Домашнее задание 3
Машинное обучение, БИБ20

1 Задание 1: EM-алгоритм

- Открыть датасет **sklearn.datasets.load_wine**, содержащий информацию о трех различных сортах вина (class0, class1, class2). Ответить на вопросы ниже, используя средства языка Python и необходимых библиотек;
- Используя файл **Sem5_EM.ipynb**, модифицировать алгоритм EM так, чтобы он умел распознавать три класса (в исходной реализации умеем делать только бинарную классификацию);
- Вместо переменной *steps* (указывает количество итераций алгоритма) и цикла по количеству шагов, сделать функцию **FullEM()**, которая будет выполнять функции **e_step()** и **m_step()**, пока не будет соблюдено условие сходимости (которое Вы выбираете сами). Таким образом, алгоритму не надо делать ровно 15 итераций, а их количество динамически зависит от условия сходимомсти;
- Не используя деление выборки на train-test (так как обучение без учителя), прогнать модифицированный EM-алгоритм (функцию **FullEM()**) и посчитать известные метрики точности классификации (спойлер: *не только* accuracy).
- Использовать GaussianMixture из sklearn, также посчитать метрики. Насколько точна классификация в данном случае? Какой из методов оказался точнее?

2 Задание 2: kNN

- Используем датасет с сортами вина из предыдущей задачи;
- Использовать три подхода к делению выборки на тренировочную и тестовую: KFold, LOO, Stratified KFold. Для воспроизводимости зафиксируйте параметр random_state=42;

- Для каждого из методов кросс-валидации, а также для каждого $k \in [1, 50]$ (число "соседей") прогнать алгоритм ближайших соседей (sklearn.neighbors.KNeighborsClassifier) и посчитать долю правильных ответов. Какая кросс-валидация и при каком значении k дает лучший результат?;
- Произведите масштабирование признаков с помощью функции sklearn.preprocessing.scale. Снова найдите оптимальное k на трех разных кросс-валидациях. Чему оно равно? Изменилось ли оно? Изменился ли оптимальный метод валидации?