Фамилия:.....

Имя:.....

Группа:.....

Задача №1

У вас имеются следующие данные:

Наблюдение i	1	2	3	4	5	6
Y_i	9	1	6	7	7	6
X_{1i}	0	0.125	-0.375	0	0.175	-0.125
X_{2i}	0.25	-0.75	0.5	-0.25	0	0
T_i	1	0	0	1	1	0
Разбиение выборки	I			II		

1. Для прогнозирования целевой переменной Y_i с помощью признаков X_{1i} и X_{2i} вы используете метод ближайших соседей с метрикой расстояния Манхэттэн. Используя двухчастную кросс-валидацию (разбиение указано в таблице), ориентируясь на значение MAE, сделайте выбор между 1 и 2 соседями.

2. Считайте, что признак X_{2i} более не доступен в данных и в качестве контрольной переменной вы рассматриваете лишь X_{1i}. Используя метод ближайших соседей с оптимальным числом соседей (выбранным исходя из результатов предыдущего пункта) оцените с помощью T-learner условный средний эффект воздействия переменной воздействия T_i на целевую переменную Y_i для 2-го наблюдения в выборке.
   Подсказка: При внутривыборочном прогнозировании методом ближайших соседей само наблюдение рассматривается в качестве ближайшего соседа самого себя.

Решение:

1. Определим расстояние между каждым наблюдением в первой и второй частях выборки по формуле:

$$D_{ij} = |X_{1i} - X_{1j}| + |X_{2i} - X_{2j}|.$$

Например, расстояние между первым и четвертым наблюдением будет равняться:

$$D_{14} = |0 - 0| + |0.25 - (-0.25)\,| = 0.5.$$

Представим результаты вычислений в форме таблицы, в которой в ячейках представлены значения соответствующих расстояний:

	Y_4	Y_5	Y_6
Y_1	0.5	0.425	0.375
Y_2	0.625	0.8	1
Y_3	1.125	1.05	0.75

Для первого наблюдения ближайшими соседями являются пятое и шестое наблюдения. Таким образом, при использовании метода с двумя ближайшими соседями, например, для первого наблюдения имеем:

$$\hat{Y}_1^2 = \frac{\hat{Y}_5 + \hat{Y}_6}{2} = \frac{7 + 6}{2} = 6.5$$

При одном ближайшем соседе для этого же наблюдения получаем:

$$\hat{Y}_1^1 = \hat{Y}_6 = 6.$$

По аналогии осуществляем прогнозы для остальных наблюдений по валидационным выборкам, а также считаем разницы между истинными и спрогнозированными значениям. Через Δ_i^1 и Δ_i^2 обозначим соответствующие разницы i-го наблюдения, посчитанные методами с 1 и 2 соседями соответственно. Например, для первого наблюдения и метода с двумя соседями получаем:

$$\Delta_1^2 = |Y_i - \hat{Y}_i^2| = |9 - 6.5| = 2.5$$

Агрегируем результаты в таблицу:

	Y_i	$\hat{Y}_i^2$	$\hat{Y}_i^1$	Δ_i^2	Δ_i^1
Y_1	9	6.5	6	2.5	3
Y_2	1	7	7	6	6
Y_3	6	6.5	6	0.5	0
Y_4	7	5	9	2	2
Y_5	7	5	9	2	2
Y_6	6	7.5	9	1.5	3

Таким образом, валидационное значение метрики по модели с двумя и одним ближайшими соседями равняются соответственно:

$$\text{MAE}_2 = \frac{1}{2}\left(\frac{\Delta_1^2 + \Delta_2^2 + \Delta_3^2}{3} + \frac{\Delta_4^2 + \Delta_5^2 + \Delta_6^2}{3}\right) = \frac{1}{2}\left(\frac{2.5 + 6 + 0.5}{3} + \frac{2 + 2 + 1.5}{3}\right) \approx 2.42;$$

$$\text{MAE}_1 = \frac{1}{2}\left(\frac{\Delta_1^1 + \Delta_2^1 + \Delta_3^1}{3} + \frac{\Delta_4^1 + \Delta_5^1 + \Delta_6^1}{3}\right) = \frac{1}{2}\left(\frac{3 + 6 + 0}{3} + \frac{2 + 2 + 3}{3}\right) \approx 2.67.$$

Поскольку $\text{MAE}_2 < \text{MAE}_1$, то метод ближайших соседей с двумя соседями является более предпочтительным согласно данному критерию.

2. Оценим условные математические ожидания $\hat{E}\left(Y_2 | X_{12}, T = 1\right)$ и $\hat{E}\left(Y_2 | X_{12}, T = 0\right)$ по группе воздействия и контрольной группе по отдельности. Так, среди получшивших воздействие двумя ближайшими соседями для второго наблюдения будут являться первое, четвертое и пятое (поскольку в условии не указана конкретная процедура выбора между равноудаленными наблюдениями, на свое усмотрение возьмем из них четвертое и пятое), а среди не получивших – второе и шестое. Таким образом получаем:

$$\hat{E}\left(Y_2 | X_{12}, T = 1\right) = \frac{7 + 7}{2} = 7,$$

$$\hat{E}\left(Y_2 | X_{12}, T = 0\right) = \frac{1 + 6}{2} = 3.5.$$

Тогда оценка условного среднего эффекта воздействия будет равняться:

$$\widehat{\text{CATE}}_2 = 7 - 3.5 = 3.5$$

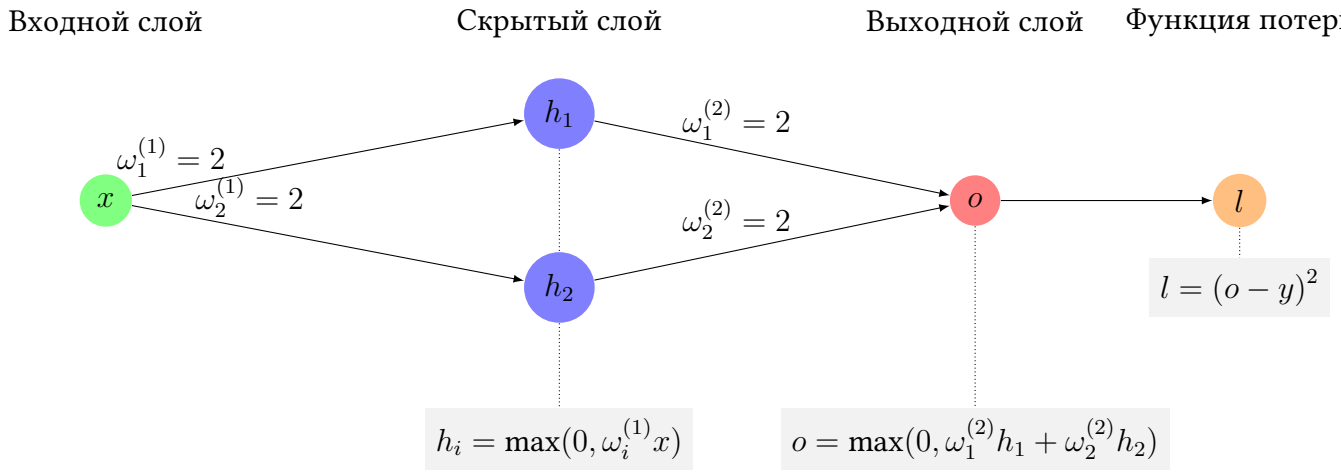
Задача №2

Имеется нейросеть, включающая всего одно наблюдение по одному признаку $x = 1$. Значение целевой переменной равняется $y = 10$. Имеется лишь один скрытый слой с двумя нейронами. В качестве функции активации в скрытом и выходном слоях используется ReLU. Применяется квадратичная функция потерь. В нейросети нет смещений (констант) и все ее параметры равняются 2.

1. Изобразите графически описанную нейросеть.
2. Рассчитайте значение функции потерь данной нейросети при заданных значениях весов.
3. Рассмотрим вес, с которым признак входного слоя входит в первый нейрон. Найдите значение данного веса после одной итерации алгоритма градиентного спуска со скоростью обучения $\alpha = 0.125$.
4. Вы добавили в скрытый слой исключение (dropout). Нейроны отключаются независимо друг от друга с вероятностью 0.5. Повторите предыдущий пункт, определяя математическое ожидание обновленного веса, с которым признак входит в первый нейрон.

Решение:

1. Описанную нейросеть можно представить в виде следующей иллюстрации.



2. Последовательно рассчитаем значения в различных нейронах:

$$h_1 = \max(0, 2 \times 1) = 2 \qquad h_2 = \max(0, 2 \times 1) = 2$$
$$o = \max(0, 2 \times 2 + 2 \times 2) = 8 \qquad l = (8 - 10)^2 = 4$$

3. Используя метод обратного распространения ошибки получаем значение производной:

$$\frac{\partial l}{\partial \omega_1^{(1)}} = 2 \times (o - y) \times \omega_1^{(2)} \times x = 2 \times (8 - 10) \times 2 \times 1 = -8$$

Применяя градиентный спуск с заданной скоростью обучения рассчитаем обновленный вес:

$$\tilde{\omega}_1^{(1)} = \omega_1^{(1)} - \alpha \frac{\partial l}{\partial \omega_1^{(1)}} = 2 - 0.125 \times (-8) = 3$$

4. Обратим внимание, что при исключении нейрона h_1 производная l по $\omega_1^{(1)}$ обнуляется. Также, если h_2 обнуляется, а h_1 нет, то мы получаем:

$$o = \max(0, 2 \times 2 + 0) = 4 \qquad l = (4 - 10)^2 = 36$$
$$\frac{\partial l}{\partial \omega_1^{(1)}} = 2 \times (4 - 10) \times 2 \times 1 = -24$$

Рассматривая данную производную как случайную величину, получаем ее распределение:

$$P\left(\frac{\partial l}{\partial \omega_1^{(1)}} = 0\right) = P(h_1 = 0) = 0.5$$
$$P\left(\frac{\partial l}{\partial \omega_1^{(1)}} = -8\right) = P(h_1 \neq 0, h_2 \neq 0) = 0.25$$
$$P\left(\frac{\partial l}{\partial \omega_1^{(1)}} = -24\right) = P(h_1 \neq 0, h_2 = 0) = 0.25$$

Следовательно, применяя формулу полного математического ожидания получаем:

$$\mathrm{E}(\tilde{\omega}_1^{(1)}) = 2 - (0.5 \times 0.125 \times 0 + 0.25 \times 0.125 \times (-8) + 0.25 \times 0.125 \times (-24)) = 3.$$

Задача №3

У вас имеются следующие данные:

Наблюдение i	1	2	3	4
Y_i	0	12	24	36
X_{1i}	1	1	0	1
X_{2i}	1	1	1	0

Дана квадратичная функция потерь $L\left(Y_i, \hat{Y}_i\right) = \left(Y_i - \hat{Y}_i\right)^2$. Для прогнозирования вы используете градиентный бустинг со скоростью обучения 0.5 и одной итерацией. В качестве базовой модели используется метод наименьших квадратов лишь с константой. Для прогнозирования градиентов используется регрессионное дерево глубины 1, в качестве критерия разбиения использующего средневзвешенную дисперсию (среднеквадратическая ошибка).

1. Запишите прогнозы базовой модели.
2. С помощью одной итерации градиентного бустинга спрогнозируйте значения целевой переменной для всех наблюдений в выборке.
3. Вы решили рассмотреть скорость обучения $\alpha \in R$ обученного в предыдущем пункте градиентного бустинга в качестве гиперпараметра. Подберите оптимальное значение скорости обучения (тюнинг), ориентируясь на критерий MSE, рассчитанный на следующей валидационной выборке:

Наблюдение i	1	2
Y_i	18	58
X_{1i}	1	1
X_{2i}	1	0

Решение:

1. Метод наименьших квадратов, в котором в качестве единственного оцениваемого параметра выступает константа, всегда прогнозирует выборочное среднее:

$$F_0(x) = \overline{Y} = \frac{0 + 12 + 24 + 36}{4} = 18, \text{ где } x \in R^2$$

2. Для каткости введем обозначение $X_i = (X_{1i}, X_{2i})$.

   На протяжении решения задачи для удобства рекомендуется заполнять следующую таблицу:

Наблюдение i	1	2	3	4
Y_i	0	12	24	36
X_{1i}	1	1	0	1
X_{2i}	1	1	1	0
r_{1i}	-36	-12	12	36
$h_1(X_i)$	-12	-12	-12	36
$\hat{Y}_i = F_1(X_i)$	12	12	12	36

   Разберемся с тем, как были получены значения в данной таблице, **последняя строка** которой содержит искомые прогнозы.

   Производная функции потерь по прогнозам равняется $(\nabla L)_i = -2\left(Y_i - F_0(X_i)\right)$. В результате получаем остатки:

$$r_{11} = -(\nabla L)_1 = 2 \times (0 - 18) = -36 \qquad r_{12} = -(\nabla L)_2 = 2 \times (12 - 18) = -12$$
$$r_{13} = -(\nabla L)_3 = 2 \times (24 - 18) = -12 \qquad r_{14} = -(\nabla L)_4 = 2 \times (36 - 18) = 36$$

   Далее необходимо получить прогнозы $(\nabla L)_i$ с помощью дерева глубины 1. Для этого необходимо определить оптимальное разбиение.

   Если разбиение осуществляется по признаку X_{1i}, то средневзвешенная дисперсия составит:

$$\overline{r}_{11} = \frac{-36 - 12 + 36}{3} = -4 \quad \overline{r}_{10} = 12$$
$$\widehat{\text{Var}}(r_i, X_{1i}) = \frac{(36 - (-4))^2 + (12 - (-4))^2 + (-36 - (-4))^2}{3} + \frac{(12 - 12)^2}{1} = 896$$

   По аналогии при разбиении по X_{2i} получаем:

$$\overline{r}_{21} = \frac{-36 - 12 + 12}{3} = -12 \quad \overline{r}_{20} = 36$$
$$\widehat{\text{Var}}(r_i, X_{2i}) = \frac{(-36 - (-12))^2 + (-12 - (-12))^2 + (12 - (-12))^2}{3} + \frac{(36 - 36)^2}{1} = 384$$

   Поскольку $\widehat{\text{Var}}(r_1, X_{2i}) < \widehat{\text{Var}}(r_1, X_{1i})$, то разбиение осуществляется по X_{2i}, а значит получаем следующую модель прогнозирования градиента:

$$h_1((x_1, x_2)) = \begin{cases} \frac{-36-12+12}{3}, \text{ если } x_2 = 1 \\ \frac{36}{1}, \text{ если } x_2 = 0 \end{cases} = \begin{cases} -12, \text{ если } x_2 = 1 \\ 36, \text{ если } x_2 = 0 \end{cases}$$

В результате получаем прогнозы:

$$F_1((1,1)) = F_0((1,1)) + 0.5 \times h_1((1,1)) = 18 + 0.5 \times (-12) = 12$$
$$F_1((0,1)) = F_0((0,1)) + 0.5 \times h_1((0,1)) = 18 + 0.5 \times (-12) = 12$$
$$F_1((1,0)) = F_0((1,0)) + 0.5 \times h_1((1,0)) = 18 + 0.5 \times 36 = 36$$

3. Запишем выражение среднеквадратичной ошибки на валидационной выборке:

$$\text{MSE} = \frac{(18 - F_1((1,1)))^2 + (58 - F_1((1,0)))^2}{2} =$$
$$= \frac{(18 - (18 + \alpha \times (-12)))^2 + (58 - (18 + \alpha \times 36))^2}{2} = 720\alpha^2 - 1440\alpha + 800$$

Минимизируя данное выражение по α получаем оптимальное значение $\alpha^* = 1$.