**Непараметрическая регрессия**

# **Цели работы:**

1. практика первичных навыков обработки данных: нормализация, Ohe-Hot преобразование;
2. сведение задачи классификации к задаче непараметрической регрессии;
3. реализация решения задачи непараметрической регрессии ядерным сглаживанием Надарая Ватсона;
4. практика наивного способа настройки и анализа гиперпараметров модели, решающей задачу непараметрической регрессии.

# **Набор данных**

Выберите любой понравившийся набор данных для задачи классификации из следующего списка:

1. car <https://www.openml.org/d/40975>
2. vehicle <https://www.openml.org/d/54>
3. wine <https://www.openml.org/d/187>
4. bridges <https://www.openml.org/d/327>

# **Сведение к задаче регрессии и обработка данных**

Перейдите от задачи классификации к задаче регрессии, используя [OneHot преобразование](https://ru.wikipedia.org/wiki/%D0%A3%D0%BD%D0%B8%D1%82%D0%B0%D1%80%D0%BD%D1%8B%D0%B9_%D0%BA%D0%BE%D0%B4). Вместо одного целевого признака в выбранный набор данных добавляется столько новых числовых переменных, сколько в нём содержится классов. Помимо этого, если выбранный Вами набор данных содержит нечисловые признаки, эти признаки необходимо векторизовать (перейти от категорий к числам), заполнить пропуски (если есть) и нормализовать. В наборе данных bridges также необходимо избавиться от столбца IDENTIF, поскольку он является идентификатором записи.

# **Реализация алгоритма и настройка гиперпараметров, анализ результатов**

Реализуйте алгоритм решения задачи непареметрической регрессии при помощи ядерного сглаживания Надарая-Ватсона.

Найдите лучшую комбинацию гиперпараметров алгоритма непараметрической регрессии:

* функция расстояния:
  + расстояние Евклида,
  + расстояние Манхэттена,
  + расстояние Чебышева;
* функция ядра
  + uniform:



* + triangular:



* + epanechnikov:



* + quartic:
* тип окна (окно, зависящее от количества соседей и фиксированное)
* параметр окна:
  + количество ближайших соседей от 1 до размер набора данных, является эвристикой на число ближайших соседей для метрических алгоритмов **ИЛИ**
  + размер окна, его необходимо выбирать исходя из “размеров” набора данных; хорошей практикой является настройка ширины окна на отрезке до с шагомгде — самое большого расстояния между элементами в наборе данных.

Таким образом требуется перебрать комбинаций гиперпараметров и найти лучшую.

Используйте Leave-One-Out перекрёстную проверку для настройки алгоритма.

Критерием качества является F-мера. Для её подсчёта этого потребуется определить максимальную компоненту результирующего вектора целевых признаков, полученных из One-Hot преобразования, после применения очередной конфигурации алгоритма непараметрической регрессии (алгоритм с одной из комбинаций гиперпараметров).

Для лучшей найденной комбинации гиперпараметров постройте графики зависимости F-меры от числа ближайших соседей **или** ширины окна (при фиксированных лучших значениях прочих гиперпараметров).

**Схема работы**

