Математические и инструментальные методы машинного обучения

8. Классификация

Понятие классификации

Классификация — системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства. Под классификацией будем понимать отнесение объектов (наблюдений, событий) к одному из заранее известных классов. Формально: $I=\{i1,...,in\}$, $ii = \{x1..xn, y\}$ (xi — атрибуты-независимые переменные, y — зависимая).

Понятие классификации

Классификация требует соблюдения следующих правил:

- в каждом акте деления необходимо применять только одно основание;
- деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
- члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
- деление должно быть последовательным.

Виды классификации

Различают:

- **вспомогательную** (искусственную) классификацию, которая производится по внешнему признаку и служит для упорядочивания множества предметов (процессов, явлений);
- **естественную** классификацию, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатом и важным средством научного исследования, так как предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

Виды классификации

В зависимости от выбранных признаков, их сочетания и процедуры деления понятий классификация может быть:

- **простой** — деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой классификации является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: "A и не A");
- **сложной** — применяется для деления одного понятия по разным основаниям и синтеза этих простых делений в единое целое. Примером такой классификации является периодическая система химических элементов.

Процесс классификации

Процесс классификации состоит из двух этапов: конструирования модели и ее использования.

- **Конструирование модели:**
  - описание множества предопределенных классов.
    - Каждый пример набора данных относится к одному предопределенному классу.
    - На этом этапе используется обучающее множество, на нем происходит конструирование модели.
    - Полученнаямодель может быть представлена классификационными правилами, деревом решений или математической формулой.

Процесс классификации

- **Использование модели:**
  - **классификация новых или неизвестных значений.**
  - **Оценка правильности (точности) модели.**
    - **Известные значения из тестового примера сравниваются с результатами использования полученной модели.**

Методы классификации

- классификация методами дискриминантного анализа;
- классификация с помощью деревьев решений;
- байесовская (наивная) классификация;
- классификация при помощи искусственных нейронных сетей;
- Классификация методом опорных векторов;
- Статистические методы, в частности, линейная регрессия;
- Классификация при помощи метода ближайшего соседа;
- Классификация *CBR*-методом;
- Классификация при помощи генетических алгоритмов.

1R-алгоритм

Пусть у нас есть независимые переменные $A^1 \ldots A^j \ldots A^k$, принимающие значения $< x_1^1 \ldots x_n^1 >, \ldots < x_1^j \ldots x_n^j >, \ldots < x_1^k \ldots x_n^k >$ соответственно, и зависимая переменная C, принимающая значения $c_1 \ldots c_r$. Для любого возможного значения каждой независимой переменной формируется правило, которое классифицирует объект из обучающей выборки. В если-части правила указывают значение независимой переменной (Если $A^j = x_i^j$). В то-части правила указывается наиболее часто встречающееся значение зависимой переменной у данного значения независимой переменной(то $C = c_r$). Ошибкой правила является количество объектов, имеющих данное значение рассматриваемой независимой переменной ($A^j = x_i^j$), но не имеющих наиболее часто встречающееся значение зависимой переменной у данного значения независимой переменной($C \neq c_r$). Оценив ошибки, выбирается переменная, для которой ошибка набора минимальна.

В случае непрерывных значений манипулируют промежутками. В случае пропущенных значений - достраивают. Наиболее серьезный недостаток - сверхчувствительность.

Условия для игры

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

Условия для игры. Набор правил

RuleModel

- if **Наблюдение = Облачность** *then* **Да (0 / 4)**
- if **Влажность = Норма** *and* **Ветер = Нет** *then* **Да (0 / 3)**
- if **Температура = Жарко** *then* **Нет (2 / 0)**
- if **Температура = Холодно** *then* **Нет (1 / 0)**
- if **Влажность = Норма** *then* **Да (0 / 1)**
- if **Наблюдение = Солнце** *then* **Нет (1 / 0)**
- if **Ветер = Нет** *then* **Да (0 / 1)** *else* **Нет (0 / 0)**

Метод Naive Bayes

"Наивная" классификация — достаточно прозрачный и понятный метод классификации. "Наивной" она называется потому, что исходит из предположения о взаимной независимости признаков.

Свойства наивной классификации:

- Использование всех переменных и определение всех зависимостей между ними.
- Наличие двух предположений относительно переменных:
- Все переменные являются одинаково важными;
- Все переменные являются статистически независимыми, т.е. значение одной переменной ничего не говорит о значении другой.

Метод Naive Bayes. Продолжение

Вероятность того, что некий объект i_i, относится к классу $c_r(y = c_r)$ обозначим как $P(y = c_r)$. Событие, соответствующее равенству независимых переменных определенному значению, обозначим как E, а его вероятность - P(E). Идея алгоритма в расчете условной вероятности принадлежности объекта к cr при равенстве его независимых переменных определенным значениям. Из тервера:

$$P(y = c_r|E) = \frac{P(E|y = c_r) * P(y = c_r)}{P(E)}$$

Таким образом формулируются правила, в условных частях которых сравниваются все независимые переменные с соответсввующими возможными значениями. В заключительной части - все возможные значения зависимой переменной: $x_1 = c_1^k, ..., x_n = c_n^k, y = c_r$....{и так для все наборов} Для каждого из этих правил по формуле Байеса определяется его вероятность. Так как независимые переменные независимы друг от друга, то :

$P(E|y = c_r) = P(x_1 = c_1^k|y = c_r) * ... * P(x_n = c_n^k|y = c_r)$, что подставляем в верхную формулу и получаем вероятность всего правила.

Метод Naive Bayes. Окончание

Вероятность принадлежности объекта к классу c_r при равенстве его переменной x_n определенному значению c_n^k :

$$P(x_n = c_n^k | y = c_r) = \frac{P(x_n = c_n^k \ \& \ y = c_r)}{P(y = c_r)}$$

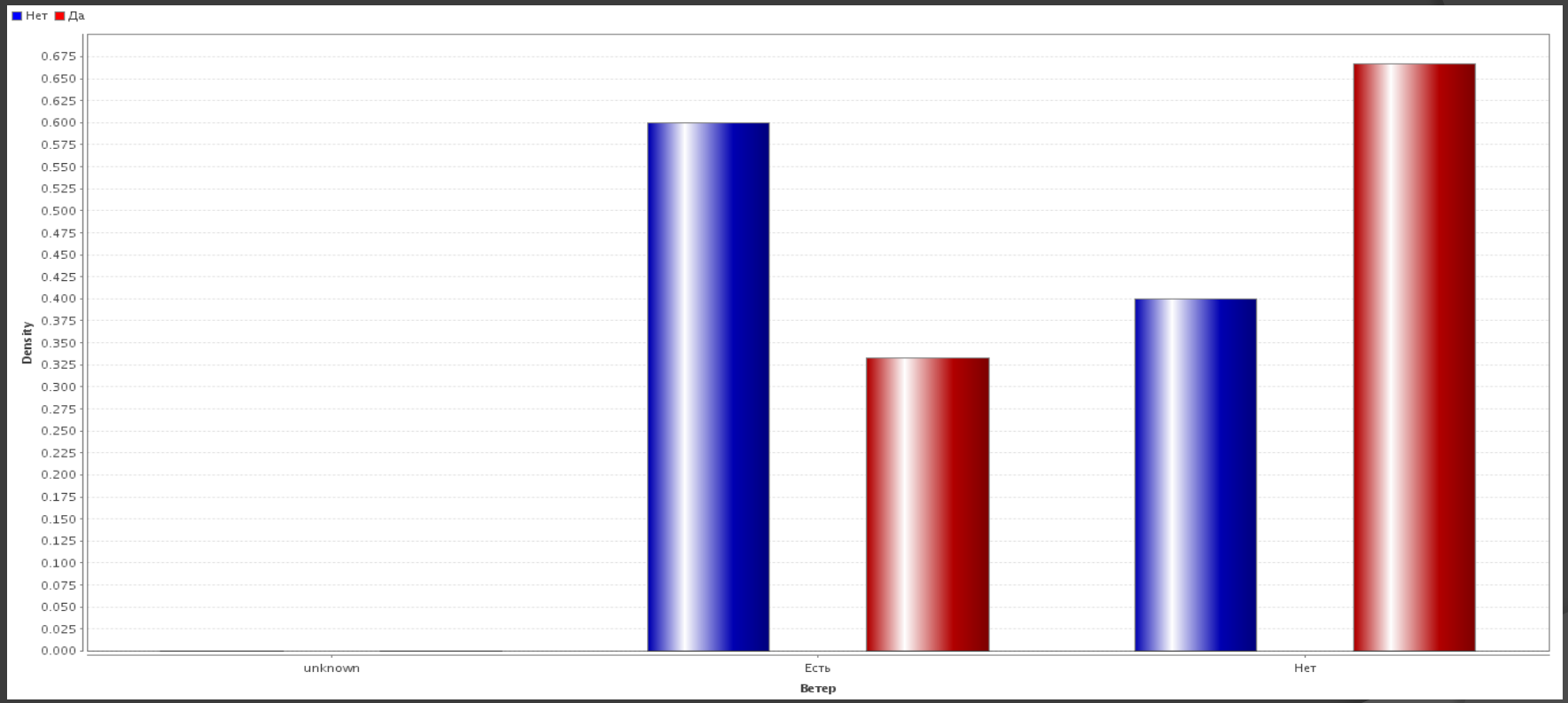
Нормализованная вероятность вычисляется по формуле:

$$P'(y = cr|E) = \frac{P(y = c_r|E)}{\sum_{c_r} P(y = c_r|E)}$$

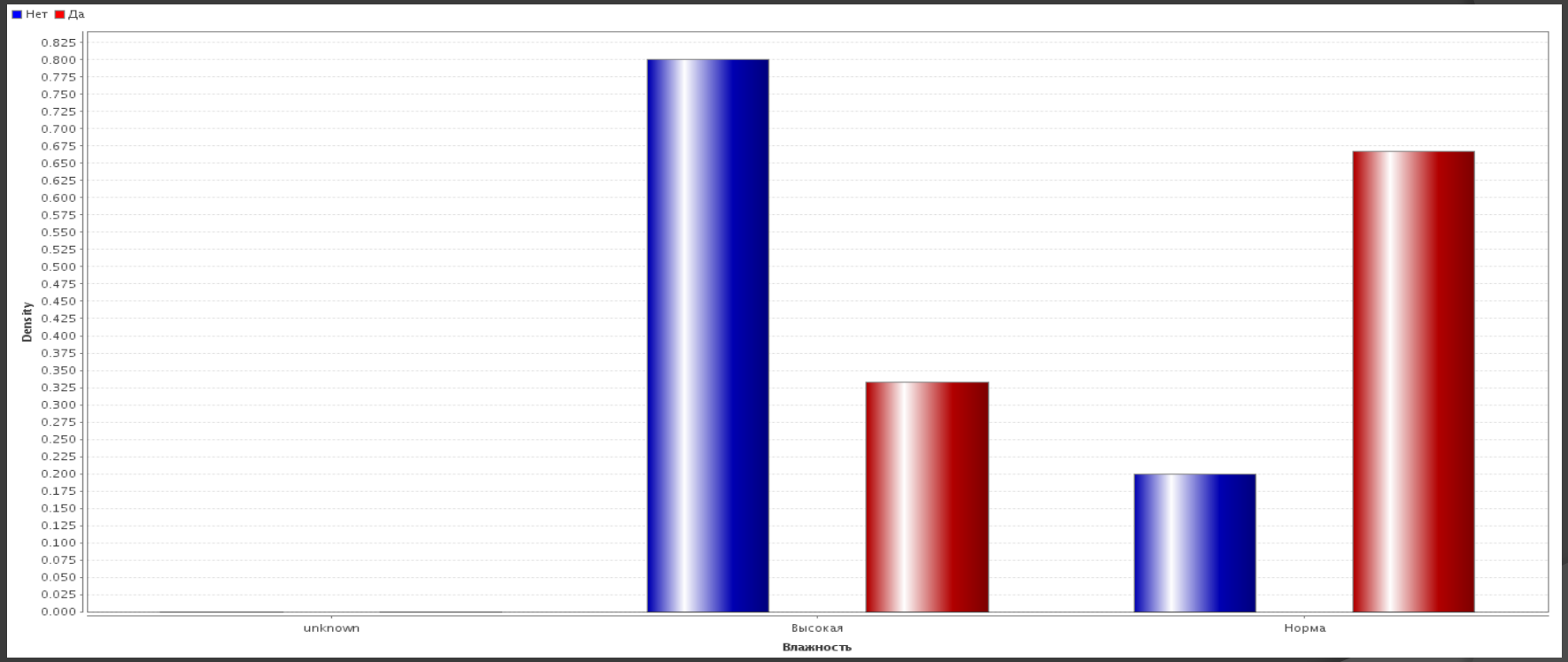
и является вероятностью наступления данного исхода вообще, а не только при E. P(E) просто сокращается.

Проблема: в обучающей выборке может не быть объекта с $x_n = c_n^k$ и при этом принадлежащему к классу c_r. Тогда вероятность равна нулю и соответственно вероятность правила равна нулю. Чтобы этого избежать, к каждой вероятности прибавляют значение, отличное от нуля. Это называется оценочной функцией Лапласа. При подсчете вероятностей тогда эти вероятности пропускаются.

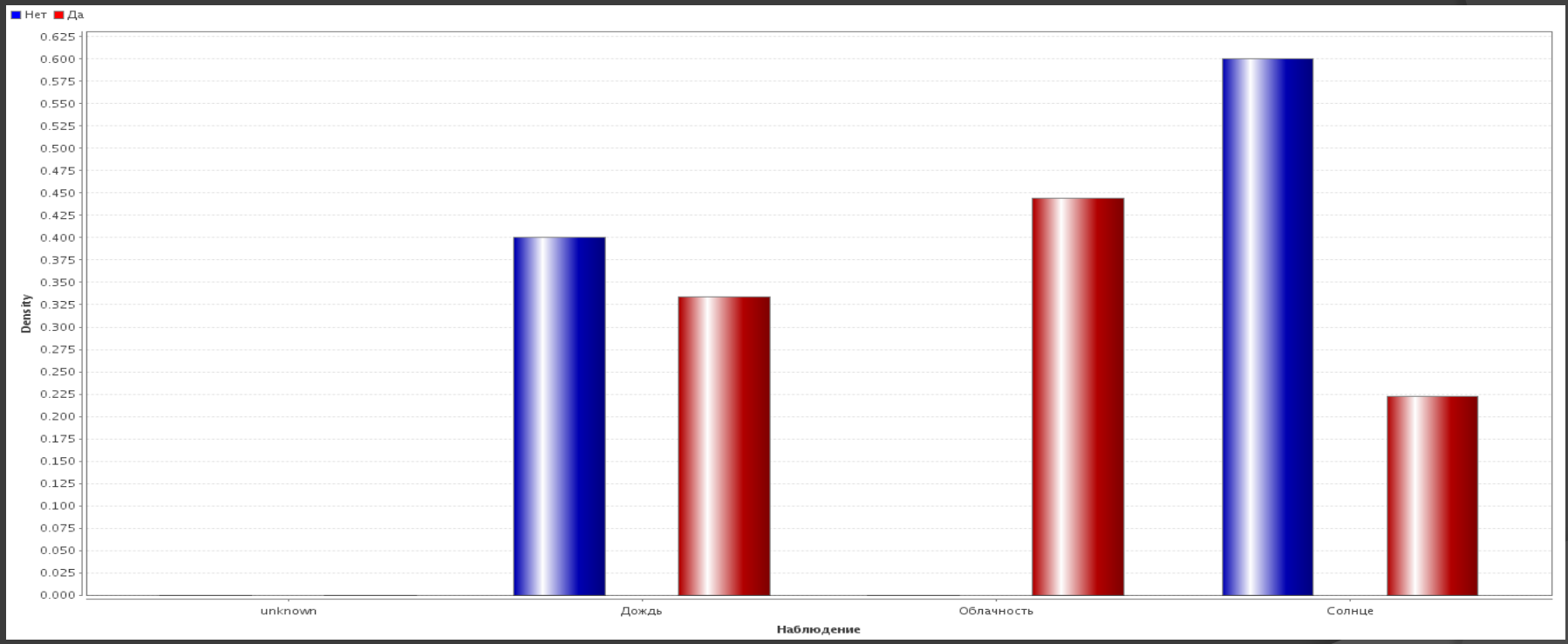
Условия для игры. Naïve Bayes



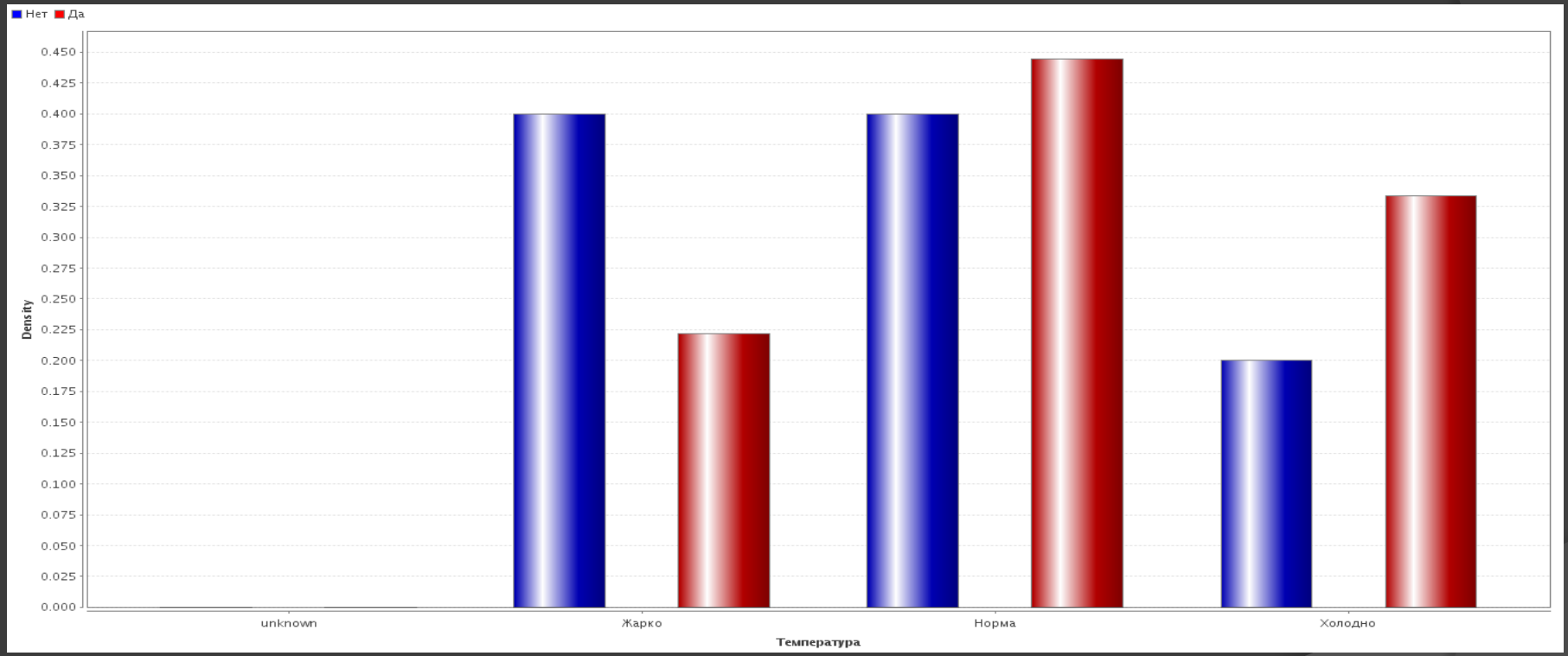
Условия для игры. Naïve Bayes



Условия для игры. Naïve Bayes



Условия для игры. Naïve Bayes



Условия для игры. Naïve Bayes

Attribute	Parameter	Нет	Да
Наблюдени	value=Солнце	0.600	0.222
Наблюдени	value=Облачность	0	0.444
Наблюдени	value=Дождь	0.400	0.333
Наблюдени	value=unknown	0	0
Температур	value=Жарко	0.400	0.222
Температур	value=Норма	0.400	0.444
Температур	value=Холодно	0.200	0.333
Температур	value=unknown	0	0
Влажность	value=Высокая	0.800	0.333
Влажность	value=Норма	0.200	0.667
Влажность	value=unknown	0	0
Ветер	value=Нет	0.400	0.667
Ветер	value=Есть	0.600	0.333
Ветер	value=unknown	0	0

Условия для игры. Naïve Bayes

ExampleSet (14 examples, 4 special attributes, 4 regular attributes)

Row No.	Игра	confidence(...	confidence(...	prediction(...	Наблюдение	Температура	Влажность	Ветер
1	Нет	0.795	0.205	Нет	Солнце	Жарко	Высокая	Нет
2	Нет	0.921	0.079	Нет	Солнце	Жарко	Высокая	Есть
3	Да	0	1	Да	Облачность	Жарко	Высокая	Нет
4	Да	0.464	0.536	Да	Дождь	Норма	Высокая	Нет
5	Да	0.067	0.933	Да	Дождь	Холодно	Норма	Нет
6	Нет	0.178	0.822	Да	Дождь	Холодно	Норма	Есть
7	Да	0	1	Да	Облачность	Холодно	Норма	Есть
8	Нет	0.660	0.340	Нет	Солнце	Норма	Высокая	Нет
9	Да	0.139	0.861	Да	Солнце	Холодно	Норма	Нет
10	Да	0.097	0.903	Да	Дождь	Норма	Норма	Нет
11	Да	0.422	0.578	Да	Солнце	Норма	Норма	Есть
12	Да	0	1	Да	Облачность	Норма	Высокая	Есть
13	Да	0	1	Да	Облачность	Жарко	Норма	Нет
14	Нет	0.722	0.278	Нет	Дождь	Норма	Высокая	Есть

Кросс-проверка

Оценка точности классификации может проводиться при помощи кросс-проверки. Кросс-проверка (Cross-validation) — это процедура оценки точности классификации на данных из тестового множества, которое также называют кросс-проверочным множеством. Точность классификации тестового множества сравнивается с точностью классификации обучающего множества. Если классификация тестового множества дает приблизительно такие же результаты по точности, как и классификация обучающего множества, считается, что данная модель прошла кросс-проверку. Разделение на обучающее и тестовое множества осуществляется путем деления выборки в определенной пропорции, например обучающее множество - две трети данных и тестовое - одна треть данных.

Условия для игры. Naïve Bayes

Table View Plot View

accuracy: 92.86%

	true Нет	true Да	class precision
pred. Нет	4	0	100.00%
pred. Да	1	9	90.00%
class recall	80.00%	100.00%	

Понятие деревьев решений

Метод деревьев решений (*decision trees*) **является одним из наиболее популярных методов решения задач классификации и прогнозирования. Иногда этот метод** *Data Mining* **также называют деревьями решающих правил, деревьями классификации и регрессии. Если зависимая, т.е. целевая переменная принимает дискретные значения, при помощи метода дерева решений решается задача классификации. Если же зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т.е. решает задачу численного прогнозирования.**

Понятие деревьев решений

Деревья решений — это способ представления классификационных правил в иерархической, последовательной структуре. Обычно каждый узел включает проверку одной независимой переменной. Иногда в узле дерева две независимые переменные сравниваются друг с другом или определяется некоторая функция от одной или нескольких переменных. Если переменная, которая проверяется в узле, принимает категориальные значения, то каждому возможному значению соответствует ветвь, выходящая из узла дерева. Если значением переменной является число, то проверяется больше или меньше это значение некоторой константы. Иногда область числовых значений разбивают на интервалы. (Проверка попадания значения в один из интервалов). Листья деревьев соответствуют значениям зависимой переменной, т.е. классам.

Методика построения деревьев

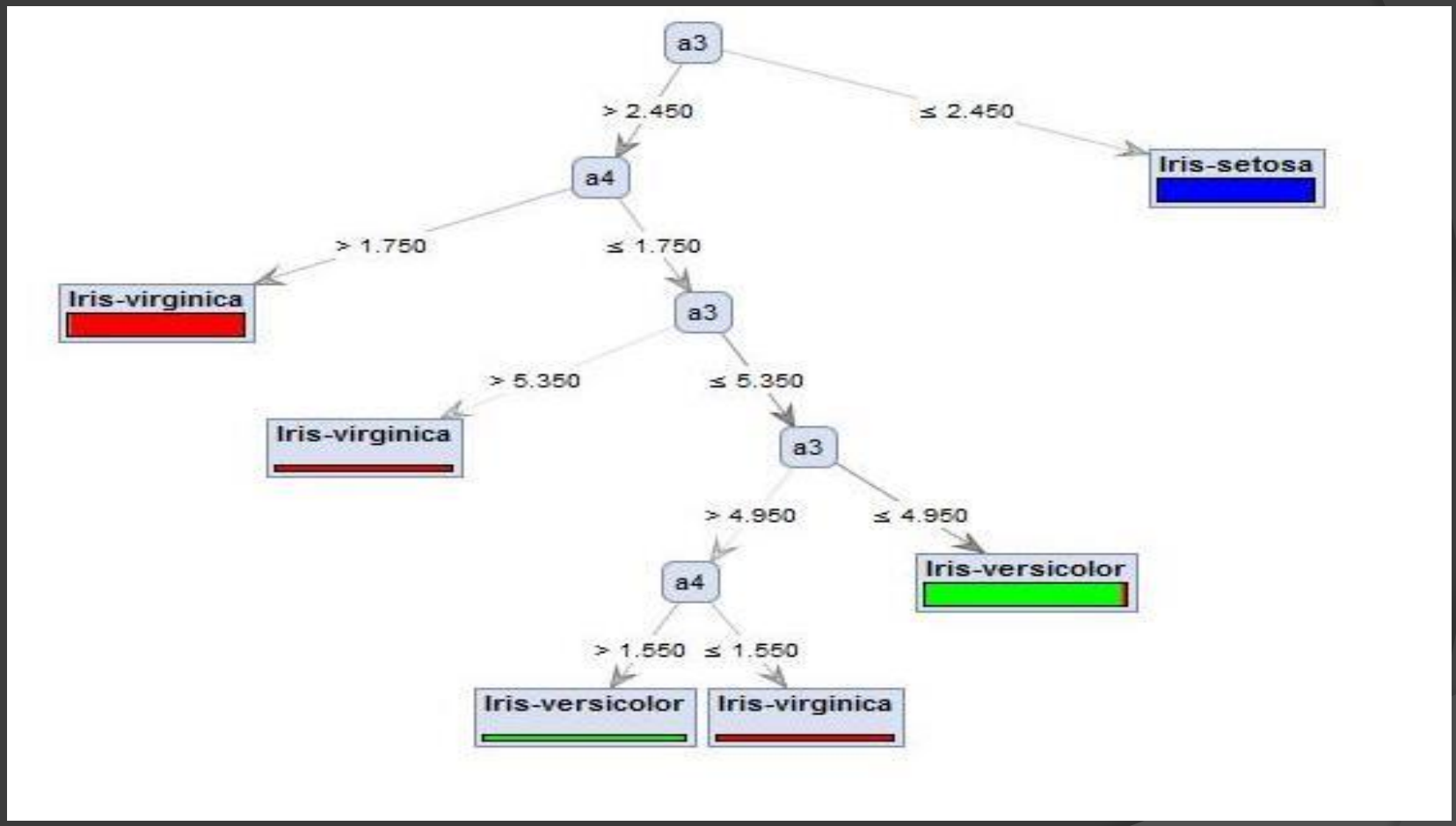
Методика основана на рекурсивном разбиении множества объектов из обучающей выборки на подмножества, содержащие объекты, относящиеся к одинаковым классам. Сперва выбирается независимая переменная, которая помещается в корень дерева. Из вершины строятся ветви, соответствующие всем возможным значениям выбранной независимой переменной. Множество объектов из обучающей выборки разбивается на несколько подмножеств в соответствии со значением выбранной независимой переменной. Таким образом, в каждом подмножестве будут находиться объекты, у которых значение выбранной независимой переменной будет одно и то же.

Методика построения деревьев

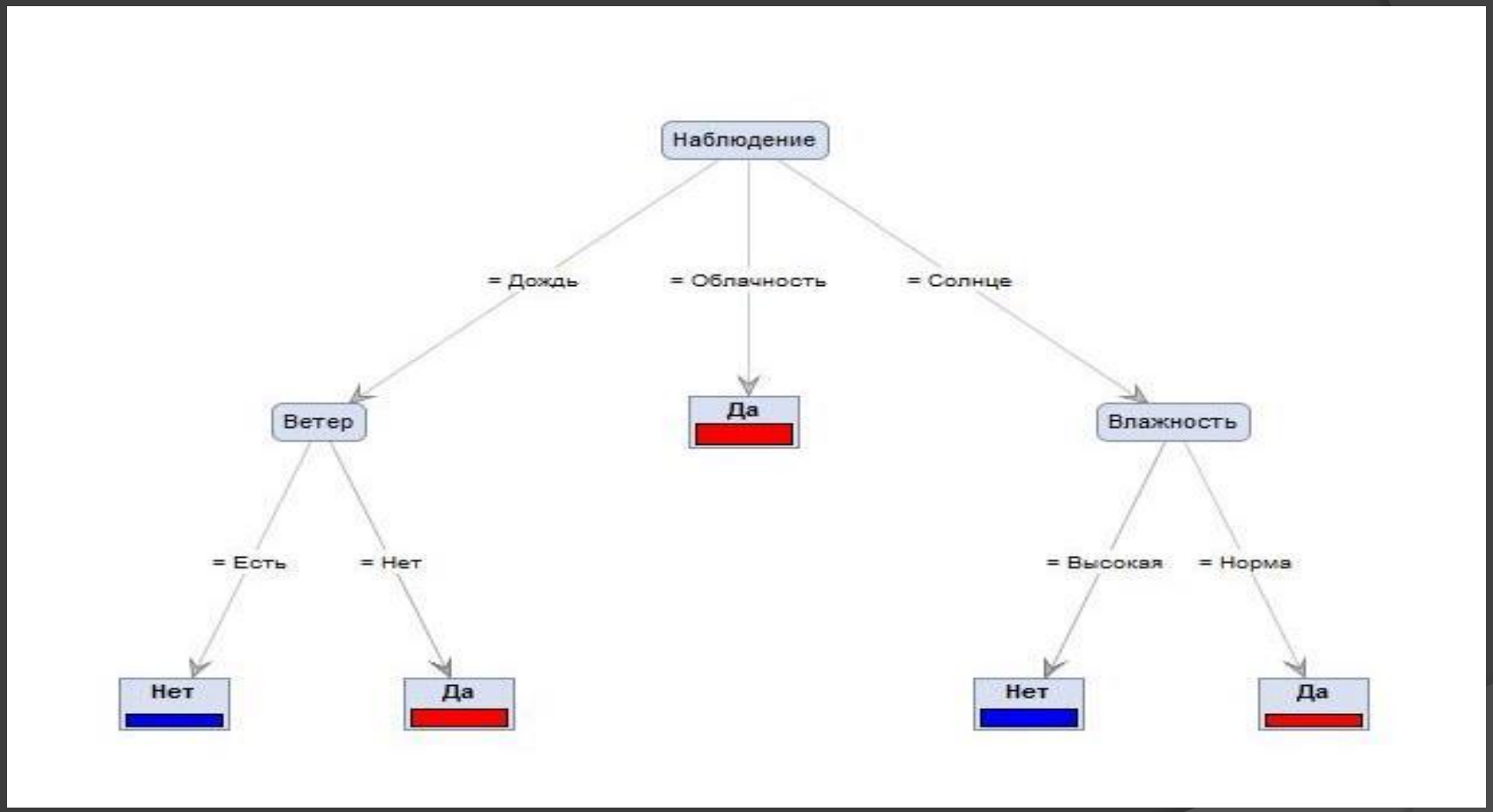
При использовании данной методики построение дерева решений будет происходить сверху вниз. Большинство алгоритмов, которые её используют, являются "жадными алгоритмами". Это значить, что если один раз переменная была выбрана и по ней было произведено разбиение, то алгоритм не может вернуться назад и выбрать другую переменную, которая дала бы лучшее разбиение. Вопрос в том, какую зависимую переменную выбрать для начального разбиения. От этого целиком зависит качество получившегося дерева.

Общее правило для выбора переменной для разбиения: выбранная переменная должны разбить множество так, чтобы получаемые в итоге подмножества состояли из объектов, принадлежащих к одному классу, или были максимально приближены к этому, т.е. чтобы количество объектов из других классов ("примесей") в каждом из этих множеств было минимальным.

Ирисы Фишера. Пример дерева классификации



Условия для игры. Пример дерева классификации



Алгоритм ID3

Рассмотрим критерий выбора независимой переменной, от которой будет строиться дерево. Полный набор вариантов разбиения $|X|$ — количество независимых переменных. Рассмотрим проверку переменой x_h, которая принимает m значений $c_h1, c_h2, \ldots, c_hm$. Тогда разбиение множества всех объектов обучающей выборке N по проверке переменной x_h даст подмножества $T_b T_2, \ldots, T_m$.

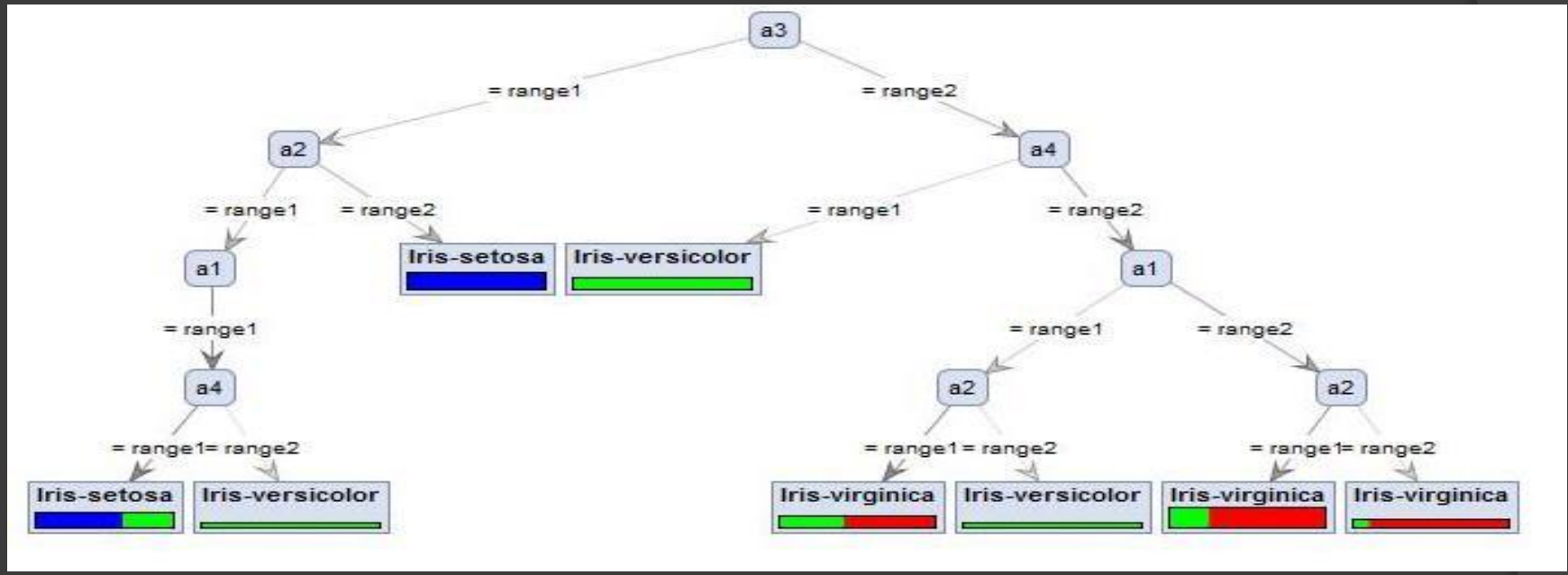
Мы ожидаем, что при разбиении исходного множества, будем получать подмножества с меньшим числом объектом, но более упорядоченные. Так, чтобы в каждом из них были по возможности объекты одного класса. Эта мера упорядоченности (неопределенности) характеризуется информацией. В контексте рассматриваемой задачи это количество информации, необходимое для того, чтобы отнести объект к тому или иному классу.

Алгоритм ID3

При разделении исходного множества на более мелкие подмножества, используя в качестве критерия для разделения значения выбранной независимой переменной, неопределённость принадлежности объектов конкретным классам будет уменьшаться.

Задача состоит в том, чтобы выбрать такие независимые переменные, чтобы максимально уменьшить эту неопределенность и в конечном итоге получить подмножества, содержащие объекты только одного класса. В последнем случае неопределенность равна нулю. Единственная доступная информация - каким образом классы распределены в множестве T и его подмножествах, получаемых при разбиении. Именно она и используется при выборе переменной.

Ирисы Фишера. Пример дерева классификации ID3



Области применения деревьев решений

Деревья решений являются прекрасным инструментом в системах поддержки принятия решений, интеллектуального анализа данных (data mining). В состав многих пакетов, предназначенных для интеллектуального анализа данных, уже включены методы построения деревьев решений. Деревья решений успешно применяются для решения практических задач в следующих областях:

- Банковское дело. Оценка кредитоспособности клиентов банка при выдаче кредитов.
- Промышленность. Контроль за качеством продукции (выявление дефектов), испытания без разрушений (например проверка качества сварки) и т.д.
- Медицина. Диагностика различных заболеваний.
- Молекулярная биология. Анализ строения аминокислот.

Преимущества использования деревьев решений

Преимущества использования деревьев решений:

- быстрый процесс обучения;
- генерация правил в областях, где эксперту трудно формализовать свои знания;
- извлечение правил на естественном языке;
- интуитивно понятная классификационная модель;
- высокая точность прогноза, сопоставимая с другими методами.

Анализ результатов классификации

- **Recall** — доля найденных объектов класса
- **Precision** — доля правильно найденных объектов класса
- **Confusion matrix** — матрица с распределением объектов по истинным классам и предсказанным классам

Confusion Matrix-Based Performance Measures

True class → / ↓ Hypothesized class	Pos	Neg
Yes	TP	FP
No	FN	TN
	P=TP+FN	N=FP+TN

- Multi-Class Focus:
  - **Accuracy** = (TP+TN)/(P+N)
- Single-Class Focus:
  - **Precision** = TP/(TP+FP)
  - **Recall** = TP/P
  - **Fallout** = FP/N
  - **Sensitivity** = TP/(TP+FN)
  - **Specificity** = TN/(FP+TN)