Домашнее задание 5

По курсу "Машинное обучение"

Аннотация
В этом задании вам нужно решить несколько задач по ансамблям моделей.

Задача 1 (2 балла)

Рассмотрим модель логистической регрессии с функцией потерь:

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i), \tag{1}$$

$$\hat{y}_i = \sigma(z_i), \tag{2}$$

$$z_i = x_i^T w. \tag{3}$$

где n - количество объектов в выборке; $\hat{y}_i$ - прогноз модели; w - веса модели; Покажите, что градиент вычисляется по формуле:

$$\frac{\partial L}{\partial w} = -\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{y}_i). \tag{4}$$

Как это выражение будет выглядеть в матричной форме?

Задача 2 (2 балла)

Рассмотрим задачу регрессии, где нам нужно предсказать значение функции $f(x)$, где x - одномерная непрерывная переменная. Пусть у нас есть M моделей регрессии, которые были обучены на случайно сгенерированных наборах данных. Прогноз каждой модели $\hat{y}_m(x)$ можем записать так:

$$\hat{y}_m(x) = f(x) + \epsilon_m(x), \tag{5}$$

где $\epsilon_m(x)$ - случайная величина со стандартным нормальным распределением $\mathcal{N}(\mu = 0, \sigma = 1)$. Рассмотрим модель бэггинга:

$$\hat{y}_{bag}(x) = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_m(x). \tag{6}$$

Обозначим среднюю ошибку всех моделей следующим образом:

$$E_{av} = \frac{1}{M}\sum_{m=1}^{M} E[(\hat{y}_m(x) - f(x))^2]. \tag{7}$$

Обозначим среднюю ошибку модели бэггинга так:

$$E_{bag} = E[(\hat{y}_{bag}(x) - f(x))^2]. \tag{8}$$

Покажите, что

$$E_{bag} = \frac{1}{M} E_{av}. \tag{9}$$

Ошибки $\epsilon_m(x)$ считайте независимыми.

Задача 3 (2 балла)

Используя неравенство Йенсена покажите, что $E_{bag} \leq E_{av}$ для любой выпуклой функции потерь, а не только для MSE.

Задача 4 (2 балла)

Рассмотрим алгоритм градиентного бустинга на решающих деревьях для задачи классификации с логистической функцией потерь:

$$L(y, \hat{y}_k(x)) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log \hat{y}_k(x_i) + (1 - y_i)\log(1 - \hat{y}_k(x_i)), \tag{10}$$

где n - количество объектов в выборке; $\hat{y}_k(x_i)$ - прогноз ансамбля из k деревьев. Покажите, что остатки (сдвиги) s вычисляются по формуле:

$$s(x_i) = \frac{y_i - \hat{y}_k(x_i)}{\hat{y}_k(x_i)(1 - \hat{y}_k(x_i))}. \tag{11}$$

Что дальше нужно делать с этими остатками? Опишите остальные шаги алгоритма построения градиентного бустинга. Какую функцию потерь нужно использовать для обучения нового дерева в бустинге?

Задача 5 (2 балла)

Рассмотрим выборку из n объектов, где n достаточно большое. Будем использовать метод бутстрапа для сэмплирования подвыборок из n объектов с повторениями. Покажите, что каждая подвыборка содержит в среднем 63% объектов исходной выборки.