# ШАД. Машинное обучение, часть 1.

## Лабораторная работа 5. Часть 1.

1. В лабораторной работе 4 вы подробно познакомились с линейными моделями, выяснили о необходимости обработки непрерывных и категориальных признаков, узнали о способах подбора гиперпараметров. Вам предлагается ответить на следующие вопросы:

   (a) Какие побочные эффекты могут возникунуть при стандартизации (нормализации) признаков с помощью `StandardScaler`, `MinMaxScaler`? Что с этим можно сделать?

   (b) Рассмотрим пример с первого занятия про модель потребления мороженного от температуры:

   $$ic = \theta_1 + \theta_2 t$$

   Предположим, что нам также известен еще один признак, отвечающий за год. Обозначим его за $y$. Пусть $y \in \{1, 2, 3\}$. Попробуем учесть влияние года двумя разными способами:

   - Модель $ic = \theta_1 + \theta_2\ t + \theta_3 y_1 + \theta_4 y_2$, где $y_1 = I\{y = 1\}, y_2 = I\{y = 2\}$.
   - Для каждого года рассматривается своя линейная зависимость $ic = \theta_1 + \theta_2\ t$.

   Объясните, в чем различие этих двух подходов?

2. Визуализируйте совместные распределения вещественных признаков и целевой переменной для данных из лабораторной работы. Что можно сказать о зависимости таргета от признаков? Сделайте вывод о том, насколько хорошо построенные модели приближают истинные зависимости.
   *Полученные графики приложите к решению теоретического задания.*

3. Выпишите формулы GD и SGD для регрессии Хьюбера. Задачу оптимизации для этой модели можно записать следующим образом:

   $$\sum_{i=1}^{n} R(Y_i - x_i^T \theta) \to \min_{\theta}$$

   $R(x) = \dfrac{x^2}{2} I\{|x| < c\} + c\left(|x| - \dfrac{c}{2}\right) I\{|x| > c\}$ – функция потерь Хьюбера.

   В чем польза выбора такой функции потерь?

4. Пусть $\widehat{\theta}$ – оценка коэффициентов линейной модели в методе ридж-регрессии. Посчитайте $MSE_{\widehat{\theta}}(\theta) = \mathsf{E}_\theta \left(\widehat{\theta} - \theta\right)^T \left(\widehat{\theta} - \theta\right)$.