НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

**Эссе на тему**

«Деревья принятия решений –

критерий информационной выгоды»

Выполнили:

студенты гр. 17-В-2

Пегов С.С.

Пудовинников П.М.

Принял:

Авербух М.Л.

Нижний Новгород

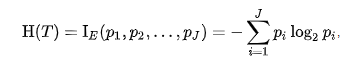
2020 г.

Деревья решений являются одним из наиболее эффективных инструментов интеллектуального анализа данных и предсказательной аналитики, которые позволяют решать задачи [классификации](https://wiki.loginom.ru/articles/classification.html) и [регрессии](https://wiki.loginom.ru/articles/regression.html).

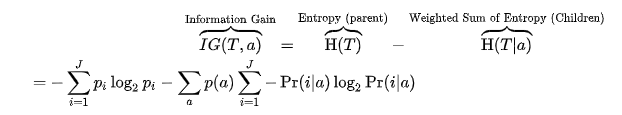
Они представляют собой иерархические древовидные структуры, состоящие из решающих правил вида «Если ..., то ...».

В алгоритмах генерации деревьев ID3 (**Алгоритм ID3** — один из алгоритмов для построения [дерева принятия решений](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9). Разработан Джоном Р. Квинланом ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *John R. Quinlan*). Впоследствии Квинлан создал усовершенствованную версию — [алгоритм C4.5](https://ru.wikipedia.org/wiki/C4.5" \o "C4.5).), используется информационная выгода, которая основывается на понятии энтропии и объема информации из теории информации.

Энтропия определяется следующим образом:



где p1,p2.. являются долями, в сумме дающими 1, которые представляют процент каждого класса, полученный от разбиения в дереве.



В формуле:

* Information Gain – Информационная выгода
* Entropy (parent) – Энтропия (родитель)
* Weighted Sum of Entropy (Children) – Взвешенная сумма энтропии (наследники)

Информационная выгода используется для принятия решения, какой признак использовать для расщепления на каждом шаге построения дерева. Простота является лучшим выбором, так что мы хотим сохранять дерево небольшим. Чтобы это сделать, на каждом шаге нам следует выбрать расщепление, которое приводит к простейшим узлам-наследникам. Обычно используемая мера простоты называется информацией, которая измеряется в [битах](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D1%82). Для каждого узла дерева значение информации «представляет ожидаемое количество, которая нужна для определения, должен ли новый объект классифицирован как да или нет, если дано, что пример достигает этого узла»".

Рассмотрим пример набора данных с четырьмя атрибутами: погода (солнечно, облачно, дождь), температура (жарко, мягкая погода, холодно), влажность (высокая, нормальная) и ветер (есть, нет) с бинарной целевой переменной (да или нет) игра с 14 точками данных. Для построения дерева решений по этим данным нам нужно сравнить информационную выгоду каждого из четырёх деревьев, на которые расщепляется согласно одному из четырёх признаков. Расщепление с максимальной информационной выгодой будет взято в качестве первого расщепления и процесс продолжается, пока все наследники не станут простыми или пока информационная выгода не станет нулём.

Расщепление, использующее признак ветер, приводит к двум узлам-наследникам, один узел для признака ветер со значением **есть** и один со значением **нет**. В этом наборе данных имеется шесть точек данных со значением **есть** для ветер, три имеют для целевого значения игра значение **да** и три имеют значение **нет**. Восемь оставшихся точек данных для параметра ветер со значением **нет** содержат два **нет** и шесть **да**. Информация ветер=да узел вычисляется с помощью уравнения для энтропии выше. Поскольку имеется равное число **да** и **нет** в этом узле, мы имеем



Для узла с ветер=нет имелось восемь точек данных, шесть с целевым значением **да** и два с **нет**. Таким образом, мы имеем



Чтобы найти информацию разбиения, вычислим взвешенное среднее этих двух чисел на основе количества наблюдений, попавших в каждый узел.



Чтобы найти информационную выгоду расщепления с помощью ветер, мы должны вычислить информацию в данных до расщепления. Исходные данные содержали девять **да** и пять **нет**.



Теперь мы можем вычислить информационную выгоду, получаемый расщеплением по признаку *ветер*.



Чтобы построить дерево, нужно вычислить информационную выгоду каждого возможного первого расщепления. Лучшее первое расщепление, это то, которое даёт наибольший информационную выгоду. Этот процесс повторяется для каждого узла (со смешанными признаками), пока дерево не будет построено.