*Теоретическое домашнее задание не сдается и не проверяется. Вместо проверки **в начале** каждого семинара будет устраиваться самостоятельная работа, которая будет включать в себя набор задач из домашнего задания. Время, отведенное на выполнение самостоятельной работы, будет невелико (около 10 мин). Использование чего-либо, кроме ручки и выданных листов бумаги, запрещено. В случае обнаружения списывания, подлога и т.д. выставляется оценка "0".*

Теоретическое домашнее задание №2

Лед тронулся

Задача 1.
Найдите константу C, решающую следующую задачу для $\tau = 0.6$ и $y_i = \{10, 5, 3, 8, 6\}$:

$$\sum_{i=1}^{l} \rho_\tau(y_i - C) \to \min_C,$$

$$\rho_\tau(x) = \begin{cases} \tau x, & x > 0 \\ (\tau - 1)x, & x \leq 0 \end{cases}$$

Задача 2. Мальчик Петя мечтает стать дата сайентистом, когда вырастет. Поэтому пока, он ещё не вырос, он тренируется строить модели для прогноза дождя, используя "бабушкино обучение". В течение года он с бабушкой много раз выходил на прогулку и записывал в свой дневник наблюдений следующие факты:

- Зима на улице, или нет.
- Есть ли на небе облака или нет.
- Взяла ли бабушка зонт или нет.
- Идёт ли дождь или нет.

В итоге, в **таблицу 1 (а)** он записал все возможные случаи, которые происходили на его прогулках (неважно, сколько раз). Бабушкино обучение состоит в следующем. Сначала она ищет и выписывает неопровергаемые отношения типа $A \implies Y$, где Y соответствует утверждению о состоянии целевой переменной, а A - утверждению о состоянии **одного** признака. Далее, когда бабушка посмотрит на все такие отношения, она перейдёт к рассмотрению влияния состояний одновременно любых двух признаков на Y. Например: $((B \cap C \implies Y)$. Однако, если на предыдущем шаге уже было построено отношение, для $(B \implies Y)$ или $(C \implies Y)$, то отношение от двух признаков не строится за ненадобностью. Если случай не подпадает под построенное отношение, то для него не производится прогноз. Например, когда бабушке говорят, что в течение 3-х прогулок дождь шёл всегда, когда она брала зонт, то она формулирует отношнение: "Зонт $\implies$ Дождь" и строит модель:

Таблица 1: Global caption

(a) Петин дневник

	Зима	Облака	Зонт	Дождь
1	0	0	0	0
2	0	1	1	1
3	0	1	1	0
4	0	1	0	1
5	0	0	1	1
6	0	1	0	0
7	0	0	1	0
8	1	0	0	0

(b) Пример

	A	B	Y
1	0	1	1
2	1	1	0
3	0	0	1
4	1	0	0
5	1	0	1

- Если зонт взят, то предсказать дождь.
- Если зонт не взят, то дождь может пойти или не пойти.

После этого можно проверить метрику качества модели на некой выборке:

$$Q = \frac{(\text{число верных прогнозов}) - (\text{число ошибочных прогнозов})}{\text{число объектов в выборке}},$$

Приведем также более формальный пример. Смотря на **таблицу 1 (b)**, бабушка найдёт следующие отношения:

- not A $\implies Y$
- $A \cap B \implies$ not Y

И метрика качества на всей таблице 1 (b) будет равна: $Q = \frac{3-0}{5} = 0.6$

1. Пусть бабушке Петя для прогноза дождя дал только значение признака ”Зима” для всех наблюдений. Какие отношения между целевой переменной и обучающей выборкой найдёт бабушка? Каково значение метрики?
2. После Петя попросил бабушку построить модель по признакам ”Облака” и ”Зонт”. Какие на сей раз будут выявлены отношения и метрика?
3. Теперь же пусть признаки ”Облака” и ”Зонт” даны бабушке только для первых 5 наблюдений. Как изменится набор выявленных отношений и метрика по сравнению с предыдущим пунктом? Почему появились новые отношения? Какова метрика на первых 5 наблюдениях?
4. Какова будет метрика качества модели из предыдущего пункта на тех данных, которые не использовались для её построения? Что тянет метрику вниз?

5. Является ли наблюдение за зонтом бабушки полезным занятием? Если бы бабушка строила модель для первых 5 наблюдений только по облакам, модель была бы качественнее? Сравните метрику такой модели для обучающей выборки и всей таблицы Пети.

Убедитесь, что Вы знаете ответы на следующие вопросы:

- Почему L_1-регуляризация производит отбор признаков?
- Почему коэффициент регуляризации нельзя подбирать по обучающей выборке?
- Что такое кросс-валидация, чем она лучше использования отложенной выборки?
- Как функция ShuffleSplit(n_splits=5, test_size=0.5) разбивать выборку?