Машинное обучение.
Домашнее задание №5

Задача 1. Пусть подвыборка $\tilde{X}^\ell$ генерируется с помощью бустрэппинга из выборки X^ℓ размера ℓ. Найдите вероятность того, что фиксированный объект $x \in X^\ell$ попадет в подвыборку $\tilde{X}^\ell$. Чему равна эта вероятность, если $\ell \to \infty$?

Задача 2. Известно, что бэггинг плохо работает, если в качестве базовых классификаторов взять методы ближайшего соседа. Попробуем понять причины на простом примере.

Пусть дана выборка X^ℓ из ℓ объектов с ответами из множества $\mathbb{Y} = \{-1, +1\}$. Будем рассматривать классификатор одного ближайшего соседа в качестве базового алгоритма. Построим с помощью бэггинга композицию длины N:

$$a_N(x) = \text{sign} \sum_{n=1}^{N} b_n(x).$$

Оцените вероятность того, что ответ композиции на произвольном объекте x будет отличаться от ответа одного классификатора ближайшего соседа, обученного по всей выборке. Покажите, что эта вероятность стремится к нулю при $N \to \infty$.

Подсказка: ответ композиции на x может отличаться от ответа одного алгоритма только в том случае, если ближайший сосед x попал в обучение для менее чем половины базовых алгоритмов.

Задача 3. Пусть $x_1, \ldots, x_N$ — одинаково распределенные случайные величины с дисперсией σ^2. Если они независимы, то дисперсия их среднего равна σ^2/N. Покажите, что если корреляция между любой парой этих величин равна $\rho > 0$, то дисперсия среднего вычисляется по формуле

$$\mathbb{D}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] = \rho\sigma^2 + \frac{1-\rho}{N}\sigma^2.$$