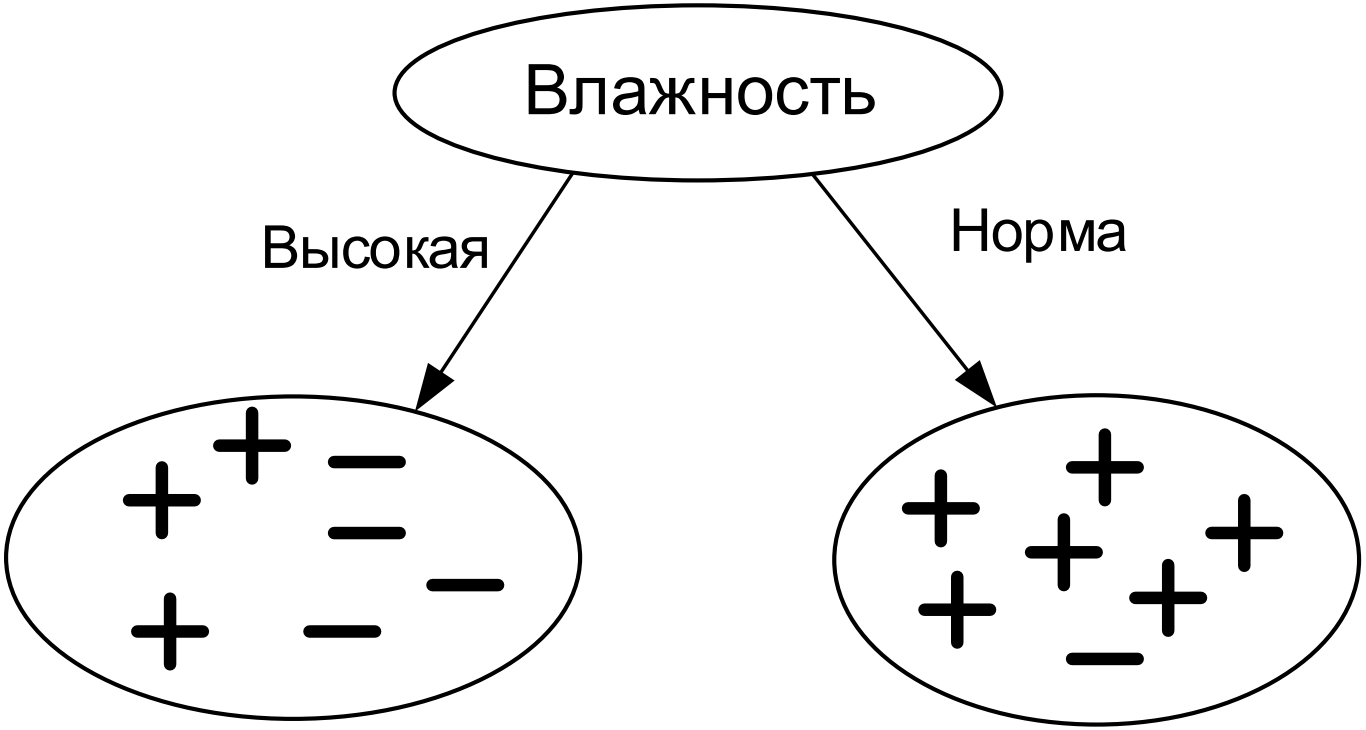
Алгоритм ID3



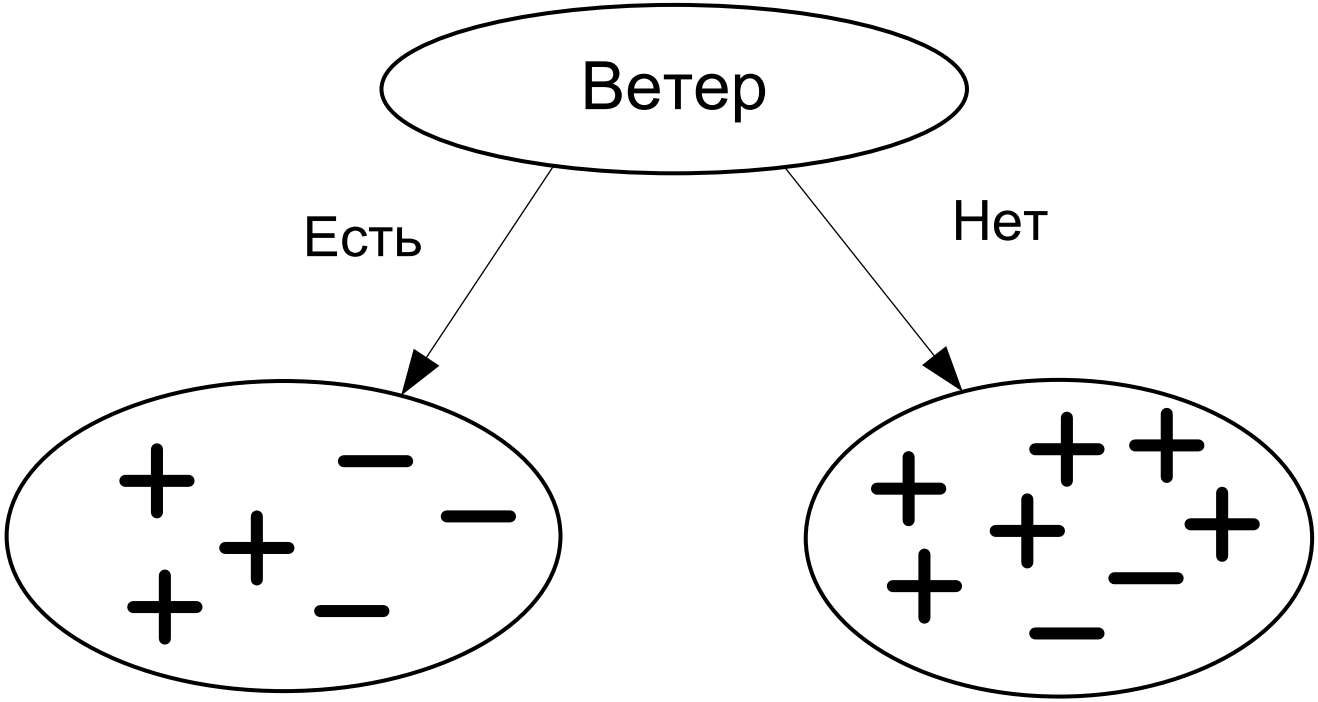
Алгоритм ID3



Алгоритм ID3

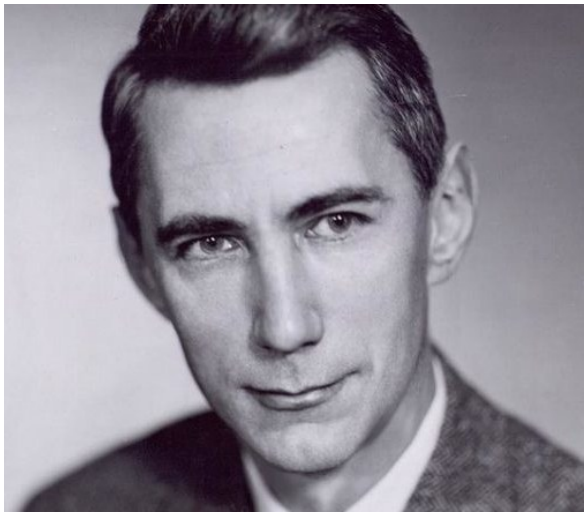


Алгоритм ID3



Алгоритм ID3

- Как формально выбрать признак f_i?
- Нужен некий критерий полезности выбора признака
- В алгоритме ID3 используется теория информации (Клод Ше́ннон, 1948 г.)
- Подходящим критерием является ожидаемый объем информации, предоставляемый признаком



Количество информации

- *Количество (объем) информации* (по Клоду Шеннону):
  - имеется множество возможных событий, вероятности осуществления которых $p_1, \dots, p_n$
  - эти вероятности известны, но это – всё, что нам известно относительно того, какое событие произойдет
- Можно ли найти меру того, насколько велик «выбор» из такого набора событий или сколь неопределенен для нас его исход?

Количество информации

- Количество информации (*энтропия множества вероятностей*) находится по формуле:

$$Info(p_1, p_2, \dots, p_n) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

- Количество информации измеряется в *битах*

Количество информации

- ***Пример***. В результате подбрасывания монеты возможны два события с вероятностями 0,5

$$Info\left(\frac{1}{2},\frac{1}{2}\right) = -\frac{1}{2}\cdot\log_2\frac{1}{2} - \frac{1}{2}\cdot\log_2\frac{1}{2} = 1 \text{ бит}$$

- *Бит* – максимальное количество информации, которое можно получить при ответе на вопрос в форме «да» – «нет»
- *Энтропия* (др.-греч. *поворот, превращение*) – мера неопределённости (неупорядоченности) некоторой системы
- Чем меньше порядок, предсказуемость, тем выше энтропия

Количество информации

- ***Пример***. Пусть возможны два события с вероятностями 1 и 0:

$$Info(1,0) = -1 \cdot \log_2 1 - 0 \cdot \log_2 0 = 0 \text{ бит}$$

Алгоритм ID3

- Выбор признака f_i при построении дерева решений нужно осуществлять так, чтобы приращение информации в результате выбора было максимальным
- Обозначим:
  - $Info_a$ – количество информации, необходимое для определения класса объекта из множества T (до разбиения)
  - $Info_b(f_i)$ – количество информации, необходимое для определения класса объекта после разбиения по признаку f_i
  - $Gain(f_i)$ – приращение информации при разбиении по признаку f_i:

$$Gain(f_i) = Info_a - Info_b(f_i)$$

Алгоритм ID3

- Рассмотрим пример с двумя классами (см. пример «Состоится ли игра?»):

$$Info_a = Info(p^+, p^-) = -p^+ \cdot \log_2 p^+ - p^- \cdot \log_2 p^- ,$$

где p^+ – вероятность отнесения примера к классу «+»

p^- – вероятность отнесения примера к классу «–»

Алгоритм ID3

- Вычислим вероятности и количество информации:

$$p^+ = \frac{9}{14}, \qquad p^- = \frac{5}{14}$$

$$Info\left(\frac{9}{14}, \frac{5}{14}\right) = -\frac{9}{14} \cdot \log_2 \frac{9}{14} - \frac{5}{14} \cdot \log_2 \frac{5}{14} = 0{,}94 \text{ бит}$$

- Таким образом:
  - в наилучшем случае при разбиении по признаку f_i объем информации составит 0.94 бита
  - в наихудшем случае при разбиении по признаку f_i объем информации составит 0 бит
  - В остальных случаях объем информации будет находиться в этих пределах

Алгоритм ID3

- Признак $f_i = \{d_1, \dots, d_k, \dots, d_n\}$ делит множество T на n подмножеств:

$$T_1, \dots, T_k, \dots, T_n$$

- Для каждого из подмножеств T_k количество информации, необходимое для определения класса примера, вычисляется по формуле:

$$Info(p_k^+, p_k^-) = -p_k^+ \cdot \log_2 p_k^+ - p_k^- \cdot \log_2 p_k^-$$

Алгоритм ID3

- Для примера возьмем признак «*наблюдение*»:
  - $p_1^+ =$, $p_1^- =$ $(d_1 = \text{«солнце»})$
  - $p_2^+ =$, $p_2^- =$ $(d_2 = \text{«облачно»})$
  - $p_3^+ =$, $p_3^- =$ $(d_3 = \text{«дождь»})$

$p_1^+ =$, $p_1^- =$ $(d_1 = \text{«солнце»})$

$p_2^+ =$, $p_2^- =$ $(d_2 = \text{«облачно»})$

$p_3^+ =$, $p_3^- =$ $(d_3 = \text{«дождь»})$

Наблюдение	Температура	Влажность	Ветер	Игра
Солнце	Жарко	Высокая	Нет	Нет
Солнце	Жарко	Высокая	Есть	Нет
Облачность	Жарко	Высокая	Нет	Да
Дождь	Норма	Высокая	Нет	Да
Дождь	Холодно	Норма	Нет	Да
Дождь	Холодно	Норма	Есть	Нет
Облачность	Холодно	Норма	Есть	Да
Солнце	Норма	Высокая	Нет	Нет
Солнце	Холодно	Норма	Нет	Да
Дождь	Норма	Норма	Нет	Да
Солнце	Норма	Норма	Есть	Да
Облачность	Норма	Высокая	Есть	Да
Облачность	Жарко	Норма	Нет	Да
Дождь	Норма	Высокая	Есть	Нет

Алгоритм ID3

- Для примера возьмем признак «*наблюдение*»:
  - $p_1^+ = 2/5\,,\ p_1^- = 3/5\ (d_1 = \text{«солнце»})$
  - $p_2^+ = 4/4\,,\ p_2^- = 0/4\ (d_2 = \text{«облачно»})$
  - $p_3^+ = 3/5\,,\ p_3^- = 2/5\ (d_3 = \text{«дождь»})$

Алгоритм ID3

$$Info(p_1^+, p_1^-) = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = 0.971$$

$$Info(p_2^+, p_2^-) = -\frac{4}{4} \cdot \log_2 \frac{4}{4} - \frac{0}{4} \cdot \log_2 \frac{0}{4} = 0$$

$$Info(p_3^+, p_3^-) = -\frac{3}{5} \cdot \log_2 \frac{3}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5} = 0.971$$

Алгоритм ID3

- Чтобы найти $Info_b(f_i)$ – суммарное количество информации, необходимое для определения класса объекта после разбиения по признаку «*наблюдение*», вычислим взвешенную сумму:

$$Info_b(f_i) = \\ = \frac{5}{14} \cdot Info(p_1^+, p_1^-) + \frac{4}{14} \cdot Info(p_2^+, p_2^-) + \frac{5}{14} \cdot Info(p_3^+, p_3^-)$$

$$Info_b(f_i) = \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0.971 = 0.693$$

- Таким образом, после разбиения по признаку «*наблюдение*», останется неизвестным 0.693 бит информации

Алгоритм ID3

- Приращение информации:

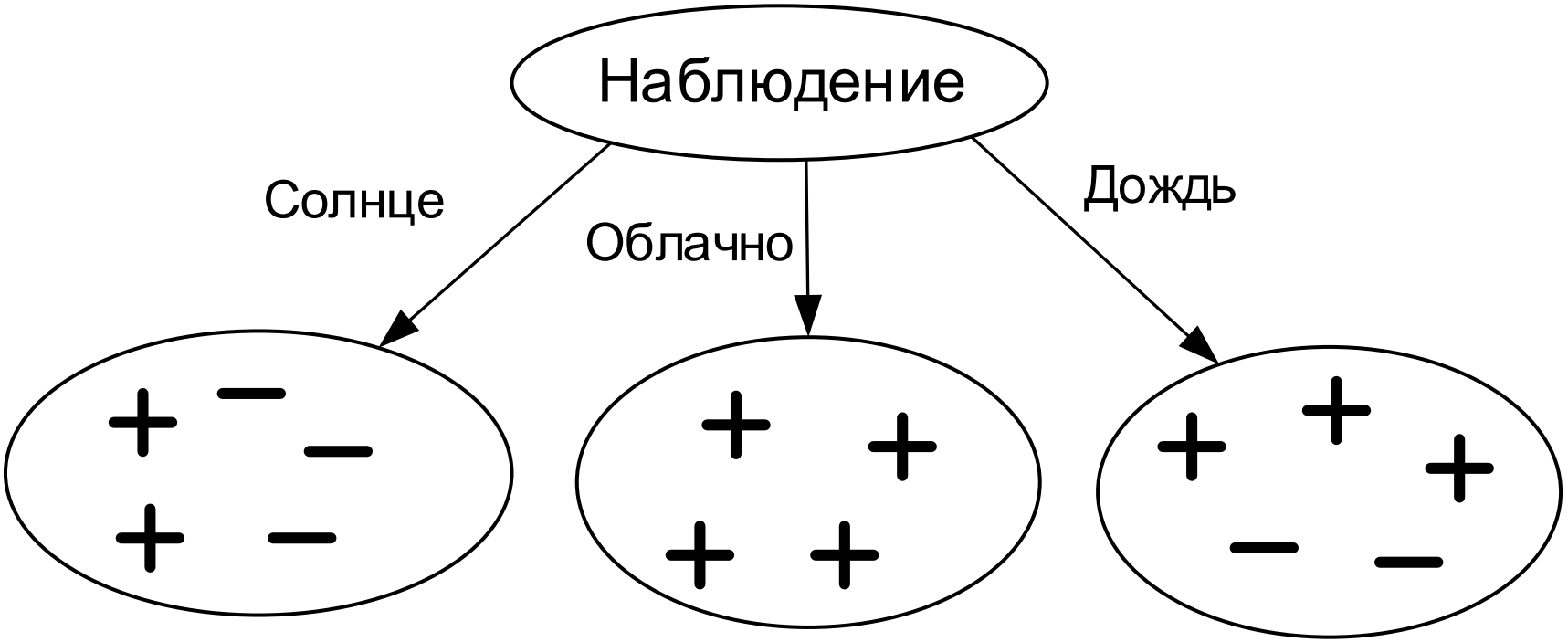
  Gain(«наблюдение») = 0.94 – 0.693 = 0.247 бит

  Gain(«температура») = 0.94 – 0.911 = 0.029 бит

  Gain(«влажность») = 0.94 – 0.788 = 0.152 бит

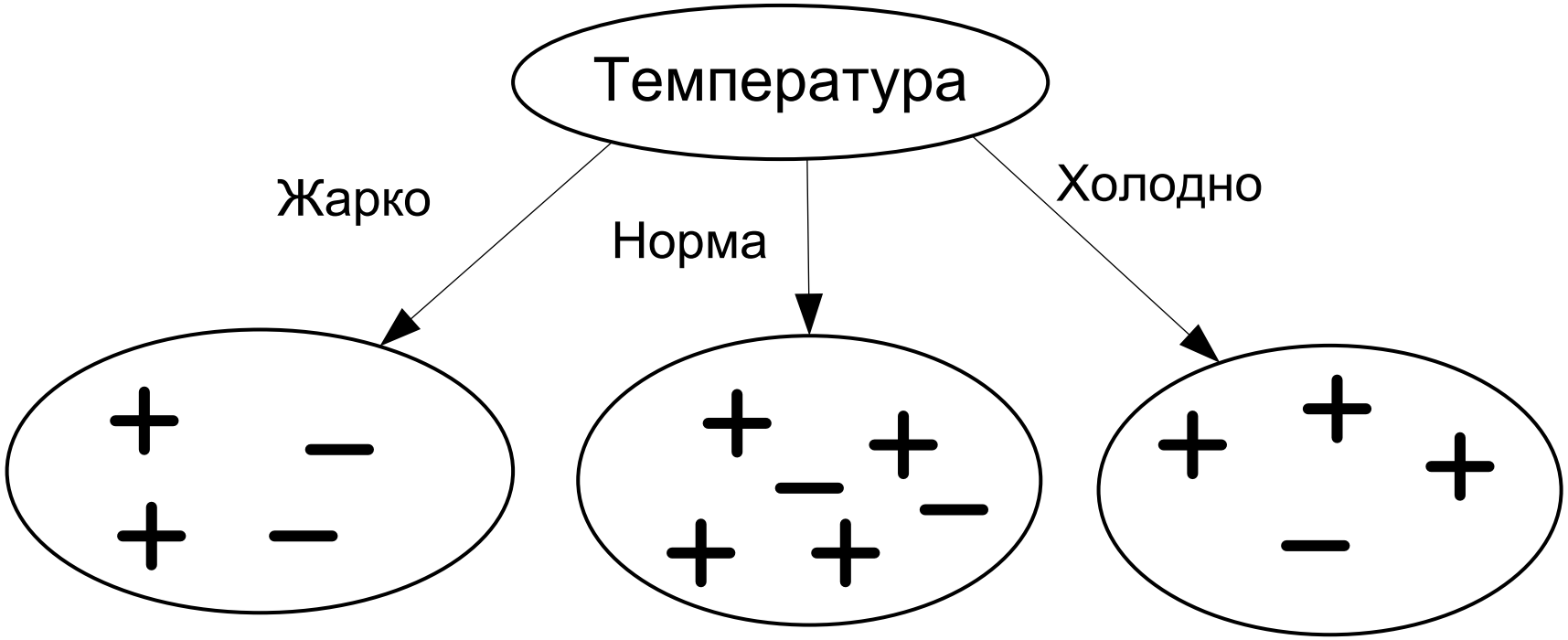
  Gain(«ветер») = 0.94 – 0.892 = 0.048 бит

Алгоритм ID3



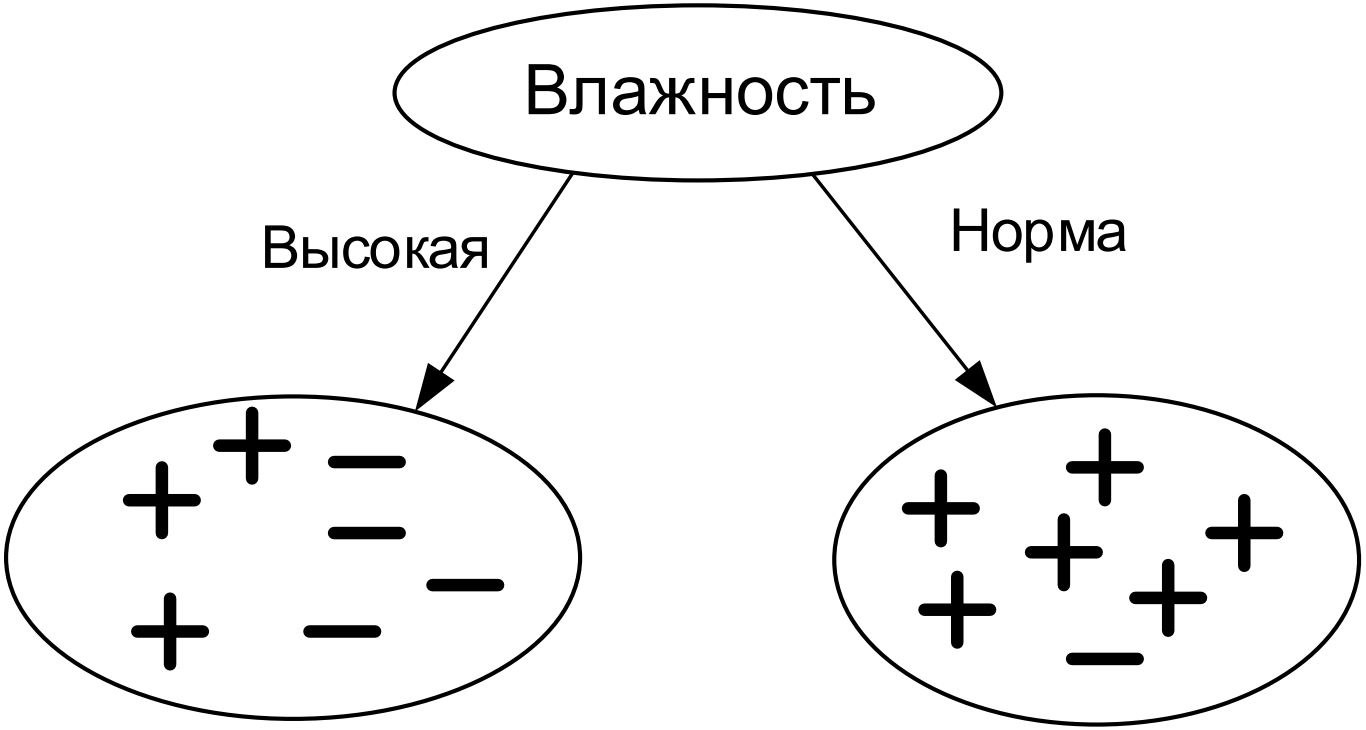
Gain(«наблюдение») = 0.247 бит

Алгоритм ID3



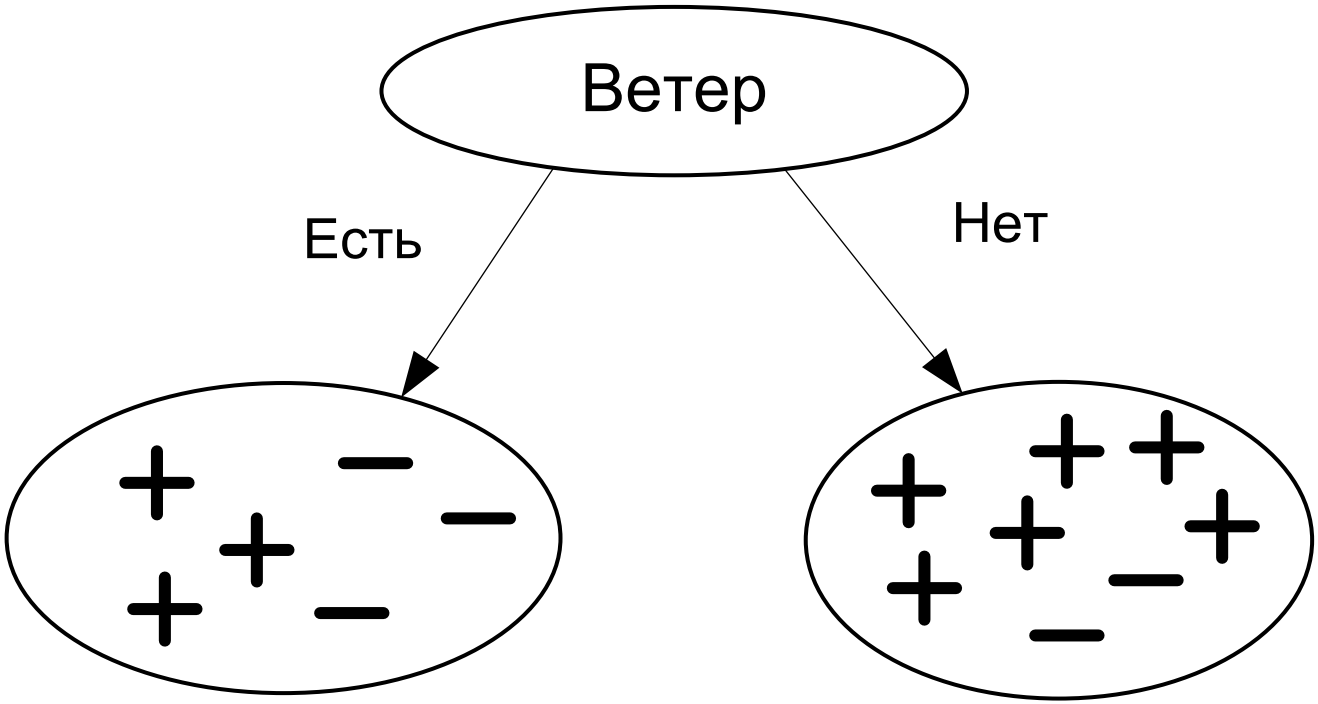
Gain(«температура») = 0.029 бит

Алгоритм ID3



Gain(«влажность») = 0.152 бит

Алгоритм ID3

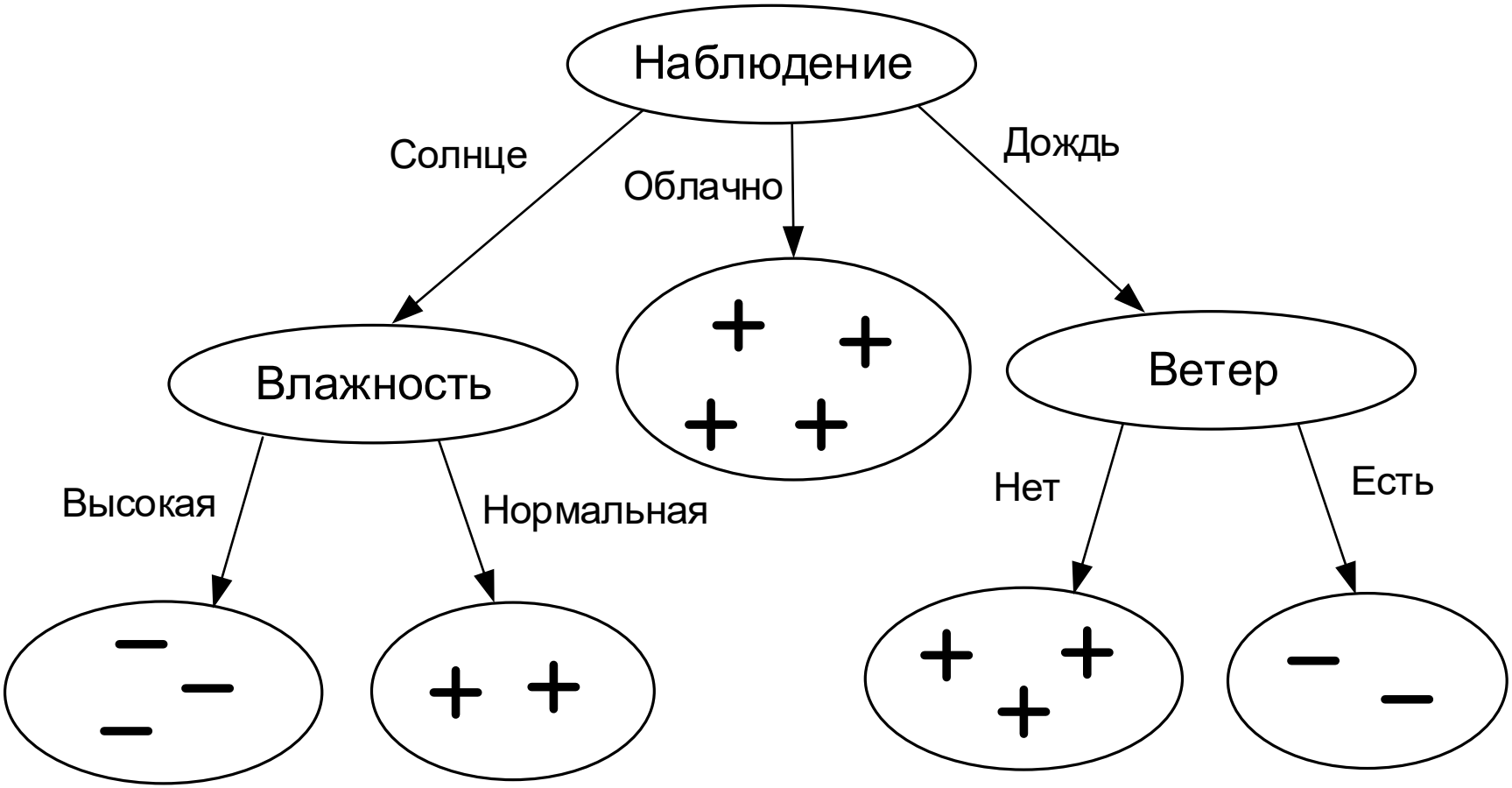


Gain(«ветер») = 0.048 бит

Алгоритм ID3

- Выбирается признак с максимальным значением приращения *Gain*
  - В примере выбирается признак «*Наблюдение*»
- Затем процесс рекурсивно повторяется для каждого из подмножеств T_k
- Если в процессе работы алгоритма получен узел с пустым множеством T_k, узел считается листом, а в качестве решения листа выбирается наиболее часто встречающийся класс у родителя данного листа

Алгоритм ID3



Индекс Джини

- Индекс Джини (Gini Index, Gini Impurity) показывает вероятность ошибочной классификации при случайном назначении меток классов:

$$I_G(k) = \sum_{i=1}^{C} \left(p_i \sum_{j \neq i} p_j \right) \sum_{i=1}^{C} p_i (1 - p_i) = \sum_{i=1}^{C} p_i - \sum_{i=1}^{C} p_i^2 = 1 - \sum_{i=1}^{C} p_i^2 ,$$

  где p_i – вероятности классов в вершине k

- Gini Index для некоторого разбиения с K вершинами:

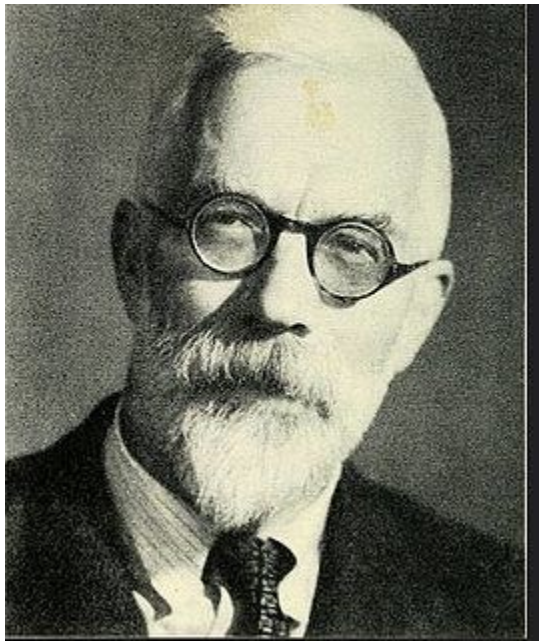
$$I(G) = \sum_{k=1}^{K} w_k I_G(k) ,$$

  где w_k – доля примеров в k-й вершине

- Выбирают разбиение с минимальным индексом Джини

Пример: ирисы Фишера

- Рональд Фишер, 1936 год
- Задача классификации
- 3 вида ирисов:
  - ирис виргинский (Iris virginica)
  - ирис щетинистый (Iris setosa)
  - ирис разноцветный (Iris versicolor)
- 150 примеров

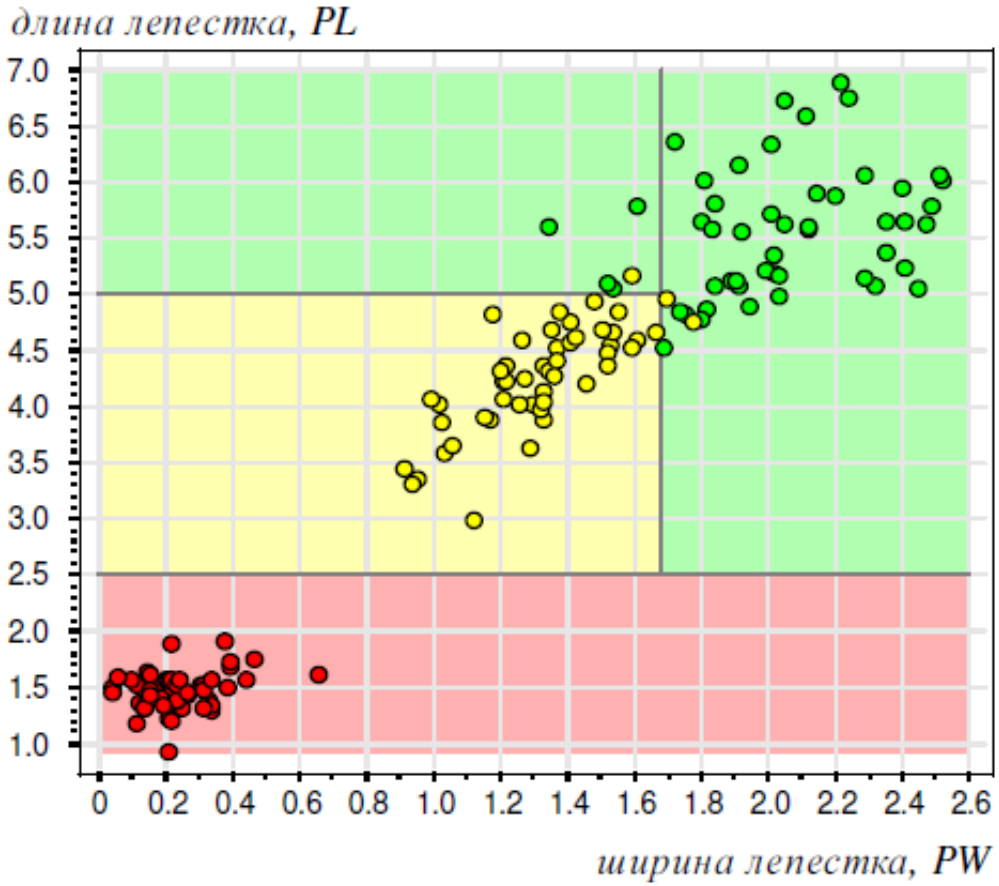


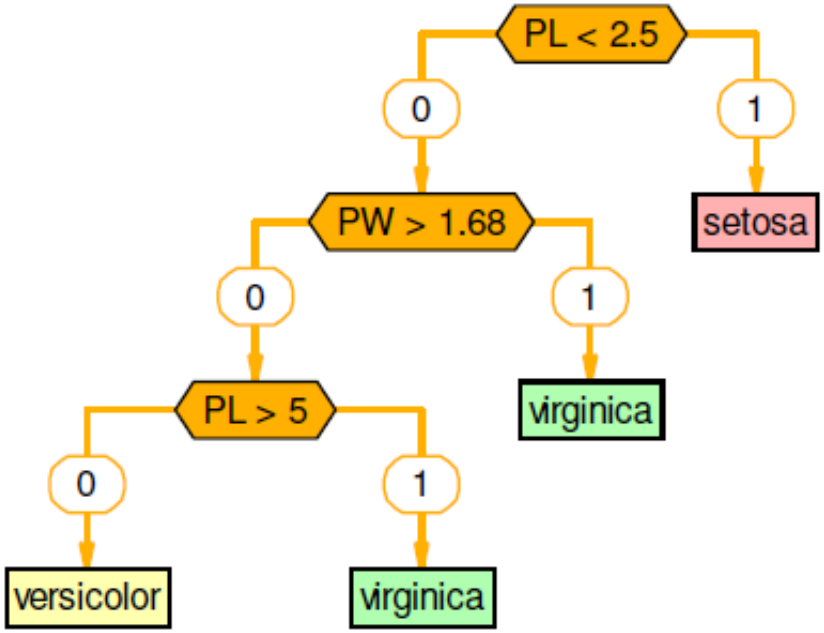
Пример: ирисы Фишера

Признаки:

- Длина чашелистика (Sepal Length, SL)
- Ширина чашелистика (Sepal Width, SW)
- Длина лепестка (Petal Length, PL)
- Ширина лепестка (Petal Width, PW)

Пример: ирисы Фишера





Правила остановки разбиения

- Разбивать дальше узел или отметить его как лист?
  - ограничение высоты дерева
  - нетривиальное разбиение – узлы должны содержать не менее заданного числа примеров
- Часто алгоритмы дают сложные деревья, имеющие много узлов и ветвей
- В идеале: дерево с малым количеством узлов, содержащих примерно одинаковое число примеров

Проблема сложности дерева

- Первое решение проблемы сложности: перебрать все возможные деревья и выбрать дерево с наименьшей глубиной
  - НО: доказано, что данная задача NP-полная
- Второе решение: *использование отсечения* (pruning)
  - Двигаясь снизу вверх, отсекаем или заменяем листами те ветви, которые не приводят к возрастанию общей ошибки

Достоинства ID3

- Позволяет восстанавливать любые нелинейные зависимости
- Интерпретируемость
- Низкая трудоемкость
- Простота алгоритма
- Не бывает отказов от классификации
  (в отличие от решающих списков)

Недостатки ID3

- Жадность – локально-оптимальный выбор признака для разбиения не всегда оказывается глобально-оптимальным
- Высокая чувствительность к выборке – изменение 1-2 примеров может привести к полному изменению дерева
- Склонность к переобучению (излишне сложное дерево с низкой обобщающей способностью)