Деревья принятия решения

Выполнила
Дейнега Л.А.
532 группа

Преподаватель
Малинина М.А.

Дерево (структура данных)

Дерево — одна из наиболее широко распространённых структур данных в информатике, эмулирующая древовидную структуру в виде набора связанных узлов. Является связанным графом, не содержащим циклы.

Определения

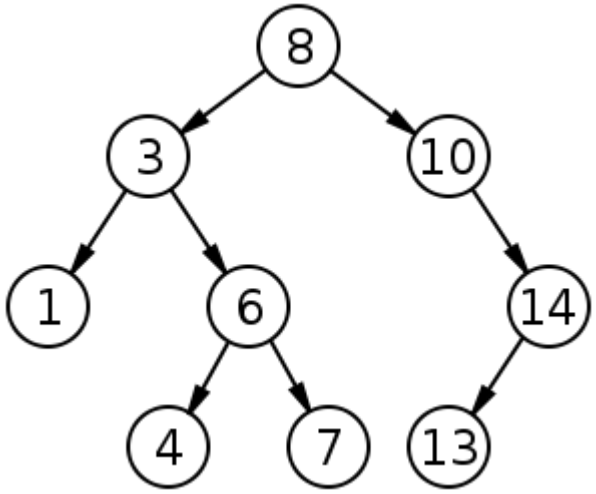


Рис.1. Пример дерева

- *Корневой узел* — самый верхний узел дерева (узел 8);
- *Лист* — узел, не имеющий дочерних элементов (узлы 1, 4, 7, 13);
- *Внутренний узел* — любой *узел* дерева, имеющий *потомков*(3, 6, 10, 14).

Дерево принятия решений

- **Дерево принятия решений** — это дерево, в листьях которого стоят значения целевой функции, а в остальных узлах — условия перехода, определяющие по какому из ребер идти.
- Если для данного наблюдения условие истинно(true), то осуществляется переход по левому ребру, если же ложно(false) — по правому.

Понятие энтропии

- **Энтропия** означает меру неупорядоченности системы; чем меньше элементы системы подчинены какому-либо порядку, тем выше энтропия.

- *Энтропия Шеннона*:

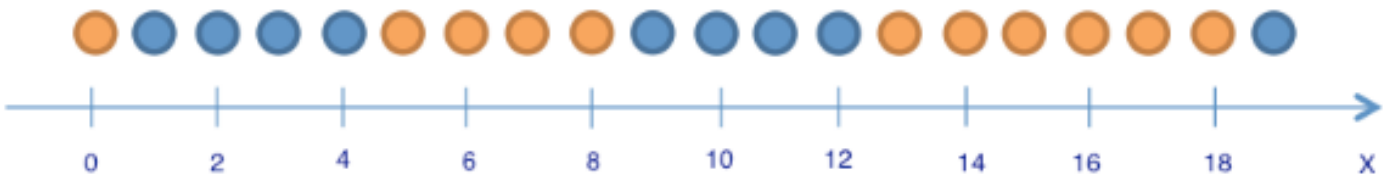
$$S = -\sum p_i * \ln(p_i)$$

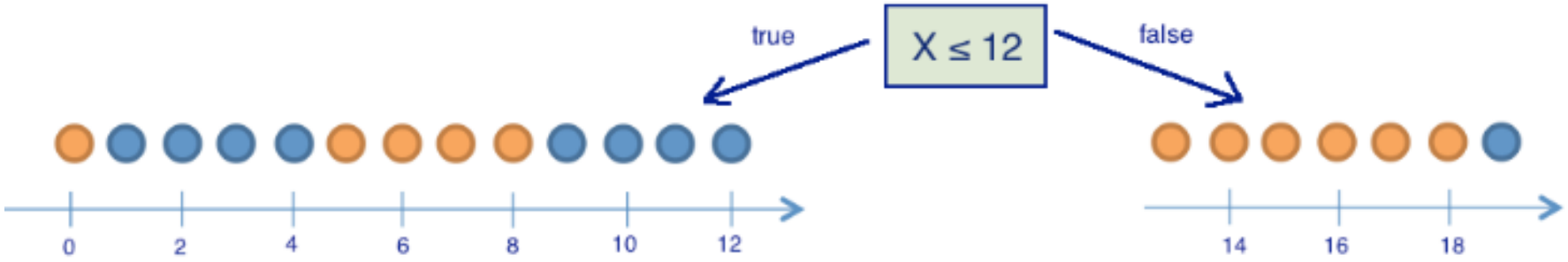
$$p_i = \frac{N_i}{N}$$

Пример

- Рассмотрим множество двухцветных шариков, в котором цвет шарика зависит только от координаты *x* *(при расчетах используем энтропию Шеннона)*:

$$S_0 = -\left(\frac{9}{20}\right)*\mathrm{Ln}\left(\frac{9}{20}\right) - \left(\frac{11}{20}\right)*\mathrm{Ln}\left(\frac{11}{20}\right) \approx 0{,}69$$

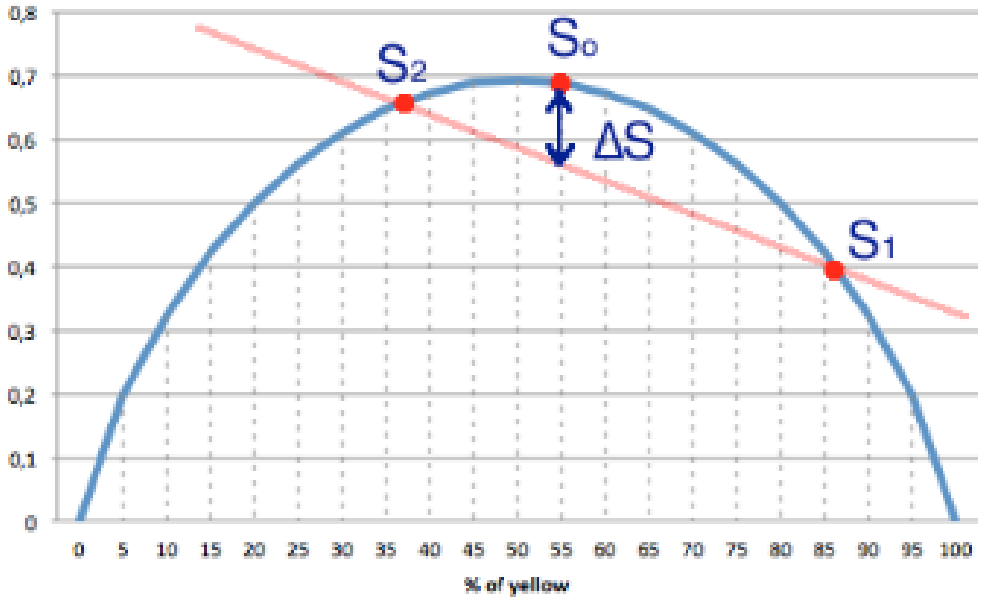




$$S_2 = -\left(\frac{8}{13}\right)*\mathrm{Ln}\left(\frac{8}{13}\right) - \left(\frac{5}{13}\right)*\mathrm{Ln}\left(\frac{5}{13}\right) \approx 0{,}66$$

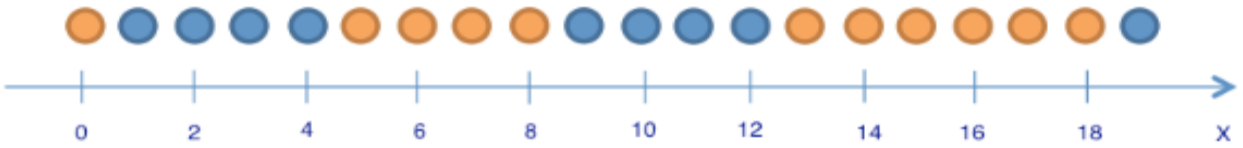
$$S_1 = -\left(\frac{6}{7}\right)*\mathrm{Ln}\left(\frac{6}{7}\right) - \left(\frac{1}{7}\right)*\mathrm{Ln}\left(\frac{1}{7}\right) \approx 0{,}4$$

Пример

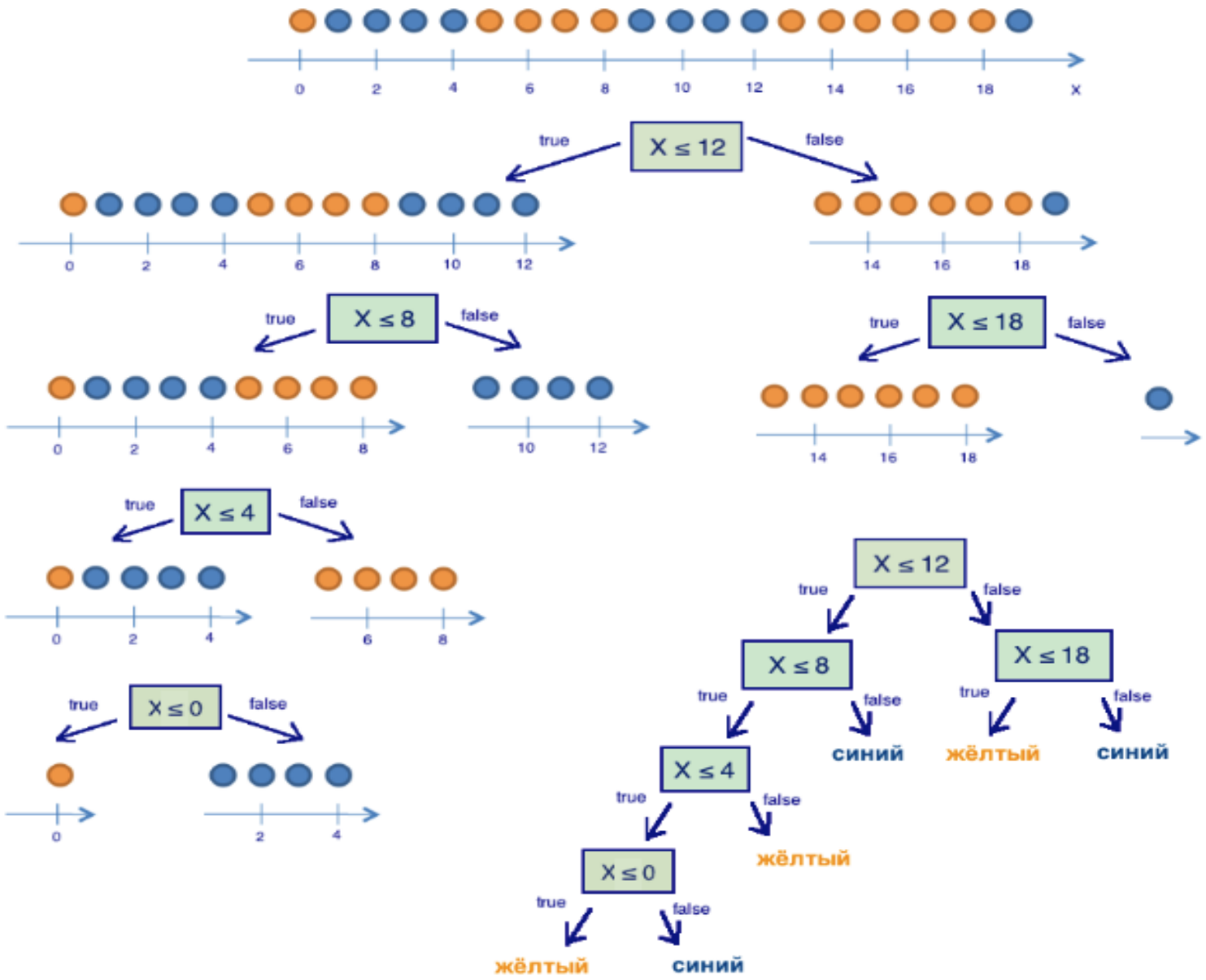


Вывод: если таким образом разделить множество на две части, то средняя энтропия будет меньше исходной на ΔS.

Это значит, что данный предикат обобщает некоторую информацию о данных.



Пример


0 2 4 6 8 10 12 14 16 18 X
X ≤ 12
true
false
X ≤ 8
X ≤ 18
X ≤ 4
X ≤ 0
синий
жёлтый
синий
жёлтый
жёлтый
синий

Алгоритм построения дерева принятия решений

s0 = вычисляем энтропию исходного множества

Если s0 == 0 значит:

Все объекты исходного набора, принадлежат к одному классу

Сохраняем этот класс в качестве листа дерева

Если s0 != 0 значит:

Перебираем все элементы исходного множества:

Для каждого элемента перебираем все его атрибуты:

На основе каждого атрибута генерируем предикат, который разбивает исходное множество на два подмножества

Рассчитываем среднее значение энтропии

Вычисляем ΔS

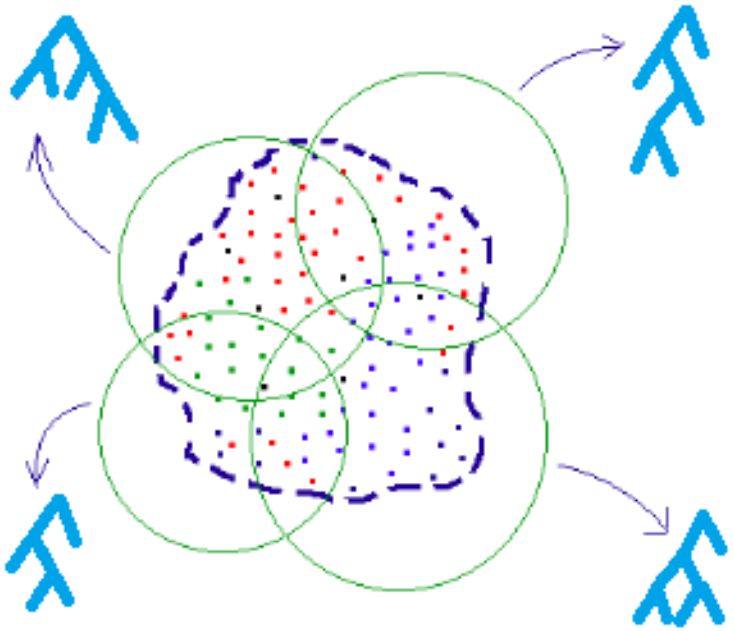
Нас интересует предикат, с наибольшим значением ΔS

Найденный предикат является частью дерева принятия решений, сохраняем его

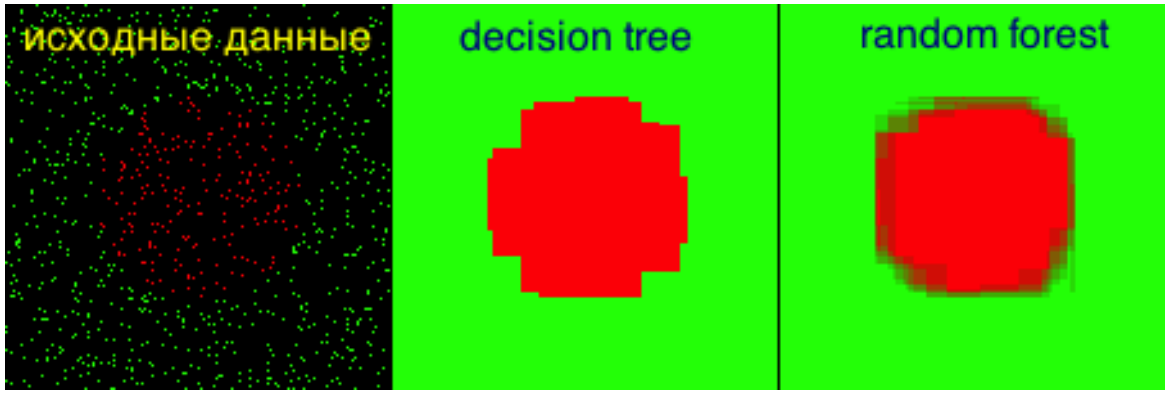
Разбиваем исходное множество на подмножества, согласно предикату

Повторяем данную процедуру рекурсивно для каждого подмножества

Random forest



- Выбираем случайные подмножества из обучающей выборки,
- *Для каждого подмножества строим своё дерево принятия решений.*
- Каждое дерево «голосует» за принадлежность объекта к определённому классу.
- Определяем с какой вероятностью объект принадлежит к какому-либо классу



Практика

- Постройте дерево принятия решения для набора данных, представленных в таблице

Соперник	Играем	Лидеры	Дождь	Победа
Выше	Дома	На месте	Да	Нет
Выше	Дома	На месте	Нет	Да
Выше	Дома	Пропускают	Нет	Да
Ниже	Дома	Пропускают	Нет	Да
Ниже	В гостях	Пропускают	Нет	Нет
Ниже	Дома	Пропускают	Да	Да
Выше	В гостях	На месте	Да	Нет
Ниже	В гостях	На месте	Нет	Да

Спасибо за внимание!