Метод ближайших соседей

<https://classroom.github.com/a/JsMt3CJl>

# Алгоритм

Реализуйте метод ближайших соседей:

* Алгоритм должен работать с окнами фиксированного и нефиксированного размера.
* Алгоритм должен работать с различными ядрами. Не менее 4 штук. Обязательно должно быть Равномерное и Гауссово ядро. Желательно, чтобы было Гауссово и общее ядро вида (1 - |*u*|*ᵃ*)*ᵇ*.
* Алгоритм должен работать с различными метриками. Не менее 3 штук. Обязательно должно быть Косинусное расстояние. Желательно, чтобы было Косинусное расстояние и расстояние Минковского Lp.
* Алгоритм должен работать с априорными весами.
* Разрешается использовать готовую реализацию алгоритма поиска ближайших объектов.

# Набор данных

* Выберите любой набор данных для задачи классификации. Желательно использовать с предыдущей лабораторной работы.
* Преобразуйте его в числовой вид и нормализуйте.
* Разбейте его на тренировочную и тестовую часть.
* Выберите целевую функцию ошибки или качества.
* Если в наборе данных очень много объектов, можно выбрать случайное подмножество.
* Если в наборе данных очень много признаков, можно выбрать случайное подмножество, умножить на случайную матрицу, использовать выбор или извлечение признаков (если знаете что это и как).

# Гиперпараметры

Найдите лучшие гиперпараметры:

* Переберите возможные комбинации гиперпараметров: расстояний, ядер, окон, соседей или радиусов.
* Тестовое множество не должно использоваться для валидации при поиске.
* Выведите лучшие значения гиперпараметров.
* Постройте график зависимости целевой функции ошибки/качества на тестовом и тренировочном множестве в зависимости от числа соседей или ширины окна (смотря какая функция оказалась лучше). Остальные гиперпараметры должны быть зафиксированы.

# Поиск аномалий

* Реализуйте алгоритм поиска аномалий LOWESS.
* Взвесьте объекты из тренировочного множества.
* Вычислите результат валидации реализованного алгоритма на тестовом множестве до и после взвешивания.

# Примеры вопросов на теормин

1. Чем обучение с учителем отличается от обучения без учителя. Какое отношение оно имеет к кластеризации, классификации, регрессии, ранжирования, генерации
2. Что такое мягкая классификация
3. Что такое бейзлайн? Что такое наивный алгоритм? Какие примеры наивных алгоритмов можно привести для задачи классификации, регрессии, генерации, рекомендации
4. Линейная и нелинейная зависимость. Как выглядит мультимодальное распределение, чем оно неудобно для ML-алгоритмов. Придумайте пример признака, имеющего мультимодальное распределение (например из своего датасета из 1 лабы)
5. Как и зачем дискретизировать, бинаризовать, нормализовывать и взвешивать признаки?
6. Как и зачем делать one-hot encoding? Можно ли это делать функцией pandas.get\_dummies? Чем one-hot отличается от binary-encoding?
7. Как вычисляется cross-entropy и почему она cross?
8. Зачем делить данные на train, test, val? Чем val отличается от test? Кросс-валидация, ее виды. Что такое data leak
9. Модель kNN, **ядра kNN**