Фамилия:...

Имя:...

Группа:...

Задача №1

У вас имеется выборка из $n = 100$ наблюдений, таких, что $X_i = 2i$ и $Y_i = i$, то есть $X = (2, 4, 6, ..., 200)$ и $Y = (1, 2, ..., 100)$. На этой выборке вы обучили градиентный бустинг, где в качестве базовой модели использовалось выборочное среднее, скорость обучения равнялась 0.1, функция потерь была квадратичной (без деления на n), а градиент прогнозировался с помощью метода ближайших соседей с 3-мя ближайшими соседями с расстоянием Манхэттен.

1. С помощью одного шага данного градиентного бустинга спрогнозируйте y, если $x = 2025$. **(10 баллов)**

2. С помощью двух шагов данного градиентного бустинга спрогнозируйте y, если $x = 2025$. **(10 баллов)**

3. Определите, чему на 2025-м шаге будет равняться разница прогнозов градиентного бустинга при $x = 200$ и $x = 2025$. Ответ подробно обоснуйте. **(5 баллов)**

4. Повторите первый пункт, заменив градиентный спуск на метод Ньютона. Сделайте вывод о целесообразности применения метода Ньютона по сравнению с градиентным спуском в данном случае. Уточните, справедлив ли ваш вывод для альтернативных функций потерь. **(5 баллов)**

5. Вернемся к исходному градиентному бустингу из первого пункта – с 1-й итерацией и градиентным спуском. Вы разбили выборку на 2 части: с 1-го по 50-е наблюдения и с 51-го по 100-е. Рассчитайте MAE 2-х частной кросс-валидации. **(10 баллов)**

Подсказска: среднее членов арифметической прогрессии от 1 до m считается как:

$$\frac{1}{m}\sum_{t=1}^{m} t = \frac{1+m}{2}$$

Решение

1. Для удобства запишем функцию потерь и ее производную:

$$\mathrm{L}(y, F(x)) = (F(x) - y)^2$$
$$\mathrm{L}'(y, F(x)) = 2\,(F(x) - y)$$

Обратим внимание, что ближайшими соседями $x = 2025$ будут $X_{98} = 196$, $X_{99} = 198$ и $X_{100} = 200$.

Рассчитаем остатки (отрицательные градиенты) для каждого из этих наблюдений:

$$r_{98} = -2 \times (50.5 - 98) = 95$$
$$r_{99} = -2 \times (50.5 - 99) = 97$$
$$r_{100} = -2 \times (50.5 - 100) = 99$$

В результате получаем прогноз остатка:

$$h_1(2025) = \frac{95 + 97 + 99}{3} = 97$$

Таким образом, прогноз градиентного бустинга имеет вид:

$$F_1(2025) = 50.5 + 0.1 \times 97 = 60.2$$

2. По аналогии с предыдущим пунктом получаем, что:

$$F_1(196) = 50.5 + 0.1 \times \frac{-2 \times (50.5 - 97 + 50.5 - 98 + 50.5 - 99)}{3} = 60$$
$$F_1(198) = 50.5 + 0.1 \times \frac{-2 \times (50.5 - 98 + 50.5 - 99 + 50.5 - 100)}{3} = 60.2$$
$$F_1(200) = 50.5 + 0.1 \times \frac{-2 \times (50.5 - 98 + 50.5 - 99 + 50.5 - 100)}{3} = 60.2$$

Следовательно, остатки второго шага имеют вид:

$$r^*_{98} = -2 \times (F_1(196) - 98) = -2 \times (60 - 98) = 78$$
$$r^*_{99} = -2 \times (F_1(198) - 99) = -2 \times (60.2 - 99) = 77.6$$
$$r^*_{100} = -2 \times (F_1(200) - 100) = -2 \times (60.2 - 100) = 79.6$$

Отсюда получаем:

$$h_2(2025) = \frac{78 + 77.6 + 79.6}{3} = 78.4$$

В итоге имеем:

$$F_2(2025) = 60.2 + 0.1 \times 78.4 = 68.04$$

3. Поскольку наблюдения $x = 200$ и $x = 2025$ имеют одних и тех же ближайших соседей, то независимо от шага алгоритма разница прогнозов будет равняться 0.

4. Обратим внимание, что:

$$\mathrm{L}''(y, F(x)) = 2$$

Следовательно, остатки будут считаться по формуле:

$$r_i = -\frac{\mathrm{L}'(Y_i, F(X_i))}{\mathrm{L}''(Y_i, F(X_i))} = -\frac{2\,(F(X_i) - Y_i)}{2} = Y_i - F(X_i)$$

Рассчитаем соответствующие остатки:

$$r_{98} = 98 - 50.5 = 47.5$$
$$r_{99} = 99 - 50.5 = 48.5$$
$$r_{100} = 100 - 50.5 = 49.5$$

В результате получаем прогноз остатка:

$$h_1(2025) = \frac{47.5 + 48.5 + 49.5}{3} = 48.5$$

Таким образом, прогноз градиентного бустинга имеет вид:

$$F_1(2025) = 50.5 + 0.1 \times 48.5 = 55.35$$

5. Обратим внимание, что при прогнозировании на 1-й части с использованием классификатора, обученного на 2-й, ближайшими соседями всегда будут наблюдения X_{51}, X_{52} и X_{53}. Следовательно, для всех наблюдений будет использоваться один и тот же прогноз, то есть при $x \in \{X_1, ..., X_{50}\}$ получаем:

$$F_1(x) = \frac{51 + 100}{2} + 0.1 \times \frac{-2 \times \left(\frac{51+100}{2} - 51 + \frac{51+100}{2} - 52 + \frac{51+100}{2} - 53\right)}{3} = 70.8$$

Отсюда получаем:

$$\mathrm{MAE}_1 = \frac{|70.8 - 1| + ... + |70.8 - 50|}{50} = \frac{69.8 + 70.8 + ... + 20.8}{50} = \frac{69.8 + 20.8}{2} = 45.3$$

По аналогии при прогнозировании на 2-й части с использованием классификатора, обученного на 1-й, ближайшими соседями всегда будут наблюдения X_{50}, X_{49} и X_{48}. Следовательно, для всех наблюдений будет использоваться один и тот же прогноз, то есть при $x \in \{X_{51}, ..., X_{100}\}$ получаем:

$$F_1(x) = \frac{1 + 50}{2} + 0.1 \times \frac{-2 \times \left(\frac{1+50}{2} - 50 + \frac{1+50}{2} - 49 + \frac{1+50}{2} - 48\right)}{3} = 30.2$$

Отсюда получаем:

$$\mathrm{MAE}_2 = \frac{|30.2 - 51| + ... + |30.2 - 100|}{50} = \frac{20.8 + 21.8 + ... + 69.8}{50} = \frac{20.8 + 69.8}{2} = 45.3$$

Таким образом, получаем:

$$\mathrm{MAE}_{\mathrm{CV}} = \frac{\mathrm{MAE}_1 + \mathrm{MAE}_2}{2} = \frac{45.3 + 45.3}{2} = 45.3$$

Задача №2

Решите следующие задачи.

1. Известно, что $\text{pAUC}(0.5, 1) = 2\text{pAUC}(0, 0.5) = 0.5$. Определите, чему равняется AUC. (**5 баллов**)

2. Вы оцениваете условный средний эффект воздействия T_i на Y_i с помощью S-learner и имеете в распоряжении одну контрольную переменную X_i. В качестве метода оценивания условного математического ожидания используется нейросеть без смещений (констант), 1-м скрытым слоем, 2-мя нейронами, функцией активации ReLU как в скрытом слое, так и в выходном. После обучения нейросети оказалось, что все веса равняются 0.5. Оцените условный средний эффект воздействия при $X_i = 2$. (**10 баллов**)

3. Рассмотрим ансамбль из k решающих деревьев, основанный на бэггинге. Корреляция между прогнозами решающих деревьев равняется 0.6. Определите, при каком количестве решающих деревьев дисперсия прогноза ансамбля окажется ровно в 1.25 раза меньше дисперсии прогноза одного решающего дерева. (**5 баллов**)

4. В добавок к ансамблю из предыдущего пункта на тех же данных с использованием бэггинга оценили еще один аналогичный ансамбль с таким же количеством деревьем (значение k, найденное в предыдущем пункте). Найдите корреляцию между прогнозами этих двух ансамблей. (**5 баллов**)
   Подсказка: корреляция между случайными величинами X и Y, считается по формуле:
   $$\text{Cor}\left(X, Y\right) = \frac{\text{Cov}\left(X, Y\right)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

5. Определите, к чему будет стремиться корреляция между прогнозами ансамблей из предыдущего пункта (с равным количеством деревьев) по мере стремления числа деревьев k к бесконечности. Сделайте вывод о том, насколько вероятно то, что при очень большом числе деревьев k эти ансамбли дадут существенно различающиеся прогнозы. Ответ подробно обоснуйте. (**5 баллов**)

Решение

1. Используя свойства интегралов получаем, что:

$$\text{AUC} = \text{pAUC}(0, 0.5) + \text{pAUC}(0.5, 1) = 0.5\text{pAUC}(0.5, 1) + \text{pAUC}(0.5, 1) = 0.5 \times 0.5 + 0.5 = 0.75$$

2. Обратим внимание, что:

$$\hat{\text{E}}\left(Y_i|X_i, T_i\right) = w_1^{(2)}\left(w_{11}^{(1)}X_i + w_{12}^{(1)}T_i\right) + w_2^{(2)}\left(w_{21}^{(1)}X_i + w_{22}^{(1)}T_i\right) =$$
$$0.5 \times (0.5 \times X_i + 0.5 \times T_i) + 0.5 \times (0.5 \times X_i + 0.5 \times T_i) = \frac{X_i + T_i}{2}$$

Отсюда получаем, что при $X_i = 2$ имеем:

$$\widehat{\text{CATE}} = \hat{\text{E}}\left(Y_i|X_i = 2, T_i = 1\right) - \hat{\text{E}}\left(Y_i|X_i = 2, T_i = 0\right) =$$
$$\frac{2+1}{2} - \frac{2+0}{2} = 0.5$$

3. Необходимо найти такое k, что:

$$\left(\rho\sigma^2 + \frac{(1-\rho)\sigma^2}{k}\right)/\sigma^2 = \rho + \frac{1-\rho}{k} = \frac{1}{1.25} = 0.8$$

Учитывая, что $\rho = 0.6$, решая соответствующее равенство, получаем $k = 2$.

4. Обозначим через $\hat{Y}_{1i}$ и $\hat{Y}_{2i}$ прогнозы первого ансамбля, а через $\hat{Z}_{1i}$ и $\hat{Z}_{2i}$ – прогнозы второго ансамбля. Сперва рассмотрим ковариацию:

$$\text{Cov}\left(\frac{\hat{Y}_{1i} + \hat{Y}_{2i}}{2}, \frac{\hat{Z}_{1i} + \hat{Z}_{2i}}{2}\right) = 0.25\text{Cov}\left(\hat{Y}_{1i} + \hat{Y}_{2i}, \hat{Z}_{1i} + \hat{Z}_{2i}\right) =$$
$$= 0.25\left(\text{Cov}\left(\hat{Y}_{1i}, \hat{Z}_{1i}\right) + \text{Cov}\left(\hat{Y}_{1i}, \hat{Z}_{2i}\right) + \text{Cov}\left(\hat{Y}_{2i}, \hat{Z}_{1i}\right) + \text{Cov}\left(\hat{Y}_{2i}, \hat{Z}_{2i}\right)\right) =$$
$$= 0.25\left(\rho\sigma^2 + \rho\sigma^2 + \rho\sigma^2 + \rho\sigma^2\right) = \rho\sigma^2 = 0.6\sigma^2$$

Отсюда получаем корреляцию:

$$\text{Cor}\left(\frac{\hat{Y}_{1i} + \hat{Y}_{2i}}{2}, \frac{\hat{Z}_{1i} + \hat{Z}_{2i}}{2}\right) = \frac{0.6\sigma^2}{\sqrt{0.6\sigma^2 + \frac{(1-0.6)\sigma^2}{2}}\sqrt{0.6\sigma^2 + \frac{(1-0.6)\sigma^2}{2}}} = \frac{0.6}{0.6 + \frac{0.4}{2}} = 0.75$$

5. Обратим внимание, что независимо от k ковариация останется прежней, поскольку:

$$\text{Cov}\left(\frac{\hat{Y}_{1i} + \ldots + \hat{Y}_{ki}}{k}, \frac{\hat{Z}_{1i} + \ldots + \hat{Z}_{ki}}{k}\right) =$$
$$= \frac{1}{k^2}\left(\underbrace{\text{Cov}\left(\hat{Y}_{1i}, \hat{Z}_{1i}\right) + \ldots + \text{Cov}\left(\hat{Y}_{ki}, \hat{Z}_{ki}\right)}_{k^2 \text{ возможных комбинаций}}\right) =$$
$$= \frac{1}{k^2}k^2\rho\sigma^2 = 0.6\sigma^2$$

Однако, корреляция будет стремиться к 1 по мере увеличения числа деревьев, поскольку:

$$\text{Cor}\left(\frac{\hat{Y}_{1i}+...+\hat{Y}_{ki}}{k},\frac{\hat{Z}_{1i}+...+\hat{Z}_{ki}}{k}\right)=\frac{0.6}{0.6+\frac{0.4}{k}}=\frac{3k}{3k+2}\xrightarrow[k\to\infty]{}1$$

Следовательно, при достаточно большом числе деревьев k, учитывая крайне высокую корреляцию прогнозов и их одинаковое математическое ожидание, оба ансамбля будут давать практически идентичные результаты. По аналогии нетрудно показать, что этот результат сохранится при любом ρ, а также при различном, но стремящемся к бесконечности числе решающих деревьев.

Задача №3

По выборке из независимых и одинаково распределенных наблюдений вы хотите оценить квадратичный средний эффект воздействия (square average treatment effect) и условный квадратичный средний эффект воздействия (термины были придуманы специально для этой задачи):

$$\text{SATE}=\text{E}\left((Y_{1i}-Y_{0i})^2\right)\qquad \text{CSATE}_i=\text{E}\left((Y_{1i}-Y_{0i})^2|X_i\right)$$

Предположим, что при фиксированных контрольных переменных X_i потенциальные исходы Y_{1i}, Y_{0i} и переменная воздействия T_i попарно независимы (все три случайные величины независимы между собой при условии контрольных переменных). Также допустим, что $0\notin\text{supp}(Y_{0i})$.

1. Предложите и опишите аналог S-learner для оценивания CSATE_i. Обоснуйте состоятелость оценки описанного вами метода. **(10 баллов)**

2. Опираясь на результаты предыдущего пункта и предпосылки, описанные в условии задачи, предложите и опишите аналог S-learner для оценивания SATE. **(10 баллов)**

3. Допустим, что предпосылка о независимости Y_{1i} и Y_{0i} при условии X_i не выполняется. Предложите менее сильную предпосылку, при которой предложенная вами оценка CSATE_i будет состоятельной (ответ подробно обоснуйте). **(10 баллов)**

Решение

1. Поскольку Y_{1i} и Y_{0i} независимы при фиксированном X_i, то:

$$\text{CSATE}_i=\text{E}\left((Y_{1i}-Y_{0i})^2|X_i\right)=\text{E}\left(Y_{1i}^2|X_i\right)+\text{E}\left(Y_{0i}^2|X_i\right)-2\text{E}\left(Y_{1i}|X_i\right)\text{E}\left(Y_{0i}|X_i\right)$$

В силу того, что $T_i^2=T_i$ и при фиксированных X_i потенциальные исходы Y_{1i} и Y_{0i} не зависят от T_i, получаем:

$$\begin{gathered}\text{E}(Y_{1i}|X_i)=\text{E}(Y_{1i}|X_i,T_i=1)=\text{E}(T_iY_{1i}+(1-T_i)Y_{0i}|X_i,T_i=1)=\text{E}(Y_i|X_i,T_i=1)\\ \text{E}(Y_{1i}^2|X_i)=\text{E}(Y_{1i}^2|X_i,T_i=1)=\text{E}((T_iY_{1i}+(1-T_i)Y_{0i})^2|X_i,T_i=1)=\\ =\text{E}(T_iY_{1i}^2+(1-T_i)Y_{0i}^2+2\underbrace{T_i(1-T_i)}_{0}Y_{1i}Y_{0i}|X_i,T_i=1)=\text{E}(Y_i^2|X_i,T_i=1)\end{gathered}$$

По аналогии можно показать, что:

$$\begin{gathered}\text{E}(Y_{0i}|X_i)=\text{E}(Y_i|X_i,T_i=0)\\ \text{E}(Y_{0i}^2|X_i)=\text{E}(Y_i^2|X_i,T_i=0)\end{gathered}$$

Отсюда следует, что:

$$\text{CSATE}_i = \text{E}\left(Y_i^2|X_i, T_i = 1\right) + \text{E}\left(Y_i^2|X_i, T_i = 0\right) - 2\text{E}\left(Y_i|X_i, T_i = 1\right)\text{E}\left(Y_i|X_i, T_i = 0\right)$$

Полученное выражение мотивирует следующую процедуру оценивания. Сперва методами машинного обучения по всей выборке оцениваются условные математические ожидания $\text{E}(Y_i|X_i, T_i)$ и $\text{E}(Y_i^2|X_i, T_i)$. Затем оценки этих условных математических ожиданий используются для оценивания условного квадратичного среднего эффекта воздействия:

$$\widehat{\text{CSATE}}_i = \hat{\text{E}}\left(Y_i^2|X_i, T_i = 1\right) + \hat{\text{E}}\left(Y_i^2|X_i, T_i = 0\right) - 2\hat{\text{E}}\left(Y_i|X_i, T_i = 1\right)\hat{\text{E}}\left(Y_i|X_i, T_i = 0\right)$$

2. Состоятельную оценку $\widehat{\text{SATE}}$ можно получить, усреднив состоятельные оценки $\widehat{\text{CSATE}}_i$:

$$\widehat{\text{SATE}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\text{SATE}}_i \xrightarrow{p} \text{SATE}$$

3. Обратим внимание, что:

$$\text{Cov}\left(Y_{1i}, Y_{0i}|X_i\right) = \text{E}\left(Y_{1i}Y_{0i}|X_i\right) - \text{E}\left(Y_{1i}|X_i\right)\text{E}\left(Y_{0i}|X_i\right)$$

Следовательно, достаточно предположить, что Y_{1i} и Y_{0i} не коррелированны при условии X_i.