Машинное обучение Теоретическое домашнее задание №3

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. Запишите оценку сложности построения одного решающего дерева в зависимости от размера обучающей выборки ℓ , числа признаков d, максимальной глубины дерева D. В качестве предикатов используются пороговые функции $[x_j > t]$. При выборе предиката в каждой вершине перебираются все признаки, а в качестве порогов рассматриваются величины t, равные значениям данного признака на объектах, попавших в текущую вершину. Считайте сложность вычисления критерия информативности константной.