**Домашнее задание по машинному обучению №8**

Дедлайн на полный балл – **4 мая**

Дедлайн на половинный балл – **11 мая**

Датасеты (разделить на **train (80%)** и **val (20%)** с сохранением пропорций классов):

**spam.csv, smsspam.csv, noisysine.csv, hydrodynamics.csv**

Наивный Байес. Во всех задачах помимо значения точности результатом должна быть также ROC-кривая на val датасете, построенная по полученной нормированной вероятности.

1. **(6)** Реализуйте алгоритм Gaussian naive Bayes для датасета **spam.csv**.

2. **(6)** Реализуйте алгоритм Multinomial naive Bayes для количества различных слов в датасете **smsspam.csv**. Т.е., если у вас в словаре N слов, то у каждой точки будет N признаков. Stemming крайне рекомендуем.

Регрессия. Во всех задачах нужно будет на val датасете визуализировать получившуюся функцию на датасете **noisysine.csv** и посчитать R-score для датасетов **noisysine.csv hydrodynamics.csv**.

3. **(3)** Реализуйте алгоритм линейной регрессии, и полиномиальной регрессии (для датасета **noisysine** – степеней от 2 до 5, для датасета **hydrodynamics** – степени 2) без регуляризации.

4. **(2)** Реализуйте алгоритм гребневой регрессии и найдите оптимальный параметр регуляризации для случаев из задачи 3.

5. **(3) (Не рассказывался полностью, но это не мешает его уже использовать:))** Используйте алгоритм (например, из библиотеки sklearn) LASSO регрессии и оптимизируйте параметры регуляризации для случаев из задачи 3. Сколько признаков используется в лучших конфигурациях?