Машинное обучение

НИЯУ МИФИ, Кафедра финансового мониторинга

Лабораторный практикум.

В.Ю. Радыгин, Д.Ю. Куприянов

Семестр 1. Лабораторная работа 5

Лабораторная работа 5 рассчитана на два занятия и работу дома. Её целью является изучение основ классификации данных с помощью метода опорных векторов 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 какой 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.
5. На основе обучающей и тестовой выборки постройте дерево решений, рассчитайте параметры эффективности классификатора (Accuracy, Precision, Recall, ROC-AUC). 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]) обоснованно подберите наиболее удачное дерево классификации для подготовленных выборок.
6. На основе обучающей и тестовой выборки постройте SVM-классификатор, рассчитайте параметры эффективности классификатора (Accuracy, Precision, Recall, ROC-AUC). Меняя значение параметров kernel, gamma, coef0, degree, C (на основе вариантов, представленных в лекции 5) обосновано подберите наиболее удачное дерево классификации для подготовленных выборок.
7. Применить метод главных компонент. Заново построить дерево и svm-классификатор с выбранным кол-вом классов. Проанализировать результаты.
8. Используя валидационную выборку рассчитайте для лучшего дерева решений и лучшего SVM-классификатора параметры эффективности (Accuracy, Precision, Recall, ROC-AUC). Обоснуйте какой из двух классификаторов и когда лучше.

   Примечания:
   - до лекции по метрикам ROC-AUC можно не делать;
   - параметров gamma (лучше scale и auto), coef0, degree (лучше не брать >4), C - брать разумные числа.

Варианты

Задание 1

1. https://sci2s.ugr.es/keel/dataset.php?cod=210
2. https://sci2s.ugr.es/keel/dataset.php?cod=209
3. https://sci2s.ugr.es/keel/dataset.php?cod=107
4. https://sci2s.ugr.es/keel/dataset.php?cod=72