ДИСЦИПЛИНА	Прикладная математика
ИНСТИТУТ	ИПТИП
КАФЕДРА	Индустриального программирования
ВИД УЧЕБНОГО МАТЕРИАЛА	Методические указания по дисциплине
ПРЕПОДАВАТЕЛЬ	Астафьев Рустам Уралович
СЕМЕСТР	1 семестр, 2025/2026 уч. год

Ссылка на материал:
https://github.com/astafiev-rustam/applied-mathematics/tree/lecture-1-6

Лекция №6: Решающие деревья

Введение в решающие деревья

Решающие деревья являются одним из наиболее интуитивно понятных и широко применяемых алгоритмов машинного обучения. Они используются как для задач классификации, так и для регрессии. Модель, построенная этим алгоритмом, представляет собой древовидную структуру, которая имитирует процесс принятия решений человеком. Каждый внутренний узел дерева соответствует проверке значения одного из признаков, каждая ветвь результату этой проверки, а каждый лист дерева присвоенному классу или прогнозируемому значению. Интерпретируемость решающих деревьев является их key преимуществом по сравнению со многими другими сложными моделями.

Понятие решающих деревьев

Построение решающего дерева происходит сверху вниз. Алгоритм начинает с корневого узла, который содержит всю обучающую выборку. Затем он выбирает признак и пороговое значение для него, которые наилучшим образом разделяют данные на две более однородные подгруппы. Этот процесс рекурсивно повторяется для каждой образовавшейся подгруппы до тех пор, пока не будет выполнено условие остановки. Таким условием может быть достижение максимальной глубины дерева, недостаточное количество объектов в узле для дальнейшего разделения или отсутствие значимого улучшения однородности.

Измерение объема информации

Ключевым моментом в построении дерева является критерий выбора наилучшего признака для разделения. Этот выбор основан на понятиях из теории информации, таких как энтропия и индекс Джини. Энтропия количественно измеряет степень неопределенности или беспорядка в наборе данных. Если все объекты в узле принадлежат к одному классу, энтропия равна нулю. Если распределение классов равномерно, энтропия максимальна. Алгоритм стремится найти такое разбиение, которое максимально уменьшит энтропию, то есть увеличит информационный выигрыш. Индекс Джини измеряет вероятность неправильной классификации случайно выбранного объекта из узла, если бы мы присвоили ему метку в соответствии с распределением классов в этом узле.

Методы построения решающих деревьев

Различные алгоритмы построения деревьев используют разные критерии и стратегии. Алгоритм ID3 использует максимизацию информационного выигрыша и не работает с непрерывными признаками. Его преемник, алгоритм C4.5, устраняет этот недостаток и включает методы борьбы с переобучением, такие как упрощение дерева. Алгоритм CART является еще одним популярным вариантом, который может строить деревья как для классификации, так и для регрессии, используя для этого индекс Джини и дисперсию соответственно. Все эти алгоритмы реализуют жадную стратегию, выбирая локально оптимальное разбиение на каждом шаге, что не гарантирует глобального оптимума, но является computationally эффективным.

Сильные стороны и ограничения

Решающие деревья не требуют предварительной нормализации данных и могут работать как с числовыми, так и с категориальными признаками. Они способны моделировать нелинейные зависимости и устойчивы к выбросам. Однако они склонны к переобучению, особенно когда дерево становится слишком глубоким и complex. Для смягчения этой проблемы применяется упрощение дерева. Другим мощным методом является использование ансамблей деревьев, таких как случайные леса и градиентный бустинг, которые комбинируют прогнозы множества деревьев для получения более точного и устойчивого результата.

Примеры и реализация

Рассмотрим примеры по теме лекционного занятия:

Пример 1

Пример 2

Пример 3