Домашнее задание по машинному обучению №4

Дедлайн на полный балл – **23 марта**

Дедлайн на половинный балл – **30 марта**

Датасеты – spam.csv и cancer.csv

1. (5) Реализуйте алгоритм Random Forest с критерием Джини с выбором класса по сумме (или среднему) вероятностей в листьях. Постройте ROC кривую (и выведите AUC и accuracy) для полученной вероятности для 20 деревьев глубины 3.

2. (4) Оптимизируйте по AUC на кроссвалидации (размер валидационной выборки - 20%) параметры Random Forest: максимальную глубину деревьев из [2, 3, 5, 7, 10], количество деревьев из [5, 10, 20, 30, 50, 100]. Постройте ROC кривую (и выведите AUC и accuracy) для лучшего варианта.

Подсказка: можно построить сразу 100 деревьев глубины 10, а потом убирать деревья и глубину.

Дополнительные задачи:

3. (0) Реализуйте алгоритм адаптивного бустинга (AdaBoost) для деревьев с критерием Джини.