Введение.

Деревья решений — один из методов автоматического анализа данных. Разбираем общие принципы работы и области применения.

Деревья решений являются одним из наиболее эффективных инструментов интеллектуального анализа данных и предсказательной аналитики, которые позволяют решать задачи классификации и регрессии.

Они представляют собой иерархические древовидные структуры, состоящие из решающих правил вида «Если ..., то ...». Правила автоматически генерируются в процессе обучения на обучающем множестве и, поскольку они формулируются практически на естественном языке (например, «Если объём продаж более 1000 шт., то товар перспективный»), деревья решений как аналитические модели более вербализуемы и интерпретируемы, чем, скажем, нейронные сети.

Поскольку правила в деревьях решений получаются путём обобщения множества отдельных наблюдений (обучающих примеров), описывающих предметную область, то по аналогии с соответствующим методом логического вывода их называют индуктивными правилами, а сам процесс обучения — индукцией деревьев решений.

В обучающем множестве для примеров должно быть задано целевое значение, т.к. деревья решений являются моделями, строящимися на основе обучения с учителем. При этом, если целевая переменная дискретная (метка класса), то модель называют деревом классификации, а если непрерывная, то деревом регрессии.

Основополагающие идеи, послужившие толчком к появлению и развитию деревьев решений, были заложены в 1950-х годах в области исследований моделирования человеческого поведения с помощью компьютерных систем. Среди них следует выделить работы К. Ховеленда «Компьютерное моделирование мышления»[1] и Е. Ханта и др. «Эксперименты по индукции»[2].

Дальнейшее развитие деревьев решений как самообучающихся моделей для анализа данных связано с именами Джона Р. Куинлена[3], который

разработал алгоритм ID3 и его усовершенствованные модификации C4.5 и C5.0, а так же Лео Бреймана[4], который предложил алгоритм CART и метод случайного леса.

Терминология

Введем в рассмотрение основные понятия, используемые в теории деревьев решений.

Название	Описание
Объект	Пример, шаблон, наблюдение
Атрибут	Признак, независимая переменная, свойство
Целевая переменная	Зависимая переменная, метка класса
Узел	Внутренний узел дерева, узел проверки
Корневой узел	Начальный узел дерева решений
Лист	Конечный узел дерева, узел решения, терминальный узел
Решающее правило	Условие в узле, проверка

Структура дерева решений

Собственно, само дерево решений — это метод представления решающих правил в иерархической структуре, состоящей из элементов двух типов — узлов (node) и листьев (leaf). В узлах находятся решающие правила и производится проверка соответствия примеров этому правилу по какому-либо атрибуту обучающего множества.

В простейшем случае, в результате проверки, множество примеров, попавших в узел, разбивается на два подмножества, в одно из которых попадают примеры, удовлетворяющие правилу, а в другое — не удовлетворяющие.

Затем к каждому подмножеству вновь применяется правило и процедура рекурсивно повторяется пока не будет достигнуто некоторое условие остановки алгоритма. В результате в последнем узле проверка и разбиение не производится и он объявляется листом. Лист определяет решение для каждого попавшего в него примера. Для дерева классификации — это класс, ассоциируемый с узлом, а для дерева регрессии — соответствующий листу модальный интервал целевой переменной.

Таким образом, в отличие от узла, в листе содержится не правило, а подмножество объектов, удовлетворяющих всем правилам ветви, которая заканчивается данным листом.

Очевидно, чтобы попасть в лист, пример должен удовлетворять всем правилам, лежащим на пути к этому листу. Поскольку путь в дереве к каждому листу единственный, то и каждый пример может попасть только в один лист, что обеспечивает единственность решения.

Задачи

Основная сфера применения деревьев решений — поддержка процессов принятия управленческих решений, используемая в статистике, анализе данных и машинном обучении. Задачами, решаемыми с помощью данного аппарата, являются:

Классификация — отнесение объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.

Регрессия (численное предсказание) — предсказание числового значения независимой переменной для заданного входного вектора.

Описание объектов — набор правил в дереве решений позволяет компактно описывать объекты. Поэтому вместо сложных структур, описывающих объекты, можно хранить деревья решений.

Пример использования

В данной статье мы на примере рассмотрим использование метода построения дерева решений для анализа рисковых явлений.

Предположим, некая компания, назовем ее «Robots Ltd», решила инвестировать часть имеющегося у нее инвестиционного фонда в проект по разработке роботизированной техники, направленной на помощь в проведении космических исследований.

В данном конкретном случае компания будет проводить инвестирования в три этапа:

На первом этапе компания выделит 500 000 долларов для проведения маркетинговых исследований существующего рынка роботизированной техники. При успешных результатах исследования, проведенного на первом этапе, когда оценка привлекательности рынка будет достаточно велика, «Robots Ltd» выделит еще один миллион долларов на проведение работ по созданию опытных образцов. Образцы будут продемонстрированы специалистам в области космических исследований, чтобы они приняли решение о необходимости заказа роботов у компании-производителя.
Если центр космических решений, где были представлены опытные образцы, решает, что он нуждается в данном виде роботизированной техники, компаний совершает очередной инвестиционный транш в размере 10 миллионов долларов для строительства завода, где начнется серийный выпуск роботов. По данным экспертов и аналитиков, новый завод еще в течение нескольких лет будет генерировать потоки инвестиционных вливаний; объемы денежных инвестиций будут характеризоваться привлекательностью новой продукции на потребительском рынке.
Стоит отметить, что на каждом из трех этапов, если в ходе его реализации результаты оказались неудовлетворительными, компания может свернуть инвестиционную программу. Именно для анализа подобных решений, имеющих не одну ступень, используется метод построения дерева решений.

Вероятность каждого положительного исхода оценивается экспертами в процентном соотношении и, исходя из полученной величины, определяется дальнейшее развитие событий.

Прекращение реализации проекта может быть инициировано компанией в любой момент времени. Издержки от отказа можно сократить, если компания имеет альтернативные способы применения имеющихся в ее руках активов. В нашем случае проект легко закрывается в том случае, когда оборудование для производства роботизированной техники может быть использовано для выпуска

роботов другого назначения.

Метод построения дерева решений весьма распространен, но для его использования проектные руководители (менеджеры) должны обладать достаточным опытом и компетенцией в области проектного управления.