Машинное обучение
Теоретическое домашнее задание №5

Задача 1. На лекциях говорилось, что критерий информативности для набора объектов R вычисляется на основе того, насколько хорошо их целевые переменные предсказываются константой (при оптимальном выборе этой константы):

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c),$$

где $L(y, c)$ — некоторая функция потерь. Соответственно, чтобы получить вид критерия при конкретной функции потерь, необходимо аналитически найти оптимальное значение константы и подставить его в формулу для $H(R)$.

Выведите критерии информативности для следующих функций потерь:

1. $L(y, c) = (y - c)^2$;
2. $L(y, c) = \sum_{k=1}^{K} (c_k - [y = k])^2$;
3. $L(y, c) = \sum_{k=1}^{K} [y = k] \log c_k$.

У вас должны получиться дисперсия, критерий Джини и энтропийный критерий соответственно.

Задача 2. Ответьте на вопросы:

1. Что такое решающее дерево? Как по построенному дереву найти прогноз для объекта?
2. Зачем в вершинах нужны предикаты? Какие типы предикатов вы знаете? Приведите примеры.
3. Почему для любой выборки можно построить решающее дерево, имеющее нулевую ошибку на ней?
4. Почему не рекомендуется строить небинарные деревья (т.е. имеющие больше двух потомков у каждой вершины)?
5. Как устроен жадный алгоритм построения дерева? Какие у него параметры?
6. Зачем нужны критерии информативности?
7. Как задается критерий ошибки классификации? Критерий Джини? Энтропийный критерий? Какой у них смысл?

8. Как задается критерий информативности, основанный на среднеквадратичной ошибке, в задачах регрессии?

9. Какие критерии останова вы знаете?

10. Что такое стрижка дерева?

11. Какие методы обработки пропущенных значений вы знаете?

12. Как можно учитывать категориальные признаки в решающем дереве?

13. Как можно свести задачу перебора всех разбиений категориального признака к задаче поиска оптимального разбиения для вещественного признака?