

# Машинное обучение.

## Домашнее задание №1

**Задача 1.** Ответьте на вопросы:

1. Как формально записывается алгоритм  $k$  ближайших соседей для классификации? Для регрессии?
2. Как в kNN можно добавить поддержку весов? Какие способы задания весов вы знаете?
3. Какие способы поиска ближайших соседей вы знаете? Какая у них сложность?
4. Как записывается метрика Минковского? На что влияет ее параметр  $p$ ?
5. Как записывается косинусное расстояние? Расстояние Джаккарда?
6. Как выглядит семейство хэш-функций для расстояния Джаккарда? Косинусного расстояния? Евклидова расстояния?
7. Как определяется отступ объекта? Как отступ можно использовать для отбора объектов?

**Задача 2.** Покажите, что расстояние Джаккарда является метрикой.

Подсказка: может пригодиться доказанное на занятии утверждение о том, что коэффициент Джаккарда равен кое-какой вероятности.

**Задача 3.** Пусть зависимость вероятности коллизии от расстояния между объектами имеет вид  $p = 1 - d$  (для исходного семейства хэш-функций). После построения композиции над этим семейством вероятность примет вид  $1 - (1 - (1 - d)^m)^L$ . Данное выражение как функция от  $d$  имеет вид сигмоиды. Найдите координату точки ее перегиба, и исследуйте ее поведение при стремлении  $L$  к бесконечности при фиксированном  $m$  (и наоборот, при стремлении  $m$  к бесконечности при фиксированном  $L$ ). Проинтерпретируйте результаты.

**Задача 4.** Рассмотрим универсальное множество слов

$$U = \{\text{зависимость, вероятность, коллизия, испанская, инквизиция, расстояние}\}.$$

Вычислите MinHash для множества

$$A = \{\text{зависимость, вероятность, инквизиция}\}$$

для перестановок  $\pi_1, \pi_2$ :

---

|         | зависимость | вероятность | коллизия | испанская | инквизиция | расстояние |
|---------|-------------|-------------|----------|-----------|------------|------------|
| $\pi_1$ | 4           | 2           | 6        | 1         | 3          | 5          |
| $\pi_2$ | 3           | 6           | 2        | 5         | 6          | 1          |