Домашнее задание по машинному обучению №3

Стоимость задач указана в скобках

Дополнительных задач в этом домашнем задании нет, но дедлайн продлён на неделю.

Дедлайн на полный балл – **5 апреля**

Дедлайн на половинный балл – **12 апреля**

Датасеты – **spamdataset_old.csv + spamdataset_new.csv**

Для заданий 1 и 2 используем датасет **spamdataset_old.**

1. **(2)** Постройте Precision-recall и ROC кривые для правила типа threshold ($x_i \geq a$) по трем последним признакам (capital_run_length_average, capital_run_length_longest, capital_run_length_total).

2. **(2)** Посчитайте AUC для правил типа threshold для всех признаков и найдите 10 лучших.

   Для заданий 3-7 мы предполагаем, что для построения классификатора у нас есть данные **spamdataset_old (обучающая выборка)**, а **spamdataset_new** – это новые письма, которые получает наш классификатор спама, и на нем мы будем проверять его работу **(тестовая выборка)**.

3. **(5)** Реализуйте алгоритмы построения дерева с критерием информационного выигрыша и критерием Джини и определению класса по мажоритарному классу в листе. Найдите оптимальную глубину дерева в обоих случаях (в отрезке 2-10).

4. **(1)** В задаче 3 используйте только 10 лучших признаков из задачи 2.

5. **(5)** Реализуйте алгоритм Random Forest для любого типа критерия (информационный выигрыш или Джини) с выбором класса по сумме вероятностей. Постройте Precision-recall и ROC кривую для полученной вероятности на тестовой выборке.

6. **(3)** Оптимизируйте по AUC на тестовой выборке параметры Random Forest: максимальную глубину деревьев (в отрезке 2-10), количество деревьев (5, 10, 20, 30, 50, 100, 200, 300). В этой задаче разумно зафиксировать Random Seed. Постройте Precision-recall и ROC кривую для лучшего варианта.