Деревья принятия решений

Основные элементы

Решающее дерево – ациклический граф, основными элементами которого являются:

- вершины: предикаты – функции, принимающие значения $\{0,1\}$;
- листья (терминальные вершины): содержат значения целевой переменной;
- ребра: значения предикатов, из которых выходит ребро.

Ответ алгоритма: соответствующее предикатам объекта значение целевой переменной в терминальной вершине.

Пример

Стоит задача предсказать, выиграет ли «Зенит» свой следующий матч.

Список бинарных признаков:

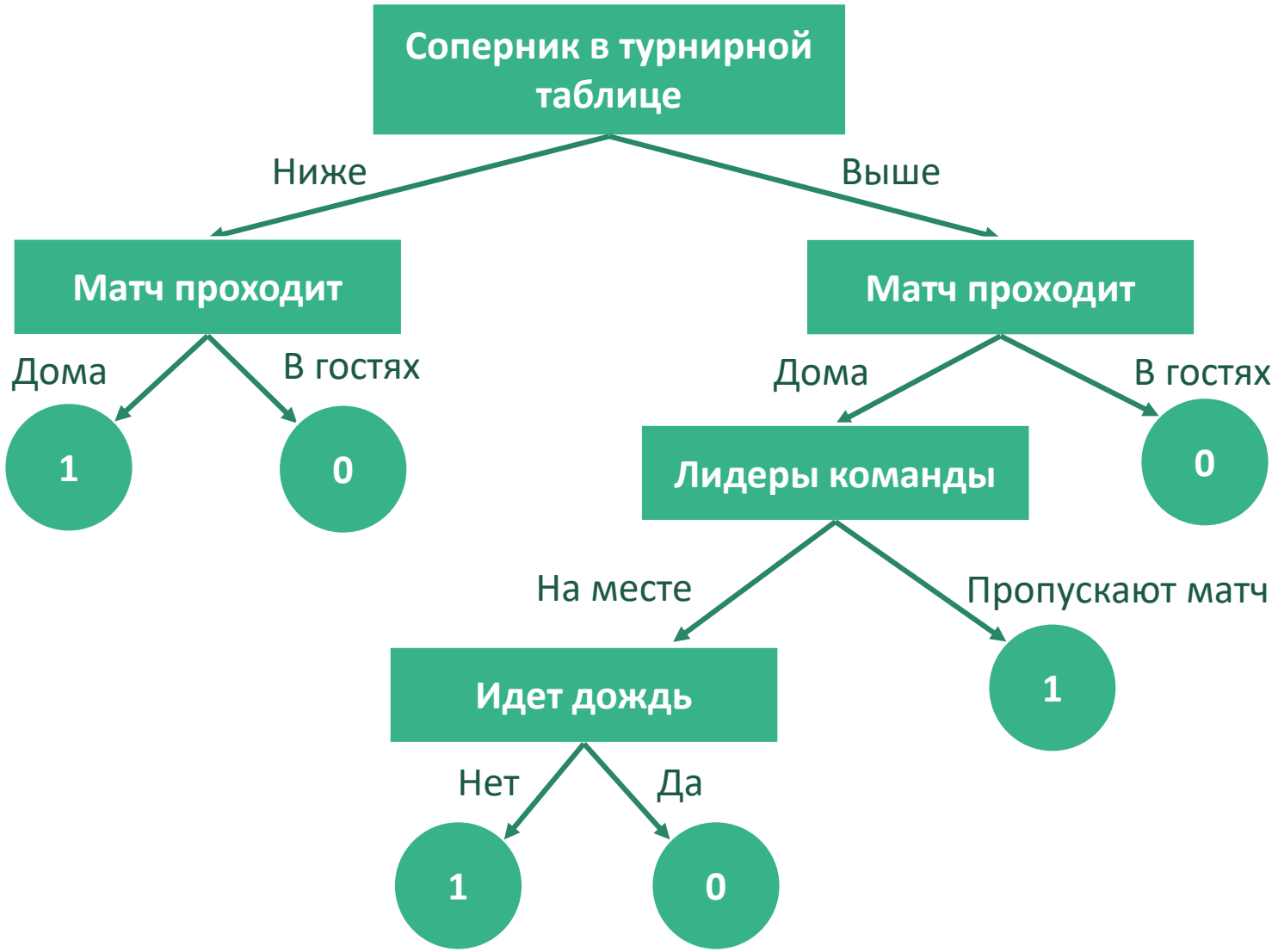
- находится ли соперник выше в турнирной таблице;
- проходит ли матч на домашнем стадионе;
- пропускают ли матч лидеры команды-соперника;
- обещают ли в этот день дождь.

Зная историю проведенных «Зенитом» игр, можно попробовать предсказать результат предстоящего матча.

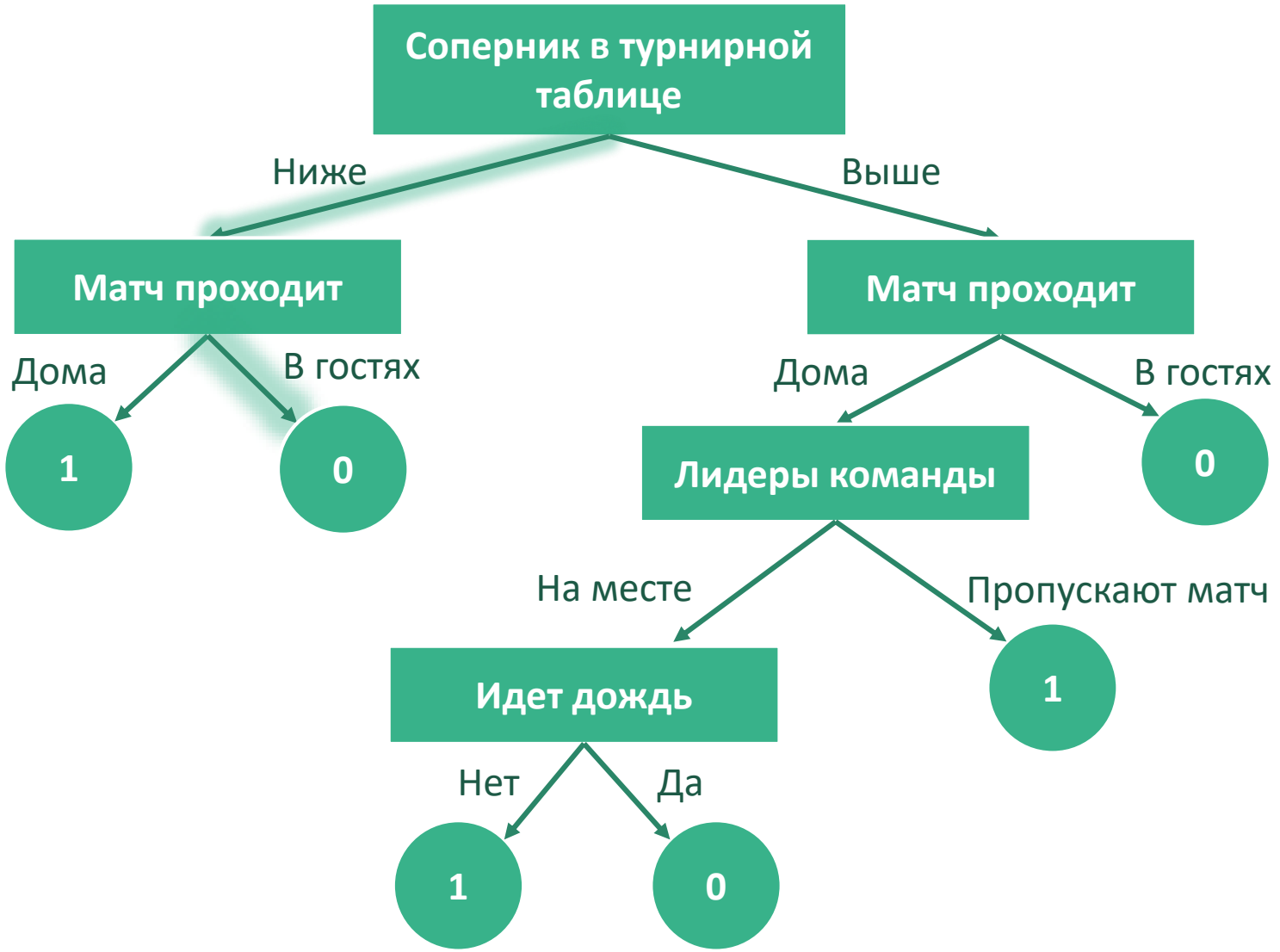
История матчей «Зенита»

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	?

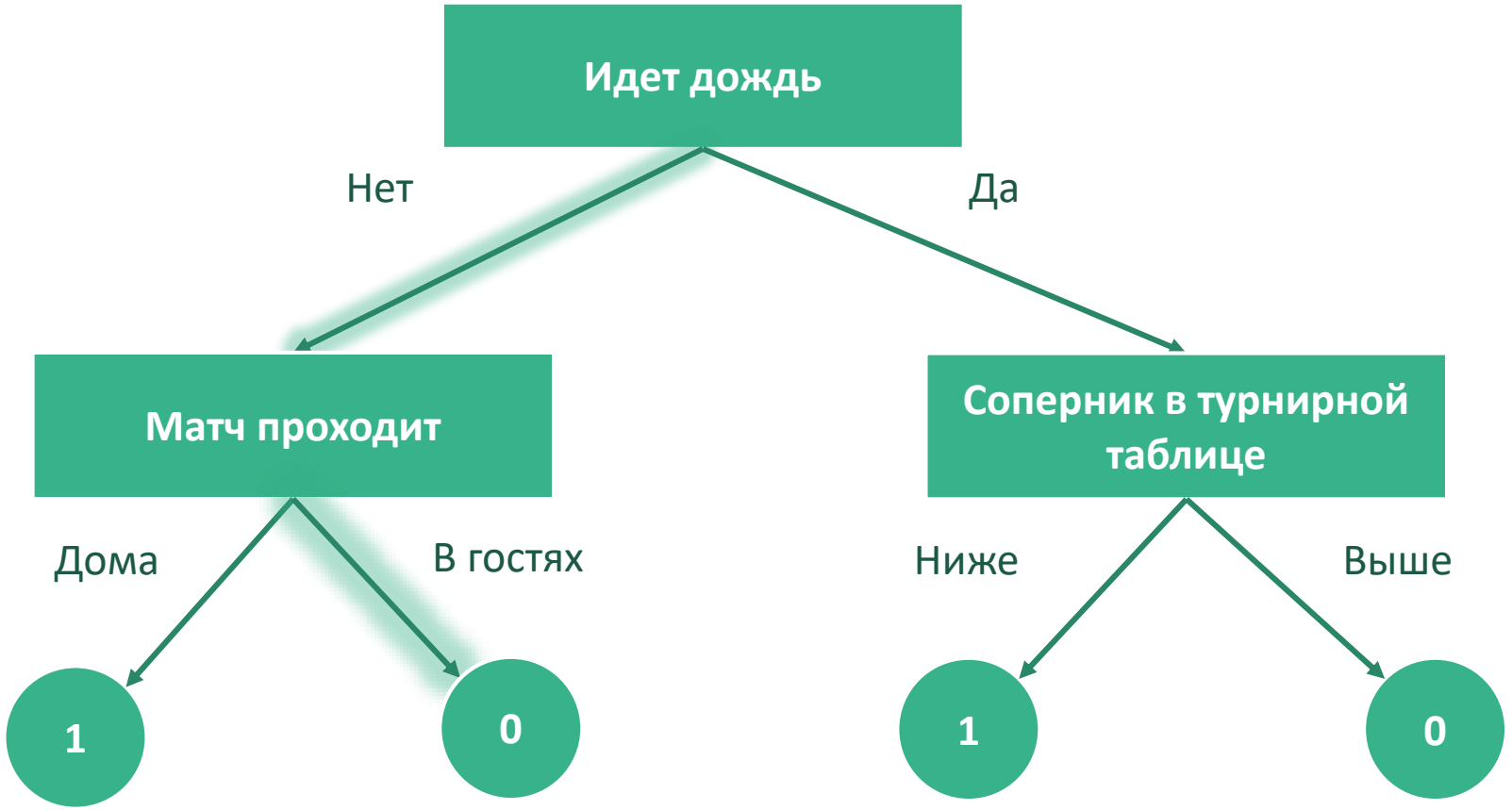
Пример решающего дерева



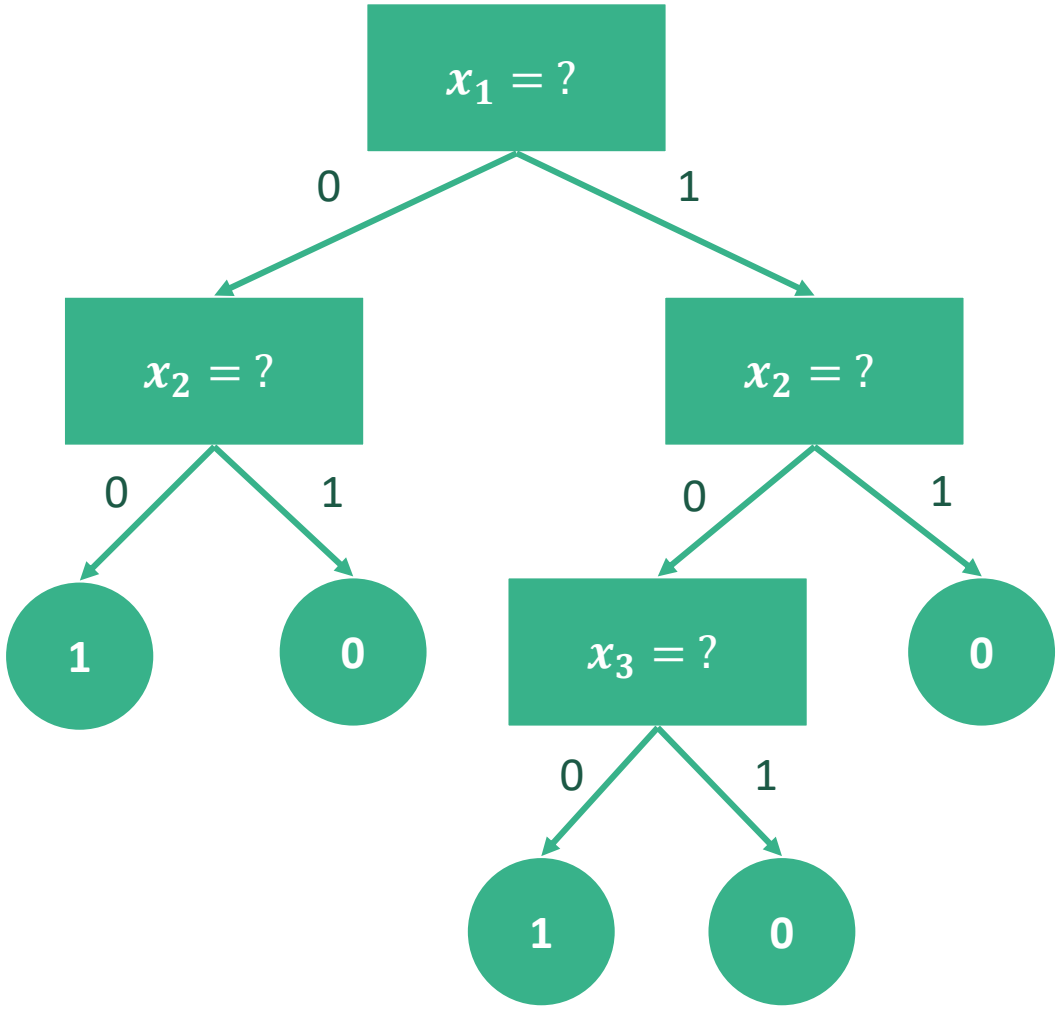
Пример решающего дерева



Оптимальное дерево



Деревья и булевы функции

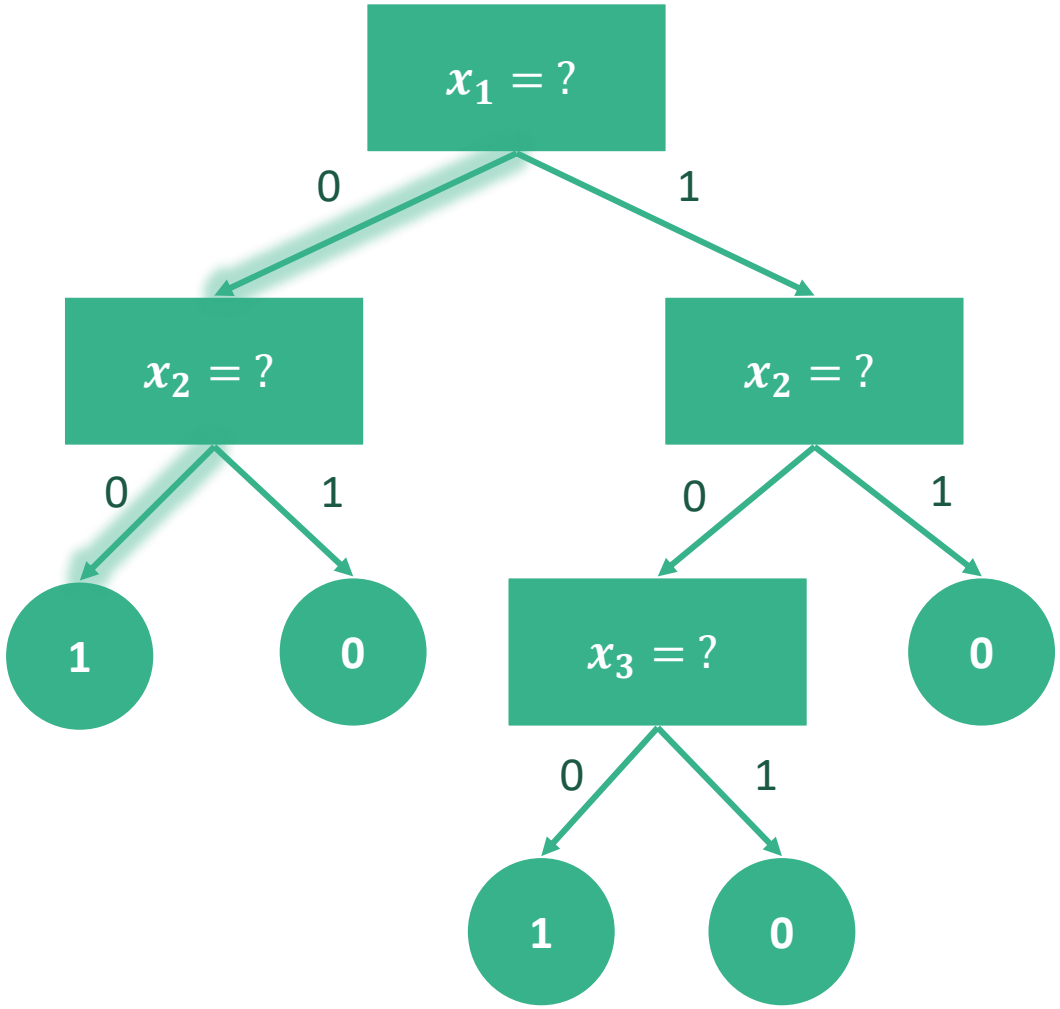


Дерево решений может быть представлено в виде булевой функции в ДНФ.

Например, дерево на рисунке соответствует функции:

$$f(x_1, x_2, x_3) = (\overline{x_1} \wedge \overline{x_2}) \vee (x_1 \wedge \overline{x_2} \wedge \overline{x_3})$$

Деревья и булевы функции

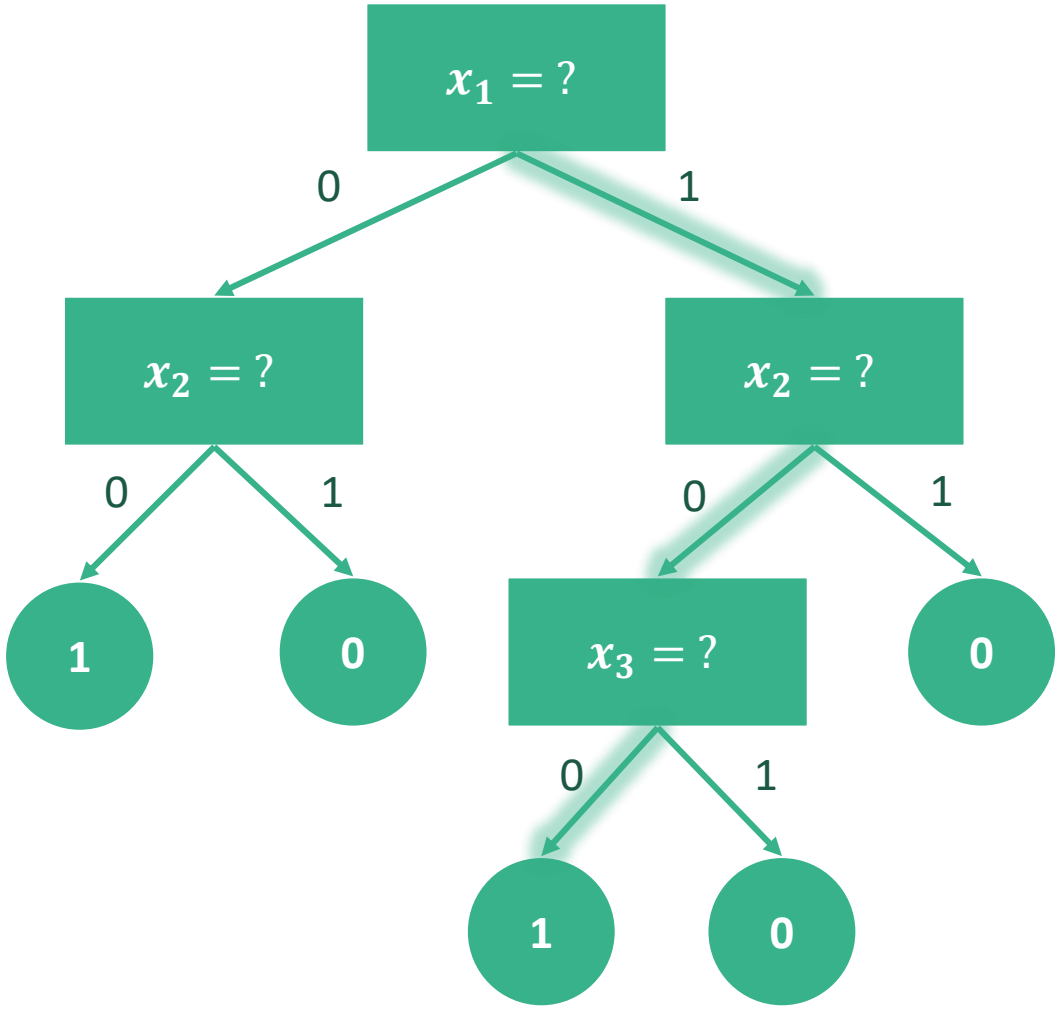


Дерево решений может быть представлено в виде булевой функции в ДНФ.

Например, дерево на рисунке соответствует функции:

$$f(x_1, x_2, x_3) = (\overline{x_1} \wedge \overline{x_2}) \vee (x_1 \wedge \overline{x_2} \wedge \overline{x_3})$$

Деревья и булевы функции



Дерево решений может быть представлено в виде булевой функции в ДНФ.

Например, дерево на рисунке соответствует функции:

$$f(x_1, x_2, x_3) = (\overline{x_1} \wedge \overline{x_2}) \vee (x_1 \wedge \overline{x_2} \wedge \overline{x_3})$$

Алгоритм ID3

Вход: LearnID3 ($U \subseteq X^l$)

если все объекты из U лежат в одном классе $c \in Y$ то

    вернуть новый лист $v, c_v := c$;

найти предикат с максимальной информативностью (*Information Gain*):

    $\beta := \arg\max_{\beta \in \mathcal{B}} IG(\beta, U)$;

разбить выборку на две части $U = U_0 \sqcup U_1$ по предикату β:

    $U_0 := \{x \in U: \beta(x) = 0\}$;

    $U_1 := \{x \in U: \beta(x) = 1\}$;

если $U_0 = \emptyset$ или $U_1 = \emptyset$ то

    вернуть новый лист $v, c_v :=$ Мажоритарный класс (U);

создать новую внутреннюю вершину v: $\beta_v := \beta$;

построить левое поддерево $L_v :=$ LearnID3 (U_0);

построить правое поддерево $R_v :=$ LearnID3 (U_1);

вернуть v

Критерии информативности

Энтропия Шеннона

При N возможных классах:

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i \,,$$

где p_i – вероятность принадлежности к классу i. В случае бинарной классификации:

$$S = -p \log_2 p - (1-p) \log_2 (1-p).$$

Уменьшение энтропии называют приростом информации (*Information gain*):

$$IG(\beta) = S_0 - \sum_i \frac{N_i}{N} S_i \,.$$

История матчей «Зенита»

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	?

Энтропия

Из 7 матчей «Зенит» три проиграл и четыре выиграл. Исходная энтропия:

$$S_0 = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} \approx 0.9852$$

Вычислим прирост информации по признаку «Соперник в турнирной таблице»:

$$IG(\text{Соперник}) = S_0 - \frac{4}{7}S_1 - \frac{3}{7}S_2$$

$$\approx 0.9852 - \frac{4}{7}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{3}{7}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) \approx \mathbf{0.0202}$$

Энтропия

Для определения оптимального предиката вычислим значения энтропийного критерия для всех возможных предикатов:

$$IG(\text{Соперник}) \approx 0.0202$$

$$IG(\text{Играем}) \approx 0.4696$$

$$IG(\text{Лидеры}) \approx 0.1281$$

$$IG(\text{Дождь}) \approx 0.1281$$

Согласно энтропийному критерию в корень дерева необходимо поместить предикат «домашний матч или гостевой».

Критерии информативности

Критерий Джини

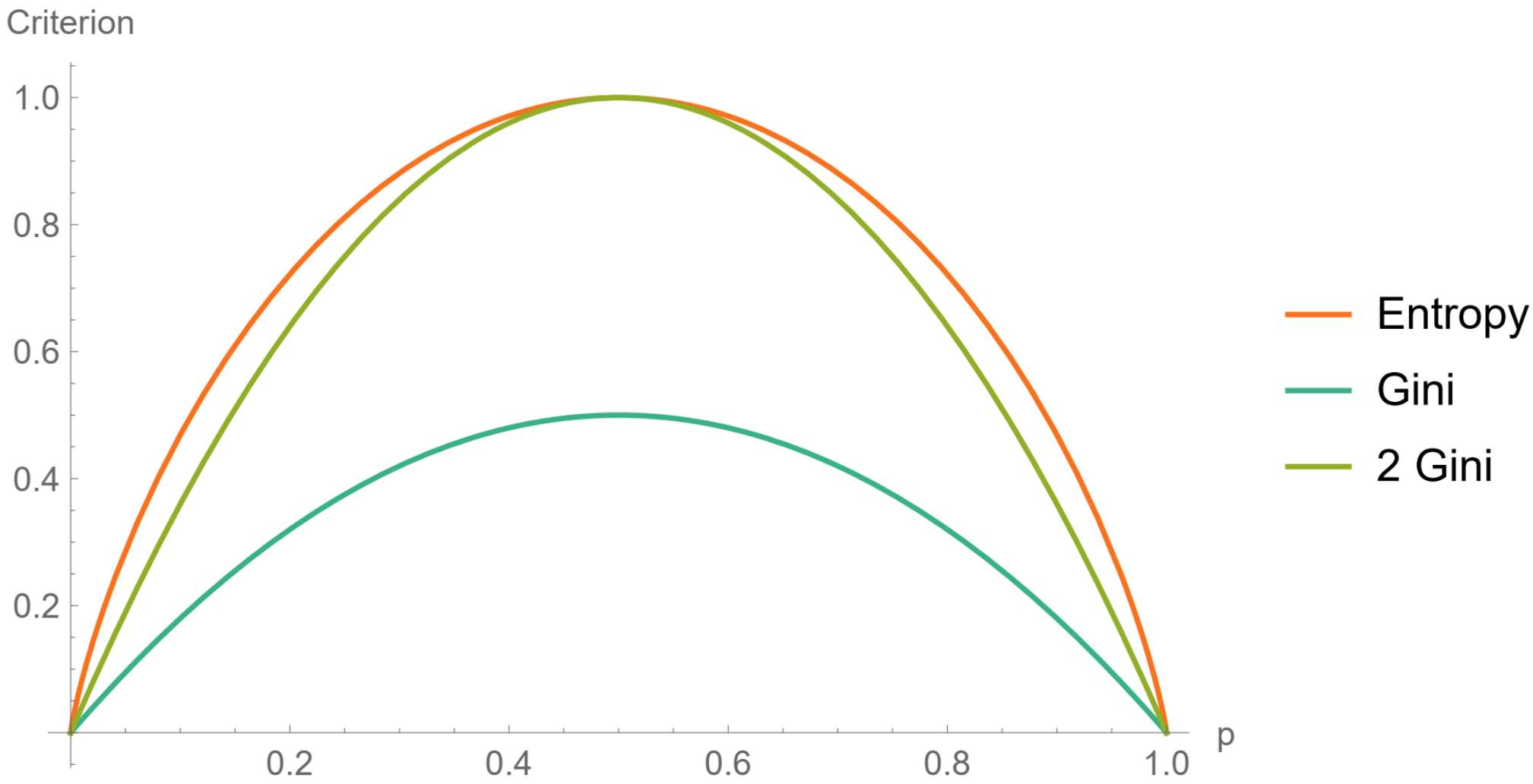
Максимизирует число пар объектов одного класса, оказавшихся в одном поддереве:

$$G = 1 - \sum_{i=1}^{N} p_i^2 ,$$

В случае бинарной классификации:

$$G = 1 - p^2 - (1 - p)^2 = 2p(1 - p).$$

Сравнение критериев информативности



Критерии в задачах регрессии

Дисперсионный критерий

$$D(U) = \frac{1}{|U|}\sum_{x_j \in X}\left(y_j - \frac{1}{|U|}\sum_{x_i \in X} y_i\right)^2,$$

где $|U|$ – число объектов в листе, y_i – значение целевой переменной.

$$IG(\beta, U) = D(U) - \frac{|U_0|}{|U|}D(U_0) - \frac{|U_1|}{|U|}D(U_1).$$

Необходимо выбрать такой критерий, при котором дисперсия как в левом, так и в правом поддеревьях значительно уменьшится.