Машинное обучение, ФКН ВШЭ
Домашнее задание №7

Задача 1. Убедитесь, что знаете ответы на следующие вопросы:

1. По какой формуле вычисляется значение алгоритма в бустинге после обучения?
2. По какому критерию выбираются параметры очередного базового алгоритма?
3. По какой переменной берется производная функции потерь при вычислении сдвигов?

Задача 2. Найдите значения сдвигов для следующих функций потерь:

1. Экспоненциальная $L(y, z) = e^{-yz}$
2. Сигмоидная $L(y, z) = 2\sigma(-yz)$.

Задача 3. Предположим, что в градиентном бустинге мы используем дифференцируемые алгоритмы $b_n(x) = f(x; w_n)$, и хотим обучать их очень грубо: одним шагом градиентного спуска из стартового значения параметров $w_n^{(0)} = 0$. Считаем, что функция потерь $L(y, z)$ произвольная.

1. Найдите, чему равно $b_n(x)$.
2. Положим теперь $f(x; w) = w^T x$. Покажите, что итоговая композиция будет эквивалентена градиентному спуску для одного базового алгоритма (с точностью до длины шага).

Задача 4. Исследуем возможность использования градиентного бустинга для многоклассовой классификации. Конечно, можно разбить ее на несколько задач бинарной классификации, но интереснее было бы использовать возможность многоклассовой классификации у базовых алгоритмов (например, решающие деревья такую возможность предоставляют). Выполните ряд задач ниже, чтобы прийти к этому алгоритму.

1. Рассмотрим случай бинарной классификации. Обозначим $y^1 = 1, y^2 = -1$. Покажите, что отступ алгоритма b на объекте (x, y^k) равен

$$y^k b(x) = \frac{1}{2}(y^k b(x) - \max_{l \neq k} y^l b(x))$$

2. Рассмотрим случай K-классовой классификации. Положим

$$y^1 = (1, 0 \ldots, 0), y^2 = (0, 1, 0, \ldots, 0), \ldots, y^K = (0, \ldots, 0, 1).$$

   Также считаем, что алгоритм $b(x)$ возвращает на объекте вектор вероятностей классов длины K. По аналогии с предыдущим пунктом определите отступ $M(y^k, b(x))$ алгоритма на объекте (x, y^k).

3. Предположим, что функция потерь зависит только от отступа: $L(y^k, b(x)) = \varphi(M(y^k, b(x)))$. Сформулируйте, как теперь выглядит задача градиентного бустинга. Какая с ней возникла проблема?

4. Рассмотрим функцию SmoothMax:

$$S_\alpha(x_1, \dots, x_n) = \frac{\sum_{i=1}^n x_i e^{\alpha x_i}}{\sum_{i=1}^n e^{\alpha x_i}}$$

Изучите свойства этой функции и решите с помощью нее проблему предыдущего пункта. Сформулируйте теперь окончательный алгоритм многоклассового градиентного бустинга.