Лабораторная работа 2 (2 семестр)

Лабораторная работа 2 рассчитана на два занятия и работу дома. Её целью является изучение основ классификации данных с помощью метода случайный лес и расчёта характеристик качества классификатора.

Задание 1

1. Загрузите с сайта https://sci2s.ugr.es/keel/datasets.php набор статистических данных, указанный в вашем варианте. Разберитесь, какие данные приведены в наборе и каой атрибут является меткой класса.
2. На основе загруженного файла создайте Pandas DataFrame, подобрав правильные типы данных столбцов.
3. Выполните стандартизацию полученного дата фрейма.
4. Разделите дата фрейм на обучающую, тестовую и валидационную выборки в соотношении 5 / 3 / 2 с применением стратификации.
5. На основе обучающей и тестовой выборки постройте дерево решений. Меняя значение параметра альфа ([0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.2, 0.8]) и критерий классификации ([Entropy, Gini]) подберите наиболее удачное по макро усреднённому параметру ROC-AUC дерево классификации для подготовленных выборок.
6. На основе обучающей и тестовой выборки постройте SVM-классификатор. Меняя значение параметров kernel, gamma, coef0, degree, C (на основе вариантов, представленных в лекции 1 второго семестра) обосновано подберите наиболее удачное дерево по макро усреднённому параметру ROC-AUC классификации для подготовленных выборок.
7. На основе обучающей и тестовой выборки постройте Random Forest-классификатор. Меняя значение параметра критерий классификации ([Entropy, Gini]), а также число генерируемых деревьев и число используемых полей подберите наиболее удачный по макро усреднённому параметру ROC-AUC лес для подготовленных выборок.
8. Выполните обогащение выборки и повторите шаги 5, 6, 7. Сравните с помощью ROC-AUC-критерия и валидационной выборки, полученные 6 классификаторов и выберите лучший.

Варианты

Задание 1

1. https://sci2s.ugr.es/keel/dataset.php?cod=153
2. https://sci2s.ugr.es/keel/dataset.php?cod=155
3. https://sci2s.ugr.es/keel/dataset.php?cod=156