МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ им. Р.Е.АЛЕКСЕЕВА

Институт радиоэлектроники и информационных технологий

Кафедра «Вычислительные системы и технологии»

Реферат

По теме «Деревья принятия решений – критерий Джини»

по дисциплине «Методы и средства обработки сигналов»

ПРОВЕРИЛ:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Авербух М.Л.

СТУДЕНТЫ:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Караганов И.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Перминов Я.

17-В-2

Работа защищена «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_

С оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Нижний Новгород

2020

Дерево принятия решений (также может называться деревом классификации или регрессионным деревом) — средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике. Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.  
 Каждый лист представляет собой значение целевой переменной, изменённой в ходе движения от корня по листу. Каждый внутренний узел соответствует одной из входных переменных. Дерево может быть также «изучено» разделением исходных наборов переменных на подмножества, основанные на тестировании значений атрибутов. Это процесс, который повторяется на каждом из полученных подмножеств. Рекурсия завершается тогда, когда подмножество в узле имеет те же значения целевой переменной, таким образом, оно не добавляет ценности для предсказаний. Процесс, идущий «сверху вниз», индукция деревьев решений , является примером поглощающего «жадного» алгоритма, и на сегодняшний день является наиболее распространённой стратегией деревьев решений для данных, но это не единственная возможная стратегия. В интеллектуальном анализе данных, деревья решений могут быть использованы в качестве математических и вычислительных методов, чтобы помочь описать, классифицировать и обобщить набор данных, которые могут быть записаны следующим образом:   
  
  
 Зависимая переменная *Y* является целевой переменной, которую необходимо проанализировать, классифицировать и обобщить. Вектор *x* состоит из входных переменных *x1, x2, x3* и т. д., которые используются для выполнения этой задачи.

Алгоритмы построения деревья решений обычно работают сверху вниз путём выбора переменной на каждом шаге, которая лучшим образом разбивает множество элементов. Разные алгоритмы используют различные метрики для измерения «лучшего» решения. Они обычно измеряют однородность целевой переменной на подмножествах. Эти метрики применяются к каждому подмножеству и получающиеся значения комбинируются (например, вычисляется среднее) для получения меры качества разбиения.

**Алгоритм CART**

Алгоритм CART предназначен для построения бинарного дерева решений. Бинарные деревья также называют двоичными, значит, что каждый узел дерева при разбиении имеет только двух потомков. Для алгоритма CART «поведение» объектов выделенной группы означает долю модального значения выходного признака. Выделенные группы — те, для которых эта доля достаточно высока. На каждом шаге построения дерева правило, формируемое в узле, делит заданное множество примеров на две части — часть, в которой выполняется правило (потомок — right) и часть, в которой правило не выполняется (потомок — left).

Преимуществом алгоритма CART является определенная гарантия того, что если искомые детерминации существуют в исследуемой совокупности, то они будут выявлены. Кроме того, CART позволяет не «замыкаться» на единственном значении выходного признака, а искать все такие его значения, для которых можно найти соответствующее объясняющее выражение.

Метод CART применяется для **номинальных** (обычно двухуровневых) и **порядковых** предикторных переменных. В этом методе перебираются все возможные варианты ветвления для каждого узла, и выбирается та предикторная переменная, при которой оценочная функция дает наилучший показатель.

### **Примесь (критерий) Джини**

Используемый в алгоритме деревьев классификации и регрессии ([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) *classification and regression tree*, CART) критерий Джини является мерой, насколько часто случайно выбранный элемент из набора неверно помечается, если он случайным образом помечается согласно распределению меток в подмножестве. Критерий Джини может быть вычислен путём суммирования вероятности ***pi***элемента с выбранной меткой *i*, умноженной на вероятность



ошибки категоризации этого элемента. Критерий принимает минимум (нуль), когда все случаи в узле попадают в одну целевую категорию.

Для вычисления критерия Джини для набора элементов с ***J*** классами, предположим, что, и пусть ***pi*** будет долей элементов, помеченных классом ***i***  в наборе.



**Неопределенность Джини (Gini impurity):**



Максимизацию этого критерия можно интерпретировать как максимизацию числа пар объектов одного класса, оказавшихся в одном поддереве.

В случае задачи бинарной классификации (p+ – вероятность объекта иметь метку +) энтропия и неопределенность Джини примут следующий вид:

